Simultaneous Blackwell Approachability and Applications to Multiclass Omniprediction

Lunjia Hu LUNJIA@ALUMNI.STANFORD.EDU
Northeastern University

Kevin Tian KJTIAN@CS.UTEXAS.EDU
University of Texas at Austin

Chutong Yang CYANG98@CS.UTEXAS.EDU
University of Texas at Austin

Editors: Steve Hanneke and Tor Lattimore

Abstract

Omniprediction is a learning problem that requires suboptimality bounds for each of a family of losses $\mathcal{L}$ against a family of comparator predictors $\mathcal{C}$. We initiate the study of omniprediction in a multiclass setting, where the comparator family $\mathcal{C}$ may be infinite. Our main result is an extension of the recent binary omniprediction algorithm of (Okoroafor et al., 2025) to the multiclass setting, with sample complexity (in statistical settings) or regret horizon (in online settings) $\approx \varepsilon^{-(k+1)}$, for ε-omniprediction in a k-class prediction problem. En route to proving this result, we design a framework of potential broader interest for solving Blackwell approachability problems where multiple sets must simultaneously be approached via coupled actions.

Keywords: Calibration, Omniprediction, Online Learning, Blackwell Approachability

1. Introduction

Omniprediction is a powerful definition of learning introduced recently by (Gopalan et al., 2022). Consider a standard supervised learning task: we receive i.i.d. samples $(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}$, where $\mathbf{x} \in \mathbb{R}^d$ are the features and $\mathbf{y}$ is the label, and we wish to build a predictor $\mathbf{p}(\mathbf{x}) \approx \mathbb{E}[\mathbf{y} \mid \mathbf{x}]$. In omniprediction, a family of losses $\mathcal{L}$ is fixed, as well as a family of comparator predictors $\mathcal{C}$. The goal is then to satisfy the simultaneous loss minimization guarantee, for $\varepsilon > 0$ and a predictor $\mathbf{p}$:

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[\ell\left(\mathbf{k}^\star_\ell(\mathbf{p}(\mathbf{x})), \mathbf{y}\right)\right] \le \min_{\mathbf{c}\in\mathcal{C}} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[\ell\left(\mathbf{c}(\mathbf{x}), \mathbf{y}\right)\right] + \varepsilon, \text{ for all } \ell \in \mathcal{L}. \tag{1}$$

Here, $\mathbf{k}^\star_\ell$ is the *ex ante optimum* mapping for a particular loss $\ell \in \mathcal{L}$, defined in (3).

The formulation (1) effectively decouples the tasks of *prediction* and *action*: once the learner has decided on a predictor $\mathbf{p}$, the decision maker who wishes to minimize a particular $\ell \in \mathcal{L}$ then takes the action $\mathbf{k}^\star_\ell \circ \mathbf{p}$. This property is particularly useful when e.g., losses can depend on parameters unknown at training time (such as a market price), or robustness to a range of loss hyperparameters is desirable. Because (1) applies to a family of losses, the predictor $\mathbf{p}$ can be viewed as a "supervised sufficient statistic" that goes beyond single loss minimization. This perspective built upon work in algorithmic fairness (Hébert-Johnson et al., 2018), and has intimate connections to indistinguishability arguments in pseudorandomness (Gopalan et al., 2023a; Gopalan and Hu, 2025).

By now, there is a rich body of work on omniprediction in statistical and online learning settings (as well as beyond) (Gopalan et al., 2022, 2023a; Hu et al., 2023; Gopalan et al., 2023b; Garg et al.,

2024; Hu et al., 2025; Dwork et al., 2025; Okoroafor et al., 2025). However, essentially all prior works focused on binary classification, where labels $\mathbf{y}$ live in the set $\{0,1\} \equiv \partial\Delta^2$ (the boundary of the simplex; see Section 2.1 for notation). This is a rather stringent restriction in the context of real-world supervised learning, which is often used for multiclass tasks, e.g., (Deng et al., 2009; Maas et al., 2011; Deng, 2012). Even the ability to handle labels $\mathbf{y} \in \partial\Delta^k \equiv [k]$, for k a constant number of classes, would substantially extend the applicability of omnipredictors.

To our knowledge, the problem of multiclass omniprediction has only been studied recently by (Noarov et al., 2025; Lu et al., 2025). These papers focused on a setting motivated by economics, where the family of comparators (viewed as an action space) is finite; the former's multiclass omniprediction result (Theorem 6.5, (Noarov et al., 2025)) is restricted to ℓ that independently decompose coordinatewise. On the other hand, Corollary 6, (Lu et al., 2025) gives a more general statement for multiclass omniprediction, but again the result is for finite $\mathcal{C}$, and incurs an $\approx \varepsilon^{-4k-2}$ overhead in the sample complexity for achieving (1) (without the consideration of runtime).

The main motivation of our work is to bridge this gap, by developing multiclass omnipredictors with guarantees closer to the state-of-the-art in binary omniprediction. Indeed, there has been substantial recent progress on improving the sample complexity and runtime of binary omniprediction for concrete pairs $(\mathcal{C}, \mathcal{L})$. For example, in the *generalized linear model* (GLM) setting, where $\mathcal{C}$ is bounded linear predictors and $\mathcal{L}$ is appropriate convex losses (cf. (5)), (Hu et al., 2025; Okoroafor et al., 2025) developed end-to-end efficient algorithms using $\approx \varepsilon^{-2}$ samples.[1] In fact, (Okoroafor et al., 2025) gave a substantial generalization, showing how to reduce binary omniprediction for arbitrary pairs of $(\mathcal{C}, \mathcal{L})$ to online learning tasks against appropriate function classes.

1.1. Our results

Our approach to multiclass omniprediction is based on the framework of (Okoroafor et al., 2025). Both (Hu et al., 2025; Okoroafor et al., 2025), as well as many prior results on binary omniprediction, leverage a reduction from (Gopalan et al., 2023a). This reduction (Proposition 7) shows that (1) is satisfied for predictors $\mathbf{p}$ satisfying appropriate notions of *multiaccuracy* (Definition 3) and *calibration* (Definiton 4), concepts we review in Section 2.2. Intuitively, these properties guarantee that our predictor $\mathbf{p}(\mathbf{x})$ passes certain statistical tests against the ground truth $\mathbf{p}^\star(\mathbf{x}) := \mathbb{E}[\mathbf{y} \mid \mathbf{x}]$, induced by the particular pair $(\mathcal{C}, \mathcal{L})$ of interest.

As in the binary case, learning multiclass predictors that satisfy multiaccuracy and calibration individually is well-studied. We discuss the former in Appendix D.4, and the latter is possible in $\approx \varepsilon^{-(k+1)}$ timesteps (in the online setting) and samples (in the statistical setting), as shown by seminal work of (Foster and Vohra, 1998) (see also (Mannor and Stoltz, 2010)). However, it is less clear how to achieve both simultaneously. In the binary setting, (Okoroafor et al., 2025) leveraged an existing calibration algorithm from (Abernethy et al., 2011) based on *Blackwell approachability*, and augmented it to also guarantee multiaccuracy. Their analysis used several important facts about binary losses, e.g., existence of an approximate basis for proper losses (Lemma 21), and a custom "halfspace satisfiability oracle" specialized to their application (Algorithm 3). Unfortunately, the natural extension of these tools to $k > 2$ classes both provably fail, necessitating a stronger framework capable of handling the multiclass setting.

1. The (Hu et al., 2025) omnipredictor requires $\mathcal{L}$ to be well-conditioned, but outputs a mixture of proper hypotheses from $\mathcal{C}$; the (Okoroafor et al., 2025) omnipredictor is improper but holds for general bounded losses. The sample complexity of $\approx \varepsilon^{-2}$ is tight for GLMs, even for a single loss (Shamir, 2015). See also (Dwork et al., 2025), who gave a similar result in a RKHS setting.

Simultaneous Blackwell approachability. Our starting point is to isolate a key technical primitive needed in the (Okoroafor et al., 2025) algorithm, and study sufficient conditions for it in greater generality. This is the focus of Section 3; here, we provide a brief overview of the technique.

The standard setting of Blackwell approachability (reviewed in Section 3.1) generalizes von Neumann's minimax theorem to vector-valued games. Consider a bilinear, vector-valued function $\mathbf{v} : \mathcal{A} \times \mathcal{B}$, and a set $\mathcal{V}$, living in the same space $\mathcal{H}$. Unlike the scalar setting, the following are not equivalent: for all $\mathbf{b} \in \mathcal{B}$ there exists $\mathbf{a} \in \mathcal{A}$ such that $\mathbf{v}(\mathbf{a}, \mathbf{b}) \in \mathcal{V}$ ("response satisfiability"), and there exists $\mathbf{a} \in \mathcal{A}$ such that for all $\mathbf{b} \in \mathcal{B}$, $\mathbf{v}(\mathbf{a}, \mathbf{b}) \in \mathcal{V}$ ("satisfiability"). Blackwell approachability (Blackwell, 1956) is an elegant compromise: whenever response satisfiability holds, we can choose $\{\mathbf{a}_t\}_{t\in[T]}$ in an online manner (before the corresponding $\{\mathbf{b}_t\}_{t\in[T]}$ is revealed), such that $\lim_{T\to\infty} \frac{1}{T} \sum_{t\in[T]} \mathbf{v}(\mathbf{a}_t, \mathbf{b}_t) \to \mathcal{V}$. This strategy has intimate connections to calibration: since (Foster, 1999) many researchers have used ideas from approachability to design calibration algorithms.

In Problem 1, we propose a *simultaneous* variant of Blackwell approachability, where there are m pairs of vector-valued functions $\mathbf{v}^{(i)}$ and sets $\mathcal{V}^{(i)}$. The goal is to choose a sequence of $\{\mathbf{a}_t\}_{t\in[T]}$ (responding online to $\{\mathbf{b}_t\}_{t\in[T]}$) such that $\lim_{T\to\infty} \frac{1}{T} \sum_{t\in[T]} \mathbf{v}^{(i)}(\mathbf{a}_t, \mathbf{b}_t) \to \mathcal{V}^{(i)}$, simultaneously for all $i \in [m]$. This primitive has clear connections to omniprediction, as both (binary and multiclass) multiaccuracy and calibration can be written in the language of Blackwell approachability.

Simultaneous Blackwell approachability can naturally be cast as a (standard) Blackwell approachability instance, by lifting the vectors and sets into a product space $\mathcal{H}^{(1)} \times \mathcal{H}^{(2)} \times \ldots \times \mathcal{H}^{(m)}$. However, we find the perspective in Problem 1 useful, as the sufficient condition of response satisfiability does not lift cleanly. We show in Lemma 13 that even when $m = 2$, there are two one-dimensional subsets $\mathcal{V}^{(1)}$, $\mathcal{V}^{(2)}$ and corresponding vector-valued functions, that are both response satisfiable (and hence approachable in isolation), but not simultaneously approachable.

Our main contribution in Section 3 is a sufficient condition for simultaneous Blackwell approachability, stated in the form of an oracle requirement (Definition 11). Our oracle is natural in the context of (Blackwell, 1956), who gave an alternate characterization of approachability: every halfspace containing $\mathcal{U}$ should be satisfiable. This statement was later made algorithmic by (Abernethy et al., 2011) using online learning techniques. Our sufficient condition can then be cleanly stated as: for any halfspaces each containing one $\mathcal{U}^{(i)}$, and any specified convex combination $\mathbf{w} \in \Delta^m$, the halfspaces should be satisfiable *on average* (with respect to $\mathbf{w}$). We further build upon (Abernethy et al., 2011) to leverage existence of such an oracle to solve simultaneous Blackwell approachability with an explicit rate (Theorem 14).

While our reduction is a relatively straightforward extension of (Abernethy et al., 2011) (and indeed, (Okoroafor et al., 2025) also implicitly gave a variant of Theorem 14), we believe that explicitly isolating this sufficient condition will prove useful to the community. To ease applications, we show that Theorem 14 holds in much greater generality, e.g., statistical and contextual settings. We provide a high-probability bound capable of flexibly handling these extensions (Corollary 18).

Multiclass omniprediction. Our simultaneous Blackwell approachability framework reduces omniprediction to implementing an appropriate mixture linear optimization oracle (MLOO, Definition 11), and to designing appropriate online learners for each of two sets $\mathcal{V}^{(i)}$ in isolation (corresponding to calibrated and multiaccurate predictors). While an appropriate MLOO was explicitly given in (Okoroafor et al., 2025) for the binary omniprediction setting, it is unclear how to generalize their strategy (based on an algorithmic Sperner's lemma) to hold in higher dimensions.

Towards leveraging our results for omniprediction, in Appendix D.1, we give a meta-result for designing MLOOs when all $\mathbf{v}^{(i)}$ share a common structure. Roughly speaking, we require each $\mathbf{v}^{(i)}$ to take as input a prediction $\mathbf{p}$ and a label $\mathbf{y}$, and to be linear in the prediction error $\mathbf{p} - \mathbf{y}$ (a more formal statement is in (39)). Under these assumptions, we show how to use the minimax theorem and linear programming to generically design MLOOs compatible with our simultaneous Blackwell approachability framework, which extends to future potential applications.

By combining our framework, our new MLOO construction, and known online learners, we obtain our multiclass omnipredictors in Appendix D. The following is a representative result.

Theorem 1 (Informal, see Theorem 43) *Let $\mathcal{L}$ be the family of multiclass GLM losses (5) and let $\mathcal{C}$ be the family of bounded $k \times d$ linear classifiers (49). Then given T i.i.d. samples $(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}$ for*

$$T = k \cdot \Omega\left(\frac{1}{\varepsilon}\right)^{k+1}, \tag{2}$$

we return an ε-omnipredictor in time $O(dkT) + O(\frac{1}{\varepsilon})^{2k}\mathrm{poly}(k, \log \frac{1}{\varepsilon})$, with high probability.

We pause to make some remarks about Theorem 43, which is specialized to the benchmark class of multiclass GLM losses, an expressive family that includes all proper losses after reparameterization (cf. Lemma 6). First, it is fully explicit and does not rely on any computationally-infeasible oracles. Second, although its sample complexity scales exponentially in the number of classes k, this growth is relatively mild for small constant k, and the bound is independent of the ambient dimension d of the features $\mathbf{x}$. Third, because our approach is based on the indistinguishability argument of (Gopalan et al., 2023a) (i.e., it goes through calibration and multiaccuracy), the exponential dependence on k is inevitable due to a lower bound from Theorem 1.12, (Hu and Vadhan, 2025). Indeed, our bound (2) recovers the same dependence on k as existing algorithms for the simpler task of multiclass calibration,[2] and improves by a quartic factor over the prior work (Lu et al., 2025) (while also handling infinite $\mathcal{C}$).

We prove our formal variant of Theorem 1 in Appendix D, as well as extensions to online omniprediction, and general families of multiclass losses and comparators (Theorem 46).

Other consequences. As a warmup, in Appendix C, we rederive the main results of (Okoroafor et al., 2025) in the binary setting by way of our new formalism. We believe this may be useful to the community, as it cleanly separates out the requirements of each online learner. For example, Theorem 30, our specialization of Theorem 1 to binary omniprediction, uses $\approx \frac{1}{\varepsilon^2}$ samples and gives an end-to-end construction of an omnipredictor for binary GLMs. This removes the well-conditioning requirement from (Hu et al., 2025), and does not rely on computationally-infeasible halfspace optimization oracles (i.e., ERM for linear thresholds) as in (Okoroafor et al., 2025).[3] During the preparation of this manuscript, the third arXiv version of (Okoroafor et al., 2025) independently noted this requirement is removable (see their updated Theorem 7.1); our modular framework makes this point transparent, which we believe will prove useful in similar future applications.

2. Recent works by (Peng, 2025; Fishelson et al., 2025) on multiclass omniprediction have traded off the exponential dependence on k for an exponential dependence on $\frac{1}{\varepsilon}$. We discuss these works in greater detail in Section 1.2.

3. This requirement is stated in Theorem 5, (Okoroafor et al., 2025), where ERM access for the composition of thresholding with the comparator family is assumed. Even when the comparator family is linear functions, this requires implementing a halfspace ERM oracle, a well-known NP-hard problem in computational learning theory (Johnson and Preparata, 1978; Ben-David et al., 2003).

Our framework has additional implications for the theory of calibration and omniprediction. For example, in Appendix E, we show that our construction directly extends to omnipredicting against *unions of comparators*, i.e., the best comparator in any of m families $\{\mathcal{C}^{(i)}\}_{i\in[m]}$, as long as we can omnipredict against each family individually. This simple extension was previously unknown, and is made possible by the generality of our construction in Appendix D.1. We are optimistic that our pipeline for constructing omnipredictors will have future consequences for related problems.

1.2. Related work

Multiclass calibration. Multiclass calibration has seen a resurgence of interest recently due to its use in evaluating modern classifiers in machine learning (Guo et al., 2017). A range of works have proposed new algorithms and relaxations of this problem (Kull and Flach, 2015; Kull et al., 2019; Zhao et al., 2021; Gopalan et al., 2024a).

Of particular note, recent works (Peng, 2025; Fishelson et al., 2025) gave algorithms with horizons $\approx k^{\text{poly}(\varepsilon^{-1})}$ for ε-multi-class calibration, which is polynomial in k for constant ε. This improves upon the classical $\varepsilon^{-(k+1)}$ rate for multiclass calibration (Foster and Vohra, 1998) (as in Theorem 1) in some parameter regimes. However, these results are not obtained through Blackwell approachability, and thus it seems difficult to incorporate a multiaccuracy component directly as would be required for omniprediction using the indistinguishability framework of (Gopalan et al., 2023a). Further, the $\exp(\Omega(k))$ lower bound in Theorem 1.12, (Hu and Vadhan, 2025) effectively rules out the use of these results within the framework.

Finally, as discussed earlier, (Noarov et al., 2025; Lu et al., 2025) are the primary works that have studied multiclass omniprediction; we compared our Theorem 1 to (Lu et al., 2025) earlier. Regarding (Noarov et al., 2025), their omniprediction result only applies to a restricted family of multiclass GLM losses, namely those which decompose coordinatewise in their argument (see Definition 6.12). This makes it incompatible with several common GLM losses in the (standard) setting of Theorem 1. For example, consider the *cross entropy* $\ell(\mathbf{p}, \mathbf{y}) = -\mathbb{E}_{i\sim\mathbf{y}}[\log \mathbf{p}_i]$ popularly used in machine learning evaluations. This GLM loss falls into the framework of (Lu et al., 2025), but once it is cast as a GLM learning problem, the correct parameterization is in the unlinked space (where linear comparator predictors live), for which the corresponding loss is $\ell(\mathbf{t}, \mathbf{y}) = \log(\sum_{i\in[d]} \exp(\mathbf{t}_i)) - \langle \mathbf{t}, \mathbf{y}\rangle$, which is not coordinatewise separable. For more discussion on this point, see Lemma 6 and Section 2.2, (Hu et al., 2025).

Finally, we note that while (Gopalan et al., 2024b) focuses on omnipredictors for one-dimensional regression tasks, they provide a sufficient statistics framework for omniprediction that could plausibly extend to multiclass omniprediction. However, it is unclear what the precise exponential dependence on k would be in the complexities resulting from this strategy, whereas a focus of our paper is obtaining a relatively sharp constant of 1 in the exponent.

Multiclass learning. Multiclass learning is a well-studied topic in learning theory in general. For example, classical works by (Natarajan, 1989; Ben-David et al., 1992) proposed various statistical dimensions characterizing the sample complexity of multiclass PAC learning. More recently, (Daniely et al., 2015; Brukhim et al., 2022) shows that in multiclass setting, empirical risk minimization does not provide a uniform bound on sample complexity, and proved that a quantity known as the DS dimension does tightly characterizes multiclass PAC learnability. A follow-up work by (Charikar and Pabbaraju, 2023) shows that k-DS dimension, a generalization of DS dimension,

characterizes k-list learnability. Most of these works primarily consider the sample complexity of multiclass learning rather than end-to-end efficient algorithms.

Blackwell approachability. Various works have generalized Blackwell approachability and applied it to problems with a similar spirit to our simultaneous approachability setting in Problem 1. However, to our knowledge none of them directly study achievability and algorithms for approaching multiple sets. The works most closely-related to our setup include (Mannor et al., 2014) who study Blackwell approachability for unknown games, i.e., where the structure of the game or target is unknown; (Fournier et al., 2021) who study Blackwell approachability with additional constraints for the payoff space; and (Lee et al., 2022) who study multiclass calibration and calibeating.

2. Preliminaries

In Section 2.1, we define notation used throughout the paper, and state helper results from the online learning literature. We then introduce preliminaries for omniprediction in Section 2.2.

2.1. Notation

We denote vectors in lowercase boldface and matrices in uppercase boldface. For $n \in \mathbb{N}$ we let $[n] := \{i \in \mathbb{N} \mid i \le n\}$. We let $\mathbf{1}_d$ and $\mathbf{0}_d$ denote the all-ones and all-zeroes vectors in $\mathbb{R}^d$. For $i \in [d]$, we let $\mathbf{e}_i \in \mathbb{R}^d$ denote the i^{th} standard basis vector when the dimension d is clear from context. When $\mathcal{E}$ is some event, we let $\mathbb{I}_{\mathcal{E}}$ to denote the 0-1 indicator variable of the event.

When $\|\cdot\|$ is a norm on $\mathbb{R}^d$, we let $\|\cdot\|_*$ denote its dual norm. For $p \in \mathbb{R}_{\ge 1} \cup \{\infty\}$ we let $\|\cdot\|_p$ denote the ℓ_p norm of a vector argument. Note that when $\|\cdot\| = \|\cdot\|_p$ for some $p \in \mathbb{R}_{\ge 1} \cup \{\infty\}$, then $\|\cdot\|_* = \|\cdot\|_q$ for the value of $q \in \mathbb{R}_{\ge 1} \cup \{\infty\}$ satisfying $\frac{1}{p} + \frac{1}{q} = 1$. We say that a vector-valued function $\mathbf{v} : \mathbb{R}^d \to \mathbb{R}^k$ is β-Lipschitz in $\|\cdot\|$ if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, $\|\mathbf{v}(\mathbf{x}) - \mathbf{v}(\mathbf{y})\|_* \le \beta \|\mathbf{x} - \mathbf{y}\|$.

For $\bar{\mathbf{x}} \in \mathbb{R}^d$ and $r > 0$ we define $\mathbb{B}_p^d(\bar{\mathbf{x}}, r) := \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x} - \bar{\mathbf{x}}\|_p \le r\}$ to be an ℓ_p ball centered at $\bar{\mathbf{x}}$. When $\bar{\mathbf{x}}$ is omitted, $\bar{\mathbf{x}} = \mathbf{0}_d$. For a compact set $\mathcal{K} \subseteq \mathbb{R}^d$ we let $\mathbf{\Pi}_{\mathcal{K}}(\mathbf{v}) := \arg\min_{\mathbf{x} \in \mathcal{K}} \|\mathbf{v} - \mathbf{x}\|_2$ denote the Euclidean projection. For $p, q \ge 1$ and $\mathbf{M} \in \mathbb{R}^{n \times d}$, we denote

$$\|\mathbf{M}\|_{p \to q} := \max_{\mathbf{v} \in \mathbb{B}_p^d(1)} \|\mathbf{M}\mathbf{v}\|_q .$$

We let $\Delta^k := \{\mathbf{v} \in \mathbb{R}^k_{\ge 0} \mid \|\mathbf{v}\|_1 = 1\}$ denote the probability simplex in dimension k. When S is a set, we overload notation and let $\mathbf{v} \in \mathbb{R}^S$ be a vector with coordinates indexed by elements $s \in S$, and we similarly define $\mathbf{1}_S$, $\mathbf{0}_S$, Δ^S, etc. For convex $\mathcal{K} \subseteq \mathbb{R}^d$ we use $\partial\mathcal{K}$ to denote the *boundary* of $\mathcal{K}$, i.e., all $\mathbf{v} \in \mathcal{K}$ that cannot be written as a convex combination of other points in $\mathcal{K}$. For example, $\partial\Delta^k$ is the set of standard basis vectors $\{\mathbf{e}_i\}_{i \in [k]}$. For a distribution $\mathcal{P}$ supported on Ω we write $\omega \sim \mathcal{P}$ to mean a sample from the distribution, and when $\mathbf{p} \in \Delta^k$ we overload notation and let $i \sim \mathbf{p}$ (resp. $\mathbf{y} \sim \mathbf{p}$) mean a sample that takes on the value i (resp. $\mathbf{y} = \mathbf{e}_i$) with probability $\mathbf{p}_i$. We say that $\mathcal{N}$ is an *ε-net in* $\|\cdot\|$ for $\mathcal{K}$ if for all $\mathbf{x} \in \mathcal{K}$, there exists $\mathbf{x}' \in \mathcal{N}$ such that $\|\mathbf{x} - \mathbf{x}'\| \le \varepsilon$. When $\|\cdot\|$ is omitted, we let $\|\cdot\| = \|\cdot\|_1$ by default. The following construction is standard.[4]

Fact 1 *For all $k \in \mathbb{N}$ and $\varepsilon \in (0, 1)$, there exists $\mathcal{N}$, an ε-net of Δ^k, satisfying $|\mathcal{N}| \le (\frac{5}{\varepsilon})^{k-1}$.*

4. This result is stated in (Vershynin, 2018), Corollary 4.2.11 for the ℓ_2 ball, but the same construction works for the ℓ_1 ball in dimension $k - 1$. Projection onto the subspace $\mathbf{1}_k^\top \mathbf{v} = 1$ at most doubles the ℓ_1 distance, so we adjusted the constant.

We say that $\mathcal{H} \subseteq \mathbb{R}^d$ is a *halfspace* if for some $(\mathbf{v}, c) \in \mathbb{R}^d \times \mathbb{R}$, $\mathcal{H} = \{\mathbf{x} \mid \mathbf{v} \cdot \mathbf{x} \leq c\}$. For a sequence $\{\mathbf{v}_t\}_{t\in\mathbb{N}} \subseteq \mathbb{R}^d$ and a set $\mathcal{V} \subseteq \mathbb{R}^d$, we write $\lim_{t\to\infty} \mathbf{v}_t \to \mathcal{V}$ to mean that for any $\varepsilon > 0$, there is some $T(\varepsilon)$ such that for all $t \geq T(\varepsilon)$, $\mathbf{v}_t$ is within ε of the set $\mathcal{V}$ in Euclidean distance.[5]

We often refer to sequences of vectors, e.g., $\{\mathbf{x}_t\}_{t\in[T]}$, by indexing the set of indices, e.g., as $\mathbf{x}_{[T]}$. We use $f \circ g$ to denote the composition of two functions f and g.

2.2. Omniprediction

We consider two *supervised learning* problem settings in k-class prediction. In the following discussion, let $\ell : \Omega \times \partial\Delta^k \to \mathbb{R}$ be a loss function that evaluates predictions (in a set Ω) and labels.

Online setting. There is a sequence of examples $\{(\mathbf{x}_t, \mathbf{y}_t)\}_{t\in[T]}$ presented to us in an online fashion. Our goal is to predict $\{\mathbf{p}_t \in \Omega\}_{t\in[T]}$ where $\mathbf{p}_t$ can only depend on previous examples $\{(\mathbf{x}_s, \mathbf{y}_s)\}_{s<t}$ and the current features $\mathbf{x}_t$, in a way that approximately minimizes $\sum_{t\in[T]} \ell(\mathbf{p}_t, \mathbf{y}_t)$.

Statistical setting. There is a distribution $\mathcal{D}$ over $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^d \times \partial\Delta^k$. We refer to a pair $(\mathbf{x}, \mathbf{y})$ as an *example*, and we refer to the first marginal $\mathbf{x} \in \mathbb{R}^d$ as the *features* (distributed $\sim \mathcal{D}_{\mathbf{x}}$) and $\mathbf{y} \in \partial\Delta^k$ as the *label* of the example. A label $\mathbf{y} = \mathbf{e}_i$ represents that the example belongs to class $i \in [k]$. For example, when $k = 2$ this is the setting of binary classification. Our goal is to learn a predictor $\mathbf{p} : \mathbb{R}^d \to \Omega$ that approximately minimizes the population loss, $\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[\ell(\mathbf{p}(\mathbf{x}), \mathbf{y})]$.

Two natural questions arise from these problem formulations: what loss ℓ should we consider, and what benchmark should we use to measure approximate optimality? The recently-introduced notion of omniprediction (Gopalan et al., 2022) captures both facets of the problem simultaneously.

Fix a family of loss functions $\mathcal{L}$ such that each $\ell \in \mathcal{L}$ sends $\Omega \times \partial\Delta^k \to \mathbb{R}$, and fix a family of *comparator* predictors $\mathcal{C}$ such that each $\mathbf{c} \in \mathcal{C}$ sends features $\mathbf{x} \in \mathbb{R}^d$ to a prediction $\mathbf{c}(\mathbf{x}) \in \Omega$. We also define, for each ℓ, an associated *ex ante optimum* mapping $\mathbf{k}^\star_\ell : \Delta^k \to \Omega$,

$$\mathbf{k}^\star_\ell(\mathbf{p}) := \arg\min_{\mathbf{k}^\star\in\Omega} \mathbb{E}_{i\sim\mathbf{p}} \left[\ell(\mathbf{k}^\star, \mathbf{e}_i)\right], \tag{3}$$

with arbitrary tie-breaking. To interpret (3), fix some distribution over labels $\mathbf{p} \in \Delta^k$. Then $\mathbf{k}^\star_\ell$ maps $\mathbf{p}$ to the best possible prediction (according to ℓ), had labels actually been generated $\sim \mathbf{p}$.

A particularly desirable set of losses ℓ is those which permit taking $\mathbf{k}^\star_\ell(\mathbf{p}) = \mathbf{p}$, i.e., where the best post-processing is the identity function (in this case, the first argument of ℓ lives in $\Omega = \Delta^k$). Such losses are called *proper*, and we denote the family of all proper loss functions by $\mathcal{L}_{\text{prop}}$.

We are now ready to define omniprediction in the online and statistical settings.

Definition 2 (Omniprediction) *In the online setting, we call $\mathbf{p}_{[T]} \in (\Delta^k)^T$ an ε-omnipredictor for $(\mathbf{x}_{[T]}, \mathbf{y}_{[T]}, \mathcal{L}, \mathcal{C})$, where $\mathcal{L}$ is a family of loss functions, and $\mathcal{C}$ is a family of comparator predictors, if*

$$\frac{1}{T}\sum_{t\in[T]} \ell(\mathbf{k}^\star_\ell(\mathbf{p}_t), \mathbf{y}_t) \leq \frac{1}{T}\sum_{t\in[T]} \ell(\mathbf{c}(\mathbf{x}_t), \mathbf{y}_t) + \varepsilon, \text{ for all } \ell \in \mathcal{L},\ \mathbf{c} \in \mathcal{C}.$$

In the statistical setting, we call $\mathbf{p} : \mathbb{R}^d \to \Delta^k$ an ε-omnipredictor for $(\mathcal{D}, \mathcal{L}, \mathcal{C})$, where $\mathcal{D}$ is a distribution over $\mathbb{R}^d \times \partial\Delta^k$, if

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}} \left[\ell(\mathbf{k}^\star_\ell(\mathbf{p}(\mathbf{x})), \mathbf{y})\right] \leq \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}} \left[\ell(\mathbf{c}(\mathbf{x}), \mathbf{y})\right] + \varepsilon, \text{ for all } \ell \in \mathcal{L},\ \mathbf{c} \in \mathcal{C}.$$

5. In $\mathbb{R}^d$, all norms are equivalent up to universal constants, so using Euclidean distance is without loss of generality.

If $\mathbf{p}$ *is a randomized function from* $\mathbb{R}^d \to \Delta^k$*, then we call it an* ε*-omnipredictor for* $(\mathcal{D}, \mathcal{L}, \mathcal{C})$ *if the above display holds taking expectations over* $\mathbf{p}$ *as well.*

We design omnipredictors via indistinguishability, a recipe pioneered by (Gopalan et al., 2023a).

Definition 3 ($\mathcal{F}$-multiaccuracy) *Let* $\mathcal{F}$ *be a family of functions* $\mathbf{f} : \mathbb{R}^d \to \Omega$ *where* $\Omega \subseteq \mathbb{R}^k$*. In the online setting, we say that* $\mathbf{p}_{[T]} \in (\Delta^k)^T$ *satisfies* ε*-*$(\mathbf{x}_{[T]}, \mathbf{y}_{[T]}, \mathcal{F})$*-multiaccuracy if*

$$\frac{1}{T}\sum_{t\in[T]} \langle \mathbf{p}_t - \mathbf{y}_t, \mathbf{f}(\mathbf{x}_t)\rangle \le \varepsilon, \text{ for all } \mathbf{f} \in \mathcal{F}.$$

In the statistical setting, we say that $\mathbf{p} : \mathbb{R}^d \to \Delta^k$ *satisfies* ε*-*$(\mathcal{D}, \mathcal{F})$*-multiaccuracy if*

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[\langle \mathbf{p}(\mathbf{x}) - \mathbf{y}, \mathbf{f}(\mathbf{x})\rangle\right] \le \varepsilon, \text{ for all } \mathbf{f} \in \mathcal{F}.$$

Definition 4 ($\mathcal{W}$-calibration) *Let* $\mathcal{W}$ *be a family of weight functions* $\mathbf{w} : \Delta^k \to \mathbb{R}^k$*. In the online setting, we say that* $\mathbf{p}_{[T]} \in (\Delta^k)^T$ *satisfies* ε*-*$(\mathbf{y}_{[T]}, \mathcal{W})$*-calibration if*

$$\frac{1}{T}\sum_{t\in[T]} \langle \mathbf{p}_t - \mathbf{y}_t, \mathbf{w}(\mathbf{p}_t)\rangle \le \varepsilon, \text{ for all } \mathbf{w} \in \mathcal{W}.$$

In the statistical setting, we say that $\mathbf{p} : \mathbb{R}^d \to \Delta^k$ *satisfies* ε*-*$(\mathcal{D}, \mathcal{W})$*-calibration if*

$$\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[\langle \mathbf{p}(\mathbf{x}) - \mathbf{y}, \mathbf{w}(\mathbf{p}(\mathbf{x}))\rangle\right], \text{ for all } \mathbf{w} \in \mathcal{W}.$$

To connect Definitions 3 and 4 to omniprediction, we use the following equivalence.

Lemma 5 *Let* $\ell : \Omega \times \partial\Delta^k$ *be a loss function, and define its* discrete derivative $\mathbf{d}_\ell : \Omega \to \mathbb{R}^k$*:*

$$[\mathbf{d}_\ell(\mathbf{t})]_i := \ell(\mathbf{t}, \mathbf{e}_i) \text{ for all } i \in [k]. \tag{4}$$

Then for any $\mathbf{t} \in \Omega$*,* $\mathbf{p} \in \Delta^k$*, and* $\mathbf{q} \in \Delta^k$*,* $\mathbb{E}_{\mathbf{y}\sim\mathbf{p}}[\ell(\mathbf{t}, \mathbf{y})] - \mathbb{E}_{\mathbf{y}\sim\mathbf{q}}[\ell(\mathbf{t}, \mathbf{y})] = \langle \mathbf{d}_\ell(\mathbf{t}), \mathbf{p} - \mathbf{q}\rangle$*. Moreover, the same is true if we redefine* $\mathbf{d}_\ell(\mathbf{t}) \leftarrow \mathbf{d}_\ell(\mathbf{t}) - \alpha(\mathbf{t})\mathbf{1}_k$ *for any* $\alpha(\mathbf{t}) \in \mathbb{R}$*.*

We defer a proof of Lemma 5 to Appendix A. One notable loss family is $\mathcal{L}_{\text{GLM}}$, the family of *generalized linear model* (GLM) losses:

$$\mathcal{L}_{\text{GLM}} := \left\{\ell : \mathbb{R}^k \times \partial\Delta^k \to \mathbb{R} \mid \ell(\mathbf{t}, \mathbf{y}) = \omega(\mathbf{t}) - \langle \mathbf{t}, \mathbf{y}\rangle,\ \omega : \mathbb{R}^k \to \mathbb{R} \text{ convex with } \nabla\omega : \mathbb{R}^k \to \Delta^k\right\}. \tag{5}$$

This is because by taking $\alpha(\mathbf{t}) = \omega(\mathbf{t})$ in Lemma 5, we may choose

$$\mathbf{d}_\ell(\mathbf{t}) = -\mathbf{t}, \text{ for all } \ell \in \mathcal{L}_{\text{GLM}}. \tag{6}$$

A result of (Gneiting and Raftery, 2007) shows $\mathcal{L}_{\text{GLM}} \equiv \mathcal{L}_{\text{prop}}$ up to reparameterization.

Lemma 6 (Theorem 1, (Gneiting and Raftery, 2007)) *Let* $\ell : \Delta^k \times \partial\Delta^k \to \mathbb{R}$ *be a loss function. Then* $\ell \in \mathcal{L}_{\text{prop}}$ *iff there exists a convex function* $\psi : \Delta^k \to \mathbb{R}$*, such that* $\ell(\mathbf{p}, \mathbf{y}) = -\psi(\mathbf{p}) + \langle \partial\psi(\mathbf{p}), \mathbf{p} - \mathbf{y}\rangle$*. Taking* $\omega := \psi^*$*, we have that* $\ell(\partial\psi(\mathbf{p}), \mathbf{y}) = \omega(\partial\psi(\mathbf{p})) - \langle \partial\psi(\mathbf{p}), \mathbf{y}\rangle$ *is a GLM loss in the argument* $\mathbf{t} := \partial\psi(\mathbf{p})$*.*

We now state our omnipredictor recipe, extending (Gopalan et al., 2023a) to multiclass setting.

Proposition 7 *Following notation from Definition 2 and (4), in the online setting, if* $\mathbf{p}_{[T]}$ *satisfies* ε_1*-*$(\mathbf{x}_{[T]}, \mathbf{y}_{[T]}, \mathcal{F})$*-multiaccuracy and* ε_2*-*$(\mathbf{y}_{[T]}, \mathcal{W})$*-calibration for*

$$\mathcal{F} := \{\mathbf{d}_\ell \circ \mathbf{c}\}_{\ell \in \mathcal{L}, \mathbf{c} \in \mathcal{C}}\,,\ \mathcal{W} := \{-\mathbf{d}_\ell \circ \mathbf{k}_\ell^\star\}_{\ell \in \mathcal{L}}\,,$$

then $\mathbf{p}_{[T]}$ *is an* $(\varepsilon_1 + \varepsilon_2)$*-omnipredictor for* $(\mathbf{x}_{[T]}, \mathbf{y}_{[T]}, \mathcal{L}, \mathcal{C})$*.*

In the statistical setting, if $\mathbf{p} : \mathbb{R}^d \to \Delta^k$ *satisfies* ε_1*-*$(\mathcal{D}, \mathcal{F})$*-multiaccuracy and* ε_2*-*$(\mathcal{D}, \mathcal{W})$*-calibration, then* $\mathbf{p}$ *is an* $(\varepsilon_1 + \varepsilon_2)$*-omnipredictor for* $(\mathcal{D}, \mathcal{L}, \mathcal{C})$*.*

As the proof of Proposition 7 only slightly modifies the binary case, we defer it to Appendix A.

3. Simultaneous Blackwell Approachability

In this section, we develop our main framework, which solves a simultaneous variant of classical Blackwell approachability (Blackwell, 1956) with $m \geq 1$ convex sets. We review the standard variant in Section 3.1. We then give our algorithm in Section 3.2, which builds upon (Abernethy et al., 2011) to reduce simultaneous approachability problems to implementing a certain oracle.

3.1. Blackwell approachability

Let $\mathcal{A} \subseteq \mathbb{R}^a$ and $\mathcal{B} \subseteq \mathbb{R}^b$ be compact, convex sets. The classical minimax theorem of von Neumann implies that if $f(\mathbf{a}, \mathbf{b})$ is a bilinear function of $\mathbf{a} \in \mathcal{A}$ and $\mathbf{b} \in \mathcal{B}$, then

$$\min_{\mathbf{a} \in \mathcal{A}} \max_{\mathbf{b} \in \mathcal{B}} f(\mathbf{a}, \mathbf{b}) = \max_{\mathbf{b} \in \mathcal{B}} \min_{\mathbf{a} \in \mathcal{A}} f(\mathbf{a}, \mathbf{b}). \tag{7}$$

(Blackwell, 1956) considered the following generalization to vector-valued functions.

Definition 8 (Satisfiability) *Let* $\mathcal{A} \subseteq \mathbb{R}^a, \mathcal{B} \subseteq \mathbb{R}^b$ *be compact and convex, let* $\mathcal{V} \subseteq \mathbb{R}^d$ *be convex, and let* $\mathbf{v} : \mathcal{A} \times \mathcal{B} \to \mathbb{R}^d$ *be a bilinear, vector-valued function.*

- *We call* $\mathcal{V}$ satisfiable *if there exists* $\mathbf{a} \in \mathcal{A}$ *such that* $\mathbf{v}(\mathbf{a}, \mathbf{b}) \in \mathcal{V}$ *for all* $\mathbf{b} \in \mathcal{B}$*.*
- *We call* $\mathcal{V}$ response-satisfiable *if for all* $\mathbf{b} \in \mathcal{B}$ *there exists* $\mathbf{a} \in \mathcal{A}$ *such that* $\mathbf{v}(\mathbf{a}, \mathbf{b}) \in \mathcal{V}$*.*
- *We call* $\mathcal{V}$ halfspace-satisfiable *if all halfspaces containing* $\mathcal{V}$ *are satisfiable.*

Definition 8 suggests a natural generalization of (7): are all response-satisfiable sets also satisfiable? This holds when $d = 1$ by taking $\mathcal{V}$ to be a sublevel set $(\infty, v]$: in this case, $\mathcal{V}$ is both satisfiable and response-satisfiable iff v is at least the value of the game (7). Unfortunately, simple counterexamples (e.g., $\mathcal{V} = \{(x, x) \mid x \in [0, 1]\}, \mathbf{v}(a, b) = (a, b)$) preclude this equivalence for $d > 1$. The main result of (Blackwell, 1956) (see also a more modern exposition by (Abernethy et al., 2011)) is that a different equivalence holds.

Proposition 9 ((Blackwell, 1956)) *In the setting of Definition 8, the following are equivalent.*

- $\mathcal{V}$ *is response-satisfiable.*
- $\mathcal{V}$ *is halfspace-satisfiable.*

- $\mathcal{V}$ *is* approachable*, i.e., for any sequence* $\{\mathbf{b}_t\}_{t\in\mathbb{N}}$*, there is a choice of* $\{\mathbf{a}_t\}_{t\in\mathbb{N}}$ *such that* $\mathbf{a}_t$ *depends only on* $\mathbf{b}_{[t-1]}$*, and such that*

$$\lim_{t\to\infty}\frac{1}{t}\sum_{s\in[t]}\mathbf{v}(\mathbf{a}_s,\mathbf{b}_s)\to\mathcal{V}.$$

A quantitative, algorithmic variant of Proposition 9 was given by (Abernethy et al., 2011). To explain its relevance to Section 3.2, we specialize our attention to sets $\mathcal{V}$ induced via a convex set of distinguishers $\mathcal{U}\subseteq\mathbb{R}^d$, and a scalar $\rho\in\mathbb{R}$. Specifically, we are interested in sets of the form

$$\mathcal{V}:=\left\{\mathbf{v}\in\mathbb{R}^d\mid\sup_{\mathbf{u}\in\mathcal{U}}\langle\mathbf{u},\mathbf{v}\rangle\le\rho\right\}. \tag{8}$$

The function $\sup_{\mathbf{u}\in\mathcal{U}}\langle\mathbf{u},\mathbf{v}\rangle$ is called the *support function* of $\mathcal{U}$ in convex analysis. Intuitively, we can view each member $\mathbf{u}\in\mathcal{U}$ as a distinguisher that tests whether a given $\mathbf{v}\in\mathbb{R}^d$ satisfies $\langle\mathbf{u},\mathbf{v}\rangle\le\rho$. If $\mathbf{v}$ passes all tests given by $\mathcal{U}$, then we can certify $\mathbf{v}\in\mathcal{V}$ as in (8).

Focusing on (8), i.e., sublevel sets of support functions, may seem restrictive. We first mention that this specialization captures an important family of sets.

Lemma 10 *If* $\mathcal{V}\subseteq\mathbb{R}^d$ *is closed and convex,* $\mathbf{0}_d\in\mathcal{V}$*, and* $\rho>0$*, there exists* $\mathcal{U}$ *so that* (8) *holds.*

We defer the proof to Appendix A.

In fact, specializing to (8) is the first step in the reduction of (Abernethy et al., 2011) (see their Proposition 2). The key observation of (Abernethy et al., 2011) is that if $\mathcal{V}$ is a *convex cone* (a set closed under nonnegative linear combinations), then it satisfies (8) for $\rho=0$ and $\mathcal{U}$ taken to be the *dual cone* to $\mathcal{V}$. Moreover, any convex set $\mathcal{V}\subseteq\mathbb{R}^d$ can be lifted to a convex cone in $\mathbb{R}^{d+1}$, by defining the set

$$\mathrm{clift}(\mathcal{V}):=\left\{(\mathbf{v},c)\in\mathbb{R}^d\times\mathbb{R}_{>0}\mid\frac{\mathbf{v}}{c}\in\mathcal{V}\right\}\cup\{\mathbf{0}_{d+1}\}.$$

Intuitively, the $c=1$ "slice" of $\mathrm{clift}(\mathcal{V})$ projects to $\mathcal{V}$ in the first d dimensions, and the slice at an arbitrary $c\ge0$ projects to $c\mathcal{V}$. By converting between the distance of a point in $\mathbb{R}^{d+1}$ to $\mathrm{clift}(\mathcal{V})$, and the distance of its projection in $\mathbb{R}^d$ to $\mathcal{V}$ (paying an overhead of $\approx\mathrm{diam}(\mathcal{V})$), (Abernethy et al., 2011) show that approaching sets (8) suffices to derive a general algorithmic variant of Proposition 9.

3.2. Framework

We next provide a generalization of Algorithm 2, (Abernethy et al., 2011) that simultaneously approaches a collection of m convex sets of the form (8). In fact, under our oracle abstraction (to be introduced in Definition 11), the convex sets we approach need not live in a finite-dimensional space, and can come from an arbitrary Hilbert space. We formalize our problem setting here.

Problem 1 (Simultaneous Blackwell approachability) *Let* $a,b,m\in\mathbb{N}$*, let* $\rho,\varepsilon\ge0$*, and let* $\mathcal{A}\subseteq\mathbb{R}^a$ *and* $\mathcal{B}\subseteq\mathbb{R}^b$ *be compact and convex. For all* $i\in[m]$*, let* $\mathcal{V}^{(i)},\mathcal{U}^{(i)}\subseteq\mathcal{H}^{(i)}$ *satisfy*

$$\mathcal{V}^{(i)}:=\left\{\mathbf{v}\in\mathcal{H}^{(i)}\mid\sup_{\mathbf{u}\in\mathcal{U}^{(i)}}\langle\mathbf{u},\mathbf{v}\rangle\le\rho\right\}, \tag{9}$$

where $\mathcal{H}^{(i)}$ is a Hilbert space. Let $\mathbf{v}^{(i)} : \mathcal{A} \times \mathcal{B} \to \mathcal{H}^{(i)}$ be a bilinear, vector-valued function for all $i \in [m]$. Our goal is to observe a sequence $\mathbf{b}_{[T]} \in \mathcal{B}^T$, and to choose $\mathbf{a}_{[T]} \in \mathcal{A}^T$ so that $\mathbf{a}_t$ depends on $\mathbf{b}_{[t-1]}$ for all $t \in [T]$, and

$$\max_{i \in [m]} \sup_{\mathbf{u}^{(i)} \in \mathcal{U}^{(i)}} \left\langle \mathbf{u}^{(i)}, \frac{1}{T} \sum_{t \in [T]} \mathbf{v}^{(i)}(\mathbf{a}_t, \mathbf{b}_t) \right\rangle \leq \rho + \varepsilon. \tag{10}$$

When $m = 1$ and $\varepsilon \to 0$, the bound (10) implies that $\frac{1}{T} \sum_{t \in [T]} \mathbf{v}^{(1)}(\mathbf{a}_t, \mathbf{b}_t)$ approaches the set $\mathcal{V}^{(1)}$, because it passes all tests induced by $\mathcal{U}^{(1)}$. Equivalences between the regret bound (10) and distance to $\mathcal{V}^{(1)}$ are standard, e.g., Lemma 3 of (Abernethy et al., 2011). In Problem 1, we pose a generalization allowing for m approachability instances, each with their own associated set $\mathcal{V}^{(i)}$, distinguishers $\mathcal{U}^{(i)}$, and function $\mathbf{v}^{(i)}$. The goal (10) then asks to approach all m sets simultaneously.

To achieve simultaneous approachability, we assume existence of the following type of oracle.

Definition 11 (Mixture linear optimization oracle) *In the setting of Problem 1, we call $\mathcal{O}$ an ε-*mixture linear optimization oracle *(MLOO) if on inputs $\mathbf{w} \in \Delta^m$, and $\{\mathbf{u}^{(i)}\}_{i \in [m]} \in \prod_{i \in [m]} \mathcal{U}^{(i)}$, the oracle outputs $\mathbf{a} \in \mathcal{A}$ satisfying*

$$\sum_{i \in [m]} \mathbf{w}_i \left\langle \mathbf{u}^{(i)}, \mathbf{v}^{(i)}(\mathbf{a}, \mathbf{b}) \right\rangle \leq \rho + \varepsilon \text{ for all } \mathbf{b} \in \mathcal{B}. \tag{11}$$

We briefly interpret Definition 11 in the finite-dimensional setting. When the input $\mathbf{w}$ is a point mass $\mathbf{e}_i$, the oracle definition is equivalent to finding a distribution $\mathcal{P}$ such that

$$\left\langle \mathbf{u}^{(i)}, \mathbf{v}^{(i)}(\mathbf{a}, \mathbf{b}) \right\rangle \leq \rho + \varepsilon \text{ for all } \mathbf{b} \in \mathcal{B}. \tag{12}$$

Fortunately, we know (12) is achievable (even when $\varepsilon = 0$) whenever $\mathcal{V}^{(i)}$ is approachable, because for any $\mathbf{u}^{(i)} \in \mathcal{U}^{(i)}$, the set $\{\mathbf{v} \mid \left\langle \mathbf{u}^{(i)}, \mathbf{v} \right\rangle \leq \rho\}$ is a halfspace containing $\mathcal{V}^{(i)}$. Proposition 9 shows $\mathcal{V}^{(i)}$ is approachable iff it is halfspace-satisfiable, a.k.a. there exists $\mathbf{a}$ achieving (12).

Remark 12 *There is a natural equivalent formulation of Problem 1 that shed lights on how Definition 11 naturally arises.*[6] *In particular, Problem 1 can be viewed as an instance of (standard) Blackwell approachability in the setting of (8). Consider the family of all tests*

$$(\mathbf{a}, \mathbf{b}) \to \left\langle \mathbf{u}^{(i)}, \mathbf{v}^{(i)}(\mathbf{a}, \mathbf{b}) \right\rangle, \text{ for all } i \in [m],\ \mathbf{u}^{(i)} \in \mathcal{U}^{(i)}.$$

By bilinearity of each $\mathbf{v}^{(i)}$, each such test is also bilinear. Thus the goal of simultaneous Blackwell approachability is simply to approach the scalar set $\{c \mid c \leq \rho\}$ with respect to this family of tests. Each linear test passed by Definition 11 belongs to the convex hull of the tests that parameterize this Blackwell approachability instance.

We next observe that in general, achievability of Definition 11 (and our goal (10)) are strictly stronger requirements than each $\mathcal{V}^{(i)}$ being individually approachable.

6. We thank the anonymous COLT reviewer for raising this comment.

Lemma 13 *There exists a simultaneous Blackwell approachability instance (Problem 1) where each individual set $\mathcal{V}^{(i)}$ is approachable, but a MLOO does not exist, and the simultaneous approachability bound (10) is impossible to achieve, for sufficiently small $\varepsilon > 0$.*

Proof We take $m = 2$, $\rho = 0, \mathcal{U}^{(1)} = \mathcal{U}^{(2)} = \{1\}$, $\mathcal{V}^{(1)} = \mathcal{V}^{(2)} = \{c \in \mathbb{R} \mid c \le 0\}$, $\mathcal{A} := \Delta^2$, and

$$\mathbf{v}^{(1)}(\mathbf{a}, \mathbf{b}) := \mathbf{a}_1, \ \mathbf{v}^{(2)}(\mathbf{a}, \mathbf{b}) := \mathbf{a}_2.$$

Clearly these are bilinear functions, and in fact independent of $\mathbf{b} \in \mathcal{B}$. Moreover, $\mathcal{V}^{(1)}$ and $\mathcal{V}^{(2)}$ are both approachable, the former by repeatedly playing $\mathbf{a}_t \leftarrow \mathbf{e}_2$, and the latter by repeatedly playing $\mathbf{b}_t \leftarrow \mathbf{e}_1$. However, the mixture oracle fails for any $\varepsilon < \frac{1}{2}$ by taking $\mathbf{w} := \frac{1}{2}(\mathbf{e}_1 + \mathbf{e}_2)$, because for any $\mathbf{a} \in \mathcal{A}$ (and any $\mathbf{b} \in \mathcal{B}$), (11) would yield the false statement $\frac{1}{2} = \sum_{i \in [2]} \frac{1}{2} \cdot 1 \cdot \mathbf{a}_i \le \varepsilon$.

For this same instance, regarding the simultaneous approachability bound (10), no matter what choices of $\{\mathbf{a}_t, \mathbf{b}_t\}_{t \in [T]}$ were played, the scalars $\frac{1}{T}\sum_{t \in [T]} \mathbf{v}^{(1)}(\mathbf{a}_t, \mathbf{b}_t)$ and $\frac{1}{T}\sum_{t \in [T]} \mathbf{v}^{(2)}(\mathbf{a}_t, \mathbf{b}_t)$ must sum to 1. Thus, one of the inequalities (10), for $i \in [2]$, must be violated for $\varepsilon < \frac{1}{2}$. ■

To ease applications, we give a unified recipe for constructing MLOOs in the case of supervised multiclass prediction tasks in Appendix D.1, a setting where we show MLOOs always exist.

We are now ready to present the main result of this section. Our result reduces solving Problem 1 to the implementation of an MLOO (Definition 11) and online learners for each $\mathcal{U}^{(i)}$.

Theorem 14 (Simultaneous Blackwell approachability) *In the setting of Problem 1, assume we have access to $\mathcal{O}$, an ε-MLOO. Further, for all $i \in [m]$ and $T \in \mathbb{N}$, assume there is an online learner $\mathsf{alg}^{(i)}$ that takes inputs $(\mathbf{a}_{[T]}, \mathbf{b}_{[T]}) \in \mathcal{A}^T \times \mathcal{B}^T$, and outputs $\mathbf{u}^{(i)}_{[T]}$ such that $\mathbf{u}^{(i)}_t \in \mathcal{U}^{(i)}$ depends only on $\mathbf{a}_{[t-1]}$ and $\mathbf{b}_{[t-1]}$, and*

$$\sup_{\mathbf{u}^{(i)} \in \mathcal{U}^{(i)}} \sum_{t \in [T]} \left\langle \mathbf{v}^{(i)}(\mathbf{a}_t, \mathbf{b}_t), \mathbf{u}^{(i)} - \mathbf{u}^{(i)}_t \right\rangle \le \mathrm{reg}^{(i)}(T), \tag{13}$$

for some $\mathrm{reg}^{(i)} : \mathbb{N} \to \mathbb{R}_{\ge 0}$. Finally, assume

$$\left| \left\langle \mathbf{v}^{(i)}(\mathbf{a}, \mathbf{b}), \mathbf{u}^{(i)} \right\rangle \right| \le L \tag{14}$$

for all $i \in [m]$, $(\mathbf{a}, \mathbf{b}) \in \mathcal{A} \times \mathcal{B}$, and $\mathbf{u}^{(i)} \in \mathcal{U}^{(i)}$. Then, for any $\mathbf{b}_{[T]} \in \mathcal{B}^T$, Algorithm 1 produces $\mathbf{a}_{[T]} \in \mathcal{A}^T$ such that $\mathbf{a}_t$ depends only on $\mathbf{b}_{[t-1]}$, and

$$\sup_{\mathbf{u}^{(i)} \in \mathcal{U}^{(i)}} \left\langle \mathbf{u}^{(i)}, \frac{1}{T} \sum_{t \in [T]} \mathbf{v}^{(i)}(\mathbf{a}_t, \mathbf{b}_t) \right\rangle \le \rho + \varepsilon + \frac{\mathrm{reg}^{(i)}(T) + L\sqrt{2T \log(m)}}{T} \quad \text{for all } i \in [m].$$

We defer the proof of Theorem 14 to Appendix B. We note that Theorem 14 extends to substantially more general settings, including ones where $\mathbf{u}^{(i)}$ is a function parameterized by a context, and we can only access $\mathbf{v}^{(i)}$ through an unbiased estimator. We state these extensions in Problem 2 and Corollary 18, for ease of use in our applications as well as future work.

4. Organization of Appendices

Due to space limitations, we defer the applications of our simultaneous Blackwell approachability framework to binary omniprediction and multiclass omniprediction, respectively, to Appendices C and D. Here, we overview the main results of these sections, and highlight a key technical contribution (Appendix D.1) that allows us to leverage Theorem 14 in our omniprediction applications.

Binary omniprediction. As a warmup, we rederive the main results of (Okoroafor et al., 2025) for binary omniprediction in Appendix C. Using Lemma 5, we recall how to express $\mathcal{F}$-multiaccuracy (Definition 3) and $\mathcal{W}$-calibration (Definition 4) in terms of Blackwell approachability: we use approachibility sets and payoffs given by

$$\mathcal{U}^{(1)} := \Delta^{\mathcal{N}}, \quad \mathbf{v}^{(1)}(\mathbf{a}, b) := \left\{\mathbb{E}_{p\sim\mathbf{a}}\left[(p-b)\operatorname{sign}(p-s)\right]\right\}_{s\in\mathcal{N}},$$
$$\mathcal{U}^{(2)} := \left\{\mathrm{d}_\ell \circ c\right\}_{\ell\in\mathcal{L}, c\in\mathcal{C}}, \quad v^{(2)}(\mathbf{a}, b) := \mathbb{E}_{p\sim\mathbf{a}}\left[p-b\right].$$

Note that our definition of $\mathcal{U}^{(2)}$ above is exactly the set $\mathcal{F}$ in Proposition 7, and our definition of $\mathbf{v}^{(1)}$ implements distinguishers induced by particular elements of the set $\mathcal{W}$ in Proposition 7, which form an approximate basis for $\mathcal{W}$ via Lemma 21.

Importantly, (Okoroafor et al., 2025) gave an explicit construction of an MLOO in the binary setting; combining this oracle with an appropriate calibration algorithm, as well as online learners against comparator classes, then yields our main results. We give an end-to-end efficient omnipredictor for GLMs in Theorem 30, as well as extensions to more general settings in Theorem 35.

Multiclass omniprediction. A challenge in extending (Okoroafor et al., 2025) to the multiclass setting is that their MLOO construction does not appear to generalize to high dimensions, necessitating a new approach. To enable the use of our framework, in Appendix D.1, we give a generic oracle construction for applications of simultaneous approachability to prediction tasks. This construction crucially uses that the vector-valued $\mathbf{v}^{(i)}$ in certifying calibration and multiaccuracy are bounded linear transforms of the same function. We show how to apply the minimax theorem and linear programming to construct an MLOO for any such bounded linear transforms.

Combining this result with known characterizations of multiclass GLMs (i.e., Lemma 6) enables us to prove Theorem 1 (cf. Theorem 43), and we similarly give a general extension in Theorem 46. Notably, our set $\mathcal{U}^{(1)}$ used to certify $\mathcal{W}$-calibration becomes substantially larger compared to its $k = 2$ counterpart, Theorem 30: we take it to be $[-1, 1]^{\mathcal{N}\times k}$ for $\mathcal{N}$ an ε-net of Δ^k in the ℓ_1 norm, whose size leads to our exponential dependence on k in our final results.

We are optimistic about the applicability of our techniques to future multiobjective optimization problems. To give an example of such potential applications, in Appendix E, we overview how to extend our omniprediction results to omnipredict against multiple comparator families, which has previously been left largely undiscussed by the literature. Indeed, very little about our framework changes in this application due to the flexibility of our MLOO construction.

Acknowledgments

Chutong Yang was supported by an Amazon AI Fellowship.

References

Jacob D. Abernethy, Peter L. Bartlett, and Elad Hazan. Blackwell approachability and no-regret learning are equivalent. In *COLT 2011 - The 24th Annual Conference on Learning Theory*, volume 19 of *JMLR Proceedings*, pages 27–46. JMLR.org, 2011.

Shai Ben-David, Nicolo Cesa-Bianchi, and Philip M Long. Characterizations of learnability for classes of {O,..., n}-valued functions. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 333–340, 1992.

Shai Ben-David, Nadav Eiron, and Philip M Long. On the difficulty of approximately maximizing agreements. *Journal of Computer and System Sciences*, 66(3):496–514, 2003.

David Blackwell. An analog of the minimax theorem for vector payoffs. 1956.

Nataly Brukhim, Daniel Carmon, Irit Dinur, Shay Moran, and Amir Yehudayoff. A characterization of multiclass learnability. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 943–955. IEEE, 2022.

Niko Brummer and Johan du Preez. The PAV algorithm optimizes binary proper scoring rules. *arXiv preprint arXiv:1304.2331*, 2013.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

Moses Charikar and Chirag Pabbaraju. A characterization of list learnability. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1713–1726, 2023.

Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the erm principle. *J. Mach. Learn. Res.*, 16(1):2377–2404, 2015.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE signal processing magazine*, 29(6):141–142, 2012.

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

Maxwell Fishelson, Noah Golowich, Mehryar Mohri, and Jon Schneider. High-dimensional calibration from swap regret. *CoRR*, abs/2505.21460, 2025.

Dean P Foster. A proof of calibration via blackwell's approachability theorem. *Games and Economic Behavior*, 29(1-2):73–78, 1999.

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

Gaëtan Fournier, Eden Kuperwasser, Orin Munk, Eilon Solan, and Avishay Weinbaum. Approachability with constraints. *Eur. J. Oper. Res.*, 292(2):687–695, 2021.

Sumegha Garg, Christopher Jung, Omer Reingold, and Aaron Roth. Oracle efficient online multicalibration and omniprediction. In *Proceedings of the 2024 ACM-SIAM Symposium on Discrete Algorithms, SODA 2024*, pages 2725–2792. SIAM, 2024.

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 and Lunjia Hu. Calibration through the lens of indistinguishability. *CoRR*, abs/2509.02279, 2025.

Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In *13th Innovations in Theoretical Computer Science Conference, ITCS 2022*, volume 215 of *LIPIcs*, pages 79:1–79:21. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022.

Parikshit Gopalan, Lunjia Hu, Michael P. Kim, Omer Reingold, and Udi Wieder. Loss minimization through the lens of outcome indistinguishability. In *14th Innovations in Theoretical Computer Science Conference, ITCS 2023*, volume 251 of *LIPIcs*, pages 60:1–60:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2023a.

Parikshit Gopalan, Michael P. Kim, and Omer Reingold. Swap agnostic learning, or characterizing omniprediction via multicalibration. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*, 2023b.

Parikshit Gopalan, Lunjia Hu, and Guy N. Rothblum. On computationally efficient multi-class calibration. In *The Thirty Seventh Annual Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 1983–2026. PMLR, 2024a.

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

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 2017.

Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1944–1953. PMLR, 2018.

Lunjia Hu and Salil Vadhan. Generalized and unified equivalences between hardness and pseudoentropy. In *Theory of Cryptography: 23rd International Conference, TCC 2025, Aarhus, Denmark, December 1–5, 2025, Proceedings, Part IV*, page 258–288, Berlin, Heidelberg, 2025. Springer-Verlag. ISBN 978-3-032-12289-6. doi: 10.1007/978-3-032-12290-2_9. URL https://doi.org/10.1007/978-3-032-12290-2_9.

Lunjia Hu, Inbal Rachel Livni Navon, Omer Reingold, and Chutong Yang. Omnipredictors for constrained optimization. In *International Conference on Machine Learning, ICML 2023*, volume 202 of *Proceedings of Machine Learning Research*, pages 13497–13527. PMLR, 2023.

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

David S. Johnson and Franco P Preparata. The densest hemisphere problem. *Theoretical Computer Science*, 6(1):93–107, 1978.

Adam Tauman Kalai and Ravi Sastry. The isotron algorithm: High-dimensional isotonic regression. In *COLT 2009 - The 22nd Conference on Learning Theory*, 2009.

Bobby Kleinberg, Renato Paes Leme, Jon Schneider, and Yifeng Teng. U-calibration: Forecasting for an unknown agent. In *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023*, volume 195 of *Proceedings of Machine Learning Research*, pages 5143–5145. PMLR, 2023.

Meelis Kull and Peter A. Flach. Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part I*, volume 9284 of *Lecture Notes in Computer Science*, pages 68–85. Springer, 2015.

Meelis Kull, Miquel Perelló-Nieto, Markus Kängsepp, Telmo de Menezes e Silva Filho, Hao Song, and Peter A. Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, pages 12295–12305, 2019.

Daniel Lee, Georgy Noarov, Mallesh M. Pai, and Aaron Roth. Online minimax multiobjective optimization: Multicalibeating and other applications. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, 2022.

Jiuyao Lu, Aaron Roth, and Mirah Shi. Sample efficient omniprediction and downstream swap regret for non-linear losses. In *The Thirty Eighth Annual Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 3829–3878. PMLR, 2025.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150, 2011.

Shie Mannor and Gilles Stoltz. A geometric proof of calibration. *Mathematics of Operations Research*, 35(4):721–727, 2010.

Shie Mannor, Vianney Perchet, and Gilles Stoltz. Approachability in unknown games: Online learning meets multi-objective optimization. In *Proceedings of The 27th Conference on Learning Theory, COLT 2014*, volume 35 of *JMLR Workshop and Conference Proceedings*, pages 339–355. JMLR.org, 2014.

Balas K Natarajan. On learning sets and functions. *Machine Learning*, 4(1):67–97, 1989.

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

Princewill Okoroafor, Robert Kleinberg, and Michael P. Kim. Near-optimal algorithms for omniprediction. *CoRR*, abs/2501.17205, 2025.

Binghui Peng. High dimensional online calibration in polynomial time. *CoRR*, abs/2504.09096, 2025.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning via sequential complexities. *J. Mach. Learn. Res.*, 16:155–186, 2015.

R. Tyrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970a.

Ralph Rockafellar. On the maximal monotonicity of subdifferential mappings. *Pacific Journal of Mathematics*, 33(1):209–216, 1970b.

Ohad Shamir. The sample complexity of learning linear predictors with the squared loss. *J. Mach. Learn. Res.*, 16:3475–3486, 2015.

Jan van den Brand, Yin Tat Lee, Yang P. Liu, Thatchaphol Saranurak, Aaron Sidford, Zhao Song, and Di Wang. Minimum cost flows, mdps, and ℓ_1-regression in nearly linear time for dense instances. In *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 859–869. ACM, 2021.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Shengjia Zhao, Michael P. Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. Calibrating predictions to decisions: A novel approach to multi-class calibration. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 22313–22324, 2021.

Appendix A. Deferred proofs

We prove Lemma 5 here.

Lemma 5 *Let $\ell : \Omega \times \partial\Delta^k$ be a loss function, and define its* discrete derivative $\mathbf{d}_\ell : \Omega \to \mathbb{R}^k$*:*

$$[\mathbf{d}_\ell(\mathbf{t})]_i := \ell(\mathbf{t}, \mathbf{e}_i) \text{ for all } i \in [k]. \tag{4}$$

Then for any $\mathbf{t} \in \Omega$, $\mathbf{p} \in \Delta^k$, and $\mathbf{q} \in \Delta^k$, $\mathbb{E}_{\mathbf{y}\sim\mathbf{p}}[\ell(\mathbf{t},\mathbf{y})] - \mathbb{E}_{\mathbf{y}\sim\mathbf{q}}[\ell(\mathbf{t},\mathbf{y})] = \langle \mathbf{d}_\ell(\mathbf{t}), \mathbf{p}-\mathbf{q}\rangle$. Moreover, the same is true if we redefine $\mathbf{d}_\ell(\mathbf{t}) \leftarrow \mathbf{d}_\ell(\mathbf{t}) - \alpha(\mathbf{t})\mathbf{1}_k$ for any $\alpha(\mathbf{t}) \in \mathbb{R}$.

Proof It suffices to expand definitions, e.g., $\mathbb{E}_{\mathbf{y}\sim\mathbf{p}}[\ell(\mathbf{t},\mathbf{y})] = \langle \mathbf{d}_\ell(\mathbf{t}), \mathbf{p}\rangle$, and $\mathbf{1}_k^\top \mathbf{p} = \mathbf{1}_k^\top \mathbf{q} = 1$. ■

We prove Lemma 10 here.

Lemma 10 *If $\mathcal{V} \subseteq \mathbb{R}^d$ is closed and convex, $\mathbf{0}_d \in \mathcal{V}$, and $\rho > 0$, there exists $\mathcal{U}$ so that (8) holds.*

Proof It suffices to take $\mathcal{U} = \rho\mathcal{V}^\circ$, where $\mathcal{V}^\circ := \{\mathbf{u} \in \mathbb{R}^d \mid \sup_{\mathbf{v}\in\mathcal{V}} \langle \mathbf{u}, \mathbf{v}\rangle \le 1\}$ is the polar set. That $\mathcal{V}^\circ$ satisfies (8) with $\rho = 1$ is by the well-known fact $\mathcal{V}^{\circ\circ} = \mathcal{V}$ (see e.g., Theorem 14.5, (Rockafellar, 1970a)). ■

We prove Proposition 7 here.

Proposition 7 *Following notation from Definition 2 and (4), in the online setting, if $\mathbf{p}_{[T]}$ satisfies ε_1-$(\mathbf{x}_{[T]}, \mathbf{y}_{[T]}, \mathcal{F})$-multiaccuracy and ε_2-$(\mathbf{y}_{[T]}, \mathcal{W})$-calibration for*

$$\mathcal{F} := \{\mathbf{d}_\ell \circ \mathbf{c}\}_{\ell\in\mathcal{L}, \mathbf{c}\in\mathcal{C}}, \ \mathcal{W} := \{-\mathbf{d}_\ell \circ \mathbf{k}_\ell^\star\}_{\ell\in\mathcal{L}},$$

then $\mathbf{p}_{[T]}$ is an $(\varepsilon_1+\varepsilon_2)$-omnipredictor for $(\mathbf{x}_{[T]}, \mathbf{y}_{[T]}, \mathcal{L}, \mathcal{C})$.

In the statistical setting, if $\mathbf{p} : \mathbb{R}^d \to \Delta^k$ satisfies ε_1-$(\mathcal{D}, \mathcal{F})$-multiaccuracy and ε_2-$(\mathcal{D}, \mathcal{W})$-calibration, then $\mathbf{p}$ is an $(\varepsilon_1+\varepsilon_2)$-omnipredictor for $(\mathcal{D}, \mathcal{L}, \mathcal{C})$.

Proof We begin with the statistical setting. By definition of $\mathbf{k}_\ell^\star$, for all $\ell \in \mathcal{L}$ and $\mathbf{c} \in \mathcal{C}$,

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\mathbb{E}_{\mathbf{y}\sim\mathbf{p}(\mathbf{x})}\left[\ell\left(\mathbf{k}_\ell^\star(\mathbf{p}(\mathbf{x})), \mathbf{y}\right)\right]\right] \le \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\mathbb{E}_{\mathbf{y}\sim\mathbf{p}(\mathbf{x})}\left[\ell\left(\mathbf{c}(\mathbf{x}), \mathbf{y}\right)\right]\right], \tag{15}$$

and thus

$$\begin{aligned}
\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[\ell\left(\mathbf{k}_\ell^\star(\mathbf{p}(\mathbf{x})), \mathbf{y}\right)\right] &= \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[\ell\left(\mathbf{k}_\ell^\star(\mathbf{p}(\mathbf{x})), \mathbf{y}\right)\right] - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\mathbb{E}_{\mathbf{y}\sim\mathbf{p}(\mathbf{x})}\left[\ell\left(\mathbf{k}_\ell^\star(\mathbf{p}(\mathbf{x})), \mathbf{y}\right)\right]\right] \\
&\quad + \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\mathbb{E}_{\mathbf{y}\sim\mathbf{p}(\mathbf{x})}\left[\ell\left(\mathbf{k}_\ell^\star(\mathbf{p}(\mathbf{x})), \mathbf{y}\right)\right]\right] - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\mathbb{E}_{\mathbf{y}\sim\mathbf{p}(\mathbf{x})}\left[\ell\left(\mathbf{c}(\mathbf{x}), \mathbf{y}\right)\right]\right] \\
&\quad + \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\mathbb{E}_{\mathbf{y}\sim\mathbf{p}(\mathbf{x})}\left[\ell\left(\mathbf{c}(\mathbf{x}), \mathbf{y}\right)\right]\right] - \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[\ell\left(\mathbf{c}(\mathbf{x}), \mathbf{y}\right)\right] \\
&\quad + \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[\ell\left(\mathbf{c}(\mathbf{x}), \mathbf{y}\right)\right] \\
&\le \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[\ell\left(\mathbf{k}_\ell^\star(\mathbf{p}(\mathbf{x})), \mathbf{y}\right)\right] - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\mathbb{E}_{\mathbf{y}\sim\mathbf{p}(\mathbf{x})}\left[\ell\left(\mathbf{k}_\ell^\star(\mathbf{p}(\mathbf{x})), \mathbf{y}\right)\right]\right] \\
&\quad + \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\mathbb{E}_{\mathbf{y}\sim\mathbf{p}(\mathbf{x})}\left[\ell\left(\mathbf{c}(\mathbf{x}), \mathbf{y}\right)\right]\right] - \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[\ell\left(\mathbf{c}(\mathbf{x}), \mathbf{y}\right)\right] \\
&\quad + \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[\ell\left(\mathbf{c}(\mathbf{x}), \mathbf{y}\right)\right],
\end{aligned} \tag{16}$$

where the second line in (16) was bounded by (15). Taking an expectation of Lemma 5 over $\mathbf{x} \sim \mathcal{D}_\mathbf{x}$,

$$\begin{aligned}
\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[\ell\left(\mathbf{k}_\ell^\star(\mathbf{p}(\mathbf{x})), \mathbf{y}\right)\right] - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\mathbb{E}_{\mathbf{y}\sim\mathbf{p}(\mathbf{x})}\left[\ell\left(\mathbf{k}_\ell^\star(\mathbf{p}(\mathbf{x})), \mathbf{y}\right)\right]\right] &= \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[\langle \mathbf{d}_\ell(\mathbf{k}_\ell^\star(\mathbf{p}(\mathbf{x}))), \mathbf{y} - \mathbf{p}(\mathbf{x})\rangle\right], \\
\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\mathbb{E}_{\mathbf{y}\sim\mathbf{p}(\mathbf{x})}\left[\ell\left(\mathbf{c}(\mathbf{x}), \mathbf{y}\right)\right]\right] - \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[\ell\left(\mathbf{c}(\mathbf{x}), \mathbf{y}\right)\right] &= \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[\langle \mathbf{d}_\ell(\mathbf{c}(\mathbf{x})), \mathbf{p}(\mathbf{x}) - \mathbf{y}\rangle\right],
\end{aligned}$$

and the conclusion follows by applying Definitions 3 and 4. The online setting is similar:

$$\begin{aligned}\frac{1}{T}\sum_{t\in[T]}\ell(\mathbf{k}_\ell^\star(\mathbf{p}_t),\mathbf{y}_t) &= \frac{1}{T}\sum_{t\in[T]}\left(\ell(\mathbf{k}_\ell^\star(\mathbf{p}_t),\mathbf{y}_t) - \mathbb{E}_{\mathbf{y}\sim\mathbf{p}_t}\left[\ell(\mathbf{k}_\ell^\star(\mathbf{p}_t),\mathbf{y})\right]\right)\\ &+ \frac{1}{T}\sum_{t\in[T]}\left(\mathbb{E}_{\mathbf{y}\sim\mathbf{p}_t}\left[\ell(\mathbf{k}_\ell^\star(\mathbf{p}_t),\mathbf{y})\right] - \mathbb{E}_{\mathbf{y}\sim\mathbf{p}_t}\left[\ell(\mathbf{c}(\mathbf{x}_t),\mathbf{y})\right]\right)\\ &+ \frac{1}{T}\sum_{t\in[T]}\left(\mathbb{E}_{\mathbf{y}\sim\mathbf{p}_t}\left[\ell(\mathbf{c}(\mathbf{x}_t),\mathbf{y})\right] - \ell(\mathbf{c}(\mathbf{x}_t),\mathbf{y}_t)\right) + \frac{1}{T}\sum_{t\in[T]}\ell(\mathbf{c}(\mathbf{x}_t),\mathbf{y}_t)\end{aligned}$$

at which point the conclusion again follows from Definitions 3 and 4, because for all $t\in[T]$,

$$\begin{aligned}\ell(\mathbf{k}_\ell^\star(\mathbf{p}_t),\mathbf{y}_t) - \mathbb{E}_{\mathbf{y}\sim\mathbf{p}_t}\left[\ell(\mathbf{k}_\ell^\star(\mathbf{p}_t),\mathbf{y})\right] &= \left\langle \mathbf{d}_\ell(\mathbf{k}_\ell^\star(\mathbf{p}_t)), \mathbf{y}_t - \mathbf{p}_t\right\rangle,\\ \mathbb{E}_{\mathbf{y}\sim\mathbf{p}_t}\left[\ell(\mathbf{c}(\mathbf{x}_t),\mathbf{y})\right] - \ell(\mathbf{c}(\mathbf{x}_t),\mathbf{y}_t) &= \left\langle \mathbf{d}_\ell(\mathbf{c}(\mathbf{x}_t)), \mathbf{p}_t - \mathbf{y}_t\right\rangle.\end{aligned}$$

■

Appendix B. Proofs for Simultaneous Blackwell Approachability

Before we dive into our proof of Theorem 14, to instantiate our framework, we require various online learning algorithms from the literature. We begin by stating a general fact about the generalization of stochastic mirror descent methods.[7]

Lemma 15 (Lemma 9, (Hu et al., 2025)) *Let $T\in\mathbb{N}$, $\eta>0$, $\delta\in(0,1)$, let $\mathcal{X}\subseteq\mathbb{R}^d$ have diameter $\le R$ in $\|\cdot\|$. Let $r:\mathcal{X}\to\mathbb{R}$ be 1-strongly convex in a norm $\|\cdot\|$ such that $\max_{\mathbf{x}\in\mathcal{X}} r(\mathbf{x}) - \min_{\mathbf{x}\in\mathcal{X}} r(\mathbf{x}) \le \Theta$, and let $\mathbf{x}_1 := \arg\min_{\mathbf{x}\in\mathcal{X}} r(\mathbf{x})$. For a sequence of deterministic vectors $\{\mathbf{g}_t\}_{t\in[T]}$ such that $\mathbf{g}_t$ can depend on all randomness used in iterations $t\in[T]$, let*

$$\mathbf{x}_{t+1}\leftarrow\arg\min_{\mathbf{x}\in\mathcal{X}}\left\{\langle\eta\tilde{\mathbf{g}}_t - \nabla r(\mathbf{x}_t),\mathbf{x}\rangle + r(\mathbf{x})\right\},\ \textit{where}\ \mathbb{E}\left[\tilde{\mathbf{g}}_t\mid\tilde{\mathbf{g}}_1,\ldots,\tilde{\mathbf{g}}_{t-1}\right]=\mathbf{g}_t,\ \textit{for all}\ t\in[T].$$

Further suppose $\|\tilde{\mathbf{g}}_t\|_*\le L$ deterministically. Then for some choice of η, with probability $\ge 1-\delta$,

$$\sup_{\mathbf{x}\in\mathcal{X}}\sum_{t\in[T]}\langle\mathbf{g}_t,\mathbf{x}_t-\mathbf{x}\rangle \le 4L\sqrt{T\Theta} + 16LR\sqrt{T\log\left(\frac{2}{\delta}\right)}.$$

We will apply two specializations of Lemma 15, where r is either a Euclidean regularizer (projected gradient descent), or negative entropy over the probability simplex (multiplicative weights). Finally, for deterministic applications of Lemma 15, we state the following sharper bound.

Lemma 16 (Theorem 4.2, (Bubeck, 2015)) *In the setting of Lemma 15, if $\tilde{\mathbf{g}}_t=\mathbf{g}_t$ in every iteration,*

$$\sup_{\mathbf{x}\in\mathcal{X}}\sum_{t\in[T]}\langle\mathbf{g}_t,\mathbf{x}_t-\mathbf{x}\rangle \le L\sqrt{2T\Theta}.$$

7. Lemma 9 of (Hu et al., 2025) only claims this result for an ℓ_2 setup, but the same argument extends to all stochastic mirror descent setups as the result simply bounds the random error term via martingale concentration.

Now we are ready to prove Theorem 14 here.

Theorem 14 (Simultaneous Blackwell approachability) *In the setting of Problem 1, assume we have access to $\mathcal{O}$, an ε-MLOO. Further, for all $i \in [m]$ and $T \in \mathbb{N}$, assume there is an online learner $\mathsf{alg}^{(i)}$ that takes inputs $(\mathbf{a}_{[T]}, \mathbf{b}_{[T]}) \in \mathcal{A}^T \times \mathcal{B}^T$, and outputs $\mathbf{u}^{(i)}_{[T]}$ such that $\mathbf{u}^{(i)}_t \in \mathcal{U}^{(i)}$ depends only on $\mathbf{a}_{[t-1]}$ and $\mathbf{b}_{[t-1]}$, and*

$$\sup_{\mathbf{u}^{(i)} \in \mathcal{U}^{(i)}} \sum_{t \in [T]} \left\langle \mathbf{v}^{(i)}(\mathbf{a}_t, \mathbf{b}_t), \mathbf{u}^{(i)} - \mathbf{u}^{(i)}_t \right\rangle \le \mathrm{reg}^{(i)}(T), \tag{13}$$

for some $\mathrm{reg}^{(i)} : \mathbb{N} \to \mathbb{R}_{\ge 0}$. Finally, assume

$$\left| \left\langle \mathbf{v}^{(i)}(\mathbf{a}, \mathbf{b}), \mathbf{u}^{(i)} \right\rangle \right| \le L \tag{14}$$

for all $i \in [m]$, $(\mathbf{a}, \mathbf{b}) \in \mathcal{A} \times \mathcal{B}$, and $\mathbf{u}^{(i)} \in \mathcal{U}^{(i)}$. Then, for any $\mathbf{b}_{[T]} \in \mathcal{B}^T$, Algorithm 1 produces $\mathbf{a}_{[T]} \in \mathcal{A}^T$ such that $\mathbf{a}_t$ depends only on $\mathbf{b}_{[t-1]}$, and

$$\sup_{\mathbf{u}^{(i)} \in \mathcal{U}^{(i)}} \left\langle \mathbf{u}^{(i)}, \frac{1}{T} \sum_{t \in [T]} \mathbf{v}^{(i)}(\mathbf{a}_t, \mathbf{b}_t) \right\rangle \le \rho + \varepsilon + \frac{\mathrm{reg}^{(i)}(T) + L\sqrt{2T \log(m)}}{T} \quad \text{for all } i \in [m].$$

Proof The algorithm is presented in Algorithm 1, and we follow the notation therein throughout. By observation, the algorithm computes $\mathbf{a}_t$ on Line 2 before observing $\mathbf{b}_t$. We first show

$$\max_{i \in [m]} \frac{1}{T} \sum_{t \in [T]} \left\langle \mathbf{u}^{(i)}_t, \mathbf{v}^{(i)}_t \right\rangle = \sup_{\mathbf{w} \in \Delta^m} \frac{1}{T} \sum_{t \in [T]} \langle \mathbf{w}, \mathbf{g}_t \rangle \le \rho + \varepsilon + \frac{L\sqrt{2T \log(m)}}{T}. \tag{17}$$

Indeed, this follows from

$$\begin{aligned} \sup_{\mathbf{w} \in \Delta^m} \frac{1}{T} \sum_{t \in [T]} \langle \mathbf{w}, \mathbf{g}_t \rangle &= \frac{1}{T} \sum_{t \in [T]} \langle \mathbf{w}_t, \mathbf{g}_t \rangle + \sup_{\mathbf{w} \in \Delta^m} \frac{1}{T} \sum_{t \in [T]} \langle \mathbf{w} - \mathbf{w}_t, \mathbf{g}_t \rangle \\ &\le \rho + \varepsilon + \frac{L\sqrt{2T \log(m)}}{T}. \end{aligned} \tag{18}$$

Here, we bounded the first term in the right-hand side of the first line by using the oracle guarantee (11), which holds for any $\mathbf{b}_t$ used to define each $\mathbf{g}_t$. Moreover, to bound the second term, we observe that Lines 2 and 2 are implementing multiplicative weight updates, i.e., Lemma 16 with $r(\mathbf{w}) = \sum_{i \in [m]} \mathbf{w}_i \log \mathbf{w}_i$, using feedback vectors $\mathbf{g}_t$ satisfying $\|\mathbf{g}_t\|_\infty \le L$ by assumption (14).

Next we complete the proof: given (17) and the regret bound (13), for all $i \in [m]$,

$$\begin{aligned} \sup_{\mathbf{u}^{(i)} \in \mathcal{U}^{(i)}} \left\langle \mathbf{u}^{(i)}, \frac{1}{T} \sum_{t \in [T]} \mathbf{v}^{(i)}_t \right\rangle &\le \frac{1}{T} \sum_{t \in [T]} \left\langle \mathbf{u}^{(i)}_t, \mathbf{v}^{(i)}_t \right\rangle + \sup_{\mathbf{u}^{(i)} \in \mathcal{U}^{(i)}} \frac{1}{T} \sum_{t \in [T]} \left\langle \mathbf{u}^{(i)} - \mathbf{u}^{(i)}_t, \mathbf{v}^{(i)}_t \right\rangle \\ &\le \rho + \varepsilon + \frac{\mathrm{reg}^{(i)}(T) + L\sqrt{2T \log(m)}}{T}. \end{aligned}$$

■

Algorithm 1: SimultaneousApproach($\mathbf{b}_{[T]}, \{\mathsf{alg}^{(i)}\}_{i\in[m]}, \mathcal{O}$)

Input: Online sequence $\mathbf{b}_{[T]} \in \mathcal{B}^T$, online learners $\{\mathsf{alg}^{(i)}\}_{i\in[m]}$ satisfying (13), ε-MLOO $\mathcal{O}$ (following notation in Problem 1, Definition 11)

Output: $\mathbf{a}_{[T]} \in \mathcal{A}^T$ such that each $\mathbf{a}_t$ is output before observing $\mathbf{b}_t$

$\mathbf{w}_1 \leftarrow \frac{1}{m}\mathbf{1}_m$

$\mathbf{u}_1^{(i)} \leftarrow \mathsf{alg}^{(i)}(\{\})$ for all $i \in [m]$

`// Initialize each` $\mathbf{u}_1^{(i)}$ `as` $\mathsf{alg}^{(i)}$ `does before observing any examples.`

$\mathbf{a}_1 \leftarrow \mathcal{O}(\mathbf{w}_1, \{\mathbf{u}_1^{(i)}\}_{i\in[m]})$

$\eta \leftarrow \frac{1}{L} \cdot \sqrt{2\log(m)} \cdot T^{-1/2}$

for $2 \le t \le T$ **do**

- $\mathbf{v}_{t-1}^{(i)} \leftarrow \mathbf{v}^{(i)}(\mathbf{a}_{t-1}, \mathbf{b}_{t-1})$ for all $i \in [m]$
- $\mathbf{g}_{t-1} \leftarrow$ vector in $\mathbb{R}^m$ such that $[\mathbf{g}_{t-1}]_i = \langle \mathbf{u}_{t-1}^{(i)}, \mathbf{v}_{t-1}^{(i)} \rangle$ for all $i \in [m]$
- $\mathbf{u}_t^{(i)} \leftarrow \mathsf{alg}^{(i)}(\mathbf{v}_{[t-1]}^{(i)})$ for all $i \in [m]$
- $\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} \circ \exp(\eta \mathbf{g}_{t-1})$

  `// ∘ denotes entrywise multiplication and exp is applied entrywise.`
- $\mathbf{w}_t \leftarrow \mathbf{w}_t \|\mathbf{w}_t\|_1^{-1}$
- $\mathbf{a}_t \leftarrow \mathcal{O}(\mathbf{w}_t, \{\mathbf{u}_t^{(i)}\}_{i\in[m]})$

end

Return: $\mathbf{a}_{[T]}$

Theorem 14 in fact holds in a much more general setting than captured by simultaneous Blackwell approachability (Problem 1). Indeed, both the statement of Theorem 14 and the MLOO definition do not explicitly specify sets $\mathcal{V}^{(i)}$ distinguished by the corresponding $\mathcal{U}^{(i)}$ in the sense of (9). For our applications to online omniprediction, we require a more general version of Theorem 14, where each $\mathcal{U}^{(i)}$ consists of functions $\mathbf{u}^{(i)} : \mathcal{X} \to \mathcal{H}^{(i)}$, taking as input a context $\mathbf{x}$ from some domain $\mathcal{X}$. For these uses, we generalize Problem 1 and Definition 11, and give an analog of Theorem 14.

Problem 2 (Contextual Blackwell approachability) *Let $a, b, m \in \mathbb{N}$, let $\varepsilon \geq 0$, let $\mathcal{A}$ and $\mathcal{B}$ be subsets of vector spaces, and let $\mathcal{X}$ be an abstract domain of contexts. For all $i \in [m]$, let $\mathcal{U}^{(i)}$ consist of functions $\mathbf{u}^{(i)} : \mathcal{X} \to \mathcal{H}^{(i)}$, where $\mathcal{H}^{(i)}$ is a Hilbert space. Let $\mathbf{v}^{(i)} : \mathcal{A} \times \mathcal{B} \to \mathcal{H}^{(i)}$ be a bilinear, vector-valued function for all $i \in [m]$. Our goal is to observe sequences $\mathbf{b}_{[T]} \in \mathcal{B}^T$ and $\mathbf{x}_{[T]} \in \mathcal{X}^T$, and to choose $\mathbf{a}_{[T]}$ so that $\mathbf{a}_t$ depends on $\mathbf{b}_{[t-1]}$ and $\mathbf{x}_{[t]}$ for all $t \in [T]$, and*

$$\max_{i \in [m]} \sup_{\mathbf{u}^{(i)} \in \mathcal{U}^{(i)}} \frac{1}{T} \sum_{t \in [T]} \left\langle \mathbf{u}^{(i)}(\mathbf{x}_t), \mathbf{v}^{(i)}(\mathbf{a}_t, \mathbf{b}_t) \right\rangle \leq \varepsilon.$$

When each $\mathcal{U}^{(i)}$ consists only of constant functions, i.e., each $\mathbf{u}^{(i)} \in \mathcal{U}^{(i)}$ can only take on one value, we abuse notation and let $\mathcal{U}^{(i)} \subseteq \mathcal{H}^{(i)}$ represent a set of $\mathbf{u}^{(i)} \in \mathcal{H}^{(i)}$.

Definition 17 (Contextual mixture linear optimization oracle) *In the setting of Problem 2, we call $\mathcal{O}$ an ε-*contextual mixture linear optimization oracle *(CMLOO) if on inputs $\mathbf{w} \in \Delta^m$, $\mathbf{x} \in \mathcal{X}$, and $\{\mathbf{u}^{(i)}\}_{i \in [m]} \in \prod_{i \in [m]} \mathcal{U}^{(i)}$, the oracle outputs $\mathbf{a} \in \mathcal{A}$ satisfying*

$$\sum_{i \in [m]} \mathbf{w}_i \left\langle \mathbf{u}^{(i)}(\mathbf{x}), \mathbf{v}^{(i)}(\mathbf{a}, \mathbf{b}) \right\rangle \leq \varepsilon \text{ for all } \mathbf{b} \in \mathcal{B}. \tag{19}$$

We now state our extension of Theorem 14.

In our statement, we allow for unbiased estimators of the CMLOO outputs to be played, and give a high-probability guarantee on the error. We also allow for improper learners that satisfy (20), but output hypotheses from an (appropriately bounded) different set than $\mathcal{U}^{(i)}$.

Corollary 18 *In the setting of Problem 2, assume we have access to $\mathcal{O}$, an ε-CMLOO. Further, for all $i \in [m]$ and $T \in \mathbb{N}$, assume there is an online learner $\mathsf{alg}^{(i)}$ that takes inputs $(\mathbf{a}_{[T]}, \mathbf{b}_{[T]}, \mathbf{x}_{[T]}) \in \mathcal{A}^T \times \mathcal{B}^T \times \mathcal{X}^T$, and outputs $\mathbf{u}^{(i)}_{[T]} \in ((\mathcal{U}')^{(i)})^T$ such that $\mathbf{u}^{(i)}_t$ depends only on $\mathbf{a}_{[t-1]}$, $\mathbf{b}_{[t-1]}$, and $\mathbf{x}_{[t]}$, and*

$$\sup_{\mathbf{u}^{(i)} \in \mathcal{U}^{(i)}} \sum_{t \in [T]} \left\langle \mathbf{v}^{(i)}(\mathbf{a}_t, \mathbf{b}_t), \mathbf{u}^{(i)}(\mathbf{x}_t) - \mathbf{u}^{(i)}_t(\mathbf{x}_t) \right\rangle \leq \mathrm{reg}^{(i)}(T), \tag{20}$$

for some $\mathrm{reg}^{(i)} : \mathbb{N} \to \mathbb{R}_{\geq 0}$. Finally, assume

$$\left| \left\langle \mathbf{v}^{(i)}(\mathbf{a}, \mathbf{b}), \mathbf{u}^{(i)}(\mathbf{x}) \right\rangle \right| \leq L \tag{21}$$

for all $i \in [m]$, $(\mathbf{a}, \mathbf{b}, \mathbf{x}) \in \mathcal{A} \times \mathcal{B} \times \mathcal{X}$, and $\mathbf{u}^{(i)} \in \mathcal{U}^{(i)} \cup (\mathcal{U}')^{(i)}$. Then, for any $(\mathbf{b}_{[T]}, \mathbf{x}_{[T]}) \in \mathcal{B}^T \times \mathcal{X}^T$, Algorithm 2 produces $\mathbf{p}_{[T]} \in \mathcal{A}^T$ such that $\mathbf{p}_t$ depends only on $\mathbf{b}_{[t-1]}$ and $\mathbf{x}_{[t]}$, and for

Algorithm 2: ContextualSimultaneousApproach($\mathbf{b}_{[T]}, \mathbf{x}_{[T]}, \{\mathsf{alg}^{(i)}\}_{i\in[m]}, \mathcal{O}$)

Input: Online sequences $\mathbf{b}_{[T]} \in \mathcal{B}^T$, $\mathbf{x}_{[T]} \in \mathcal{X}^T$, online learners $\{\mathsf{alg}^{(i)}\}_{i\in[m]}$ satisfying (20), ε-CMLOO $\mathcal{O}$ (following notation in Problem 2, Definition 17)
Output: $\mathbf{a}_{[T]} \in \mathcal{A}^T$ such that each $\mathbf{a}_t$ is output after observing $\mathbf{x}_t$ and before observing $\mathbf{b}_t$
$\mathbf{w}_1 \leftarrow \frac{1}{m}\mathbf{1}_m$
$\mathbf{u}_1^{(i)} \leftarrow \mathsf{alg}^{(i)}(\{\})$ for all $i \in [m]$
$\mathbf{a}_1 \leftarrow \mathcal{O}(\mathbf{w}_1, \{\mathbf{u}_1^{(i)}\}_{i\in[m]})$
$\eta \leftarrow \frac{1}{L} \cdot \sqrt{2\log(m)} \cdot (5T)^{-1/2}$
for $2 \le t \le T$ **do**
  $\mathbf{v}_{t-1}^{(i)} \leftarrow \mathbf{v}^{(i)}(\mathbf{p}_{t-1}, \mathbf{b}_{t-1})$ for all $i \in [m]$
  $\tilde{\mathbf{g}}_{t-1} \leftarrow$ vector in $\mathbb{R}^m$ such that $[\tilde{\mathbf{g}}_{t-1}]_i = \langle \mathbf{u}_{t-1}^{(i)}(\mathbf{x}_{t-1}), \mathbf{v}_{t-1}^{(i)} \rangle$ for all $i \in [m]$
  $\mathbf{u}_t^{(i)} \leftarrow \mathsf{alg}^{(i)}(\mathbf{v}_{[t-1]}^{(i)})$ for all $i \in [m]$
  $\mathbf{w}_t \leftarrow \mathbf{w}_{t-1} \circ \exp(\eta \tilde{\mathbf{g}}_{t-1})$
  $\mathbf{w}_t \leftarrow \mathbf{w}_t \|\mathbf{w}_t\|_1^{-1}$
  $\mathbf{a}_t \leftarrow \mathcal{O}(\mathbf{w}_t, \mathbf{x}_t, \{\mathbf{u}_t^{(i)}\}_{i\in[m]})$
  $\mathbf{p}_t \leftarrow$ any random element of $\mathcal{A}$ such that $\mathbb{E}[\mathbf{p}_t \mid \mathbf{a}_{[t-1]}, \mathbf{b}_{[t-1]}] = \mathbf{a}_t$
end
Return: $\mathbf{p}_{[T]}$

any $\delta \in (0,1)$,

$$\sup_{\mathbf{u}^{(i)}\in\mathcal{U}^{(i)}} \frac{1}{T}\sum_{t\in[T]} \left\langle \mathbf{u}^{(i)}(\mathbf{x}_t), \mathbf{v}^{(i)}(\mathbf{p}_t, \mathbf{b}_t) \right\rangle \le \varepsilon + \frac{\mathrm{reg}^{(i)}(T) + 28L\sqrt{T\log(\frac{4m}{\delta})}}{T} \text{ for all } i \in [m], \tag{22}$$

with probability at least $1-\delta$ *over the randomness of the* $\mathbf{p}_{[T]}$.

Proof The proof is largely analogous to Theorem 14, substituting the contexts $\mathbf{x}_{[T]}$ as necessary. The key difference is in (18), which no longer holds deterministically. We instead apply the variant in Lemma 15 with the $\tilde{\mathbf{g}}_{[T]}$ as defined in Algorithm 2. Note that this is unbiased for $\mathbf{g}_t$ with entries $\langle \mathbf{u}_t^{(i)}, \mathbf{v}^{(i)}(\mathbf{a}_t, \mathbf{b}_t)\rangle$, by linearity of each $\mathbf{v}^{(i)}$ in its first argument. Thus, in place of (18),

$$\begin{aligned}\sup_{\mathbf{w}\in\Delta^m} \frac{1}{T}\sum_{t\in[T]} \langle \mathbf{w}, \tilde{\mathbf{g}}_t\rangle &= \frac{1}{T}\sum_{t\in[T]} \langle \mathbf{w}_t, \mathbf{g}_t\rangle + \frac{1}{T}\sum_{t\in[T]} \langle \mathbf{w}_t, \tilde{\mathbf{g}}_t - \mathbf{g}_t\rangle + \sup_{\mathbf{w}\in\Delta^m} \frac{1}{T}\sum_{t\in[T]} \langle \mathbf{w} - \mathbf{w}_t, \tilde{\mathbf{g}}_t\rangle \\ &\le \varepsilon + \frac{1}{T}\sum_{t\in[T]} \langle \mathbf{w}_t, \tilde{\mathbf{g}}_t - \mathbf{g}_t\rangle + 20L\sqrt{\frac{\log(\frac{4m}{\delta})}{T}} \le \varepsilon + 28L\sqrt{\frac{\log(\frac{4m}{\delta})}{T}},\end{aligned}$$

with probability $\ge 1-\delta$. Here, the first inequality used the CMLOO guarantee to bound $\langle \mathbf{w}_t, \mathbf{g}_t\rangle \le \varepsilon$ for all $t \in [T]$, as well as Lemma 15 with failure probability $\frac{\delta}{2}$ to bound the last term. The second inequality used the Azuma-Hoeffding inequality with the fact that each $\langle \mathbf{w}_t, \tilde{\mathbf{g}}_t - \mathbf{g}_t\rangle$ is

mean-zero conditioned on the history, and bounded in $[-2L, 2L]$. We remark that the setting of η in Algorithm 2 is for Lemma 15 to hold; see Lemma 9, (Hu et al., 2025) for additional discussion. ■

Appendix C. Binary Omniprediction

As a warmup, we consider the binary omniprediction setting, where we make predictions over $k = 2$ classes. We state some general preliminaries in Appendix C.1. In Appendices C.2 and C.3 we apply our framework from Section 3 to reduce online and statistical omniprediction, respectively, to appropriate low-regret learners. We complete our binary omniprediction results by providing these low-regret learners in the linear (Appendix C.4) and general (Appendix C.5) classification settings.

For notational simplicity in this section only, we identify the binary simplex Δ^2 with the prediction interval $[0, 1]$ and the boundary $\partial\Delta^2$ with the label set $\{0, 1\}$, so e.g., in place of (5),

$$\mathcal{L}_{\text{GLM}} := \left\{\ell : \mathbb{R} \times \{0, 1\} \to \mathbb{R} \mid \ell(t, y) = \omega(t) - ty,\ \omega : \mathbb{R} \to \mathbb{R} \text{ convex with } \omega' : \mathbb{R} \to [0, 1]\right\}. \tag{23}$$

We will also use standard script (rather than boldface) to denote any scalar-valued variables. Finally, throughout the section we make the normalization assumptions that

$$\begin{aligned} \mathcal{X} \subseteq \mathbb{B}_2^d(1), \quad c(\mathbf{x}) \in [-1, 1] \text{ for all } c \in \mathcal{C},\ \mathbf{x} \in \mathcal{X}, \\ \ell(t, y) \in [-1, 1] \text{ for all } \ell \in \mathcal{L},\ (t, y) \in [-1, 1] \times \{0, 1\}. \end{aligned} \tag{24}$$

All our results generalize to generic bounds on $\mathcal{L}$ and $\mathcal{C}$ in a scale-invariant way. For example, the assumption (24) is enforced for $\mathcal{L}_{\text{GLM}}$ by requiring that $\omega'(t) \in [0, 1]$ for all $t \in [-1, 1]$.

C.1. Binary omniprediction preliminaries

We recall two useful characterizations of proper losses from prior work.

Lemma 19 (Lemma 3, (Kleinberg et al., 2023)) *Let $\ell : \Omega \times \partial\Delta^k \to \mathbb{R}$ be arbitrary. Then defining $\mathbf{k}_\ell^\star$ as in (3), the function $(\mathbf{p}, \mathbf{y}) \to \ell(k_\ell^\star(\mathbf{p}), \mathbf{y})$ is a proper loss.*

Lemma 20 (Theorem 8, (Kleinberg et al., 2023)) *For all $s \in [0, 1]$ and $(p, y) \in [0, 1]^2$, let*

$$\ell_s(p, y) := -|p - v| + (p - y)\text{sign}(p - s). \tag{25}$$

Then ℓ_s is proper for all $s \in [0, 1]$, and for every bounded proper loss $\ell : [0, 1] \times \{0, 1\} \to [-1, 1]$,

$$\ell(p, y) - \int w_\ell(v)\ell_v(p, y)\mathrm{d}v$$

is a linear function in y, for some nonnegative weights $\{w_\ell(v)\}_{v\in[0,1]}$ satisfying $\int_0^1 w_\ell(v)\mathrm{d}v \le 2$.

By combining Lemmas 19 and 20, (Okoroafor et al., 2025) showed that obtaining calibration against the family of weights $\{-d_\ell \circ k_\ell^\star\}_{\ell\in\mathcal{L}}$, as required by Proposition 7, can be reduced to calibration against an appropriate basis of weights induced by the specific proper losses in (25).

Lemma 21 *Let $\mathcal{L}$ be a family of losses over $\Omega \times \{0,1\}$ such that $\ell(\omega, y) \in [-1,1]$ for all $\ell \in \mathcal{L}$, $\omega \in \Omega$, $y \in \{0,1\}$. In the statistical setting, if $p : \mathbb{R}^d \to [0,1]$ satisfies ε-$(\mathcal{D}, \mathcal{W}_{\text{thresh}})$-calibration for*

$$\mathcal{W}_{\text{thresh}} := \{w(p) = \text{sign}(p - s)\}_{s \in [0,1]}, \tag{26}$$

it also satisfies 2ε-$(\mathcal{D}, \{-\text{d}_\ell \circ k_\ell^\star\}_{\ell \in \mathcal{L}})$-calibration. In the online setting, if $p_{[T]} \in [0,1]^T$ satisfies ε-$(y_{[T]}, \mathcal{W}_{\text{thresh}})$-calibration, it also satisfies 2ε-$(y_{[T]}, \{-\text{d}_\ell \circ k_\ell^\star\}_{\ell \in \mathcal{L}})$-calibration.

C.2. Online binary omniprediction

We first consider the online binary omniprediction setting, i.e., Definition 2 with $k = 2$, for a family of losses $\mathcal{L}$ and a family of comparators $\mathcal{C}$. Throughout, let $\mathcal{X} \subseteq \mathbb{R}^d$ be the support of our features.

Our starting point is an observation from (Okoroafor et al., 2025) that the (online variants of) Definitions 3 and 4 can naturally be framed in the context of Problem 2. For a fixed parameter $\varepsilon \in (0,1)$, we define

$$\mathcal{A} := \Delta^{\mathcal{N}}, \quad \mathcal{B} := [0,1], \tag{27}$$

where $\mathcal{N} := \{i\varepsilon\}_{i \in [\lfloor \frac{1}{\varepsilon} \rfloor]} \cup \{0,1\}$ is an ε-net for $[0,1]$, and is viewed as a set of representative thresholds. For simplicity of notation, we use $p \sim \mathbf{a}$ to mean that $p \in \mathcal{N}$ is sampled according to $\mathbf{a}$. Note that the sequence $b_{[T]} \in \mathcal{B}^T$ will eventually correspond to our (binary) online label space.

We next define the sets and payoff vectors in Problem 2, where $\text{sign}(0) = 1$ by convention:

$$\begin{aligned} \mathcal{U}^{(1)} &:= \Delta^{\mathcal{N}}, \quad \mathbf{v}^{(1)}(\mathbf{a}, b) := \{\mathbb{E}_{p \sim \mathbf{a}}[(p - b)\,\text{sign}(p - s)]\}_{s \in \mathcal{N}}, \\ \mathcal{U}^{(2)} &:= \{\text{d}_\ell \circ c\}_{\ell \in \mathcal{L}, c \in \mathcal{C}}, \quad v^{(2)}(\mathbf{a}, b) := \mathbb{E}_{p \sim \mathbf{a}}[p - b]. \end{aligned} \tag{28}$$

We remark that our definition of $\mathcal{U}^{(2)}$ is exactly the set $\mathcal{F}$ in Proposition 7.

We equip both $\mathcal{H}^{(1)} = \mathbb{R}^{\mathcal{N}}$ and $\mathcal{H}^{(2)} = \mathbb{R}$ with the standard Euclidean inner product. Note that $\mathbf{v}^{(1)}$ is a function that takes $(\mathbf{a}, b) \in \mathcal{A} \times \mathcal{B}$ to a vector in $\mathbb{R}^{\mathcal{N}}$, whose coordinates are indexed by $\mathcal{N}$. Moreover, in a slight abuse of notation (as remarked on in Problem 2), $\mathcal{U}^{(1)}$ consists of elements of $\mathcal{H}^{(1)}$, whereas $\mathcal{U}^{(2)}$ consists of functions taking contexts from our domain $\mathcal{X}$ to $\mathcal{H}^{(2)}$.

To use Corollary 18, we first instantiate the learner $\mathsf{alg}^{(1)}$ as specified in (20).

Lemma 22 *Following definitions (27), (28), there exists $\mathsf{alg}^{(1)}$ such that for any $(\mathbf{a}_{[T]}, b_{[T]}) \in \mathcal{A}^T \times \mathcal{B}^T$, $\mathsf{alg}^{(1)}$ outputs $\mathbf{u}_{[T]}^{(1)} \in (\mathcal{U}^{(1)})^T$ such that $\mathbf{u}_t^{(1)}$ depends only on $\mathbf{a}_{[t-1]}, b_{[t-1]}$, and (20) holds with*

$$\text{reg}^{(1)}(T) := \sqrt{2T \log\left(\frac{1}{\varepsilon} + 2\right)}.$$

Proof The algorithm is multiplicative weights. More precisely, for all $(\mathbf{a}_t, b_t) \in \mathcal{A} \times \mathcal{B}$,

$$\|\mathbf{v}^{(1)}(\mathbf{a}_t, b_t)\|_\infty \le 1.$$

Therefore, Lemma 16 applies with $L = 1$ and $\|\cdot\| = \|\cdot\|_1$. We choose $r(\mathbf{u}) := \sum_{s \in \mathcal{N}} \mathbf{u}_s \log \mathbf{u}_s$, at which point the standard bound $\Theta \le \log(|\mathcal{N}|)$ in Lemma 16 yields the conclusion. ■

Next, we recall a construction of a CMLOO (Definition 17) given by (Okoroafor et al., 2025).

Lemma 23 (Lemma 3.11, (Okoroafor et al., 2025)) *Following definitions (27), (28), and assuming (24), there exists $\mathcal{O}$, an ε-CMLOO.*

Proof We state the algorithm in Algorithm 3. For notational simplicity, we fix inputs $\mathbf{w} = (q, r) \in \Delta^2$, $\mathbf{x} \in \mathcal{X}$, $\mathbf{u} := \mathbf{u}^{(1)} \in \Delta^{\mathcal{N}}$, and $\mathrm{d}_\ell \circ c := \mathbf{u}^{(2)} \in \mathcal{U}^{(2)}$ to the CMLOO. Also, let

$$f(\mathbf{a}, b) := q \sum_{s \in \mathcal{N}} \mathbf{u}_s \mathbb{E}_{p \sim \mathbf{a}} \left[(p - b)\mathrm{sign}(p - s)\right] + r\mathrm{d}_\ell(c(\mathbf{x}))\mathbb{E}_{p \sim \mathbf{a}}[p - b],$$

so that following Definition 17, we wish to find $\mathbf{a} \in \mathcal{A}$ such that $f(\mathbf{a}, b) \le \varepsilon$ for every $b \in \mathcal{B}$. It is helpful to define the "pure strategy" specialization of (29), i.e., where $\mathbf{a}$ is a point mass on $s \in \mathcal{N}$:

$$h(p) := q \sum_{s \in \mathcal{N}} \mathbf{u}_s \mathrm{sign}(p - s) + r\mathrm{d}_\ell(c(\mathbf{x})). \tag{29}$$

In particular, we have $f(\mathbf{e}_p, b) = h(p)(b - p)$ in this special case. Observe that under (24), we have for any $p \in \mathcal{N}$ that $|h(p)| \le 1$, because $|c(\mathbf{x})| \le 1$, $|\sum_{s \in \mathcal{N}} \mathbf{u}_s \mathrm{sign}(p - s)| \le 1$, and $q + r = 1$.

Case 1: $h(0) \le 0$**.** In this case, Algorithm 3 outputs $\mathbf{a} = \mathbf{e}_0$, which satisfies

$$f(\mathbf{a}, b) = h(0)(b - 0) \le 0, \text{ for all } b \in \mathcal{B}.$$

Case 2: $h(1) \ge 0$**.** In this case, we similarly have for $\mathbf{a} = \mathbf{e}_1$,

$$f(\mathbf{a}, b) = h(1)(b - 1) \le 0, \text{ for all } b \in \mathcal{B}.$$

Case 3: $h(0) > 0$ **and** $h(1) < 0$**.** In this case, there are adjacent $(p, p') \in \mathcal{N} \times \mathcal{N}$ with $p \le p' \le p + \varepsilon$, $h(p) \ge 0$, and $h(p') \le 0$, and Algorithm 3 outputs $\mathbf{a} = \frac{|h(p')|}{|h(p)| + |h(p')|}\mathbf{e}_p + \frac{|h(p)|}{|h(p)| + |h(p')|}\mathbf{e}_{p'}$. Then,

$$\begin{aligned}
f(\mathbf{a}, b) &= \frac{|h(p')|}{|h(p)| + |h(p')|} \cdot h(p)(b - p) + \frac{|h(p)|}{|h(p)| + |h(p')|} \cdot h(p')(b - p') \\
&= \frac{|h(p')|}{|h(p)| + |h(p')|} \cdot h(p)(b - p) + \frac{|h(p)|}{|h(p)| + |h(p')|} \cdot h(p')(b - p) \\
&+ \frac{|h(p)|}{|h(p)| + |h(p')|} \cdot h(p')(b - p) \cdot (p - p') \\
&= \frac{|h(p)|}{|h(p)| + |h(p')|} \cdot h(p')(b - p) \cdot (p - p') \le \varepsilon,
\end{aligned}$$

for all $b \in \mathcal{B}$, where the last line used $|p - p'| \le \varepsilon$ and $|\frac{h(p)}{|h(p)| + |h(p')|} \cdot h(p')(b - p)| \le 1$. ■

We conclude the section by showing how to apply Lemmas 22 and 23 within the context of Proposition 7 to give our result on online binary omniprediction.

Corollary 24 (Online binary omniprediction) *Let $\mathcal{L}$ be a family of loss functions and $\mathcal{C}$, $\mathcal{C}'$ be families of comparators satisfying (24). Assume there exists an online learner $\mathsf{alg}^{(2)}$ that takes*

Algorithm 3: CMLOO for binary omniprediction

Input: $\mathbf{w} = (q, r) \in \Delta^2$, $\mathbf{x} \in \mathcal{X}$, $\mathbf{u} \in \Delta^{\mathcal{N}}$, $c \in \mathcal{C}$, $\ell \in \mathcal{L}$
if $h(0) \le 0$ **then**
  // Following definition (29).
  return $\mathbf{e}_0$
else if $h(1) \ge 0$ **then**
  return $\mathbf{e}_1$
else
  $(p, p') \leftarrow$ elements in $\mathcal{N} \times \mathcal{N}$ such that $p \le p' \le p + \varepsilon$, $h(p) \ge 0$, $h(p') \le 0$
  return $\frac{|h(p')|}{|h(p)|+|h(p')|}\mathbf{e}_p + \frac{|h(p)|}{|h(p)|+|h(p')|}\mathbf{e}_{p'}$

inputs $(v_{[T]}, \mathbf{x}_{[T]}) \in [-1, 1]^T \times \mathcal{X}^T$, *and outputs* $\ell_{[T]} \in \mathcal{L}^T$, $c_{[T]} \in (\mathcal{C}')^T$,[8] *such that* (ℓ_t, c_t) *depends only on* $v_{[t-1]}$, $\mathbf{x}_{[t-1]}$, *and*

$$\sup_{(\ell,c)\in\mathcal{L}\times\mathcal{C}} \sum_{t\in[T]} v_t(\mathrm{d}_\ell(c(\mathbf{x}_t)) - \mathrm{d}_{\ell_t}(c_t(\mathbf{x}_t))) \le \mathrm{reg}(T), \tag{30}$$

for $\mathrm{reg} : \mathbb{N} \to \mathbb{R}_{\ge 0}$ *such that all* $T \ge T_{\mathcal{L},\mathcal{C}}$ *satisfy* $\frac{\mathrm{reg}(T)}{T} \le \varepsilon$. *Then if* $T = \Omega(\frac{1}{\varepsilon^2}\log(\frac{1}{\delta\varepsilon})) + T_{\mathcal{L},\mathcal{C}}$, *we can produce* $p_{[T]} \in [0, 1]^T$, *a* 9ε*-omnipredictor for* $(\mathbf{x}_{[T]}, y_{[T]}, \mathcal{L}, \mathcal{C})$, *with probability* $\ge 1 - \delta$.

Proof Following notation in Proposition 7, it suffices to show how to produce $p_{[T]} \in [0, 1]^T$ satisfying 3ε-$(\mathbf{x}_{[T]}, y_{[T]}, \mathcal{F})$-multiaccuracy and 6ε-$(y_{[T]}, \mathcal{W})$-calibration, where we recall

$$\mathcal{F} := \{\mathrm{d}_\ell \circ c\}_{\ell\in\mathcal{L}, \mathbf{c}\in\mathcal{C}}\,,\ \mathcal{W} := \{-\mathrm{d}_\ell \circ k^\star_\ell\}_{\ell\in\mathcal{L}}\,.$$

We achieve this by using Corollary 18. From our definitions of $\mathcal{U}^{(1)}$, $\mathbf{v}^{(1)}$, $\mathcal{U}^{(2)}$, and $\mathbf{v}^{(2)}$ in (28), it is clear that we may take $L = 1$ in (21). Now if we can satisfy the requirements (20) of Corollary 18, playing any $p_t \sim \mathbf{a}_t$ as $\mathbf{a}_t$ is produced by the CMLOO in Lemma 23 in each iteration, gives

$$\sup_{\mathbf{u}^{(i)}\in\mathcal{U}^{(i)}} \frac{1}{T}\sum_{t\in[T]} \left\langle \mathbf{u}^{(i)}(\mathbf{x}_t), \mathbf{v}^{(i)}(\mathbf{e}_{p_t}, b_t)\right\rangle \le \varepsilon + \frac{\mathrm{reg}^{(i)}(T) + 28\sqrt{T\log(\frac{8}{\delta})}}{T} \tag{31}$$

for $i \in [2]$ with probability $\ge 1 - \delta$. Condition on this event henceforth.

When $i = 1$, the guarantee in (31) exactly corresponds to $(y_{[T]}, \mathcal{W}_{\text{thresh}})$-calibration, when comparing with Definition 4 and Lemma 21.[9] Similarly, when $i = 2$, (31) is exactly $(\mathbf{x}_{[T]}, y_{[T]}, \mathcal{F})$-multiaccuracy. Thus, it is enough to bound the right-hand side of (31) by 3ε in both cases, which also results in 6ε-$(y_{[T]}, \mathcal{W})$-calibration via Lemma 21. To do this we use the online learner from Lemma 22 as $\mathsf{alg}^{(1)}$, and the learner with guarantee (30) as $\mathsf{alg}^{(2)}$, and take T as specified. ■

We postpone discussion of the construction of online learners $\mathsf{alg}^{(2)}$ meeting the requirement (30), for both specific and general pairs of $\mathcal{L} \times \mathcal{C}$, to Appendices C.4 and C.5.

8. We include this additional flexibility for our applications in Appendix C.5, which may return improper hypotheses.
9. Our definition of $\mathcal{U}^{(1)}$ formally only yields $(y_{[T]}, \mathcal{W}_{\text{thresh}})$-calibration for the thresholds $s \in \mathcal{N}$. However, because we only play predictions in $\mathcal{N}$, all weights induced by thresholds outside $\mathcal{N}$ agree with that of some threshold in $\mathcal{N}$.

C.3. Statistical binary omniprediction

We next consider statistical omniprediction, again with $k = 2$. However, we require a different formulation of Problem 2. In this section, we let $\mathcal{H}^{(1)}$ and $\mathcal{H}^{(2)}$ be the Hilbert spaces of (norm) square-integrable functions under $\mathcal{D}$, taking $\mathcal{X} \times \{0,1\} \to \mathbb{R}^{\mathcal{N}}$ and $\mathcal{X} \times \{0,1\} \to \mathbb{R}$ respectively. The inner products of $\mathbf{u}, \mathbf{v} \in \mathcal{H}^{(1)}$ and $u, v \in \mathcal{H}^{(2)}$ are the corresponding $L^2(\mathcal{D})$ inner products:

$$\langle \mathbf{u}, \mathbf{v} \rangle := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\langle \mathbf{u}(\mathbf{x}, y), \mathbf{v}(\mathbf{x}, y) \rangle \right], \quad \langle u, v \rangle := \mathbb{E}_{(\mathbf{x},y)} \left[u(\mathbf{x}, y) v(\mathbf{x}, y) \right].$$

Note that in this instance of Problem 2, there is no additional context $\mathbf{x} \in \mathcal{X}$, as it is implicitly specified through our inner product definitions. Next, define $\mathcal{A}$ to be the set of functions taking each $\mathbf{x} \in \mathcal{X} \to \mathbf{a}(\mathbf{x}) \in \Delta^{\mathcal{N}}$, i.e.,

$$\mathcal{A} := \left\{ \mathbf{a} : \mathcal{X} \to \Delta^{\mathcal{N}} \right\}. \tag{32}$$

As before, $\mathcal{N}$ is an ε-net for $[0,1]$ that includes $\{0,1\}$, and we write $p \sim \mathbf{a}(\mathbf{x})$ to mean $p \in \mathcal{N}$ is sampled as specified by $\mathbf{a}(\mathbf{x}) \in \Delta^{\mathcal{N}}$. Also, our payoff vectors $\mathbf{v}^{(1)}$ and $\mathbf{v}^{(2)}$ will be independent of $b \in \mathcal{B}$, so we simply let $\mathcal{B} = \emptyset$ and drop the input b from $\mathbf{v}^{(1)}$, $\mathbf{v}^{(2)}$.

Finally, we let

$$\begin{aligned} \mathcal{U}^{(1)} &:= \Delta^{\mathcal{N}}, \quad \mathbf{v}^{(1)}(\mathbf{a})(\mathbf{x}, y) := \left\{ \mathbb{E}_{p\sim\mathbf{a}(\mathbf{x})} \left[(p - y)\mathrm{sign}(p - s) \right] \right\}_{s\in\mathcal{N}}, \\ \mathcal{U}^{(2)} &:= \{ \mathrm{d}_\ell \circ c \}_{\ell\in\mathcal{L}, c\in\mathcal{C}}, \quad v^{(2)}(\mathbf{a})(\mathbf{x}, y) := \mathbb{E}_{p\sim\mathbf{a}(\mathbf{x})}[p - y]. \end{aligned} \tag{33}$$

As before, $\mathcal{U}^{(1)}$ is interpreted as the family of constant functions over $\mathcal{X} \times \{0,1\}$ with range $\Delta^{\mathcal{N}}$, and $\mathrm{d}_\ell \circ c \in \mathcal{U}^{(2)}$ acts on $(\mathbf{x}, y) \in \mathcal{X} \times \{0,1\}$ by discarding y and outputting $d_\ell(c(\mathbf{x}))$. We also specify the functions $\mathbf{v}^{(1)}(\mathbf{a}) \in \mathcal{H}^{(1)}$ and $\mathbf{v}^{(2)}(\mathbf{a}) \in \mathcal{H}^{(2)}$ by their actions on an element $(\mathbf{x}, y) \in \mathcal{X} \times \{0,1\}$.

With this setup in hand, we now extend Lemmas 22 and 23 to the statistical setting.

Lemma 25 *Let $\delta \in (0,1)$. Following definitions (32), (33), there exists $\mathsf{alg}^{(1)}$ such that for any $\mathbf{a}_{[T]} \in \mathcal{A}^T$, $\mathsf{alg}^{(1)}$ outputs $\mathbf{u}^{(1)}_{[T]} \in (\mathcal{U}^{(1)})^T$ such that $\mathbf{u}^{(1)}_t$ depends only on $\mathbf{a}_{[t-1]}$, and (20) holds with*

$$\mathrm{reg}^{(1)}(T) := 20\sqrt{T \log\left(\frac{4}{\delta\varepsilon}\right)},$$

with probability $\geq 1 - \delta$, where for each $t \in [T]$, we require one i.i.d. draw $(\mathbf{x}_t, y_t) \sim \mathcal{D}$.

Proof More concretely, our goal in (20) is to have

$$\sup_{\mathbf{u}\in\Delta^{\mathcal{N}}} \sum_{t\in[T]} \left\langle \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\mathbf{v}^{(1)}(\mathbf{a}_t)(\mathbf{x}, y) \right], \mathbf{u} - \mathbf{u}_t \right\rangle \leq \mathrm{reg}^{(1)}(T).$$

For any $\mathbf{a}_t$ we have an unbiased estimator of $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathbf{v}^{(1)}(\mathbf{a}_t)(\mathbf{x}, y)]$ conditioned on the history, with entries always in $[-1, 1]$, under our assumed sample access to $\mathcal{D}$. Thus, the conclusion follows by applying the variant of multiplicative weights in Lemma 15 with $L = 1$ and $\Theta = \log(\frac{1}{\varepsilon} + 2)$. ■

Lemma 26 *Following definitions (32), (33), and assuming (24), there exists $\mathcal{O}$, an ε-CMLOO.*

Proof The CMLOO is unchanged from Lemma 23, except we now return a function $\mathbf{a} \in \mathcal{H}^{(1)}$ such that for any $\mathbf{x} \in \mathcal{X}$, we let $\mathbf{a}(\mathbf{x})$ have the same output as Algorithm 3 with context $\mathbf{x}$. Then, for any auxiliary inputs $\mathbf{w} = (q, r) \in \Delta^2$, $\mathbf{u} \in \Delta^{\mathcal{N}}$, $c \in \mathcal{C}$, and $\ell \in \mathcal{L}$, we have for all $y \in [0, 1]$,

$$q \sum_{s \in \mathcal{N}} \mathbf{u}_s \mathbb{E}_{p \sim \mathbf{a}(\mathbf{x})} \left[(p - y) \mathrm{sign}(p - s) \right] + r \mathrm{d}_\ell(c(\mathbf{x})) \mathbb{E}_{p \sim \mathbf{a}(\mathbf{x})}[p - y] \le \varepsilon, \text{ for all } \mathbf{x} \in \mathcal{X}.$$

Taking expectations over the above display over $(\mathbf{x}, y) \sim \mathcal{D}$ then yields the CMLOO guarantee

$$\mathbb{E}_{(\mathbf{x},y) \sim \mathcal{D}} \left[q \left\langle \mathbf{u}, \mathbf{v}^{(1)}(\mathbf{a}) \right\rangle + r \mathrm{d}_\ell(c(\mathbf{x})) \mathbb{E}_{p \sim \mathbf{a}(\mathbf{x})}[p - y] \right] \le 0.$$

■

We can now derive the statistical analog of Corollary 24, again postponing the discussion of online learners meeting (34) to Appendices C.4 and C.5.

Corollary 27 (Statistical binary omniprediction) *Let $\mathcal{L}$ be a family of loss functions and $\mathcal{C}$, $\mathcal{C}'$ be families of comparators satisfying (24). Assume there exists an online learner $\mathsf{alg}^{(2)}$ that takes inputs $\{v_t : \mathcal{X} \times \{0, 1\} \to \mathbb{R}\}_{t \in [T]}$, and outputs $\ell_{[T]} \in \mathcal{L}^T$, $c_{[T]} \in (\mathcal{C}')^T$, such that (ℓ_t, c_t) depends only on $v_{[t-1]}$, and*

$$\sup_{(\ell, c) \in \mathcal{L} \times \mathcal{C}} \sum_{t \in [T]} \langle v_t, \mathrm{d}_\ell(c) - \mathrm{d}_{\ell_t}(c_t) \rangle \le \mathrm{reg}(T), \tag{34}$$

for $\mathrm{reg} : \mathbb{N} \to \mathbb{R}_{\ge 0}$ such that all $T \ge T_{\mathcal{L},\mathcal{C}}$ satisfy $\frac{\mathrm{reg}(T)}{T} \le \varepsilon$ with probability $\ge 1 - \frac{\delta}{2}$. Then if $T = \Omega(\frac{1}{\varepsilon^2} \log(\frac{1}{\delta\varepsilon})) + T_{\mathcal{L},\mathcal{C}}$, we can produce $\mathbf{p} : \mathcal{X} \to [0, 1]$, a 9ε-omnipredictor for $(\mathcal{D}, \mathcal{L}, \mathcal{C})$, with probability $\ge 1 - \delta$, given T i.i.d. samples from $\mathcal{D}$.

Proof The proof is completely analogous to Corollary 24 but using (34) and Lemmas 25, 26 in place of (30) and Lemmas 22, 23. The conclusion is that we can return $\{\mathbf{p}_t : \mathcal{X} \to [0, 1]\}_{t \in [T]}$ satisfying $\mathcal{F}$-multiaccuracy and $\mathcal{W}$-calibration on average, i.e., for all $f \in \mathcal{F}$ and $w \in \mathcal{W}$ in Proposition 7,

$$\mathbb{E}_{(\mathbf{x},y) \sim \mathcal{D}} \left[\frac{1}{T} \sum_{t \in [T]} \langle \mathbf{p}_t(\mathbf{x}) - \mathbf{y}, \mathbf{f}(\mathbf{x}) \rangle \right] \le 3\varepsilon,$$

$$\mathbb{E}_{(\mathbf{x},y) \sim \mathcal{D}} \left[\frac{1}{T} \sum_{t \in [T]} \langle \mathbf{p}_t(\mathbf{x}) - \mathbf{y}, \mathbf{w}(\mathbf{p}_t(\mathbf{x})) \rangle \right] \le 6\varepsilon,$$

simultaneously except with probability δ (taking a union bound over (34) and Lemma 25 with $\delta \leftarrow \frac{\delta}{2}$). By linearity of expectation, outputting $\mathbf{p} \leftarrow \mathbf{p}_t$ for a uniformly random $t \in [T]$ satisfies $\mathcal{F}$-multiaccuracy and $\mathcal{W}$-calibration, so applying Proposition 7 to this $\mathbf{p}$ gives the result. ■

C.4. Generalized linear models

In this section, we specialize Corollary 24 and 27 to the setting of generalized linear models, where $\mathcal{L} := \mathcal{L}_{\text{GLM}}$ as defined in (5), and $\mathcal{C} := \mathcal{C}_{\text{lin}}$ where

$$\mathcal{C}_{\text{lin}} := \left\{ c(\mathbf{x}) := \langle \mathbf{c}, \mathbf{x} \rangle \mid \mathbf{c} \in \mathbb{B}_2^d(1) \right\}. \tag{35}$$

We conflate the actual linear classifier $\mathbf{c} \in \mathbb{B}_2^d(1)$ with a function $c : \mathcal{X} \to [-1, 1]$ by using boldface, so e.g., $\mathbf{c}_t \in \mathbb{B}_2^d(1)$ corresponds to the function $c_t = \langle \mathbf{c}_t, \cdot \rangle \in \mathcal{C}_{\text{lin}}$. Recalling (6), we take the discrete derivative d_ℓ to be negation for all $\ell \in \mathcal{L}_{\text{GLM}}$, so $\mathcal{F}$ in Proposition 7 is equivalent to $\mathcal{C}_{\text{lin}}$ because $\mathcal{C}_{\text{lin}}$ is closed under negation. We next require online learners satisfying (30), (34).

Lemma 28 *Assuming (24) holds, there exists* $\mathsf{alg}^{(2)}$ *such that for any* $(v_{[T]}, \mathbf{x}_{[T]}) \in [-1, 1]^T \times \mathcal{X}^T$, $\mathsf{alg}^{(2)}$ *outputs* $\mathbf{c}_{[T]} \in (\mathbb{B}_2^d(1))^T$ *in* $O(dT)$ *time, such that* $\mathbf{c}_t$ *only depends on* $v_{[t-1]}, \mathbf{x}_{[t-1]}$, *and*

$$\sup_{c \in \mathcal{C}_{\text{lin}}} \sum_{t \in [T]} v_t \langle \mathbf{x}_t, \mathbf{c} - \mathbf{c}_t \rangle \le \sqrt{T}.$$

Proof This follows from Lemma 16 with $\mathcal{X} \leftarrow \mathbb{B}_2^d(1)$ and $r(\mathbf{c}) := \frac{1}{2} \|\mathbf{c}\|_2^2$, where we take $\mathbf{g}_t := -v_t \mathbf{x}_t$ so that our application satisfies $\|\mathbf{g}_t\|_2 \le L := 1$ and $\Theta \le \frac{1}{2}$, because $\|\mathbf{x}_t\|_2 \le 1$ under (24). ■

Lemma 29 *Let* $\delta \in (0, 1)$. *Assuming (24) holds, there exists* $\mathsf{alg}^{(2)}$ *such that for any* $\{v_t : \mathcal{X} \times \{0, 1\} \to [-1, 1]\}_{t \in [T]}$, $\mathsf{alg}^{(2)}$ *outputs* $\mathbf{c}_{[T]} \in (\mathbb{B}_2^d(1))^T$ *in* $O(dT)$ *time, such that* $\mathbf{c}_t$ *only depends on* $v_{[t-1]}$, *and*

$$\sup_{c \in \mathcal{C}_{\text{lin}}} \sum_{t \in [T]} \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} \left[v_t(\mathbf{x}, y) \langle \mathbf{x}, \mathbf{c} - \mathbf{c}_t \rangle \right] \le 20 \sqrt{T \log\left(\frac{2}{\delta}\right)},$$

with probability $\ge 1 - \delta$, *where for each* $t \in [T]$, *we require one i.i.d. draw* $(\mathbf{x}_t, y_t) \sim \mathcal{D}$.

Proof The proof is identical to Lemma 28, where we use Lemma 15 in place of Lemma 16, granting us unbiased access to $\mathbf{g}_t := \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}}[v_t(\mathbf{x}, y)\mathbf{x}]$ under our sampling assumption. ■

We now combine the pieces to give our result on omnipredicting (binary) generalized linear models.

Theorem 30 (Binary generalized linear models) *Let* $\delta \in (0, 1)$, *let* $\mathcal{L} := \mathcal{L}_{\text{GLM}}$ *and* $\mathcal{C} := \mathcal{C}_{\text{lin}}$ *defined in (23), (35) respectively, and assume (24) holds. Then if*

$$T = \Omega\left(\frac{\log\left(\frac{1}{\delta\varepsilon}\right)}{\varepsilon^2} \right)$$

for an appropriate constant, in the online setting, we can output $p_{[T]} \in [0, 1]^T$, *an* ε*-omnipredictor for* $(\mathbf{x}_T, y_{[T]}, \mathcal{L}, \mathcal{C})$, *in time* $O((d + \frac{1}{\varepsilon})T)$ *with probability* $\ge 1 - \delta$. *In the statistical setting, we can output* $\mathbf{p} : \mathcal{X} \to [0, 1]$, *an* ε*-omnipredictor for* $(\mathcal{D}, \mathcal{L}, \mathcal{C})$, *in time* $O((d + \frac{1}{\varepsilon})T)$ *with probability* $\ge 1 - \delta$, *given* T *i.i.d. samples from* $\mathcal{D}$, *such that* $\mathbf{p}$ *can be evaluated in time* $O(d + \frac{1}{\varepsilon})$.

Proof The result on online omniprediction is immediate by combining Corollary 24 and Lemma 28, and adjusting $\varepsilon \leftarrow \frac{\varepsilon}{9}$. For the result on statistical omniprediction, the result similarly follows from Corollary 27 and Lemma 29. Note that we take $\mathcal{C} = \mathcal{C}'$ in these applications.

To evaluate $\mathbf{p}$ as specified in Corollary 27,[10] we store the values of all inputs to Algorithm 2 for each iteration. We can compute the function $h(s)$ in (29) for all $s \in \mathcal{N}$ in $O(\frac{1}{\varepsilon})$ time, after spending $O(d)$ time to evaluate some $c_t(\mathbf{x})$. This complexity also dominates the cost of each iteration. ■

C.5. General classifiers and losses

We finally consider Corollary 24 and Corollary 27 for general loss functions $\mathcal{L}$ and general comparators $\mathcal{C}$ that satisfy (24). To state the guarantees of our online learners satisfying (30), (34), we define the following complexity measure parameters of a function class $\mathcal{F}$.

Definition 31 (Statistical Rademacher complexity) *Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$, $\mathcal{D}$ be a distribution over $\mathcal{X}$, and $T \in \mathbb{N}$. The* statistical Rademacher complexity *of $\mathcal{F}$ is defined to be*

$$\mathsf{rad}_T(\mathcal{F}) := \mathbb{E}_{\{\mathbf{x}_t\}_{t\in[T]}\sim_{\text{i.i.d.}}\mathcal{D}}\left[\mathbb{E}_{\sigma_{[T]}\sim_{\text{unif.}}\{\pm1\}^T}\left[\sup_{f\in\mathcal{F}}\frac{1}{T}\sum_{t\in[T]}\sigma_t f(\mathbf{x}_t)\right]\right].$$

Definition 32 (Sequential Rademacher complexity) *Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$ and $T \in \mathbb{N}$. The* sequential Rademacher complexity *of $\mathcal{F}$ is defined to be*

$$\mathsf{srad}_T(\mathcal{F}) := \sup_{\{\mathbf{x}_t:\{\pm1\}^{t-1}\to\mathcal{X}\}_{t\in[T]}}\mathbb{E}_{\sigma_{[T]}\sim_{\text{unif.}}\{\pm1\}^T}\left[\sup_{f\in\mathcal{F}}\frac{1}{T}\sum_{t\in[T]}\sigma_t f(\mathbf{x}_t(\sigma_{[t-1]}))\right].$$

For the online setting, we use the following result from (Okoroafor et al., 2025).

Lemma 33 (Theorem 4.5, (Okoroafor et al., 2025)) *In the setting of Corollary 24, let $\mathcal{F} := \{\mathrm{d}_\ell \circ c\}_{\ell\in\mathcal{L},c\in\mathcal{C}}$. There exists $\mathsf{alg}^{(2)}$ such that for any $(v_{[T]}, \mathbf{x}_{[T]}) \in [-1,1]^T \times \mathcal{X}^T$, $\mathsf{alg}^{(2)}$ outputs $c_{[T]}$, such that $c_t \in \mathcal{C}'$ only depends on $v_{[t-1]}, \mathbf{x}_{[t-1]}$, and*

$$\sup_{c\in\mathcal{C}}\sum_{t\in[T]} v_t\left(\mathrm{d}_\ell(c(\mathbf{x}_t)) - \mathrm{d}_{\ell_t}(c_t(\mathbf{x}_t))\right) \le 2T\cdot\mathsf{srad}_T\left(\mathcal{F}\right).$$

We remark that Lemma 33 is based on a computationally-inefficient (indeed, nonconstructive) argument from (Rakhlin et al., 2015), but that the regret bound is known to be tight up to a constant factor (Theorem 4.5, (Okoroafor et al., 2025)). For specific pairs $(\mathcal{L}, \mathcal{C})$, e.g., the ones in Lemma 28, it is possible to design more explicit online learners, so in general the computational cost depends on the setting.

For the statistical setting, we similarly use the following result.

10. We preprocess the indices in $[T]$ so a uniform sample is attainable in $O(1)$ time, e.g., via the alias method.

Lemma 34 (Lemma 7.4, Lemma 7.6, (Okoroafor et al., 2025)) *In the setting of Corollary 27, let $\delta \in (0,1)$ and $\mathcal{F} := \{\mathrm{d}_\ell \circ c\}_{\ell \in \mathcal{L}, c \in \mathcal{C}}$. Let $\mathcal{V} \subseteq \{v : \mathcal{X} \times \{0,1\} \to [-1,1]\}$. There exists $\mathsf{alg}^{(2)}$ such that for any $v_{[T]} \in \mathcal{V}^T$, $\mathsf{alg}^{(2)}$ outputs $c_{[T]}$, making $O(T^{1.5})$ calls to an ERM oracle for $\mathcal{F}$ over T samples per iteration, such that $c_t \in \mathcal{C}'$ only depends on $v_{[t-1]}$, and for a universal constant C,*

$$\sup_{c \in \mathcal{C}} \sum_{t \in [T]} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[v_t(\mathbf{x}, y) \left(\mathrm{d}_\ell(c(\mathbf{x})) - \mathrm{d}_{\ell_t}(c_t(\mathbf{x})) \right) \right] \le C \left(\sqrt{T \cdot \log \frac{1}{\delta}} + T \cdot \mathsf{rad}_T \left(\mathcal{F} \cdot \mathcal{V} \right) \right),$$

with probability $\ge 1 - \delta$, where for each $t \in [T]$, we require T i.i.d. draws $\sim \mathcal{D}$.

In the statement of Lemma 34, we let $\mathcal{F} \cdot \mathcal{V}$ consist of functions $(\mathbf{x}, y) \mapsto f(\mathbf{x}) v(\mathbf{x}, y)$ for $f \in \mathcal{F}$ and $v \in \mathcal{V}$, and an ERM oracle for $\mathcal{F}$ finds $\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i \in [n]} \mathbf{w}_i f(\mathbf{x}_i, y_i)$ over some dataset $\{(\mathbf{x}_i, y_i)\}$ of n i.i.d. draws from $\mathcal{D}$, and some weights $\mathbf{w} \in \mathbb{R}^n$. The computational complexity of implementing such an oracle again typically depends on the specific setting.

Finally, we conclude with our main result for binary omniprediction in the general case.

Theorem 35 (General binary omniprediction) *Let $\mathcal{L}$ be a family of loss functions and $\mathcal{C}$ be a family of comparators such that (24) holds, let $\mathcal{F} := \{\mathrm{d}_\ell \circ c\}_{\ell \in \mathcal{L}, c \in \mathcal{C}}$, and let $\delta \in (0,1)$. Let $T^{\mathrm{seq}}_{\mathcal{L},\mathcal{C}}$, $T^{\mathrm{stat}}_{\mathcal{L},\mathcal{C}}$ be such that all $T \ge T^{\mathrm{seq}}_{\mathcal{L},\mathcal{C}}$ satisfy $\mathsf{srad}_T(\mathcal{F}) \le \frac{\varepsilon}{9}$, and all $T \ge T^{\mathrm{stat}}_{\mathcal{L},\mathcal{C}}$ satisfy $\mathsf{rad}_T(\mathcal{F} \cdot \mathcal{V}) \le \frac{\varepsilon}{18C}$, where $\mathcal{V}$ consists of functions $(\mathbf{x}, y) \to p(\mathbf{x}) - y$ for all possible $p : \mathcal{X} \to [0,1]$ outputted by the CMLOO in Lemma 26 given classes $\mathcal{C}, \mathcal{L}$. Then if*

$$T = \Omega\left(\frac{\log\left(\frac{1}{\delta\varepsilon}\right)}{\varepsilon^2} \right) + T^{\mathrm{seq}}_{\mathcal{L},\mathcal{C}},$$

for an appropriate constant, in the online setting, we can output $p_{[T]} \in [0,1]^T$, an ε-omnipredictor for $(\mathbf{x}_T, y_{[T]}, \mathcal{L}, \mathcal{C})$, with probability $\ge 1 - \delta$. If

$$T = \Omega\left(\frac{\log\left(\frac{1}{\delta\varepsilon}\right)}{\varepsilon^2} \right) + T^{\mathrm{stat}}_{\mathcal{L},\mathcal{C}},$$

for an appropriate constant, in the statistical setting, we can output $\mathbf{p} : \mathcal{X} \to [0,1]$, an ε-omnipredictor for $(\mathcal{D}, \mathcal{L}, \mathcal{C})$, with probability $\ge 1 - \delta$, given T^2 i.i.d. samples from $\mathcal{D}$ and $O(T^{2.5})$ calls to an ERM oracle for $\mathcal{F}$.

Proof The proof is largely the same as the proof for Theorem 30. We apply Lemma 33 within Corollary 24 for the online setting and Lemma 34 within Corollary 27 in the statistical setting. The oracle complexity comes from the cost of the online learner in Lemma 34. ■

Appendix D. Multiclass Omniprediction

We now proceed to our main result: omniprediction with $k > 2$ classes. Here, many properties specific to the binary classification setting do not hold, e.g., the generalizations of Lemmas 21 and 23. We develop a general strategy for constructing MLOOs for multiclass omniprediction in Appendix D.1. We then give the multiclass extensions of Corollaries 24 and 27 in Appendix D.2.

Finally, in Appendices D.3 and D.4, we give our full multiclass omniprediction results for linear and general classifiers.

Throughout, we make the following normalization assumptions:

$$\begin{aligned} \mathcal{X} \subseteq \mathbb{B}_2^d(1), \quad \mathbf{c}(\mathbf{x}) \in [-1,1]^k \text{ for all } \mathbf{c} \in \mathcal{C},\ \mathbf{x} \in \mathcal{X}, \\ \ell(\mathbf{t},\mathbf{y}) \in [-1,1] \text{ for all } \ell \in \mathcal{L},\ (\mathbf{t},\mathbf{y}) \in [-1,1]^k \times \partial\Delta^k. \end{aligned} \tag{36}$$

D.1. MLOOs for multiclass omniprediction

In this section, we consider a specialized application of the machinery in Section 3.2 to multiclass prediction. Specifically, suppose that we have an instance of Problem 1, where

$$\mathcal{A} := \Delta^{\mathcal{N}}, \quad \mathcal{B} = \partial\Delta^k, \tag{37}$$

and $\mathcal{N}$ is an ε-net for Δ^k. Also, define for all $\mathbf{a} \in \Delta^{\mathcal{N}}$ and $\mathbf{b} \in \mathcal{B}$,

$$\mathbf{v}(\mathbf{a},\mathbf{b}) := \{\mathbf{a}_{\mathbf{s}}(\mathbf{s}-\mathbf{b})\}_{\mathbf{s}\in\mathcal{N}} \in \mathbb{R}^{k|\mathcal{N}|}, \tag{38}$$

and suppose that for all $i \in [m]$,

$$\mathbf{v}^{(i)}(\mathbf{a},\mathbf{b}) = \mathbf{M}^{(i)}\mathbf{v}(\mathbf{a},\mathbf{b}) \tag{39}$$

for some linear operator $\mathbf{M}^{(i)} : \mathbb{R}^{k|\mathcal{N}|} \to \mathcal{H}^{(i)}$. We give a meta-result that shows how to implement an MLOO for arbitrary simultaneous Blackwell approachability instances satisfying (38), (39), whose quality scales with bounds on the $\{\mathbf{M}^{(i)}\}_{i\in[m]}$ and the $\{\mathcal{U}^{(i)}\}_{i\in[m]}$.

Lemma 36 *In the setting of Problem 1, suppose (37), (38), and (39) hold, where $\mathcal{N}$ is an ε-net for Δ^k. Further, suppose that for all $i \in [m]$, we have*

$$\left\| \left(\mathbf{M}^{(i)}\right)^* \mathbf{u}^{(i)} \right\|_\infty \le R \text{ for all } \mathbf{u}^{(i)} \in \mathcal{U}^{(i)}. \tag{40}$$

where * denotes the adjoint. We can implement a $2\varepsilon R$-MLOO with probability at least $1-\delta$ in time $O(|\mathcal{N}| \cdot \text{poly}(k, \log\frac{1}{\delta\varepsilon}))$. Similarly, in the setting of Problem 2, we can implement a $2\varepsilon R$-CMLOO with probability at least $1-\delta$ in time $O(|\mathcal{N}| \cdot \text{poly}(k, \log\frac{1}{\delta\varepsilon}))$.

Proof We first show that for all $\mathbf{w} \in \Delta^m$, $\{\mathbf{u}^{(i)}\}_{i\in[m]} \in \prod_{i\in[m]} \mathcal{U}^{(i)}$, there exists $\mathbf{a} \in \mathcal{A}$ with

$$\max_{\mathbf{b}\in\mathcal{B}} \sum_{i\in[m]} \mathbf{w}_i \left\langle \mathbf{u}^{(i)}, \mathbf{v}^{(i)}(\mathbf{a},\mathbf{b}) \right\rangle \le \varepsilon R.$$

Throughout the proof fix a set of $\mathbf{w} \in \Delta^m$ and $\{\mathbf{u}^{(i)}\}_{i\in[m]} \in \prod_{i\in[m]} \mathcal{U}^{(i)}$, and denote

$$\mathbf{f} := \sum_{i\in[m]} \left(\mathbf{M}^{(i)}\right)^* \mathbf{u}^{(i)} \in \mathbb{B}_\infty^{k|\mathcal{N}|}(R).$$

Thus, our goal is to establish

$$\min_{\mathbf{a}\in\mathcal{A}} \max_{\mathbf{b}\in\mathcal{B}} \langle \mathbf{f}, \mathbf{v}(\mathbf{a},\mathbf{b}) \rangle \le \varepsilon R. \tag{41}$$

Because $\langle \mathbf{f}, \mathbf{v}(\mathbf{a}, \mathbf{b}) \rangle$ is a bilinear function of $\mathbf{a}, \mathbf{b}$, the von Neumann minimax theorem gives

$$\begin{aligned}\min_{\mathbf{a}\in\mathcal{A}} \max_{\mathbf{b}\in\mathcal{B}} \langle \mathbf{f}, \mathbf{v}(\mathbf{a}, \mathbf{b}) \rangle &= \max_{\mathbf{q}\in\Delta^k} \min_{\mathbf{s}\in\mathcal{N}} \mathbb{E}_{\mathbf{b}\sim\mathbf{q}} \left[\langle \mathbf{f}, \mathbf{v}(\mathbf{e}_\mathbf{s}, \mathbf{b}) \rangle \right] \\ &= \max_{\mathbf{q}\in\Delta^k} \min_{\mathbf{s}\in\mathcal{N}} \mathbb{E}_{\mathbf{b}\sim\mathbf{q}} \left[\langle \mathbf{f}_\mathbf{s}, \mathbf{s} - \mathbf{b} \rangle \right] = \max_{\mathbf{q}\in\Delta^k} \min_{\mathbf{s}\in\mathcal{N}} \langle \mathbf{f}_\mathbf{s}, \mathbf{s} - \mathbf{q} \rangle ,\end{aligned}$$

where $\mathbf{e}_\mathbf{s} \in \{0,1\}^\mathcal{N}$ is the indicator vector for strategy $\mathbf{s} \in \mathcal{N}$, and $\mathbf{f}_\mathbf{s} \in \mathbb{B}_\infty^k(R)$ concatenates the corresponding coordinates of $\mathbf{f}$. Finally we claim that for any $\mathbf{q} \in \Delta^k$,

$$\min_{\mathbf{s}\in\mathcal{N}} \langle \mathbf{f}_\mathbf{s}, \mathbf{s} - \mathbf{q} \rangle \le \varepsilon R.$$

Indeed, choosing $\mathbf{s} \in \mathcal{N}$ so that $\|\mathbf{s} - \mathbf{q}\|_1 \le \varepsilon$ and applying Hölder's inequality yields this bound.

We conclude by discussing runtime. Normalize the problem by R by resetting $\mathbf{f} \leftarrow \frac{1}{R}\mathbf{f}$, so we want to solve (41) to ε additive error. Notice that (41) is of the following form:

$$\min_{\mathbf{a}\in\Delta^\mathcal{N}} \max_{\mathbf{b}\in\Delta^k} \mathbf{g}^\top \mathbf{a} - \mathbf{b}^\top \mathbf{F} \mathbf{a} = \min_{\mathbf{a}\in\Delta^\mathcal{N}} \max_{\mathbf{b}\in\Delta^k} \mathbf{b}^\top \mathbf{M} \mathbf{a},$$

where $\mathbf{F} \in \mathbb{R}^{k\times\mathcal{N}}$ horizontally stacks the values of $\mathbf{f}$, $\mathbf{g} \in \mathbb{R}^\mathcal{N}$ has coordinate $\mathbf{s} \in \mathcal{N}$ equal to $\langle \mathbf{f}_\mathbf{s}, \mathbf{s} \rangle$, and we define $\mathbf{M} := \mathbf{1}_k \mathbf{g}^\top - \mathbf{F}$. Also, we have that $\mathbf{M} \in [-2, 2]^{k\times\mathcal{N}}$. We can rewrite this as

$$\min t \text{ such that } \mathbf{A}\mathbf{a} + \mathbf{c} = t\mathbf{1}_k,\ \mathbf{1}_\mathcal{N}^\top \mathbf{a} = 1,\ \mathbf{a} \ge \mathbf{0}_\mathcal{N},\ \mathbf{c} \ge \mathbf{0}_k \text{ entrywise.}$$

We note that $\mathbf{c}$ is enforcing the inequality constraints $\mathbf{A}\mathbf{a} \le t\mathbf{1}_k$. We can trivially enforce that $t \in [-2, 2]$, $\mathbf{a} \in [0, 1]^\mathcal{N}$, and $\mathbf{c} \in [0, 4]^k$. It is enough to obtain ε additive error for this problem for our guarantees. At this point, the solver in Theorem 1.1 of (van den Brand et al., 2021) gives the claim.

■

D.2. Reducing multiclass omniprediction to low-regret learning

We give the analogs of Corollaries 24 and 27 in the multiclass setting.

Online setting. In the online setting, for a fixed parameter $\varepsilon \in (0, 1)$, we define $(\mathcal{A}, \mathcal{B})$ as in (37), where $\mathcal{N}$ is an ε-net for Δ^k of size $(\frac{5}{\varepsilon})^{k-1}$ as guaranteed by Fact 1. For some $\mathbf{a} \in \mathcal{A}$, we use $\mathbf{p} \sim \mathbf{a}$ to mean that some $\mathbf{p} \in \mathcal{N}$ is sampled according to $\mathbf{a}$.

We next define the sets and payoff vectors in Problem 2:

$$\begin{aligned}\mathcal{U}^{(1)} &:= [-1, 1]^{\mathcal{N}\times k}, \quad \mathbf{v}^{(1)}(\mathbf{a}, \mathbf{b}) := \{\mathbf{a}_\mathbf{s}(\mathbf{s} - \mathbf{b})\}_{\mathbf{s}\in\mathcal{N}} , \\ \mathcal{U}^{(2)} &:= \{\mathbf{d}_\ell \circ \mathbf{c}\}_{\ell\in\mathcal{L},\mathbf{c}\in\mathcal{C}} , \quad \mathbf{v}^{(2)}(\mathbf{a}, \mathbf{b}) := \mathbb{E}_{\mathbf{p}\sim\mathbf{a}} [\mathbf{p} - \mathbf{b}] .\end{aligned} \tag{42}$$

Note that $\mathcal{U}^{(1)}$ and $\mathbf{v}^{(1)}$ live in a vector space of dimension $k|\mathcal{N}|$, whereas $\mathcal{U}^{(2)}$ and $\mathbf{v}^{(2)}$ are functions with range in $\mathbb{R}^k$. We require the analog of Lemma 22, an online learner for $\mathcal{U}^{(1)}$.

Lemma 37 *Following definitions (37), (42), there exists* $\mathsf{alg}^{(1)}$ *such that for any* $(\mathbf{a}_{[T]}, \mathbf{b}_{[T]}) \in \mathcal{A}^T \times \mathcal{B}^T$, $\mathsf{alg}^{(1)}$ *outputs* $\mathbf{u}_{[T]}^{(1)} \in (\mathcal{U}^{(1)})^T$ *such that* $\mathbf{u}_t^{(1)}$ *depends only on* $\mathbf{a}_{[t-1]}$, $\mathbf{b}_{[t-1]}$, *and (20) holds with*

$$\mathrm{reg}^{(1)}(T) := \varepsilon T + \frac{k|\mathcal{N}|}{\varepsilon}.$$

Proof The algorithm is projected gradient descent. More precisely, because $\|\mathbf{s} - \mathbf{b}\|_1 \le 2$ for all $(\mathbf{s}, \mathbf{b}) \in \mathcal{N} \times \mathcal{B}$, we have for all $(\mathbf{a}_t, \mathbf{b}_t) \in \mathcal{A} \times \mathcal{B}$, that

$$\|\mathbf{v}^{(1)}(\mathbf{a}_t, \mathbf{b}_t)\|_2 \le \|\mathbf{v}^{(1)}(\mathbf{a}_t, \mathbf{b}_t)\|_1 \le 2.$$

Therefore, standard regret analyses of projected gradient descent with step size $\eta \leftarrow \frac{\varepsilon}{2}$ (e.g., Theorem 3.2, (Bubeck, 2015)) gives the result, because $\mathcal{U}^{(1)}$ has ℓ_2 radius at most $\sqrt{k|\mathcal{N}|}$. ■

Corollary 38 (Online multiclass omniprediction) *Let $\mathcal{L}$ be a family of loss functions and $\mathcal{C}$, $\mathcal{C}'$ be families of comparators satisfying (36). Assume there exists an online learner $\mathsf{alg}^{(2)}$ that takes inputs $(\mathbf{v}_{[T]}, \mathbf{x}_{[T]}) \in (\mathbb{B}_2^k(2))^T \times \mathcal{X}^T$, and outputs $\ell_{[T]} \in \mathcal{L}^T$, $\mathbf{c}_{[T]} \in (\mathcal{C}')^T$, such that $(\ell_t, \mathbf{c}_t)$ depends only on $\mathbf{v}_{[t-1]}$, $\mathbf{x}_{[t-1]}$, and*

$$\sup_{(\ell, \mathbf{c}) \in \mathcal{L} \times \mathcal{C}} \sum_{t \in [T]} \langle \mathbf{v}_t, \mathbf{d}_\ell(\mathbf{c}(\mathbf{x}_t)) - \mathbf{d}_{\ell_t}(\mathbf{c}_t(\mathbf{x}_t)) \rangle \le \mathrm{reg}(T), \tag{43}$$

for $\mathrm{reg} : \mathbb{N} \to \mathbb{R}_{\ge 0}$ such that all $T \ge T_{\mathcal{L},\mathcal{C}}$ satisfy $\frac{\mathrm{reg}(T)}{T} \le \varepsilon$. Then if $T = \Omega(k(\frac{1}{\varepsilon})^{k+1} + \frac{1}{\varepsilon^2}\log(\frac{1}{\delta})) + T_{\mathcal{L},\mathcal{C}}$, we can produce $p_{[T]} \in [0,1]^T$, a 12ε-omnipredictor for $(\mathbf{x}_{[T]}, \mathbf{y}_{[T]}, \mathcal{L}, \mathcal{C})$, with probability $\ge 1 - \delta$.

Proof The proof is completely analogous to Corollary 24. We substitute Lemma 37 and (43) for Lemma 22 and (30), and note that we may take $L = 2$ in (21) by the ℓ_∞-ℓ_1 Hölder's inequality. We postpone discussion of implementing the CMLOO for a moment, but suppose we have a 2ε-CMLOO. Then, Corollary 18 yields a sequence $\mathbf{p}_{[T]} \in (\Delta^k)^T$ such that with probability $\ge 1 - \delta$,

$$\sup_{\mathbf{u}^{(i)} \in \mathcal{U}^{(i)}} \frac{1}{T} \sum_{t \in [T]} \left\langle \mathbf{u}^{(i)}(\mathbf{x}_t), \mathbf{v}^{(i)}(\mathbf{e}_{\mathbf{p}_t}, \mathbf{b}_t) \right\rangle \le 2\varepsilon + \frac{\mathrm{reg}^{(i)}(T) + 28\sqrt{T\log(\frac{4}{\delta})}}{T} \tag{44}$$

for $i \in [2]$. When $i = 1$, the guarantee in (44) corresponds to calibration against the entire ℓ_∞-norm ball in dimension $k|\mathcal{N}|$, which encompasses $\mathcal{W}$-calibration for $\mathcal{W}$ in Proposition 7, under the scaling assumption (36). When $i = 2$, the guarantee in (44) corresponds to $\mathcal{F}$-multiaccuracy as required by Proposition 7. For large enough T as specified, we thus have 5ε-$\mathcal{W}$-calibration using Lemma 37 as $\mathsf{alg}^{(1)}$, and 4ε-$\mathcal{F}$-multiaccuracy using (43), and Proposition 7 gives the claim.

It remains to give a 2ε-CMLOO. For this we use Lemma 36. Comparing the definitions (39) and (42), $\mathbf{M}^{(1)}$ is simply the identity matrix in dimension $k|\mathcal{N}|$, and $\mathbf{M}^{(2)}$ is $\mathbf{1}_k \otimes \mathbf{I}_{\mathcal{N}}$, where $\otimes$ denotes the Kronecker product. This matrix has one-sparse columns, so it satisfies

$$\left\|\mathbf{M}^{(2)}\right\|_{1\to 1} = 1 \implies \left\|\left(\mathbf{M}^{(2)}\right)^*\right\|_{\infty\to\infty} = 1.$$

Hence, we may take $R = 1$ in (42), because both $\mathcal{U}^{(1)}$ and $\mathcal{U}^{(2)}$ are contained in the ℓ_∞ balls of their respective dimension. The result now follows from Lemma 36.

■

Statistical setting. We let $\mathcal{H}^{(1)}$ and $\mathcal{H}^{(2)}$ be the Hilbert spaces of (norm) square-integrable functions under $\mathcal{D}$, with ranges $\mathbb{R}^{\mathcal{N}\times k}$, $\mathbb{R}^k$, respectively, with the standard $L^2(\mathcal{D})$ inner products.

Next, we take $\mathcal{A}$ to be functions taking each $\mathbf{x} \in \mathcal{X} \to \mathbf{a}(\mathbf{x}) \in \Delta^{\mathcal{N}}$, i.e.,

$$\mathcal{A} := \left\{\mathbf{a} : \mathcal{X} \to \Delta^{\mathcal{N}}\right\}. \tag{45}$$

Our payoff vectors will again be independent of $\mathbf{b} \in \mathcal{B}$, so we omit it from our notation. Also, let

$$\begin{aligned} \mathcal{U}^{(1)} &:= \Delta^{\mathcal{N}}, \quad \mathbf{v}^{(1)}(\mathbf{a})(\mathbf{x}, \mathbf{y}) := \left\{[\mathbf{a}(\mathbf{x})]_{\mathbf{s}}(\mathbf{s} - \mathbf{y})\right\}_{\mathbf{s} \in \mathcal{N}}, \\ \mathcal{U}^{(2)} &:= \left\{\mathbf{d}_\ell \circ \mathbf{c}\right\}_{\ell \in \mathcal{L}, \mathbf{c} \in \mathcal{C}}, \quad \mathbf{v}^{(2)}(\mathbf{a})(\mathbf{x}, \mathbf{y}) := \mathbb{E}_{\mathbf{p} \sim \mathbf{a}(\mathbf{x})}[\mathbf{p} - \mathbf{y}]. \end{aligned} \tag{46}$$

We last require an online learner for $\mathcal{U}^{(1)}$ in the statistical setting.

Lemma 39 *Following definitions (45), (46), there exists* $\mathsf{alg}^{(1)}$ *such that for any* $\mathbf{a}_{[T]} \in \mathcal{A}^T$, $\mathsf{alg}^{(1)}$ *outputs* $\mathbf{u}^{(1)}_{[T]} \in (\mathcal{U}^{(1)})^T$ *such that* $\mathbf{u}^{(1)}_t$ *depends only on* $\mathbf{a}_{[t-1]}$, *and (20) holds with*

$$\mathrm{reg}^{(1)}(T) := \varepsilon T + \frac{10k|\mathcal{N}|}{\varepsilon} + 32\sqrt{T \log\left(\frac{2}{\delta}\right)},$$

with probability $\geq 1 - \delta$, *where for each* $t \in [T]$, *we require one i.i.d. draw* $(\mathbf{x}_t, \mathbf{y}_t) \sim \mathcal{D}$.

Proof We pattern our proof off of Lemma 15, although we require a few differences to obtain the specific form of regret bound here. The key observation is that for all $\mathbf{a} \in \mathcal{A}$, the definitions (46) give $|\left\langle \mathbf{u}^{(1)}, \mathbf{v}^{(1)}(\mathbf{a})\right\rangle| \leq 2$ using the ℓ_∞-ℓ_1 Hölder's inequality. Our strategy is then to play (stochastic) projected gradient descent (PGD) against the $\{\mathbf{v}^{(1)}(\mathbf{a}_t)\}_{t \in [T]}$. To simplify notation, let

$$\mathbf{g}_t := \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}}\left[\mathbf{v}^{(1)}(\mathbf{a}_t)(\mathbf{x}, \mathbf{y})\right], \quad \tilde{\mathbf{g}}_t := \mathbf{v}^{(1)}(\mathbf{a}_t)(\mathbf{x}_t, \mathbf{y}_t), \quad \mathbf{d}_t := \mathbf{g}_t - \tilde{\mathbf{g}}_t,$$

and observe that under our sampling assumptions, $\tilde{\mathbf{g}}_t$ is unbiased for $\mathbf{g}_t$ conditioned on the history of the algorithm if we use a held out i.i.d. sample $(\mathbf{x}_t, \mathbf{y}_t)$. Also, $|\left\langle \tilde{\mathbf{g}}_t, \mathbf{u}^{(1)}\right\rangle| \leq 2$ holds for all $t \in [T]$ and $\mathbf{u}^{(1)} \in \mathcal{U}^{(1)}$, and $\max_{t \in [T]} \max\{\|\mathbf{g}_t\|_1, \|\tilde{\mathbf{g}}_t\|_1\} \leq 2$.

Now, we define $\mathbf{u}_1 \leftarrow \mathbf{0}_{\mathcal{N} \times k}$, and our iterates $\mathbf{u}_t$ using PGD with step size $\eta > 0$ and the $\{-\tilde{\mathbf{g}}_t\}_{t \in [T]}$,

$$\mathbf{u}^{(1)}_t \leftarrow \mathrm{argmin}_{\mathbf{u}^{(1)} \in \mathcal{U}^{(1)}}\left\{\left\|\mathbf{u}^{(1)} - \left(\mathbf{u}^{(1)}_{t-1} + \eta \tilde{\mathbf{g}}_{t-1}\right)\right\|_2^2\right\}. \tag{47}$$

We also define a "ghost iterate" sequence of $\mathbf{w}_{[T+1]} \in (\mathcal{U}^{(1)})^{T+1}$ that sets $\mathbf{w}_1 = \mathbf{u}^{(1)}_1$, but updates using $\mathbf{d}_{t-1}$ in place of $\tilde{\mathbf{g}}_{t-1}$ in (47). Standard PGD analysis (e.g., Theorem 3.2, (Bubeck, 2015)) shows

$$\begin{aligned} \sum_{t \in [T]} \left\langle \tilde{\mathbf{g}}_t, \mathbf{u}^{(1)} - \mathbf{u}^{(1)}_t \right\rangle &\leq 2\eta T + \frac{k|\mathcal{N}|}{2\eta}, \\ \sum_{t \in [T]} \left\langle \mathbf{d}_t, \mathbf{u}^{(1)} - \mathbf{w}_t \right\rangle &\leq 8\eta T + \frac{k|\mathcal{N}|}{2\eta}, \end{aligned}$$

simultaneously hold for all $\mathbf{u}^{(1)} \in \mathcal{U}^{(1)}$. Summing and rearranging yields

$$\sum_{t \in [T]} \left\langle \mathbf{g}_t, \mathbf{u}^{(1)} - \mathbf{u}^{(1)}_t \right\rangle \leq 10\eta T + \frac{k|\mathcal{N}|}{\eta} + \sum_{t \in [T]} \left\langle \mathbf{d}_t, \mathbf{w}_t - \mathbf{u}^{(1)}_t \right\rangle.$$

Now the last term above is the sum of T conditionally mean-zero terms, each of which is bounded in $[-8, 8]$. Thus by the Azuma-Hoeffding inequality, with probability $\geq 1 - \delta$,

$$\sum_{t\in[T]} \left\langle \mathbf{g}_t, \mathbf{u}^{(1)} - \mathbf{u}_t^{(1)} \right\rangle \leq 10\eta T + \frac{k|\mathcal{N}|}{\eta} + 32\sqrt{T \log\left(\frac{2}{\delta}\right)},$$

and supremizing this over $\mathbf{u}^{(1)} \in \mathcal{U}^{(1)}$ and setting $\eta \leftarrow \frac{\varepsilon}{10}$ gives the claim. ■

Corollary 40 (Statistical multiclass omniprediction) *Let $\mathcal{L}$ be a family of loss functions and $\mathcal{C}$, $\mathcal{C}'$ be families of comparators satisfying (24). Assume there exists an online learner $\mathsf{alg}^{(2)}$ that takes inputs $\{\mathbf{v}_t : \mathcal{X} \times \partial\Delta^k \to \mathbb{B}_2^k(2)\}_{t\in[T]}$, and outputs $\ell_{[T]} \in \mathcal{L}^T$, $\mathbf{c}_{[T]} \in (\mathcal{C}')^T$, such that $(\ell_t, \mathbf{c}_t)$ depends only on $\mathbf{v}_{[t-1]}$, and*

$$\sup_{(\ell,\mathbf{c})\in\mathcal{L}\times\mathcal{C}} \sum_{t\in[T]} \langle \mathbf{v}_t, \mathbf{d}_\ell(\mathbf{c}) - \mathbf{d}_{\ell_t}(\mathbf{c}_t)\rangle \leq \mathrm{reg}(T), \tag{48}$$

for $\mathrm{reg} : \mathbb{N} \to \mathbb{R}_{\geq 0}$ such that all $T \geq T_{\mathcal{L},\mathcal{C}}$ satisfy $\frac{\mathrm{reg}(T)}{T} \leq \varepsilon$ with probability $\geq 1 - \frac{\delta}{2}$. Then if $T = \Omega(k(\frac{1}{\varepsilon})^{k+1} + \frac{1}{\varepsilon^2}\log(\frac{1}{\delta})) + T_{\mathcal{L},\mathcal{C}}$, we can produce $\mathbf{p} : \mathcal{X} \to [0, 1]$, a 9ε-omnipredictor for $(\mathcal{D}, \mathcal{L}, \mathcal{C})$, with probability $\geq 1 - \delta$, given T i.i.d. samples from $\mathcal{D}$.

Proof The proof is the same as Corollary 38 (with modifications analogous to Corollary 27 vis-à-vis Corollary 24), where we use Lemma 39 and (48) instead of Lemma 37 and (43). In our construction of the CMLOO in the statistical setting, we note that the matrices $\mathbf{M}^{(1)}$ and $\mathbf{M}^{(2)}$ in (39) are again $\mathbf{I}_{\mathcal{N}\times k}$ and $\mathbf{1}_k \otimes \mathbf{I}_{\mathcal{N}}$, so Lemma 36 again yields a 2ε-CMLOO that outputs a function $\mathbf{a}_t : \mathcal{X} \to \Delta^{\mathcal{N}}$, from which we can sample random predictions $\mathbf{p}_t : \mathcal{X} \to \Delta^{\mathcal{N}}$. As in Corollary 27, our final omnipredictor evaluates a uniform randomly sampled $\mathbf{p}_t$, over the range $t \in [T]$. ■

D.3. Generalized linear models

In this section, we specialize Corollaries 38 and 40 to the setting of multiclass generalized linear models, where $\mathcal{L} := \mathcal{L}_{\mathrm{GLM}}$ as defined in (5), and $\mathcal{C} := \mathcal{C}_{\mathrm{lin}}$ where

$$\mathcal{C}_{\mathrm{lin}} := \left\{\mathbf{c}(\mathbf{x}) := \mathbf{C}\mathbf{x} \mid \mathbf{C} \in \mathbb{R}^{k\times d},\ \|\mathbf{C}\|_{2\to\infty} \leq 1\right\}. \tag{49}$$

In other words, $\mathbf{C}$ has sub-unit norm rows. This is the natural family of classifiers because it takes $\mathbf{x} \in \mathcal{X}$ to $\mathbf{c}(\mathbf{x}) \in [-1, 1]^k$, under the scaling bounds in (36). Analogously to Appendix C.4, we conflate a function $\mathbf{c} \in \mathcal{C}_{\mathrm{lin}}$ with the associated linear classifier $\mathbf{C} \in \mathbb{R}^{k\times d}$ via capitalization. We again observe that because $\mathbf{d}_\ell$ is negation for all $\ell \in \mathcal{L}_{\mathrm{GLM}}$ by (6), and $\mathcal{C}_{\mathrm{lin}}$ is closed under negation, we can equivalently set $\mathcal{F} \leftarrow \mathcal{C}_{\mathrm{lin}}$ in applications of Proposition 7.

Our last ingredients are online learners satisfying (43), (48).

Lemma 41 *Assuming (36) holds, there exists $\mathsf{alg}^{(2)}$ such that for any $(\mathbf{v}_{[T]}, \mathbf{x}_{[T]}) \in (\mathbb{B}_2^k(2))^T \times \mathcal{X}^T$, $\mathsf{alg}^{(2)}$ outputs $\mathbf{C}_{[T]} \in (\mathbb{B}_{2\to\infty}^{k\times d}(1))^T$ in $O(dkT)$ time, such that $\mathbf{C}_t$ only depends on $\mathbf{v}_{[t-1]}$, $\mathbf{x}_{[t-1]}$, and*

$$\sup_{\mathbf{C}\in\mathcal{C}_{\mathrm{lin}}} \sum_{t\in[T]} \langle \mathbf{v}_t \otimes \mathbf{x}_t, \mathbf{C} - \mathbf{C}_t\rangle \leq 2\sqrt{kT}.$$

Proof This follows from Lemma 16 with $\mathcal{X} \leftarrow \mathbb{B}_{2\to\infty}^{k\times d}$ and $r(\mathbf{C}) := \frac{1}{2}\|\mathbf{C}\|_{\mathrm{F}}^2$ (i.e., half the squared entrywise ℓ_2 norm). Note that for all $t \in [T]$, because $\mathbf{v}_t \otimes \mathbf{x}_t$ is rank-one,

$$\|\mathbf{v}_t \otimes \mathbf{x}_t\|_{\mathrm{F}} = \|\mathbf{v}_t \otimes \mathbf{x}_t\|_{\mathrm{op}} = \|\mathbf{v}_t\|_2 \|\mathbf{x}_t\|_2 \le 2.$$

Thus, the only adjustments compared to Lemma 28 is that now we have $L = 2$ and $\Theta \le \frac{k}{2}$. ■

Lemma 42 *Let $\delta \in (0,1)$. Assuming (36) holds, there exists $\mathsf{alg}^{(2)}$ such that for any $\{\mathbf{v}_t : \mathcal{X} \times \partial\Delta^k \to \mathbb{B}_2^k(2)\}_{t\in[T]}$, $\mathsf{alg}^{(2)}$ outputs $\mathbf{C}_{[T]} \in (\mathbb{B}_{2\to\infty}^{k\times d}(1))^T$ in $O(dkT)$ time, such that $\mathbf{C}_t$ only depends on $\mathbf{v}_{[t-1]}$, and*

$$\sup_{\mathbf{C}\in\mathcal{C}_{\mathrm{lin}}} \sum_{t\in[T]} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}} \left[\langle \mathbf{v}_t(\mathbf{x},\mathbf{y}) \otimes \mathbf{x}, \mathbf{C} - \mathbf{C}_t\rangle\right] \le 40\sqrt{kT\log\left(\frac{2}{\delta}\right)},$$

with probability $\ge 1-\delta$, where for each $t \in [T]$, we require one i.i.d. draw $(\mathbf{x}_t, \mathbf{y}_t) \sim \mathcal{D}$.

Proof The proof is identical to Lemma 41, where we use Lemma 15 in place of Lemma 16. ■

We conclude with our main result on omnipredicting multiclass generalized linear models.

Theorem 43 (Multiclass generalized linear models) *Let $\delta \in (0,1)$, let $\mathcal{L} := \mathcal{L}_{\mathrm{GLM}}$ and $\mathcal{C} := \mathcal{C}_{\mathrm{lin}}$ defined in (5), (49) respectively, and assume (36) holds. Then if*

$$T = k\left(\Omega\left(\frac{1}{\varepsilon}\right)^{k+1} + \Omega\left(\frac{\log\left(\frac{1}{\delta}\right)}{\varepsilon^2}\right)\right)$$

for an appropriate constant, in the online setting, we can output $\mathbf{p}_{[T]} \in (\Delta^k)^T$, an ε-omnipredictor for $(\mathbf{x}_T, \mathbf{y}_{[T]}, \mathcal{L}, \mathcal{C})$, in time $O(dkT) + O(\frac{1}{\varepsilon})^{2k}\mathrm{poly}(k, \log\frac{1}{\delta\varepsilon})$ with probability $\ge 1-\delta$. In the statistical setting, we can output $\mathbf{p} : \mathcal{X} \to \Delta^k$, an ε-omnipredictor for $(\mathcal{D}, \mathcal{L}, \mathcal{C})$, in time $O(dkT) + O(\frac{1}{\varepsilon})^{2k}\mathrm{poly}(k, \log\frac{1}{\delta\varepsilon})$ with probability $\ge 1-\delta$, given T i.i.d. samples from $\mathcal{D}$, such that $\mathbf{p}$ can be evaluated in time $O(dk) + O(\frac{1}{\varepsilon})^{k+1}\mathrm{poly}(k, \log\frac{1}{\delta\varepsilon})$ on any $\mathbf{x} \in \mathcal{X}$ with probability $\ge 1-\delta$.

Proof For the online omniprediction result, we combine Lemma 41 and Corollary 38, and for the statistical omniprediction result, we combine Lemma 42 and Corollary 40. We note that the runtime cost of each iteration is dominated by the $O(dk)$ time for computing $\mathbf{c}_t(\mathbf{x}_t)$, and the cost of Lemma 36. This also applies to the cost of evaluating $\mathbf{p}$ on a fresh sample. ■

D.4. General classifiers and losses

In this section, we specialize Corollaries 38 and 40 to the setting of general multiclass models. Analogously to Appendix D.3, we consider general loss functions $\mathcal{L}$ and general function class $\mathcal{C}$ that satisfy (36). We again require online learners satisfying (43), (48).

Our multiclass online learning results apply the binary online learners from Appendix C.5 in a black-box way. It is possible that tighter characterizations in the multiclass setting are possible (e.g., in the dependence on k), especially for specific structured $(\mathcal{C}, \mathcal{L})$. We demonstrated an example of

this in Appendix D.3, and leave a more general theory to future work. This section is included primarily to highlight how to apply our techniques in a general setting, as our paper's focus is developing the omniprediction framework rather than multiclass learning for specific comparators.

Applying Theorem 4.5 of (Okoroafor et al., 2025) coordinatewise, we obtain the following lemma.

Lemma 44 *In the setting of Corollary 38, let* $\mathcal{F} := \{\mathbf{d}_\ell \circ \mathbf{c}\}_{\ell \in \mathcal{L}, \mathbf{c} \in \mathcal{C}}$. *Assuming (36) holds for family of loss functions* $\mathcal{L}$ *and families of comparators* $\mathcal{C}$ *and* $\mathcal{C}'$, *there exists* $\mathsf{alg}^{(2)}$ *such that for any* $(\mathbf{v}_{[T]}, \mathbf{x}_{[T]}) \in (\mathbb{B}_2^k(2))^T \times \mathcal{X}^T$, $\mathsf{alg}^{(2)}$ *outputs* $\mathbf{c}_{[T]} \in \mathcal{C}'$, *such that* $\mathbf{c}_t$ *only depends on* $\mathbf{v}_{[t-1]}$, $\mathbf{x}_{[t-1]}$, *and*

$$\sup_{\mathbf{c} \in \mathcal{C}} \sum_{t \in [T]} \langle \mathbf{v}_t, \mathbf{d}_\ell(\mathbf{c}(\mathbf{x})) - \mathbf{d}_{\ell_t}(\mathbf{c}_t(\mathbf{x})) \rangle \le T \cdot \sum_{i \in [k]} \mathsf{srad}_T(\mathcal{F}_i),$$

where $\mathcal{F}_i$ *consists of functions* $\mathbf{x} \mapsto [\mathbf{f}(\mathbf{x})]_i$ *for* $\mathbf{f} \in \mathcal{F}$, *with* $[\mathbf{f}(\mathbf{x})]_i$ *being the* i^{th} *coordinate of* $\mathbf{f}(\mathbf{x})$.

Similarly, applying Lemma 7.4 and Lemma 7.6 of (Okoroafor et al., 2025) coordinatewise yields the following.

Lemma 45 *In the setting of Corollary 40, let* $\delta \in (0,1)$ *and* $\mathcal{F} := \{\mathbf{d}_\ell \circ \mathbf{c}\}_{\ell \in \mathcal{L}, \mathbf{c} \in \mathcal{C}}$. *Let* $\mathcal{V}$ *be a family of functions* $\mathbf{v} : \mathcal{X} \times \partial\Delta^k \to \mathbb{B}_2^k(2)$. *There exists* $\mathsf{alg}^{(2)}$ *such that for any* $\mathbf{v}_{[T]} \in \mathcal{V}^T$, $\mathsf{alg}^{(2)}$ *outputs* $\mathbf{c}_{[T]}$, *making* $O(T^{1.5})$ *calls to an ERM oracle for each* $\mathcal{F}_i$ *over* T *samples per iteration, such that* $\mathbf{c}_t \in \mathcal{C}'$ *only depends on* $\mathbf{v}_{[t-1]}$, *and for a universal constant* C,

$$\sup_{c \in \mathcal{C}} \sum_{t \in [T]} \mathbb{E}_{(\mathbf{x},\mathbf{y}) \sim \mathcal{D}} \left[\langle \mathbf{v}_t(\mathbf{x}, \mathbf{y}), \mathbf{d}_\ell(\mathbf{c}(\mathbf{x})) - \mathbf{d}_{\ell_t}(\mathbf{c}_t(\mathbf{x})) \rangle \right] \le C \left(k \sqrt{T \cdot \log \frac{k}{\delta}} + T \cdot \sum_{i \in [k]} \mathsf{rad}_T(\mathcal{F}_i \cdot \mathcal{V}_i) \right),$$

with probability $\ge 1 - \delta$, *where for each* $t \in [T]$, *we require* T *i.i.d. draws* $(\mathbf{x}_t, \mathbf{y}_t) \sim \mathcal{D}$.

In the statement of Lemma 34, the class $\mathcal{F}_i \cdot \mathcal{V}_i$ consists of functions $(\mathbf{x}, \mathbf{y}) \mapsto [\mathbf{f}(\mathbf{x})]_i [\mathbf{v}(\mathbf{x}, \mathbf{y})]_i$ for $\mathbf{f} \in \mathcal{F}$ and $\mathbf{v} \in \mathcal{V}$, with $[\mathbf{f}(\mathbf{x})]_i$, $[\mathbf{v}(\mathbf{x}, \mathbf{y})]_i$ being the i^{th} coordinates of $\mathbf{f}(\mathbf{x})$, $\mathbf{v}(\mathbf{x}, \mathbf{y})$, respectively.

We conclude with our main result on multiclass omniprediction in the general setting.

Theorem 46 (General multiclass omnprediction) *Let* $\mathcal{L}$ *be a family of loss functions and* $\mathcal{C}$ *be a family of comparators such that (36) holds, let* $\mathcal{F} := \{\mathbf{d}_\ell \circ \mathbf{c}\}_{\ell \in \mathcal{L}, \mathbf{c} \in \mathcal{C}}$, *and let* $\delta \in (0,1)$. *Let* $T^{\text{seq}}_{\mathcal{L},\mathcal{C}}, T^{\text{stat}}_{\mathcal{L},\mathcal{C}}$ *be such that all* $T \ge T^{\text{seq}}_{\mathcal{L},\mathcal{C}}$ *satisfy* $\sum_{i=1}^k \mathsf{srad}_T(\mathcal{F}_i) \le \frac{\varepsilon}{9}$, *and all* $T \ge T^{\text{stat}}_{\mathcal{L},\mathcal{C}}$ *satisfy* $\sum_{i=1}^k \mathsf{rad}_T(\mathcal{F}_i \cdot \mathcal{V}_i) \le \frac{\varepsilon}{18C}$, *where* $\mathcal{V}$ *consists of functions* $(\mathbf{x}, \mathbf{y}) \to \mathbf{p}(\mathbf{x}) - \mathbf{y}$ *for all possible* $\mathbf{p} : \mathcal{X} \to \Delta^k$ *outputted by the CMLOO in Lemma 36 given classes* $\mathcal{C}, \mathcal{L}$. *Then if*

$$T = \Omega\left(k \left(\frac{1}{\varepsilon} \right)^{k+1} + \frac{\log \frac{1}{\delta}}{\varepsilon^2} \right) + T^{\text{seq}}_{\mathcal{L},\mathcal{C}},$$

for an appropriate constant, in the online setting, we can output $\mathbf{p}_{[T]} \in (\Delta^k)^T$, *an* ε*-omnipredictor for* $(\mathbf{x}_{[T]}, \mathbf{y}_{[T]}, \mathcal{L}, \mathcal{C})$, *with probability* $\ge 1 - \delta$. *If*

$$T = \Omega\left(k \left(\frac{1}{\varepsilon} \right)^{k+1} + \frac{k^2 \log \frac{1}{\delta}}{\varepsilon^2} \right) + T^{\text{stat}}_{\mathcal{L},\mathcal{C}},$$

for an appropriate constant, in the statistical setting, we can output $\mathbf{p} : \mathcal{X} \to \Delta^k$*, an* ε*-omnipredictor for* $(\mathcal{D}, \mathcal{L}, \mathcal{C})$*, with probability* $\geq 1 - \delta$*, given* T^2 *i.i.d. samples from* $\mathcal{D}$ *and* $O(T^{2.5})$ *calls to the ERM oracle for each* $\mathcal{F}_i$*.*

Proof The proof is largely the same as the proof for Theorem 43. For the online omniprediction result, we combine Lemma 44 and Corollary 38, and for the statistical omniprediction result, we combine Lemma 45 and Corollary 40. The oracle complexity comes from Lemma 45. ■

Appendix E. Unions of Comparators

In this section, we showcase the flexibility of our framework by applying it to omniprediction against a *union of comparators*. Let $\mathcal{L}$ be a family of losses $\ell : [-1, 1]^k \times \partial\Delta^k \to [-1, 1]$, and let $\mathcal{C}^{(i)}$ be a comparator family satisfying (36) for all $i \in [m]$. Our goal is to learn an $(\mathcal{L}, \mathcal{C})$-omnipredictor for

$$\mathcal{C} := \bigcup_{i \in [m]} \mathcal{C}_i.$$

In other words, we wish to be competitive against the best $\mathbf{c}$ in any $\mathcal{C}_i$. For simplicity, here we focus on the online setting, although similar extensions for statistical omniprediction are straightforward.

To design an online omnipredictor against $(\mathcal{L}, \mathcal{C})$, we define a simultaneous approachability instance as follows: we define $(\mathcal{A}, \mathcal{B})$ as in (37), and let

$$\mathcal{U}^{(i)} := \{\mathbf{d}_\ell \circ \mathbf{c}_i\}_{\ell \in \mathcal{L}, \mathbf{c} \in \mathcal{C}^{(i)}}, \ \mathbf{v}^{(i)}(\mathbf{a}, \mathbf{b}) := \mathbb{E}_{\mathbf{p} \sim \mathbf{a}}[\mathbf{p} - \mathbf{b}], \text{ for all } i \in [m],$$
$$\mathcal{U}^{(m+1)} := [-1, 1]^{\mathcal{N} \times k}, \ \mathbf{v}^{(m+1)}(\mathbf{a}, \mathbf{b}) := \{\mathbf{a}_\mathbf{s}(\mathbf{s} - \mathbf{b})\}_{\mathbf{s} \in \mathcal{N}}.$$

In other words, there is one approachability set $\mathcal{U}^{(i)}$ for each comparator class $\mathcal{C}^{(i)}$, and the $(m+1)^{\text{th}}$ approachability set is defined analogously to $\mathcal{U}^{(1)}$ in (42).

Theorem 47 *Let* $\mathcal{L}$ *be a family of loss functions and* $\mathcal{C}^{(i)}$ *be a family comparators for all* $i \in [m]$*, such that* (36) *holds for* $\mathcal{L}$ *and every* $\mathcal{C} \leftarrow \mathcal{C}^{(i)}$*, let* $\mathcal{F}^{(i)} := \{\mathbf{d}_\ell \circ \mathbf{c}\}_{\ell \in \mathcal{L}, \mathbf{c} \in \mathcal{C}^{(i)}}$*, and let* $\delta \in (0, 1)$*. Let* $T^{\text{seq}}_{\mathcal{L},\mathcal{C}}$ *be such that all* $T \geq T^{\text{seq}}_{\mathcal{L},\mathcal{C}}$ *and* $i \in [m]$ *satisfy*

$$\sum_{j \in [k]} \mathsf{srad}_T(\mathcal{F}^{(i)}_j) \leq \frac{\varepsilon}{9}.$$

Then if

$$T = \Omega\left(k\left(\frac{1}{\varepsilon}\right)^{k+1} + \frac{\log \frac{m}{\delta}}{\varepsilon^2}\right) + T^{\text{seq}}_{\mathcal{L},\mathcal{C}}$$

for an appropriate constant, in the online setting, we can output $\mathbf{p}_{[T]} \in (\Delta^k)^T$*, an* ε*-omnipredictor for* $(\mathbf{x}_{[T]}, \mathbf{y}_{[T]}, \mathcal{L}, \bigcup_{i \in [m]} \mathcal{C}^{(i)})$*, with probability* $\geq 1 - \delta$*.*

Proof The proof is almost exactly identical to Theorem 43, save for two changes. First, the additive regret term in Corollary 18 now scales with $\log(\frac{m}{\delta})$ (as there are $m + 1$ approachability sets). Second, the CMLOO in Lemma 36 now must hold for $m + 1$ inputs. However, when applying Lemma 36 (specifically following the notation (39)), every $\mathbf{M}^{(i)}$ is identical for $i \in [m]$, and we

bounded the quantity (40) for $\mathbf{M}^{(m+1)}$ already in Corollary 38. Thus, the same proof holds and we simply adjust the logarithmic term in the T lower bound. ■

We remark that all of our main results generalize to unions of comparators; indeed, the binary omniprediction CMLOO construction in Lemma 23 also has a simple extension to this setting. Our framework is even capable of handling unions of *loss families* in much the same way, where we define an approachability set to ensure multiaccuracy for each pairing of a loss family and a comparator class, although we omit this extension to avoid tedium.

Appendix F. Counterexample for Multiclass Isotonic Regression

Our framework for multiclass omniprediction was based on the construction of (Okoroafor et al., 2025) in the binary setting. Concurrently, another construction of $\approx \varepsilon^{-2}$-sample complexity binary omnipredictors was given by (Hu et al., 2025) for GLMs. It is thus natural to ask whether the construction in (Hu et al., 2025) has a multiclass extension. In this section, we show a barrier to such a generalization.

The (Hu et al., 2025) construction was based on the *Isotron* algorithm (Kalai and Sastry, 2009), which alternates online gradient descent with isotonic regression. In particular, the isotonic regression problem that Isotron repeatedly solves is, for input labels $\{y_i\}_{i\in[n]}$, and some proper loss function $\ell : [0,1]^2 \to \mathbb{R}$,

$$\min_{\{p_i\}_{i\in[n]}\in[0,1]^n} \sum_{i\in[n]} \ell(p_i, y_i), \text{ subject to } p_i \le p_{i+1} \text{ for all } i \in [n-1]. \tag{50}$$

In other words, Isotron finds the best-fitting *monotone* sequence of $\{p_i\}_{i\in[n]}$ with respect to $\{y_i\}_{i\in[n]}$, as measured by ℓ. The monotonicity requirement comes from p_i being induced by the gradient of a one-dimensional convex function (for more on this relationship, see Lemma 6 and Section 2.2, (Hu et al., 2025)). Crucially, in the binary setting the optimal choice of $\{p_i\}_{i\in[n]}$ is independent of the choice of proper loss ℓ in (50) (Corollary 9, (Hu et al., 2025); see also (Brummer and Preez, 2013)). This omniprediction property of isotonic regression (50) is then inherited by the overall Isotron framework.

We next state the natural generalization of (50) to the multiclass setting. As (Gneiting and Raftery, 2007) shows, again any proper loss induces predictions via the gradient of a convex function. A vector field is the gradient of a convex function iff it is *cyclically monotone* (Theorem B, (Rockafellar, 1970b)), and we can capture this high-dimensional condition via the following extension of (50).

Problem 3 *Given $\{(\mathbf{v}_i, \mathbf{y}_i)\}_{i\in[T]} \subset \mathbb{R}^k \times \partial\Delta^k$, we define the following isotonic regression problem for a proper loss $\ell : \Delta^k \times \Delta^k \to \mathbb{R}$:*

$$\begin{aligned} \{\mathbf{p}_i^\star, f_i^\star\}_{i\in[n]} &:= \text{argmin}_{\{\mathbf{p}_i, f_i\}_{i\in[n]}\in(\Delta^k\times\mathbb{R})^n} \sum_{i\in[n]} \ell(\mathbf{p}_i, \mathbf{y}_i), \\ &\textit{subject to } \langle \mathbf{p}_j, \mathbf{v}_i - \mathbf{v}_j\rangle \le f_i - f_j \textit{ for all } (i,j) \in [n]\times[n]. \end{aligned} \tag{51}$$

Here, the $\{\mathbf{v}_i\}_{i\in[n]}$ should be interpreted as the "unlinked" predictors in a GLM, and the monotonicity condition in (51) is equivalent to $\mathbf{p}_i = \nabla\omega(\mathbf{v}_i)$, $f_i = \omega(\mathbf{v}_i)$ for all $i \in [n]$. This parameterization is implicit in (50) as well, where the input v_i are first sorted to define the indexing.

For the strategy in (Hu et al., 2025) to generalize to high dimensions, a reasonable necessary condition is for the same omniprediction property to hold for (51), i.e., that its minimizing $\{\mathbf{p}_i^\star\}_{i\in[n]}$ does not depend on the choice of proper loss ℓ. We give a simple numerical counterexample. Define:

$$\ell_{\mathrm{sq}}(\mathbf{p},\mathbf{y}) := \frac{1}{2}\left\|\mathbf{p}-\mathbf{y}\right\|_2^2, \quad \ell_{\log}(\mathbf{p},\mathbf{y}) := -\sum_{i\in[k]} \log(\mathbf{p}_i)\mathbb{I}_{\mathbf{y}=\mathbf{e}_i}.$$

We minimize (51) with respect to these two proper losses, and the following choices of $\{\mathbf{v}_i,\mathbf{y}_i\}_{i\in[2]}$:

$$\{\mathbf{v}_i\}_{i\in[2]} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \{\mathbf{y}_i\}_{i\in[2]} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}. \tag{52}$$

Lemma 48 *The minimizer of* (51) *with* $\ell \leftarrow \ell_{\mathrm{sq}}$ *and inputs* (52) *is*

$$\{\mathbf{p}_i^\star\}_{i\in[2]} = \begin{bmatrix} \frac{3}{7} & \frac{3}{7} \\ \frac{2}{7} & \frac{4}{7} \\ \frac{2}{7} & 0 \end{bmatrix}. \tag{53}$$

Proof For $\mathbf{v}_1 - \mathbf{v}_2 = -\mathbf{e}_1$, the constraints in (51) are equivalent to

$$[\mathbf{p}_1]_1 \le [\mathbf{p}_2]_1.$$

Our goal is to minimize $\|\mathbf{p}_1-\mathbf{e}_1\|_2^2 + \|\mathbf{p}_2-\mathbf{e}_2\|_2^2$ subject to this constraint. It is clear that the constraint is tight, because otherwise $\mathbf{p}_2$ would put any excess mass on the second coordinate. Thus the minimizing $\mathbf{p}_1$ and $\mathbf{p}_2$ are of the form

$$\mathbf{p}_1 = \begin{bmatrix} t \\ \frac{1-t}{2} \\ \frac{1-t}{2} \end{bmatrix}, \quad \mathbf{p}_2 = \begin{bmatrix} t \\ 1-t \\ 0 \end{bmatrix}.$$

The former claim is because Jensen's inequality implies $\mathbf{p}_1$ should spread all remaining mass over the last two coordinates, and the latter is because $\mathbf{p}_2$ has no incentive to place any mass on the third coordinate. The conclusion follows by solving for t that minimizes $(1-t)^2 + 2\cdot\frac{1}{4}(1-t)^2 + 2t^2$. ■

Lemma 49 *The minimizer of* (51) *with* $\ell \leftarrow \ell_{\log}$ *and inputs* (52) *is not* (53).

Proof It suffices to check that the following choices attain better function value:

$$\mathbf{p}_1 = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{4} \\ \frac{1}{4} \end{bmatrix}, \quad \mathbf{p}_2 = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 0 \end{bmatrix}.$$

This is because $-\log(\frac{12}{49}) \ge -\log(\frac{1}{4})$, and the constraint $[\mathbf{p}_1]_1 \le [\mathbf{p}_2]_1$ is satisfied. ■