The Sample Complexity of Multi-Class Omniprediction

Joey Rivkin
Cornell University
jbrivkin@cs.cornell.edu

Princewill Okoroafor
Harvard University
princewill_okoroafor@seas.harvard.edu

Michael P. Kim
Cornell University
mpk@cs.cornell.edu

Abstract

An omnipredictor is a single prediction function that competes with the optimal hypothesis from a class $\mathcal{H}$ for every loss ℓ within a broad collection of loss functions, e.g., all Lipschitz losses $\mathcal{L}_{\text{Lip}}$ or all proper losses $\mathcal{L}_{\text{prop}}$. Recent works establish, in the binary prediction setting, omniprediction for broad loss classes (including $\mathcal{L}_{\text{Lip}}$ and $\mathcal{L}_{\text{prop}}$) can be achieved at essentially the same rate as minimizing a single loss function. This work investigates the complexity of omniprediction in the multi-class prediction setting. We show, in stark contrast to the binary case, for all $K \geq 3$, K-class omniprediction requires strictly more samples than loss minimization. Specifically, we characterize the sample complexity of omniprediction with respect to $\mathcal{L}_{\text{Lip}}$, showing upper and lower bounds that are tight in ε up to logarithmic factors, with an ε-dependence scaling exponentially in K:

$$n_K \sim \begin{cases} \tilde{\Theta}\left(\varepsilon^{-5/2}\right) & K = 3 \\ \tilde{\Theta}\left(\varepsilon^{-(K-1)}\right) & K \geq 4 \end{cases}$$

These bounds follow from a new online learning algorithm for multi-class omniprediction, improving the prior upper bound of $O(\varepsilon^{-(K+1)})$ [HTY26], and new constructions of lower bounds for omniprediction, giving qualitative and quantitative improvements to prior constructions [HV25]. Our techniques are general and imply a number of other related results. In particular, we show that multi-class omniprediction with respect to proper losses $\mathcal{L}_{\text{prop}}$ is statistically *impossible*, no matter how many samples are provided, drawing a sharp line between binary and $K \geq 3$ classes.

1 Introduction

Omniprediction [Gop+22] is a powerful framework for supervised learning. Rather than optimizing a decision rule according to a single loss function, an omnipredictor is a predictor that encodes the optimal decision rule for a huge collection of loss functions.

Definition 1.1 (Omnipredictor)**.** Fix a distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, a class of loss functions $\mathcal{L} \subseteq \{\ell : \mathcal{A} \times \mathcal{Y} \to [-1, 1]\}$, and a hypothesis class $\mathcal{H} \subseteq \{h : \mathcal{X} \to \mathcal{A}\}$. A predictor $p : \mathcal{X} \to \Delta_K$ is an $(\mathcal{L}, \mathcal{H}, \varepsilon)$-omnipredictor if for every $\ell \in \mathcal{L}$,

$$\mathbb{E}\left[\ell(\mathrm{BR}_\ell \circ p(x), y)\right] \leq \min_{h \in \mathcal{H}} \mathbb{E}\left[\ell(h(x), y)\right] + \varepsilon.$$

In other words, $\mathrm{BR}_\ell \circ p$ is a loss minimizer with respect to the hypothesis class $\mathcal{H}$, for every loss in the class $\mathcal{L}$. Here, $\mathrm{BR}_\ell : \Delta_K \to \mathcal{A}$ is the "best-response" function, which takes in an estimate of the probability on outcomes $q \in \Delta_K$, and outputs the action that minimizes the expected loss under q. That is,

$$\mathrm{BR}_\ell(q) = \arg\min_{a \in \mathcal{A}} \mathbb{E}_{\tilde{y} \sim q}[\ell(a, \tilde{y})].$$

In this way, omniprediction separates the tasks of learning to predict (i.e., estimating the probabilities of $y|x$) and making decisions (i.e., given a loss function and probability for $y|x$, choose an optimal action). The appeal of omniprediction is apparent: rather than learning separate predictors for many different loss functions, we can execute learning once—to learn an omnipredictor p—then post-process p with BR_ℓ to obtain an optimal predictor according to any desired loss $\ell \in \mathcal{L}$.

After comprehending the definition of omniprediction, it is straightforward to see that omnipredictors exist, because the Bayes optimal predictor $p^*(x) = \Pr[y \mid x]$ satisfies the guarantee for every $\mathcal{L}$ and $\mathcal{H}$. In particular, for every loss ℓ, $\mathrm{BR}_\ell \circ p^*$ is the statistically optimal predictor of y given x. Thus, the salient question is not whether omniprediction is possible, but whether *efficient* omniprediction is possible. *Can omnipredictors be identified in finite samples? If so, can we get our hands on an omnipredictor of bounded complexity (independent of the complexity of p^*)?*

Starting with the work of Gopalan, Kalai, Reingold, Sharan, and Wieder [Gop+22], there has been a flurry of results establishing the efficiency of omniprediction, focused primarily on the binary prediction setting where $\mathcal{Y} = \{0, 1\}$. Much of the progress stems from a follow-up work of Gopalan, Hu, Kim, Reingold, and Wieder [Gop+23], which reframed the goal of omniprediction through the lens of Outcome Indistinguishability [Dwo+21]. Outcome indistinguishability (OI) compares statistical tests in the "real" world $(x, y) \sim \mathcal{D}$ and a "modeled" world $(x, \tilde{y}) \sim \mathcal{D}(p)$ where outcomes $\tilde{y} \sim p(x)$ are resampled according to the predictor p. The OI framing provides a general recipe for achieving omniprediction for any class of losses, through the following two types of indistinguishability tests.

$$\begin{aligned} \textbf{Hypothesis OI:} \quad & \mathbb{E}[\ell(h(x), y)] \approx_\varepsilon \mathbb{E}[\ell(h(x), \tilde{y})] && \forall \ell \in \mathcal{L}, h \in \mathcal{H} \\ \textbf{Decision OI:} \quad & \mathbb{E}[\ell(\mathrm{BR}_\ell \circ p(x), y)] \approx_\varepsilon \mathbb{E}[\ell(\mathrm{BR}_\ell \circ p(x), \tilde{y})] && \forall \ell \in \mathcal{L} \end{aligned}$$

A predictor $p : \mathcal{X} \to \Delta_K$ that satisfies both OI conditions is an $(\mathcal{L}, \mathcal{H}, 2\varepsilon)$-omnipredictor, because

$$\mathbb{E}[\ell(\mathrm{BR}_\ell \circ p(x), y)] \approx_\varepsilon \mathbb{E}[\ell(\mathrm{BR}_\ell \circ p(x), \tilde{y})] \leq \mathbb{E}[\ell(h(x), \tilde{y})] \approx_\varepsilon \mathbb{E}[\ell(h(x), y)] \tag{1.1}$$

where the two approximate equalities hold by Decision OI and Hypothesis OI, respectively, and the inequality holds because $\mathrm{BR}_\ell \circ p$ is the Bayes optimal predictor *in the modeled world* where outcomes are sampled according to $\tilde{y} \sim p(x)$.

Given the connections between OI and Multicalibration [Héb+18], the OI framework immediately suggests an algorithm for learning omnipredictors for rich classes of loss functions. In particular, Gopalan, Hu, Kim, Reingold, and Wieder [Gop+23] demonstrated how to implement Hypothesis OI as a Multiaccuracy condition [KGZ19] and used ℓ_1-calibration to implement Decision OI. Subsequent works have investigated how to achieve these OI conditions more efficiently, particularly focused on eliminating the need for full ℓ_1-calibration. This line has culminated in a number of independent works [Gar+24; OKK25; Dwo+25; HTY25] establishing that, in the binary setting, omniprediction is possible at essentially the same rate[1] $\widetilde{O}(1/\varepsilon^2)$ as loss minimization for a single loss function. In particular, this near-optimal rate is attainable even for omniprediction with respect to the essentially any learnable class of loss functions [OKK25], including, e.g., the class $\mathcal{L}_{\text{Lip}}$ of bounded losses that are 1-Lipschitz in the action, and the class $\mathcal{L}_{\text{prop}}$ of all bounded proper losses.

Proper Calibration. The omnipredictors of Okoroafor, Kleinberg, and Kim [OKK25] achieve near-optimal rates by improving over the use of ℓ_1-calibration to implement Decision OI with respect to all loss functions $\mathcal{L}_{\text{All}}$. They observed that, even though Decision OI is defined in terms of arbitrary losses, it suffices to focus on tests defined by proper[2] losses. Specifically, for any loss ℓ, we can define a "properized" version $\ell^{\text{pr}}(q, y) = \ell(\text{BR}_\ell(q), y)$ for all $q \in [0, 1]$ and $y \in \{0, 1\}$. This observation motivated their definition of *proper calibration* (which they show is equivalent to Decision OI) in terms of tests defined by the discrete derivatives of proper losses. That is, for any loss $\ell : \mathcal{A} \times \{0, 1\} \to [-1, 1]$,

$$\begin{aligned}\mathbb{E}[\ell(\text{BR}_\ell \circ p(x), y)] - \mathbb{E}[\ell(\text{BR}_\ell \circ p(x), \tilde{y})] &= \mathbb{E}[\ell^{\text{pr}}(p(x), y)] - \mathbb{E}[\ell^{\text{pr}}(p(x), \tilde{y})] \\ &= \mathbb{E}[\Delta\ell^{\text{pr}}(p(x)) \cdot (y - p(x))]\end{aligned}$$

where $\Delta\ell(p) := \ell(p, 1) - \ell(p, 0)$. Leveraging known structure of proper losses (specifically, the V-shaped loss basis [Li+22; Kle+23]), Okoroafor, Kleinberg, and Kim [OKK25] developed a strategy for achieving Decision OI—certifying that $\mathbb{E}[\Delta\ell(p(x)) \cdot (y - p(x))]$ is small for all proper losses $\ell \in \mathcal{L}_{\text{prop}}$—along with Hypothesis OI, to obtain the near-optimal statistical rate $\widetilde{O}(1/\varepsilon^2)$ for omniprediction.

With such a clear picture of omniprediction in the binary prediction setting, research in the area has turned towards the multi-class prediction setting, where $\mathcal{Y} = \{0, \ldots, K - 1\}$ for $K \in \mathbb{N}$. Recent work of Hu, Tian, and Yang [HTY26] demonstrated that, in principle, it is possible to generalize the algorithmic framework of Okoroafor, Kleinberg, and Kim [OKK25] to the multi-class setting. In particular, they observe that the argument that the discrete derivatives of properized losses capture decision calibration carries through in the multi-class setting. They give an omnipredictor construction for linear hypotheses for losses derived from Generalized Linear Models $\mathcal{L}_{\text{GLM}}$, at a statistical rate of $O(1/\varepsilon^{K+1})$.

Our Question. Optimizing a multi-class predictor for a single loss function is achievable at rate $O(1/\varepsilon^2)$. In contrast, the best upper bound for multi-class omniprediction with respect to any rich loss class achieves rate scaling as $O(1/\varepsilon^{K+1})$. In this work, we further investigate the complexity of multi-class omniprediction, aiming to pin down the right statistical rate as a function of $1/\varepsilon$ and K.

[1] In the introduction, we focus on the dependence on ε, and drop terms related to the hypothesis class $\mathcal{H}$.

[2] A loss function ℓ whose actions are predictions is *proper* if $p \in \text{BR}_\ell(p)$ for all $p \in [0, 1]$, that is, if reporting the true label distribution minimizes expected loss. In the K-class setting, the same definition applies with $p \in \Delta_K$.

1.1 Main Results

We obtain novel upper and lower bounds for the statistical rates of omniprediction in the multi-class setting. In particular, we show that—in stark contrast to the binary prediction setting—*multi-class omniprediction cannot generically be achieved at the rate of optimizing a single loss function.*

Lipschitz Omnipredictors. Our primary result is a bound, tight in ε up to logarithmic factors, on the statistical complexity of Omniprediction with respect to all Lipschitz loss functions.

Theorem 1.2 (Sample Complexity of Lipschitz Omniprediction, Informal)**.** *Fix the class $\mathcal{L}_{\mathrm{Lip}}$ of bounded loss functions $\ell : \Delta_K \times \mathcal{Y} \to [-1, 1]$ that are 1-Lipschitz in the action with respect to the Euclidean norm. For all $K \geq 2$, there is an algorithm for K-class prediction that for any hypothesis class $\mathcal{H}$ and $\varepsilon > 0$, learns an $(\mathcal{L}_{\mathrm{Lip}}, \mathcal{H}, \varepsilon)$-omnipredictor using sample complexity*

$$n_K \leq \tilde{O}\left(\frac{C_K}{\varepsilon^{\max\{K-1,(K+2)/2\}}} + \frac{d_{\mathcal{H}}}{\varepsilon^2}\right)$$

where $C_K > 0$ is a constant depending only on K and $d_{\mathcal{H}}$ is the fat-shattering dimension of $\mathcal{H}$ at scale ε/C_K. Further, there exists a hypothesis class $\mathcal{H}$, such that for sufficiently small $\varepsilon > 0$, any algorithm that learns a $(\mathcal{L}_{\mathrm{Lip}}, \mathcal{H}, \varepsilon)$-omnipredictor requires nearly matching sample complexity

$$n_K \geq \Omega\left(\frac{c_K}{\varepsilon^{\max\{K-1,(K+2)/2\}}}\right)$$

where $c_K > 0$ is a constant depending only on K.

Note that the fat-shattering dimension $d_{\mathcal{H}}$ (at scale ε/C_K) of a multi-class hypothesis class $\mathcal{H}$ is at most its pseudo-dimension and at most $\log|\mathcal{H}|$ for finite $\mathcal{H}$. The sample complexity of our algorithm depends explicitly on $d_{\mathcal{H}}/\varepsilon^2$, in particular, to enforce Hypothesis OI. The lower bound already holds when the benchmark class consists of the single identity hypothesis. Thus, supporting all bounded Lipschitz losses can require learning substantial information about the conditional label distribution even when the benchmark decision rule is known in advance. The hypothesis-dependent term is also necessary in the worst case, already for classes of binary classifiers, that is, hypotheses with values in $\{e_0, e_1\}$ (Section 3.5.2). As such, neither term of our upper bound can be improved beyond logarithmic factors.

In all, we characterize the sample complexity of omniprediction for Lipschitz losses, for all discrete outcome sizes: for binary ($K = 2$), we recover the $\sim 1/\varepsilon^2$ dependence on ε, for $K \geq 4$, we establish a tight rate of $\sim 1/\varepsilon^{K-1}$, and curiously, for $K = 3$, the rate goes as $\sim 1/\varepsilon^{5/2}$.

Impossibility of Proper Omnipredictors. Our second main result establishes that our focus on omniprediction for Lipschitz losses is essential. Specifically, we show that learning omnipredictors for the class of (possibly non-Lipschitz) proper losses $\mathcal{L}_{\mathrm{prop}}$ is simply impossible.

Theorem 1.3 (No finite sample complexity for Proper Omniprediction, Informal)**.** *For every $K \geq 3$, there exists a hypothesis class $\mathcal{H}$ such that, for every learner with a finite sample budget, there is a distribution $\mathcal{D}$ on which the learner's output has expected $(\mathcal{L}_{\mathrm{prop}}, \mathcal{H})$-omniprediction error at least $1/4$.*

This impossibility is especially surprising given that, in the binary prediction setting, omniprediction for all proper losses can be achieved at essentially the same rate as minimizing a single loss function [OKK25]. In this sense, our construction crucially leverages the $K \geq 3$ label space.

Our contributions. Our main results follow from the development of new algorithmic tools for multi-class omniprediction, as well as a new family of lower bound constructions. On the algorithmic side, we build on the frameworks of Gopalan, Hu, Kim, Reingold, and Wieder [Gop+23] and Okoroafor, Kleinberg, and Kim [OKK25], where we leverage two parallel online learners to bound the Decision OI and Hypothesis OI regret simultaneously. The resulting upper bound is generic (e.g., does not require the loss class to be Lipschitz), and scales with the complexity of each learner. Our key innovation is a novel randomized prediction strategy that allows us to achieve Decision OI for all losses surprisingly efficiently. Under the hood, we use calibration weights derived from the geometry of proper losses in high dimensions, paired with a chaining argument to achieve the $\tilde{O}(1/\varepsilon^{K-1})$ upper bound.

On the lower bound side, we develop a set of tools for showing how, given a sufficiently expressive loss class, omniprediction (or, relatedly, Decision plus Hypothesis OI) requires memorizing a large number of data points, even with a very limited hypothesis class. Note that Decision OI on its own can be achieved efficiently by simply predicting the marginal distribution on labels. Our constructions show, however, that a very simple Hypothesis OI condition is enough to eliminate the trivial strategy reporting a fixed prediction for all $x \in \mathcal{X}$. Intuitively, our constructions exploit the enormity of "simple" loss classes in high-dimensions: to give a guarantee for all losses in the class requires memorizing one of many different label patterns (each corresponding to a loss). Our results are similar in spirit to a lower bound of Hu and Vadhan [HV25], though we use different techniques and arrive at quantitatively improved bounds.

Organization of Manuscript. Next, in Section 1.2, we give a technical overview of our contributions. This overview contains more detailed statements of our results and provides intuitive overviews of our techniques and proofs. The overview can be read independently of the remainder of the manuscript. It is written for a reader who already has some familiarity with the Loss OI framework [Gop+23] in the binary setting, and algorithms for attaining Loss OI via Blackwell Approachability [OKK25; HTY26]. Starting in Section 2, we give formal preliminaries, Section 3 presents our algorithmic results, and Section 4 presents our lower bounds.

1.2 Technical Overview of Contributions

We work in the Loss OI framework of Gopalan, Hu, Kim, Reingold, and Wieder [Gop+23]: a predictor that satisfies both Decision OI and Hypothesis OI is an omnipredictor, as derived in (1.1). Analogous to the binary prediction setting, each of these indistinguishability conditions equivalently asks for the prediction residual to have low correlation with a class of test functions, derived from $\mathcal{L}$ and $\mathcal{H}$.

$$\textbf{Hypothesis OI: } \textit{(Residual Form)} \qquad |\mathbb{E}[\langle \Delta\ell(h(x)), e_y - p(x)\rangle]| \le \varepsilon \qquad \forall \ell \in \mathcal{L}, h \in \mathcal{H}$$

$$\textbf{Decision OI: } \textit{(Residual Form)} \qquad |\mathbb{E}[\langle \Delta\ell^{\mathrm{pr}}(p(x)), e_y - p(x)\rangle]| \le \varepsilon \qquad \forall \ell \in \mathcal{L}$$

Here, e_y is a one-hot encoding of the outcome y. We use $\Delta\ell(a) = (\ell(a,y) - \ell(a,0))_{y=0}^{K-1}$ to represent the discrete derivative of ℓ. Importantly, $\ell^{\mathrm{pr}}(p,y) = \ell(\mathrm{BR}(p), y)$ is the properized form of ℓ, which applies the omniprediction best-response post-processing internally. Hypothesis OI, which uses the tests $\Delta\ell(h(x))$ evaluated at a benchmark action, is a form of multiaccuracy [KGZ19]. Decision OI, which uses the *decision weights* $\Delta\ell^{\mathrm{pr}}(p(x)) = \Delta\ell(\mathrm{BR}_\ell(p(x)))$ evaluated at the action chosen from the prediction itself, is a form of weighted multi-class calibration, akin to proper calibration [OKK25].

1.2.1 The algorithm

Following Okoroafor, Kleinberg, and Kim [OKK25] and Hu, Tian, and Yang [HTY26], we build an online learning algorithm based on Blackwell Approachability [Bla56; ABH11], which simultaneously controls the maximum regret for each OI condition. Specifically, we use a separate online learner for the Hypothesis and Decision OI test families and choose predictions so that the tests have non-positive combined correlation with the residual. The combined regrets of the learners then bound the OI errors, and in sum, the Omniprediction regret. The statistical rates follow from an online-to-batch conversion.

Establishing an upper bound in this framework, then, reduces to bounding the complexity of the learners for Hypothesis OI and Decision OI. It turns out that standard approaches (akin to the weak agnostic learning approaches of prior works) suffice to achieve Hypothesis OI efficiently, so much of our work is focused on achieving Decision OI at the same time.

Covering the Decision Weights. The most direct way to handle the decision weights $\{\Delta\ell : \ell \in \mathcal{L}_{\text{prop}}\}$ is to discretize predictions $p(x)$ to a grid $G_\varepsilon \subseteq \Delta_K$, whose probabilities are all multiples of ε, as is standard in algorithms for calibration and omniprediction [ABH11; Gar+24; OKK25; HTY26]. Then, to guarantee that $|\mathbb{E}[\langle \Delta\ell(p(x)), e_y - p(x)\rangle]|$ is small, we can ensure that $|\mathbb{E}[\langle w(p(x)), e_y - p(x)\rangle]|$ is small for every function $w : G_\varepsilon \to [-2,2]^K$. Rounding the value of w at each grid point to an ε-net of $[-2,2]^K$ gives a uniform ε-cover of these functions, with log covering number $O_K(|G_\varepsilon| \log(1/\varepsilon)) = \tilde{O}_K(\varepsilon^{-(K-1)})$. Running exponential weights with this cover attains regret scaling as $\tilde{O}_K(\sqrt{\varepsilon^{-(K-1)}/n})$, so $n = \tilde{O}_K(\varepsilon^{-(K+1)})$ rounds suffice, roughly matching the $O(\varepsilon^{-(K+1)})$ bound of Hu, Tian, and Yang [HTY26].

Note, however, that this approach ignores any structure in the collection of decision weights. Our approach exploits this structure. The decision weight functions arise from the discrete derivatives of properized loss functions. An alternative view (developed in Section 2.3) casts the weight functions as the (super)gradients of the *Bayes risk* of a given ℓ. For a proper loss $\ell \in \mathcal{L}_{\text{prop}}$, the Bayes risk $\underline{L}_\ell(p)$ characterizes the best possible loss achievable under a fixed label distribution $p \in \Delta_K$; that is,

$$\underline{L}_\ell(p) = \mathbb{E}_{y\sim p}[\ell(p, y)].$$

Importantly, the Bayes risk is a concave function on the simplex, so we can appeal to known covers for concave functions in building a cover for the decision weights. Concave functions have small uniform covers at coarse accuracy (Lemma 3.10), suggesting that we could cover the decision weights by first covering the risks.

A problem with this approach, however, presents itself immediately: two Bayes risks can be uniformly close without having close gradients. The risks in Figure 1(a) differ by at most δ everywhere, but their gradients differ by 2 at the marked prediction in (b). Thus, using a cover of the Bayes risks to get to the decision weights feels doomed: a cover of the risks gives no cover of the weights at any nontrivial accuracy. Similarly, even for a single loss, an arbitrarily small change in the prediction can cause a large change in the gradient, so some randomization strategy is necessary to guarantee low regret.

Smoothing through randomized predictions. We mitigate both difficulties by committing, before learning, to randomize every reported prediction over a small window of width r. The learner chooses a base prediction $q = q(x)$, but always reports

$$p = (1-r)q + rz, \qquad z \sim \text{Unif}(\Delta_K),$$

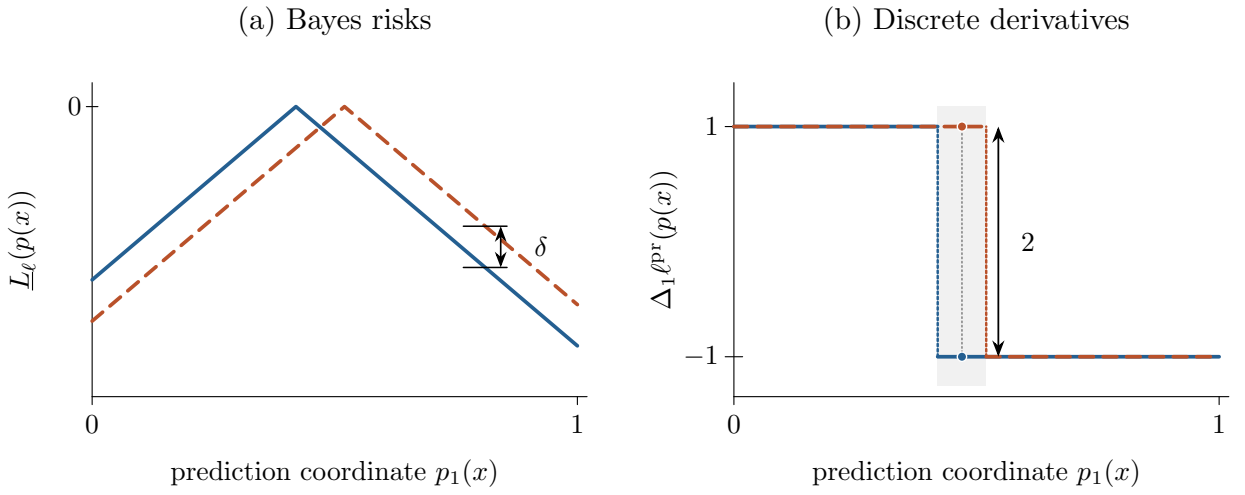


Figure 1: Bayes risks and decision weights in a binary example. The horizontal coordinate is the predicted probability $p_1(x)$ of label 1. Blue solid and orange dashed curves correspond to ℓ and ℓ', respectively, in both panels. (a) The two Bayes risks are uniformly δ-close. (b) Their label-1 decision weights differ by 2 at the marked prediction. The shaded band lies between the two risk peaks; colored dotted segments mark the jumps in the decision weights.

using fresh randomness independent of the example and base prediction. For two labels, this reports a uniformly random prediction from an interval of length r. Decision OI averages over this reporting randomness. This is what allows us to work with the averaged weights shown in Figure 2(b), rather than the pointwise weights in Figure 1(b), at the cost of an $O(r)$ error.

To see this, first replace the reported prediction p by the base prediction q in the calibration residual. Since losses are bounded and $\|p - q\|_1 \le 2r$, this costs only $O(r)$. For fixed q and label y,

$$\begin{aligned}\mathbb{E}_z[\langle \Delta\ell^{\mathrm{pr}}(p), e_y - p\rangle] &= \mathbb{E}_z[\langle \Delta\ell^{\mathrm{pr}}(p), e_y - q\rangle] + O(r) \\ &= \langle \mathbb{E}_z[\Delta\ell^{\mathrm{pr}}(p)], e_y - q\rangle + O(r).\end{aligned}$$

Once the residual is fixed at q, the expectation acts only on the decision weight. Thus the learner can use the *smoothed decision weights*

$$w_{\ell,r}(q) := \mathbb{E}_z\left[\Delta\ell^{\mathrm{pr}}\big((1-r)q + rz\big)\right].$$

Controlling their correlations at the base predictions $q(x)$ controls decision OI at the predictions $p(x)$ that we actually report, up to $O(r)$ additional error. Critically, our commitment to this reporting rule lets us *train the base predictor against these averaged tests* from the outset.

Our randomization strategy addresses the two difficulties above.

- *Nearby Bayes risks give nearby smoothed weights.* For two labels, write $\underline{L}_\ell(p)$ for the Bayes risk of ℓ when label 1 has probability p. At base prediction q, the reporting interval is $[a, a+r]$, where $a = (1-r)q_1$. The label-1 decision weight agrees with $\underline{L}'_\ell(t)$ almost everywhere, so its average is

$$\big(w_{\ell,r}(q)\big)_1 = \frac{1}{r}\int_a^{a+r} \underline{L}'_\ell(t)\,dt = \frac{\underline{L}_\ell(a+r) - \underline{L}_\ell(a)}{r}.$$

  The average slope is the rise across the window divided by its width. Changing the Bayes risk by at most δ changes each endpoint height by at most δ, so the average changes by at most $2\delta/r$. This explains the smaller separation in Figure 2(b).

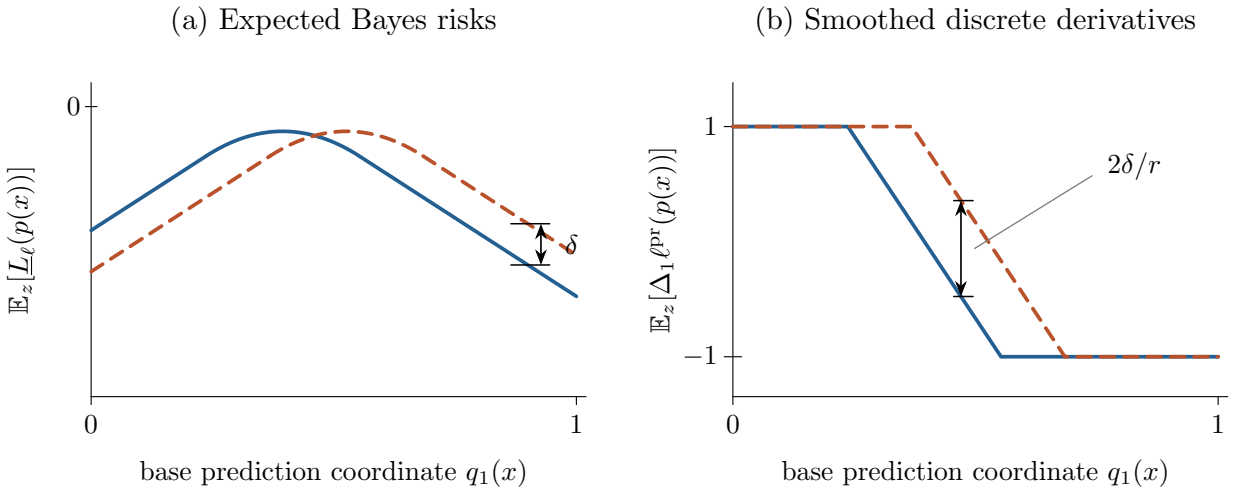


Figure 2: Expected Bayes risks and smoothed decision weights for the two binary losses in Figure 1. The horizontal coordinate is the base probability $q_1(x)$ of label 1. The reported prediction is $p(x) = (1-r)q(x) + rz$, where $z \sim \mathrm{Unif}(\Delta_2)$. Blue solid and orange dashed curves correspond to ℓ and ℓ', respectively. (a) The expected Bayes risks remain uniformly δ-close. (b) The averaged label-1 decision weights are continuous and have maximum separation $2\delta/r$ in this example, where $\delta \le r$.

For K labels, the corresponding bound is $2(K-1)\delta/r$. Consequently, a uniform cover of the Bayes risks at accuracy $r\varepsilon/[2(K-1)]$ gives an ε-cover of the smoothed decision weights. With r proportional to ε, the required accuracy is a constant multiple of ε^2 for fixed K. We can therefore construct covers using the concavity of the Bayes risks, rather than trying to approximate their discontinuous derivatives directly.

- *The smoothed weight of each loss is Lipschitz in the base prediction.* Moving q changes the averaging window, rather than evaluating a discontinuous weight at a single new point. Quantitatively, $w_{\ell,r}$ is $O_K(1/r)$-Lipschitz from ℓ_1 to ℓ_∞. The jumps in Figure 1(b) become the continuous ramps in Figure 2(b). We use this continuity when choosing the base prediction, which, unlike on the grid, need not be randomized.

  A wider window strengthens both stability bounds, but increases the $O(r)$ cost of replacing p by q in the residual. We choose r to be a sufficiently small multiple of ε to keep this cost within the target accuracy.

Covering and chaining. Smoothing alone does not improve the rate. By the concave-function covering bound (Lemma 3.10), the ε^2-accurate cover of the Bayes risks above has size M with $\log(2M) = O_K(\varepsilon^{-(K-1)})$, and it gives an ε-cover of the smoothed decision weights. A standard online-learning algorithm competing with these M bounded functions has average regret $O(\sqrt{\log(2M)/n})$, so this analysis needs n of order $\varepsilon^{-(K+1)}$ rounds, the same rate as on the grid, so it would seem we've made no progress.

On the grid $G_\varepsilon \subseteq \Delta_K$ discussed earlier, rounding the weight values at these points to accuracy u gives a cover with logarithmic size $O_K(\varepsilon^{-(K-1)} \log(2/u))$, so this bound retains the factor $\varepsilon^{-(K-1)}$ even at constant accuracy.

In contrast, the smoothed decision weights admit *much smaller covers at coarser scales*. A cover at accuracy u follows from a cover of the Bayes risks at accuracy proportional to εu. At constant

accuracy, the logarithmic cover size is only $O_K(\varepsilon^{-(K-1)/2})$. The regret bound is proportional to the magnitude of the functions being learned. We can therefore learn a coarse approximation from a smaller cover and use the larger fine covers only for small corrections. This approach is known as *chaining* [Dud67].

For each decision weight, choose approximations from fixed covers at accuracies 2ε and ε. Their difference has uniform norm $O(\varepsilon)$.

Learning these small corrections reduces the average regret from

$$O_K\left(\sqrt{\frac{\varepsilon^{-(K-1)}}{n}}\right) \qquad \text{to} \qquad O_K\left(\varepsilon\sqrt{\frac{\varepsilon^{-(K-1)}}{n}}\right).$$

For this correction, the number of rounds needed to make the regret $O(\varepsilon)$ drops from order $\varepsilon^{-(K+1)}$ to order $\varepsilon^{-(K-1)}$.

The 2ε-cover is still large, so using it in place of the ε-cover would give essentially the same bound as before. We instead compare these approximations with those at accuracy 4ε, then 8ε, and continue to constant accuracy. Each fine approximation is then its coarsest approximation plus the successive small differences. The covers get smaller at coarser accuracies, and only the coarsest approximations are learned at constant magnitude. The correlation with a fine approximation is the sum of the correlations with these pieces, so we can learn the pieces separately and add the resulting functions. Their regret bounds add as well. Writing $M(u)$ for the cover size at accuracy u, the resulting average regret is bounded by

$$O\left(r + \underbrace{\sqrt{\frac{\log(2M(1))}{n}}}_{\text{coarse approximation}} + \underbrace{\sum_{u\in\{r,2r,4r,\ldots,1/2\}} u\sqrt{\frac{\log(2M(u))}{n}}}_{\text{fine corrections}}\right).$$

The $O(r)$ term is the error of the final approximation.

The concave-function covering bound gives $\log(2M(u)) = O_K((\varepsilon u)^{-(K-1)/2})$, so the contribution at accuracy u is

$$O_K\left(\frac{\varepsilon^{-(K-1)/4}}{\sqrt{n}}\, u^{(5-K)/4}\right).$$

For $K > 5$, the finest scale $u \asymp \varepsilon$ dominates, giving the desired $\varepsilon^{-(K-1)}$ dependence. For $K = 5$, all scales contribute equally, while for $K = 3, 4$ the coarser scales dominate.

In Section 3.4.2 we show how to obtain a smaller cover at coarse scales by reducing the problem to covering concave functions in $K - 2$ variables. For $K \geq 3$, Lemma 3.11 gives

$$\log(2M(u)) = O_K\left(\frac{1}{u}(\varepsilon u)^{-(K-2)/2}\right).$$

This has the same order as the first bound at the finest scale $u \asymp \varepsilon$, but is smaller at coarser scales. The corresponding contribution to average regret is

$$O_K\left(\frac{\varepsilon^{-(K-2)/4}}{\sqrt{n}}\, u^{(4-K)/4}\right).$$

Now the finest scale dominates for every $K \geq 5$, giving $O_K(\varepsilon^{-(K-1)})$ rounds. For $K = 4$, all scales contribute equally, giving $\widetilde{O}(\varepsilon^{-3})$ rounds. For $K = 3$, the coarsest scales dominate, with total average regret $O(\varepsilon^{-1/4}/\sqrt{n})$, giving $O(\varepsilon^{-5/2})$ rounds. Together with the binary rate $\widetilde{O}(\varepsilon^{-2})$, these bounds give the first term of Theorem 1.2.

Theorem 1.4 (Decision learner, informal version of Lemma 3.14)**.** *Fix $K \geq 2$ and $0 < \varepsilon < 1$. With smoothing width $r = c_K\varepsilon$ for a sufficiently small constant $c_K > 0$ depending only on K, the smoothed decision weights $\{w_{\ell,r} : \ell \in \mathcal{L}_{\mathrm{All}}\}$ admit an online learner with average regret at most ε after*

$$\widetilde{O}_K\left(\varepsilon^{-\max\{K-1,(K+2)/2\}}\right)$$

rounds, for any sequence of base predictions and labels.

Combined with a suitable hypothesis learner and the online-to-batch conversion, this gives the first term of Theorem 1.2. We next construct the hypothesis learner for Lipschitz losses.

Hypothesis learners for Lipschitz losses. Our most general theorem (Theorem 3.8) takes as input a hypothesis learner for the loss and hypothesis classes. To obtain our bound for Lipschitz losses, we construct an explicit learner for $\Delta\mathcal{L}_{\mathrm{Lip}} \circ \mathcal{H}$.

Theorem 1.5 (Hypothesis learner for Lipschitz losses, informal version of Lemma 3.16)**.** *For finite $\mathcal{H}$, the tests $\Delta\mathcal{L}_{\mathrm{Lip}} \circ \mathcal{H}$ admit an online learner with average regret at most ε after*

$$\widetilde{O}_K\left(\varepsilon^{-\max\{2,K-1\}} + \frac{\log(2|\mathcal{H}|)}{\varepsilon^2}\right)$$

rounds.

The functions in $\Delta\mathcal{L}_{\mathrm{Lip}}$ are themselves Lipschitz, so a small change in h causes only a small change in $\Delta\ell \circ h$. This allows us to cover $\Delta\mathcal{L}_{\mathrm{Lip}}$ and $\mathcal{H}$ separately: composing the two covers gives a cover of $\Delta\mathcal{L}_{\mathrm{Lip}} \circ \mathcal{H}$.

We cover $\Delta\mathcal{L}_{\mathrm{Lip}}$ using standard covers of Lipschitz functions on the action simplex at successive accuracies. We then apply chaining to the composed covers, as we did for the decision weights. The contribution from $\Delta\mathcal{L}_{\mathrm{Lip}}$ is no larger, up to constants depending on K, than the decision learner's regret. For finite $\mathcal{H}$, we therefore obtain the decision learner's sample rates with an additional $O_K(\log(2|\mathcal{H}|)/\varepsilon^2)$ sample term. For infinite $\mathcal{H}$, we first use a sample to build a finite cover of $\mathcal{H}$ (Section 3.5.2). The hypothesis term then becomes $\widetilde{O}_K(d_{\mathcal{H}}/\varepsilon^2)$, where $d_{\mathcal{H}}$ is the fat-shattering dimension of $\mathcal{H}$ at a scale proportional to ε.

GLM losses. To isolate the improvement from our decision learner, we also consider the GLM losses and linear benchmarks studied by Hu, Tian, and Yang [HTY26]. Combining their hypothesis learner with our decision learner improves the sample bound, for fixed K, from $\widetilde{O}_K(\varepsilon^{-(K+1)})$ to $\widetilde{O}_K(\varepsilon^{-\max\{K-1,(K+2)/2\}})$. In particular, for $K \geq 4$, the exponent drops from $K+1$ to $K-1$. We give this application in Section 3.5.3.

1.2.2 Lower bounds

We now show that these rates are necessary. We prove lower bounds both for omniprediction itself, for Lipschitz losses, all proper losses and Lipschitz proper losses, and for our sufficient condition, Loss OI. All of our constructions force the predictor to memorize hidden label information at many contexts.

Recall Theorem 1.2 gives an $\Omega_K(\varepsilon^{-\max\{K-1,(K+2)/2\}})$ sample lower bound for omniprediction under bounded Lipschitz losses, even against the single identity hypothesis. This matches the hypothesis-independent term of our upper bound up to logarithmic factors. Theorem 4.8 shows

that omniprediction over all proper losses can be statistically impossible for $K \geq 3$, already for two hypotheses.

We also prove two further lower bounds: one for our sufficient condition for omniprediction, proper calibration together with multiaccuracy, and one for omniprediction with the narrower class of Lipschitz proper losses.

Theorem 1.6 (Lower bound for proper calibration and multiaccuracy, informal version of Theorem 4.7)**.** *Fix $K \geq 2$. For every sufficiently small $\varepsilon > 0$, achieving expected error at most ε for both proper calibration and the identity multiaccuracy condition, using the same possibly randomized predictor, requires*

$$\Omega_K\left(\varepsilon^{-(K+2)/2}\right)$$

independent samples in the worst case.

Theorem 1.6 isolates a statistical cost of requiring the two Loss OI conditions to hold simultaneously. For each fixed K, proper calibration alone can be achieved with $O_K(\varepsilon^{-2})$ samples by predicting the empirical marginal label distribution at every context. The identity multiaccuracy condition alone can also be achieved with $O(\varepsilon^{-2})$ samples. Yet for every $K \geq 3$, requiring the same predictor to satisfy both conditions needs $\Omega_K(\varepsilon^{-(K+2)/2})$ samples in the worst case. Thus, even adding a single fixed multiaccuracy test can make proper calibration statistically more expensive. This shows why achieving Decision OI and Hypothesis OI separately at the binary statistical rate does not suffice to extend that rate to multiclass omniprediction through Loss OI: the additional cost can arise from requiring both guarantees of the same predictor.

Theorem 1.7 (Lower bound for Lipschitz proper losses, informal version of Theorem 4.10)**.** *Fix $K \geq 2$. For every sufficiently small $\varepsilon > 0$, there are a finite context space and two hypotheses such that omniprediction for all bounded, 1-Lipschitz proper losses, with expected error at most ε, requires*

$$\Omega_K\left(\varepsilon^{-(K+1)/2}\right)$$

independent samples in the worst case, even for possibly randomized predictors.

Both proper-loss results hold even if the learner returns a separate, possibly randomized decision rule for each loss, all learned from the same sample. The guarantee must still hold simultaneously over all losses. Thus, the statistical cost persists even without a common predictor or its best-response postprocessing.

In the binary prediction setting, Okoroafor, Kleinberg, and Kim [OKK25] show that for a fixed finite benchmark class, $\widetilde{O}(1/\varepsilon^2)$ samples suffice simultaneously for all bounded Lipschitz losses, regardless of the complexity of the true distribution p^*. A nearly matching lower bound $\Omega(1/\varepsilon^2)$ already arises at a single context whose conditional label mean is either $1/2 + \varepsilon$ or $1/2 - \varepsilon$.

Our multiclass lower bounds show that an omnipredictor can be forced to memorize far more information about p^*, even when competing with only one or two hypotheses. In all of our constructions, we take a packing of the simplex (or a lower-dimensional slice) at scale ε or $\sqrt{\varepsilon}$ as our context space and sample uniformly from it. We encode independent hidden signs σ in the conditional label distributions at these contexts. At each context x, the hidden sign either specifies a deterministic binary label or determines the direction of a small perturbation of x that gives the true conditional label distribution $p^*(x)$. We show that any sufficiently accurate omnipredictor p yields an estimate $\widehat{\sigma}$ that agrees with most of the hidden signs in σ on average.

Because p only learns about σ through the sample, the sample must contain enough information for this recovery. This requires observing most contexts, with repeated observations needed when

the two possible label distributions are close. The main challenge (and the art of our lower bounds) is exhibiting a family of "simple" losses (that lie in the corresponding loss class) that force this recovery.

Lipschitz losses. We exhibit two families giving lower bounds of $\Omega_K(\varepsilon^{-(K-1)})$ and $\Omega_K(\varepsilon^{-(K+2)/2})$. In fact, we prove these lower bounds leveraging a hypothesis class $\mathcal{H} = \{h_{\mathrm{id}}\}$ with a single hypothesis, the identity function $h_{\mathrm{id}}(x) = x$. The lower bounds crucially exploit the best-response post-processing in the definition of omniprediction, and thus, non-proper Lipschitz losses. Of course, if the omnipredictor could elect to output its action directly, it could simply copy identity.

These lower bounds, in some sense, pathologically exploit the definition of omniprediction, but our constructions expose how in high dimensions, "simple" classes of losses like $\mathcal{L}_{\mathrm{Lip}}$ can be very expressive. Our first construction will exploit the following property: the Lipschitz losses $\mathcal{L}_{\mathrm{Lip}}$ contain a huge collection of losses $\mathcal{L} \subseteq \mathcal{L}_{\mathrm{Lip}}$, where every loss $\ell \in \mathcal{L}$ has the same properized loss. That is, for all $\ell, \ell' \in \mathcal{L}$, and for all $y \in \mathcal{Y}$, the properized losses $\ell(\mathrm{BR}_\ell(\cdot), y) = \ell'(\mathrm{BR}_{\ell'}(\cdot), y)$ are the same. The exact construction of this $\mathcal{L}$ is technical, but it builds a collection of losses $\mathcal{L}$ which all have the same loss "profiles" $\{\ell(p, y)_{y\in\mathcal{Y}} : p \in \Delta_K\}$, where each specific profile arises from a different loss-prediction pair. Further, each of these losses in $\mathcal{L}$ will be associated with some hidden labeling σ of some contexts packed into Δ_K.

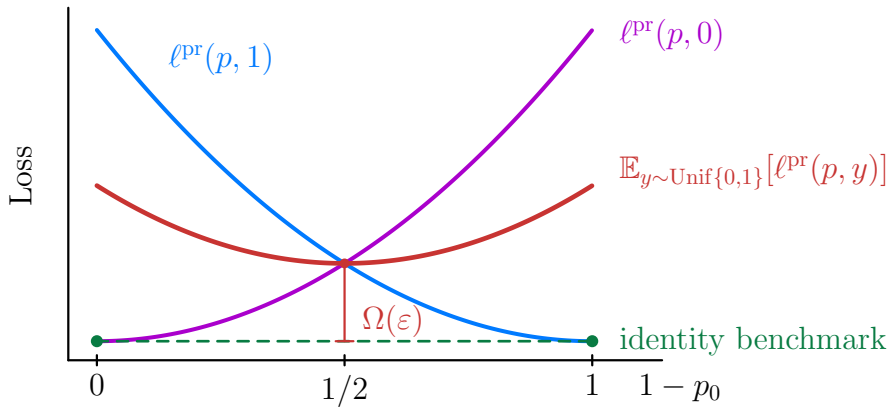


Figure 3: The purple and blue curves show the common properized losses $\ell^{\mathrm{pr}}(p,0)$ and $\ell^{\mathrm{pr}}(p,1)$, and the red curve shows their uniform average. The green points and dashed line mark the identity benchmark's loss, equal to both label-specific minima. The uniform average stays $\Omega(\varepsilon)$ above the benchmark for every prediction p.

We assign each context an independent fair label in $\{0,1\}$, fixed across samples. At an unseen context, the label remains uniform and independent of the learned prediction. The blue and purple curves show $\ell(p,1)$ and $\ell(p,0)$. The red curve records the expected properized loss of the predictor on a uniform random label: it is simply the average of the blue and purple curves. The predictor cannot hope to do better than this red curve on unseen contexts, but even at its lowest point the red curve is $\Omega(\varepsilon)$ worse than optimal.

For every hidden labeling σ, the identity hypothesis attains the label-specific minimum under ℓ_σ at every context. This benchmark is shown in green. Even the lowest point of the red curve lies $\Omega(\varepsilon)$ above it, so every unseen context contributes $\Omega(\varepsilon)$ expected excess loss. We choose the constant in this gap sufficiently large. Since identity is optimal at every context, these errors cannot

be canceled by gains elsewhere. Achieving expected regret at most ε therefore requires observing a constant fraction of the contexts.

The sample complexity is thus determined by the number of contexts. To keep the loss Lipschitz under our construction, the contexts must be spaced out by at least a multiple of ε. We can pack $\Theta_K(\varepsilon^{-(K-1)})$ contexts into Δ_K at this spacing. With contexts sampled uniformly, expected omniprediction error at most ε requires observing a constant fraction of them on average, giving the first lower bound.

Unlearnability of Proper Losses. The preceding construction uses nonproper losses to separate the benchmark action from the action selected by best responding. For proper losses, the identity map is a valid best response, so choosing these best responses lets us copy a single benchmark with zero error. With two hypotheses, however, properness alone does not ensure learnability. If we drop the Lipschitz requirement, omniprediction for all bounded proper losses is unlearnable already for three labels, even against two fixed affine hypotheses. The key to our construction is a family of proper losses that allow for independent choices of loss across all hypothesis predictions.

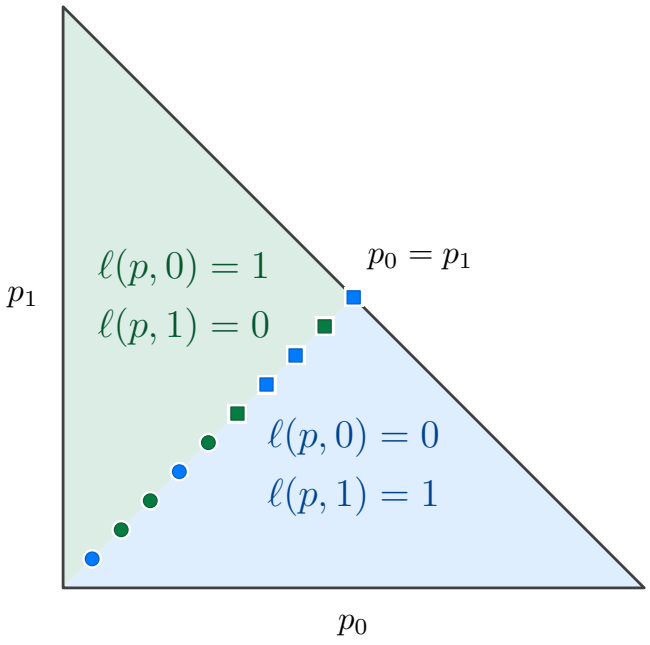


Figure 4: The simplex Δ_3 in coordinates (p_0, p_1), with $p_2 = 1 - p_0 - p_1$ and $\ell(p, 2) = 0$ throughout. On $p_0 = p_1$, either region's loss vector can be chosen independently at each report while preserving properness. Circles mark $h_0(x_i)$ and squares mark $h_1(x_i)$. All marked reports are distinct, so each marker's color can be chosen independently.

The freedom in Figure 4 comes from the ties permitted by properness. Each report assigns loss zero to one of labels 0 and 1 and loss one to the other, while label 2 always has loss zero. Hence $\ell(a, 0) + \ell(a, 1) = 1$ for every report a. For a true distribution p with $p_0 = p_1$,

$$\begin{aligned}\mathbb{E}_{y\sim p}[\ell(a, y)] &= p_0\ell(a, 0) + p_1\ell(a, 1)\\ &= p_0\big(\ell(a, 0) + \ell(a, 1)\big) = p_0.\end{aligned}$$

Every report in the simplex has this same expected loss, so truthful reporting is optimal.

Away from the line, truthful reporting assigns loss zero to the more likely of labels 0 and 1. An arbitrary report has expected loss either p_0 or p_1, while the truthful report attains the smaller:

$$\mathbb{E}_{y\sim p}[\ell(a,y)] \geq \min\{p_0, p_1\} = \mathbb{E}_{y\sim p}[\ell(p,y)].$$

Properness thus leaves the choices at different tie points completely independent. Each point may receive either $(0,1,0)$ or $(1,0,0)$, regardless of the choices at every other point; every such assignment defines a proper loss.

If all the hypothesis predictions $h_0(x_i), h_1(x_i)$ are distinct points on the tie line, their loss vectors can be assigned independently. For any fixed binary labels at the contexts, a single proper loss can therefore realize *any prescribed pattern of zero and one losses*, independently for each hypothesis at each context. For example, when the label at x_i is 1, the vectors $(1,0,0)$ and $(0,1,0)$ give $h_0(x_i)$ losses zero and one, respectively. Because all the predictions are distinct, this choice constrains none of the other hypothesis losses.

The third label supplies arbitrarily many points for these independent choices: $p_0 = p_1$ is a whole line segment in Δ_3, whereas in the binary simplex it is only $(1/2, 1/2)$. Two fixed affine hypotheses can place all their reports at distinct points on this line, for arbitrarily many contexts.

To see why a benchmark can do well while the learner cannot, choose the context labels independently and uniformly from $\{0,1\}$, fixed across samples. After seeing a small sample, most labels remain unknown. Consider one possible completion of these labels. Using the independent choices at the benchmark reports, choose a proper loss for which h_0 has loss zero and h_1 has loss one at every context under this completion.

Now flip all the unseen labels, leaving the observed labels unchanged and keeping the same loss. At every unseen context, h_1 now has loss zero and h_0 has loss one. Thus h_0 is perfect in the first case, while h_1 is perfect on all the unseen contexts in the second case.

The learner sees the same sample and is evaluated using the same loss in both cases, so it uses the same decision rule. At each unseen context, the two labels are opposite, and $\ell(a,0) + \ell(a,1) = 1$ for every report a. Its loss averaged over the two cases is therefore $1/2$, regardless of where it reports, also after randomizing over reports. In particular, copying h_0 succeeds in the first case and fails in the second; copying h_1 reverses the problem.

These paired completions are equally likely conditional on the sample. When almost all the probability mass is unseen, a benchmark is nearly perfect in each case, but the learner's loss averaged over the two cases remains close to $1/2$. Since omniprediction must compete with both hypotheses for every loss, this gives a constant expected regret lower bound.

The stronger Lipschitz bound for three labels. Returning to Lipschitz losses, our first construction gives only an $\Omega(\varepsilon^{-2})$ lower bound when $K = 3$. To obtain the $\Omega(\varepsilon^{-5/2})$ bound, we replace deterministic labels with small perturbations of the conditional label distribution, so that recovering each hidden sign requires repeated observations.

The contexts lie on the slice $x_0 = x_1$ of the simplex, which is a line segment when $K = 3$. At context x_i, the true distribution $p^*(x_i)$ is x_i perturbed by $\Theta(\varepsilon)$ in one of the two directions perpendicular to the slice. Distinguishing these two possibilities requires $\Omega(\varepsilon^{-2})$ observations at that context. We construct losses for which small omniprediction error against the identity hypothesis forces the predictor to identify the perturbation direction at most contexts.

We keep the identity action x_i optimal under $p^*(x_i)$ but add a decoy that prevents predictions on the wrong side from selecting it. The decoy beats x_i whenever the predicted distribution lies on the slice or on the wrong side of it, but costs $\Omega(\varepsilon)$ more under the true distribution. Figure 5 shows the expected losses of the identity action in green and the decoy in red.

The remaining issue is to prevent a decoy from another context from providing a cheap alternative to x_i. We score the identity actions by squared loss (Brier loss), so using x_j instead of x_i under $p^*(x_i)$ costs an amount proportional to $\|x_j - x_i\|_2^2$. The decoy at x_j can reduce this expected loss by only $O(\varepsilon)$, even when the two contexts have opposite perturbations. Spacing contexts by a sufficiently large constant multiple of $\sqrt{\varepsilon}$ makes the quadratic penalty dominate this reduction, so both x_j and its decoy remain $\Omega(\varepsilon)$ worse than x_i. Figure 6 illustrates this quadratic penalty.

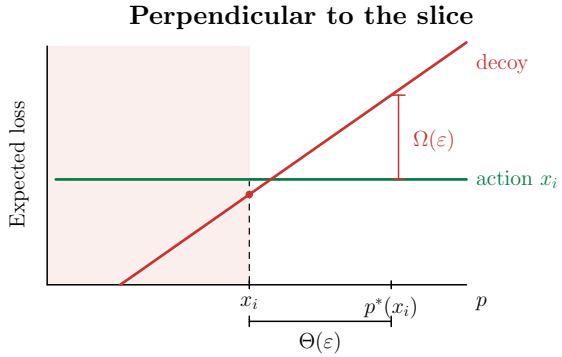


Figure 5: The red action beats the green identity action at x_i and on the shaded side. At the true distribution $p^*(x_i)$, only $\Theta(\varepsilon)$ away, the red action is $\Omega(\varepsilon)$ suboptimal.

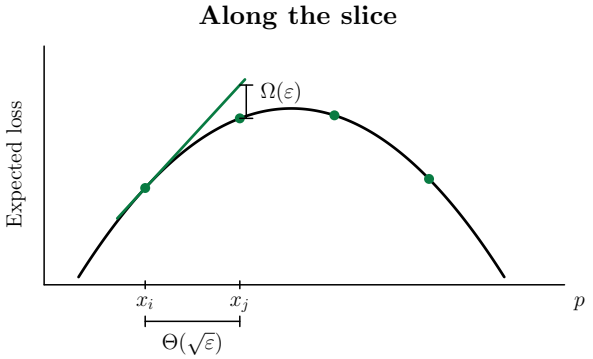


Figure 6: The black curve is optimal expected Brier loss. The green tangent gives the loss of action x_i. At another context, $\Theta(\sqrt{\varepsilon})$ away, this action is $\Omega(\varepsilon)$ suboptimal.

The full construction extends these loss assignments to all actions and rescales the losses to make them bounded and 1-Lipschitz while preserving the comparisons above. Small omniprediction error therefore forces recovery of most hidden signs on average. For $K = 3$, the line contains $\Theta(\varepsilon^{-1/2})$ contexts at the required spacing, and the $\Omega(\varepsilon^{-2})$ observations needed per context give the lower bound $\Omega(\varepsilon^{-5/2})$. More generally, the slice has dimension $K - 2$ and contains $\Theta_K(\varepsilon^{-(K-2)/2})$ contexts, giving the second Lipschitz lower bound $\Omega_K(\varepsilon^{-(K+2)/2})$.

Proper calibration and one multiaccuracy condition. The previous construction almost yields Theorem 1.6, showing that proper calibration together with a single identity multiaccuracy condition forces recovery of the hidden signs.

The key property of the second Lipschitz family is that every loss agrees with the same scaled Brier loss at the identity benchmark. Consequently, all of its hypothesis-OI tests reduce, up to a common constant factor, to the single identity correlation

$$\mathrm{MA}_{\mathrm{id}}(p) = \left|\mathbb{E}\left[\langle x, e_y - p(x)\rangle\right]\right|.$$

By (2.7), a predictor whose proper calibration and identity multiaccuracy errors are both at most ε has $O(\varepsilon)$ omniprediction error for this entire loss family. The same implication holds in expectation over the training sample and learner randomness. The second Lipschitz lower bound therefore implies that achieving expected error at most ε for both requirements needs $\Omega_K(\varepsilon^{-(K+2)/2})$ samples.

Lipschitz proper losses. The proper-loss unlearnability construction permits unit changes in loss at arbitrarily close reports. Requiring the losses to be 1-Lipschitz restores finite-sample learnability for finite benchmark classes, but the sample complexity can still have an exponent growing with K. We first explain the construction for a single shared predictor using a discontinuous proper loss,

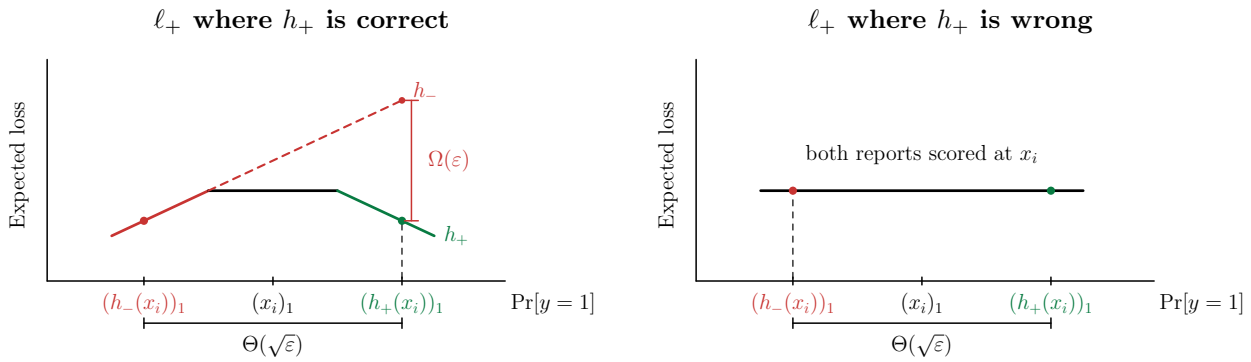
then describe the modification that makes the loss Lipschitz. The full proof in Appendix B.3 also covers a separate decision rule for each loss.

At each of many well-separated contexts x_i in the simplex, the true conditional label distribution is x_i plus or minus a small perturbation. The two hypotheses h_+ and h_- always guess the positive and negative perturbations, respectively. At each context, exactly one of the two hypotheses equals the true conditional label distribution.

We construct a loss ℓ_+ for which h_+ is optimal at every context. Where h_+ equals the true distribution, predicting the wrong perturbation incurs an $\Omega(\varepsilon)$ excess loss; at the remaining contexts, the two hypotheses tie. A second loss ℓ_- reverses the roles of the hypotheses. Because the corresponding benchmark is optimal everywhere under each loss, excess losses cannot cancel between contexts. Together, the two regret guarantees therefore force the shared predictor to recover most hidden perturbations on average.

We define each loss by choosing a set of reference predictions and evaluating squared loss at the reference nearest to the reported prediction. The loss is proper because reporting the true distribution selects the reference with the smallest expected squared loss.

The reference set for ℓ_+ contains every midpoint x_i of the two hypothesis predictions. Where h_+ is correct, we also include $h_+(x_i)$ and $h_-(x_i)$, so each hypothesis prediction is scored at itself. Where h_+ is wrong, the midpoint is the only reference point for that context, so both predictions are scored at x_i. Figure 7 shows these two cases for ℓ_+.



	h_+ correct	h_- correct
ℓ_+	h_+ strictly better	Same loss
ℓ_-	Same loss	h_- strictly better

Figure 7: The nearest-reference proper-loss construction near a context x_i. The horizontal axis varies the label distribution p along the line through $h_-(x_i)$ and $h_+(x_i)$. The solid lower envelope is the Bayes risk. Left: the truth is $p^*(x_i) = h_+(x_i)$, and the red, black, and green pieces use the references $h_-(x_i)$, x_i, and $h_+(x_i)$, respectively. The dashed red line is $\mathbb{E}_{y\sim p}[\ell_+(h_-(x_i), y)]$, which exceeds the Bayes risk by $\Omega(\varepsilon)$ at the truth. Right: the truth is $p^*(x_i) = h_-(x_i)$; both reports are scored at the reference x_i and tie. Thus h_+ is optimal at every context under ℓ_+. Red and green identify h_- and h_+, respectively. The table compares their losses under the actual truth; ℓ_- reverses the roles. Each panel suppresses a common affine term and uses schematic drawing scales.

At a context where we distinguish the hypotheses, the expected-loss gap between their predictions is proportional to their squared separation. To obtain an $\Omega(\varepsilon)$ gap, we separate $h_+(x_i)$ and $h_-(x_i)$ by a sufficiently large constant multiple of $\sqrt{\varepsilon}$. We separate the contexts by a larger constant multiple of the same scale, so each hypothesis prediction is scored using a reference point from its

own context.

The simplex contains $\Theta_K(\varepsilon^{-(K-1)/2})$ contexts at this spacing, and distinguishing the two possible truths at one context requires $\Omega(\varepsilon^{-1})$ observations. Multiplying these quantities gives the lower bound $\Omega_K(\varepsilon^{-(K+1)/2})$.

To obtain Lipschitz proper losses, we replace the selection of a single reference by a weighted mixture. Appendix B.3 gives this construction and proves that it is bounded, proper, and 1-Lipschitz, and preserves the required loss values up to a common constant factor. This yields the same sample lower bound for Lipschitz proper losses, establishing Theorem 4.10, including its stronger formulation allowing separate decision rules for individual losses.

1.3 Related Works

Omniprediction. Omniprediction was introduced by Gopalan, Kalai, Reingold, Sharan, and Wieder [Gop+22], who derived results for $\mathcal{L}_{\text{cvx}}$ the class of all Lipschitz convex loss functions. They derived the initial omniprediction result through a direct connection to Multicalibration [Héb+18]. The connection to multicalibration was developed further by Gopalan, Kim, and Reingold [GKR23], who introduced the notion of Swap Omniprediction. Subsequent work has investigated swap omniprediction in the online setting [LSS25].

A parallel line of work, initiated by Gopalan, Hu, Kim, Reingold, and Wieder [Gop+23], builds omnipredictors through Outcome Indistinguishablity [Dwo+21; Dwo+22], in both the distributional and online settings [Gar+24; OKK25; HTY25; Dwo+25; GT26]. To date, omniprediction has been considered across many prediction settings, including under constraints [Hu+23; BLR25], under performativity [KP23], and under selective prediction [CK25].

Beyond binary omniprediction. An early notable departure from binary prediction is the work on omniprediction for regression of Gopalan, Okoroafor, Raghavendra, Shetty, and Singhal [Gop+24], followed closely by the work of Lu, Roth, and Shi [LRS25] on multi-dimensional regression. Multi-class omniprediction has been considered in a few recent works [FP26; HTY26]. Farina and Perdomo [FP26] give an oracle reduction from multi-class omniprediction to online learning over the associated loss and hypothesis tests, with finite classes as one instantiation whereas Hu, Tian, and Yang [HTY26] obtain multi-class omniprediction with respect to the infinite collection of GLM losses.

Calibration and Decision OI. Decision OI was initially introduced by Zhao, Kim, Sahoo, Ma, and Ermon [Zha+21], as "decision calibration," in the context of multi-class prediction. They provided formal definitions and heuristic, practical algorithms for achieving decision calibration. Subsequent work of Gopalan, Hu, and Rothblum [GHR24] demonstrated rigorous approaches and hardness results related to decision calibration. Recognizing the connection to OI, Gopalan, Hu, Kim, Reingold, and Wieder [Gop+23] renamed the condition in the binary prediction setting to Decision OI.

As discussed, Okoroafor, Kleinberg, and Kim [OKK25] provided the first tight characterization of Decision OI in the binary prediction setting, as proper calibration. Their characterization revealed a number of equivalent ways to enforce the notion. In particular, decision OI is equivalent to a notion of calibration based on threshold test functions. Equivalent variants of this threshold-based calibration were explored in [QZ25; Ros+25].

While not directly comparable to decision calibration, U-Calibration [Kle+23], which is essentially a non-contextual, online variant of omniprediction, is a closely related concept and line of work.

Since its introduction by Kleinberg, Leme, Schneider, and Teng [Kle+23], there has been follow-up investigations, including into the multi-class setting [LSS24].

Lower bounds for calibration and multiaccuracy. Hu and Vadhan [HV25] study essentially the same combination of proper calibration and multiaccuracy, formulated through convex subgradients. We complement their $2^{\Omega(K)}$ circuit lower bound at constant accuracy with an $\Omega_K(\varepsilon^{-(K+2)/2})$ sample lower bound, which makes the accuracy dependence explicit and holds even for randomized predictors. Whereas Gibbs and Tibshirani's [GT26] binary $\Omega(\varepsilon^{-5/2})$ sample lower bound concerns full ECE calibration and multiaccuracy for deterministic predictors, our multiclass bound already holds under the weaker proper-calibration requirement.

Smoothing Predictions. The idea of adding noise to predictions to obtain omniprediction guarantees has been explored in the binary setting by Hartline, Wu, and Yang [HWY25] and Gopalan, Stavropoulos, Talwar, and Tankala [Gop+26]. A distinguishing feature of our approach is that we use the noise distribution to define the calibration conditions learned during training. Learning these conditions directly and exploiting their structure is a key ingredient in our improved learning rates for multiclass omniprediction.

Contents

1 Introduction **1**
- 1.1 Main Results . . . 3
- 1.2 Technical Overview of Contributions . . . 4
  - 1.2.1 The algorithm . . . 5
  - 1.2.2 Lower bounds . . . 9
- 1.3 Related Works . . . 16

2 Preliminaries **18**
- 2.1 Omniprediction and Loss OI . . . 19
- 2.2 Decision OI and proper calibration . . . 21
- 2.3 Geometry of the decision weights . . . 23
- 2.4 Hypothesis OI as multiaccuracy . . . 24

3 Upper bounds **24**
- 3.1 Main results . . . 25
- 3.2 Smoothed decision weights . . . 25
- 3.3 Combining decision and hypothesis learners . . . 26
- 3.4 Decision calibration learner . . . 28
  - 3.4.1 Approximating the smoothed decision weights . . . 28
  - 3.4.2 Covering the averaged weights . . . 29
  - 3.4.3 Learning from successive covers . . . 33
  - 3.4.4 The resulting regret bound . . . 35
- 3.5 Hypothesis learners . . . 35
  - 3.5.1 Lipschitz losses . . . 35
  - 3.5.2 Extending to infinite hypothesis classes . . . 38
  - 3.5.3 GLM losses . . . 40
- 3.6 From online correlations to population OI . . . 41

4 Lower bounds by forcing memorization **42**
4.1 Forcing the predictor to remember hidden signs . . . 43
4.2 Lipschitz losses . . . 44
4.2.1 Proof of the $(1/\varepsilon)^{K-1}$ lower bound . . . 45
4.2.2 Proof of the $(1/\varepsilon)^{K/2+1}$ lower bound . . . 46
4.2.3 Dependence on the benchmark class . . . 47
4.3 Proper calibration and identity multiaccuracy . . . 48
4.4 Omniprediction is unlearnable for proper losses . . . 50
4.5 Lipschitz proper losses . . . 52

A Deferred proofs from Section 3 **54**
A.1 Proof of the smoothing bounds . . . 54
A.2 Proofs of the covering bounds . . . 55
A.2.1 Simplex cover . . . 55
A.2.2 The second covering bound . . . 55
A.3 Online learning tools . . . 57
A.4 Proof of the population guarantee . . . 59
A.5 Proof of the empirical-cover lemma . . . 63
A.6 Applying the GLM hypothesis learners . . . 64
A.6.1 Normalization and the hypothesis tests . . . 64
A.6.2 Matching the borrowed learners . . . 64
A.6.3 Proof of Lemma 3.19 . . . 65
A.6.4 Combining the learners and proving the corollary . . . 66

B Deferred proofs from Section 4 **67**
B.1 Proof of the sign decoding lemma . . . 67
B.2 Proof of the ridge-loss lemma . . . 70
B.3 Lower bound for Lipschitz proper losses . . . 72
B.3.1 Hard distributions and the required loss values . . . 72
B.3.2 Recovering the hidden signs from small regret . . . 73
B.3.3 Constructing the loss family . . . 75

C Acknowledgments and AI Usage **80**

2 Preliminaries

In this section, we establish the calibration and multiaccuracy conditions that suffice for omniprediction. We first recall the Loss OI reduction (Section 2.1), then express Decision OI as weighted calibration (Section 2.2) and relate its weights to the geometry of Bayes risks (Section 2.3). We conclude with the multiaccuracy formulation of Hypothesis OI (Section 2.4).

Prediction model. We consider a multi-class prediction setting.
For $K \geq 2$, let $\mathcal{Y} = \{0, \ldots, K-1\}$ be the label set and write

$$\Delta_K := \left\{ p \in \mathbb{R}^K_{\geq 0} : \sum_{y=0}^{K-1} p_y = 1 \right\}.$$

Let e_y be the unit vector corresponding to label y. We work with a distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the context space, and write $p^*(x) = \mathbb{E}[e_y \mid x]$ for the true conditional label distribution.

A predictor reports a probability vector $p(x) \in \Delta_K$ at context x, possibly at random.[3] All expectations include prediction randomness, which is independent of the label conditional on x.

Let $\mathcal{A}$ be an action space and let $\mathcal{H} \subseteq \mathcal{A}^{\mathcal{X}}$ be a nonempty class of benchmark decision rules. Actions need not be discrete. For the Lipschitz and proper losses below, we take $\mathcal{A} = \Delta_K$, so an action can itself be viewed as a prediction of the conditional label probabilities. For each loss $\ell : \mathcal{A} \times \mathcal{Y} \to [-1, 1]$ considered, we fix a measurable *best response*

$$\mathrm{BR}_\ell(p) \in \arg\min_{a \in \mathcal{A}} \mathbb{E}_{y \sim p}\big[\ell(a, y)\big], \qquad p \in \Delta_K.$$

Throughout, $O_K(\cdot)$, $\Omega_K(\cdot)$, and $\Theta_K(\cdot)$ hide factors depending only on K, and $\widetilde{O}_K(\cdot)$ additionally hides polylogarithmic factors.

2.1 Omniprediction and Loss OI

Omniprediction asks for a single predictor whose best-response post-processing competes with every benchmark, simultaneously for all losses in a class. We consider classes of bounded losses $\mathcal{L} \subseteq \{\ell : \mathcal{A} \times \mathcal{Y} \to [-1, 1]\}$.

Definition 2.1 (Omnipredictor)**.** The omniprediction error of p is

$$\mathrm{OmniReg}_{\mathcal{L},\mathcal{H};\mathcal{D}}(p) := \sup_{\ell \in \mathcal{L},\, h \in \mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\ell(\mathrm{BR}_\ell(p(x)), y) - \ell(h(x), y)\big].$$

A predictor p is an ε-omnipredictor for $(\mathcal{L}, \mathcal{H})$ under $\mathcal{D}$ if $\mathrm{OmniReg}_{\mathcal{L},\mathcal{H};\mathcal{D}}(p) \le \varepsilon$.

We omit $\mathcal{D}$ from the notation when it is fixed and clear from context.

Notable loss classes. We use the following classes of $[-1, 1]$-valued losses.

- $\mathcal{L}_{\mathrm{All}}$ consists of all bounded decision losses admitting best responses as above, allowing arbitrary action spaces. We use this class in discussing Decision OI.
- $\mathcal{L}_{\mathrm{Lip}}$ consists of losses on $\Delta_K \times \mathcal{Y}$ that are 1-Lipschitz in the action with respect to the Euclidean norm,[4]

$$|\ell(a, y) - \ell(a', y)| \le \|a - a'\|_2 \qquad (a, a' \in \Delta_K,\ y \in \mathcal{Y}).$$

- $\mathcal{L}_{\mathrm{prop}}$ consists of proper losses on $\Delta_K \times \mathcal{Y}$, meaning that

$$\mathbb{E}_{y\sim p}\big[\ell(p, y)\big] \le \mathbb{E}_{y\sim p}\big[\ell(a, y)\big] \qquad (a, p \in \Delta_K).$$

  The minimizing action need not be unique.

[3] Formally, a randomized predictor maps $\mathcal{X}$ to the space $\Delta(\Delta_K)$ of probability distributions on Δ_K; $p(x)$ denotes the reported vector.

[4] Using the ℓ_1 or ℓ_∞ norm instead changes only K-dependent constants in our bounds.

Loss Outcome Indistinguishability. Loss OI [Gop+23] applies Outcome Indistinguishability [Dwo+21] to omniprediction. As discussed in the introduction, the selected action has no larger expected loss than any benchmark under the predictor's own label distribution. The two Loss OI conditions transfer this comparison to the true distribution.

For an example $(x, y) \sim \mathcal{D}$, the *real world* evaluates losses using the observed label y. The *modeled world* instead uses a label $\widetilde{y} \sim p(x)$.

Loss OI consists of two conditions. *Decision OI* compares the expected loss of the selected action $\mathrm{BR}_\ell(p(x))$ in the two worlds, with error

$$\mathrm{DecOI}_{\mathcal{L}}(p) := \sup_{\ell \in \mathcal{L}} \left| \mathbb{E}_{\substack{(x,y)\sim\mathcal{D} \\ \widetilde{y}\sim p(x)}} \left[\ell(\mathrm{BR}_\ell(p(x)), y) - \ell(\mathrm{BR}_\ell(p(x)), \widetilde{y}) \right] \right|.$$

Hypothesis OI makes the same comparison for each benchmark action $h(x)$, with error

$$\mathrm{HypOI}_{\mathcal{L},\mathcal{H}}(p) := \sup_{\ell \in \mathcal{L},\, h \in \mathcal{H}} \left| \mathbb{E}_{\substack{(x,y)\sim\mathcal{D} \\ \widetilde{y}\sim p(x)}} \left[\ell(h(x), y) - \ell(h(x), \widetilde{y}) \right] \right|.$$

In the modeled world, $\mathrm{BR}_\ell(p(x))$ has no larger expected loss than any benchmark action $h(x)$. Decision OI transfers the selected action's loss to this world, and Hypothesis OI transfers the benchmark's loss back. The resulting comparison in the real world gives the omniprediction guarantee, as shown in Figure 8.

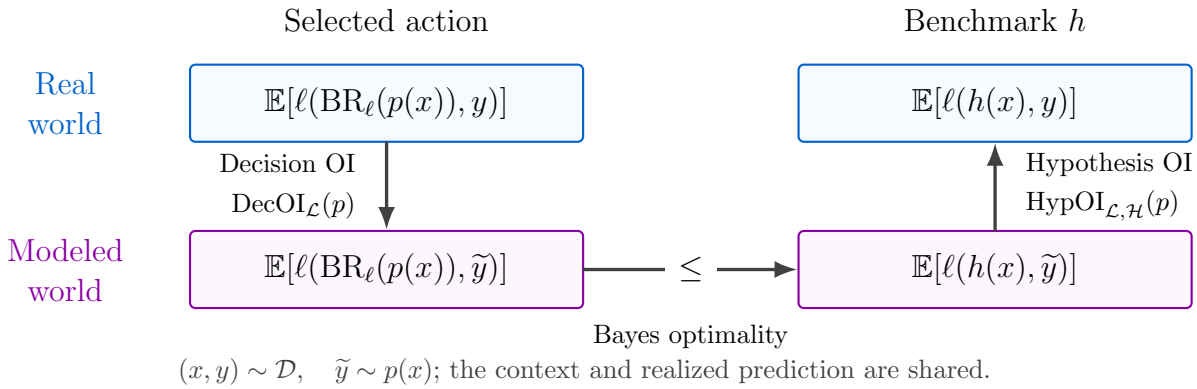


Figure 8: The Loss OI reduction [Gop+23]. Decision OI transfers the selected action's expected loss to the modeled world, where best-response optimality gives the comparison with the benchmark. Hypothesis OI transfers the benchmark's loss back to the real world. The two transfers incur the indicated additive errors.

Theorem 2.2 (Loss OI reduction [Gop+23])**.** *For every predictor p,*

$$\mathrm{OmniReg}_{\mathcal{L},\mathcal{H}}(p) \leq \mathrm{DecOI}_{\mathcal{L}}(p) + \mathrm{HypOI}_{\mathcal{L},\mathcal{H}}(p).$$

Proof. For any $\ell \in \mathcal{L}$ and $h \in \mathcal{H}$, with $\widetilde{y} \sim p(x)$,

$$\begin{aligned}
&\mathbb{E}\big[\ell(\mathrm{BR}_\ell(p(x)), y) - \ell(h(x), y)\big] \\
&= \underbrace{\mathbb{E}\big[\ell(\mathrm{BR}_\ell(p(x)), y) - \ell(\mathrm{BR}_\ell(p(x)), \widetilde{y})\big]}_{\leq \mathrm{DecOI}_{\mathcal{L}}(p)} \\
&\quad + \underbrace{\mathbb{E}\big[\ell(\mathrm{BR}_\ell(p(x)), \widetilde{y}) - \ell(h(x), \widetilde{y})\big]}_{\leq 0 \text{ by best-response optimality}} \\
&\quad + \underbrace{\mathbb{E}\big[\ell(h(x), \widetilde{y}) - \ell(h(x), y)\big]}_{\leq \mathrm{HypOI}_{\mathcal{L},\mathcal{H}}(p)}.
\end{aligned}$$

Best-response optimality holds for each context and reported prediction. Taking the supremum over ℓ and h proves the claim. □

2.2 Decision OI and proper calibration

Loss OI compares an action's expected loss under two label distributions. For a fixed action a, moving probability from label 0 to label y changes its expected loss at rate $\ell(a, y) - \ell(a, 0)$. We collect these differences in the *discrete derivative*

$$\Delta\ell(a) := \big(\ell(a, y) - \ell(a, 0)\big)_{y=0}^{K-1}.$$

We write $\Delta_y \ell(a)$ for coordinate y, so $\Delta_0 \ell(a) = 0$ and $\Delta\ell(a) \in [-2, 2]^K$.

For Decision OI, the action is selected from the prediction. We incorporate this choice into the *properized loss*

$$\ell^{\mathrm{pr}}(p, y) := \ell(\mathrm{BR}_\ell(p), y).$$

Lemma 2.3 (Decision OI as weighted calibration)**.** *For every bounded loss class $\mathcal{L}$ and predictor p,*

$$\mathrm{DecOI}_{\mathcal{L}}(p) = \sup_{\ell \in \mathcal{L}} \Big| \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\langle \Delta\ell^{\mathrm{pr}}(p(x)), e_y - p(x)\rangle\big] \Big|. \tag{2.1}$$

Every properized loss ℓ^{pr} is proper.

The functions $p \mapsto \Delta\ell^{\mathrm{pr}}(p)$ are the *decision weights*. The lemma identifies each expected weighted residual with the loss discrepancy tested by Decision OI. Since every properized loss is proper, controlling these tests for all bounded proper losses controls Decision OI for every bounded decision loss. This is the multiclass extension [HTY26] of the binary Loss OI and proper-calibration representations [Gop+23; OKK25].

Proof. For an action a and label distribution $p \in \Delta_K$, the expected loss is affine in p,

$$\mathbb{E}_{y\sim p}\big[\ell(a, y)\big] = \ell(a, 0) + \sum_{y=1}^{K-1} p_y\big(\ell(a, y) - \ell(a, 0)\big) = \ell(a, 0) + \langle \Delta\ell(a), p\rangle. \tag{2.2}$$

Consequently,

$$\begin{aligned}
\ell(a, y) - \mathbb{E}_{\widetilde{y}\sim p}\big[\ell(a, \widetilde{y})\big] &= \langle \ell(a, \cdot), e_y - p\rangle \\
&= \langle \Delta\ell(a), e_y - p\rangle.
\end{aligned} \tag{2.3}$$

Subtracting the label-0 loss from every coordinate leaves the inner product unchanged because the coordinates of $e_y - p$ sum to zero. For the selected action, $\Delta\ell(\mathrm{BR}_\ell(p(x))) = \Delta\ell^{\mathrm{pr}}(p(x))$. Taking expectations and the supremum over ℓ gives (2.1).

The properized loss is proper because, for all $p, p' \in \Delta_K$,

$$\mathbb{E}_{y\sim p}\big[\ell^{\mathrm{pr}}(p,y)\big] = \min_{a\in\mathcal{A}} \mathbb{E}_{y\sim p}\big[\ell(a,y)\big] \le \mathbb{E}_{y\sim p}\big[\ell^{\mathrm{pr}}(p',y)\big].$$

□

Definition 2.4 (Proper calibration)**.** The proper-calibration error is

$$\mathrm{PCal}_K(p) := \sup_{\ell\in\mathcal{L}_{\mathrm{prop}}} \Big|\mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\langle \Delta\ell(p(x)), e_y - p(x)\rangle\big]\Big|.$$

For $K = 2$, this is the proper-calibration error of Okoroafor, Kleinberg, and Kim [OKK25], identifying a prediction with its coordinate $p_1(x)$.

For comparison, the usual ℓ_1 expected calibration error averages the magnitude of the mean residual conditional on the realized prediction. Equivalently, it allows every bounded measurable weight depending on that prediction,

$$\begin{aligned}\mathrm{ECE}_1(p) &:= \mathbb{E}\Big[\big\|\mathbb{E}[e_y - p(x) \mid p(x)]\big\|_1\Big] \\ &= \sup_{w:\Delta_K\to[-1,1]^K} \Big|\mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\langle w(p(x)), e_y - p(x)\rangle\big]\Big|.\end{aligned}$$

The second expression follows by taking each coordinate of w to be the sign of the corresponding conditional mean residual.

Proposition 2.5 (Decision OI and calibration)**.** *For every bounded loss class $\mathcal{L}$ and predictor p,*

$$\mathrm{DecOI}_{\mathcal{L}}(p) \le \mathrm{DecOI}_{\mathcal{L}_{\mathrm{All}}}(p) = \mathrm{PCal}_K(p) \le \mathrm{ECE}_1(p).$$

Proof. The first inequality follows from inclusion of the loss classes. Every properized loss belongs to $\mathcal{L}_{\mathrm{prop}}$, so Lemma 2.3 gives $\mathrm{DecOI}_{\mathcal{L}_{\mathrm{All}}}(p) \le \mathrm{PCal}_K(p)$.

For the reverse inequality, a strictly proper perturbation removes any dependence on the choice among tied best responses. For $\ell \in \mathcal{L}_{\mathrm{prop}}$, the perturbation

$$\ell_\eta(a,y) := (1-\eta)\ell(a,y) + \frac{\eta}{2}\|a - e_y\|_2^2, \qquad 0 < \eta < 1$$

remains $[-1,1]$-valued and satisfies

$$\begin{aligned}&\mathbb{E}_{y\sim p'}\big[\ell_\eta(a,y) - \ell_\eta(p',y)\big] \\ &\quad = (1-\eta)\mathbb{E}_{y\sim p'}\big[\ell(a,y) - \ell(p',y)\big] + \frac{\eta}{2}\|a-p'\|_2^2 \ge \frac{\eta}{2}\|a-p'\|_2^2.\end{aligned}$$

Its unique best response to p' is therefore p'. The proper-calibration test for ℓ_η is consequently one of the tests in $\mathrm{DecOI}_{\mathcal{L}_{\mathrm{All}}}(p)$. Letting η tend to zero gives the same bound for ℓ. Taking the supremum over $\ell \in \mathcal{L}_{\mathrm{prop}}$ proves the reverse inequality.

Finally, subtracting the label-0 loss does not change the residual inner product,

$$\langle \Delta\ell(p(x)), e_y - p(x)\rangle = \langle \ell(p(x),\cdot), e_y - p(x)\rangle.$$

The loss vector $\ell(p(x),\cdot)$ lies in $[-1,1]^K$, so it is an allowed weight in the definition of $\mathrm{ECE}_1(p)$. Thus $\mathrm{PCal}_K(p) \le \mathrm{ECE}_1(p)$. □

Proper calibration restricts the tests in expected calibration error to weights supplied by proper losses.

2.3 Geometry of the decision weights

The decision weights come from actions that minimize expected loss at the prediction. We study their structure by comparing each action's expected loss with the minimum attainable loss as the label distribution varies.

Here $p \in \Delta_K$ is the prediction, while $p' \in \Delta_K$ is the label distribution under which the selected action is evaluated. For $p = p(x)$, the modeled loss uses $p' = p(x)$ and the real loss uses $p' = p^*(x)$.

Each action determines an affine expected-loss function of p', as in (2.2). The *Bayes risk* is the smallest expected loss attainable at each label distribution,

$$\underline{L}_\ell(p') := \min_{a \in \mathcal{A}} \mathbb{E}_{y \sim p'}\big[\ell(a, y)\big] = \mathbb{E}_{y \sim p'}\big[\ell^{\mathrm{pr}}(p', y)\big].$$

Its graph is therefore the lower envelope of these affine functions. The action selected at prediction p has expected loss

$$\mathbb{E}_{y \sim p'}\big[\ell^{\mathrm{pr}}(p, y)\big] = \underline{L}_\ell(p) + \langle \Delta\ell^{\mathrm{pr}}(p), p' - p \rangle. \tag{2.4}$$

The coefficients $\Delta\ell^{\mathrm{pr}}(p)$ in this affine function are precisely the decision weights in (2.1). At $p' = p$, the action is optimal, so the function touches the Bayes risk. It lies on or above the Bayes risk everywhere else, because the risk is the minimum over actions [GR07]. Figure 9 illustrates the binary case.

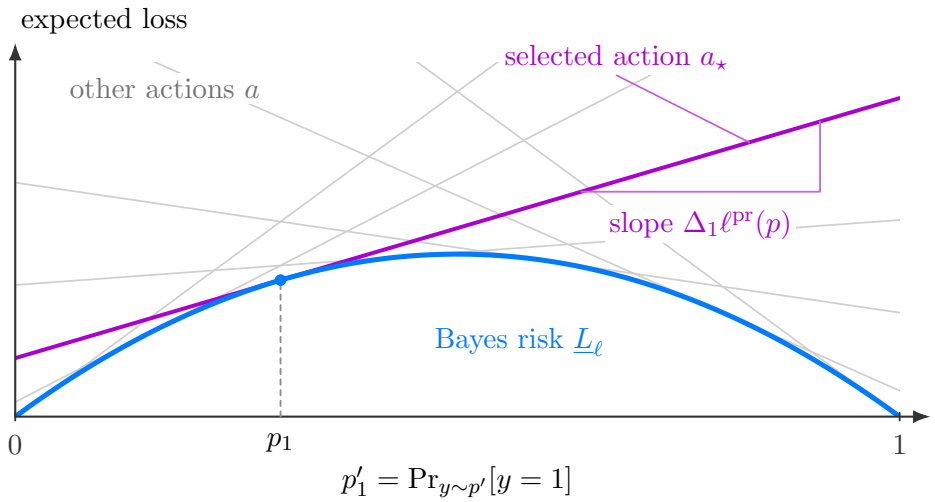


Figure 9: Decision weights for binary Brier loss. Each action determines an affine expected-loss function of the label probability p'_1. The Bayes risk is their lower envelope. At prediction p, the selected action $a_\star = \mathrm{BR}_\ell(p)$ gives a line that touches the Bayes risk at p_1 and has slope $\Delta_1\ell^{\mathrm{pr}}(p)$.

Lemma 2.6 (Bayes risk and decision weights)**.** *The Bayes risk $\underline{L}_\ell$ is concave and satisfies*

$$|\underline{L}_\ell(p)| \le 1, \qquad |\underline{L}_\ell(p') - \underline{L}_\ell(p)| \le \|p' - p\|_1 \qquad (p, p' \in \Delta_K).$$

Moreover,

$$\underline{L}_\ell(p') \le \underline{L}_\ell(p) + \langle \Delta\ell^{\mathrm{pr}}(p), p' - p \rangle. \tag{2.5}$$

Thus $\Delta\ell^{\mathrm{pr}}(p)$ is a supergradient of the Bayes risk at p.

Proof. The minimum of affine functions is concave, and expected losses lie in $[-1, 1]$. The action selected at p remains available at p', so $\underline{L}_\ell(p') \le \mathbb{E}_{y\sim p'}[\ell^{\mathrm{pr}}(p, y)]$. Together with (2.4), this gives the claimed inequality. The Lipschitz bound follows by expressing the same increment using the original loss vector,

$$\underline{L}_\ell(p') - \underline{L}_\ell(p) \le \langle \ell^{\mathrm{pr}}(p, \cdot), p' - p\rangle \le \|p' - p\|_1,$$

since $\ell^{\mathrm{pr}}(p, \cdot) \in [-1, 1]^K$. Interchanging p and p' proves the claim. □

When $K = 2$, the slope $\Delta_1\ell^{\mathrm{pr}}(p)$ is nonincreasing in p_1, since it is a selected supergradient of a concave function on an interval. This recovers the threshold structure used by Okoroafor, Kleinberg, and Kim [OKK25].

The concavity and Lipschitz bounds control the size of covers of the Bayes risks. Averaging over nearby predictions transfers these approximations to the averaged decision weights, giving the finite test families used by the decision learner in Section 3.4.

2.4 Hypothesis OI as multiaccuracy

For a loss ℓ and benchmark h, the action $h(x)$ is fixed by the context. Applying (2.3) to this action gives

$$\ell(h(x), y) - \mathbb{E}_{\widetilde{y}\sim p(x)}\big[\ell(h(x), \widetilde{y})\big] = \langle \Delta\ell(h(x)), e_y - p(x)\rangle.$$

The loss and benchmark determine the context-dependent test $x \mapsto \Delta\ell(h(x))$. We collect these tests in the class

$$\Delta\mathcal{L} \circ \mathcal{H} := \{x \mapsto \Delta\ell(h(x)) : \ell \in \mathcal{L},\ h \in \mathcal{H}\}.$$

Each test takes values in $[-2, 2]^K$. For a class $\mathcal{G}$ of maps $\mathcal{X} \to [-2, 2]^K$, the multiaccuracy error is

$$\mathrm{MA}_{\mathcal{G}}(p) := \sup_{g\in\mathcal{G}} \Big|\mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\langle g(x), e_y - p(x)\rangle\big]\Big|.$$

Hypothesis OI therefore requires multiaccuracy against every test induced by a loss and a benchmark,

$$\mathrm{HypOI}_{\mathcal{L},\mathcal{H}}(p) = \mathrm{MA}_{\Delta\mathcal{L}\circ\mathcal{H}}(p). \tag{2.6}$$

Together with Theorem 2.2 and Proposition 2.5, this yields

$$\mathrm{OmniReg}_{\mathcal{L},\mathcal{H}}(p) \le \mathrm{PCal}_K(p) + \mathrm{MA}_{\Delta\mathcal{L}\circ\mathcal{H}}(p). \tag{2.7}$$

The bound separates the universal proper-calibration requirement from the multiaccuracy requirement associated with $\mathcal{L}$ and $\mathcal{H}$. In the next section, we control both by combining learners for the two test families.

3 Upper bounds

We first state our upper bound for Lipschitz omniprediction and a general reduction for bounded losses in Section 3.1. Sections 3.2 and 3.3 then introduce smoothed decision weights and an algorithm for jointly controlling Decision OI and Hypothesis OI. Section 3.4 analyzes the decision weights using covers and chaining. We next treat the hypothesis tests for Lipschitz losses, extend the result to infinite benchmark classes, and give the GLM application. Finally, Section 3.6 converts the online guarantees to expected population OI.

3.1 Main results

Theorem 3.1 (Omniprediction for all Lipschitz Losses)**.** *Fix $K \geq 2$, $0 < \varepsilon < 1$, and a class $\mathcal{H}$ of N hypotheses $\mathcal{X} \to \Delta_K$. There is an algorithm that uses*

$$\widetilde{O}_K\left(\varepsilon^{-\max\{K-1,(K+2)/2\}} + \frac{\log(2N)}{\varepsilon^2}\right)$$

independent samples from any distribution $\mathcal{D}$ and returns a predictor p with

$$\mathbb{E}\big[\mathrm{OmniReg}_{\mathcal{L}_{\mathrm{Lip}},\mathcal{H}}(p)\big] \leq \varepsilon.$$

For $0 < \delta < 1$, replacing $\log(2N)$ by $\log(2N/\delta)$ in the sample bound gives an ε-omnipredictor with probability at least $1 - \delta$ over training.

For $K = 2$, we recover the known $1/\varepsilon^2$ dependence on accuracy for binary omniprediction, up to logarithmic factors [OKK25]. The first term grows as $\varepsilon^{-5/2}$ with three labels and as $\varepsilon^{-(K-1)}$ with $K \geq 4$ labels, again up to logarithmic factors.

For fixed K, this upper bound and Theorem 4.6 determine the minimax sample complexity for expected omniprediction error, up to logarithmic factors. The worst case ranges over context spaces, benchmark classes of size N, and data distributions. The first term is necessary even for the identity benchmark; the dependence on N already arises for a single loss. Section 3.5.2 extends Theorem 3.1 to infinite classes, replacing $\log(2N)$ by the fat-shattering dimension of $\mathcal{H}$ at scale $c_K\varepsilon$ (Theorem 3.18). In the worst case, this dependence cannot be improved beyond logarithmic factors; for classes of $\{e_0, e_1\}$-valued hypotheses, it is tight for every class.

The first $\varepsilon^{-\max\{K-1,(K+2)/2\}}$ term in Theorem 3.1 comes from the decision learner of Section 3.4, whose regret bound (Lemma 3.14) controls Decision OI simultaneously for all bounded losses, $\mathrm{DecOI}_{\mathcal{L}_{\mathrm{All}}}$. This part of the guarantee is therefore the same for every bounded loss class $\mathcal{L}$. The additional sample requirement comes from learning the hypothesis-test class

$$\Delta\mathcal{L} \circ \mathcal{H} = \{x \mapsto \Delta\ell(h(x)) : \ell \in \mathcal{L},\ h \in \mathcal{H}\}$$

from Section 2.4. We use an online learner whose accumulated correlation competes with the largest absolute accumulated correlation of any fixed test in this class. Write $\mathrm{Reg}_{\Delta\mathcal{L}\circ\mathcal{H}}(n)$ for its expected total regret, as defined in Definition 3.6.

Theorem 3.2 (General bounded losses)**.** *Let $\mathcal{L}$ be a class of bounded losses. Suppose an online learner for $\Delta\mathcal{L} \circ \mathcal{H}$ has expected total regret $\mathrm{Reg}_{\Delta\mathcal{L}\circ\mathcal{H}}(n)$ as defined in Definition 3.6.*

There are constants $0 < c_K \leq 1/2$ and $C_K > 0$ such that the following holds for every $0 < \varepsilon < 1$. Run Algorithm 1 with smoothing width $r = c_K\varepsilon$, the decision learner of Algorithm 3, and the given hypothesis learner. Using n independent examples from any distribution $\mathcal{D}$, it returns a predictor p satisfying

$$\mathbb{E}\big[\mathrm{OmniReg}_{\mathcal{L},\mathcal{H}}(p)\big] \leq \varepsilon$$

whenever

$$n \geq C_K\varepsilon^{-\max\{K-1,(K+2)/2\}}\log^2(2/\varepsilon), \qquad \frac{\mathrm{Reg}_{\Delta\mathcal{L}\circ\mathcal{H}}(n)}{n} \leq c_K\varepsilon.$$

3.2 Smoothed decision weights

Following the approach outlined in Section 1.2, we fix the reporting rule before learning. The learner chooses a base prediction $q(x)$, but reports

$$p(x) = (1 - r)q(x) + rz, \qquad z \sim \mathrm{Unif}(\Delta_K), \tag{3.1}$$

with fresh z independent of the example and base prediction. Define

$$w_{\ell,r}(q) := \mathbb{E}_z[\Delta\ell^{\mathrm{pr}}((1-r)q + rz)]\,. \tag{3.2}$$

We first record the two stability properties used to construct covers and choose base predictions. We then show that controlling correlations with these averaged weights and the original hypothesis tests at q controls the two OI errors at the reported prediction, with only $O(r)$ additional error. Throughout, $0 < r \le 1/2$; we eventually take r to be a sufficiently small constant multiple of ε.

Lemma 3.3 (Approximating the averaged weights)**.** *For $0 < r \le 1/2$, each $w_{\ell,r}$ takes values in $[-2,2]^K$ and is $2(K-1)/r$-Lipschitz from ℓ_1 to ℓ_∞. For any two bounded losses,*

$$\|w_{\ell,r} - w_{\ell',r}\|_\infty \le \frac{2(K-1)}{r}\|\underline{L}_\ell - \underline{L}_{\ell'}\|_\infty.$$

The proof is in Appendix A.1.

Lemma 3.4 (Decision OI after averaging)**.** *For the randomized prediction in* (3.1)*,*

$$\mathrm{DecOI}_{\mathcal{L}}(p) \le \sup_{\ell\in\mathcal{L}} \left|\mathbb{E}\big[\langle w_{\ell,r}(q(x)), e_y - q(x)\rangle\big]\right| + 2r.$$

The bound follows by writing the real-minus-modeled loss difference as

$$\begin{aligned}\mathbb{E}\big[\ell^{\mathrm{pr}}(p(x),y) - \underline{L}_\ell(p(x))\big] &= \mathbb{E}\big[\langle w_{\ell,r}(q(x)), e_y - q(x)\rangle\big] \\ &\quad + \mathbb{E}\big[\langle \ell^{\mathrm{pr}}(p(x),\cdot), q(x) - p(x)\rangle\big].\end{aligned}$$

The second term has absolute value at most $\mathbb{E}\big[\|p(x) - q(x)\|_1\big] \le 2r$. Appendix A.1 gives the full proof.

The hypothesis test $\Delta\ell(h(x))$ stays fixed as the prediction changes, so its smoothing cost comes entirely from the change in the residual.

Lemma 3.5 (Hypothesis OI is Lipschitz in prediction)**.** *For any two predictors coupled on the same context with predictions $p(x), q(x)$,*

$$\mathrm{HypOI}_{\mathcal{L},\mathcal{H}}(p) \le \mathrm{HypOI}_{\mathcal{L},\mathcal{H}}(q) + \mathbb{E}\big[\|p(x) - q(x)\|_1\big].$$

In particular, the prediction in (3.1) *satisfies*

$$\mathrm{HypOI}_{\mathcal{L},\mathcal{H}}(p) \le \mathrm{HypOI}_{\mathcal{L},\mathcal{H}}(q) + 2r.$$

The proof is in Appendix A.1.

Thus it remains to bound the absolute population correlations of the base residual $e_y - q(x)$ with $w_{\ell,r}(q(x))$ and with $\Delta\ell(h(x))$, uniformly over the respective test classes. The sum of the two uniform bounds, plus $4r$, bounds the omniprediction error. We first control the corresponding training correlations using online learners, then prove a population guarantee.

3.3 Combining decision and hypothesis learners

For each test class, we use an online learner that produces an adaptive sequence of tests whose cumulative correlation with the residuals $e_{y_t} - q_t$ competes, up to regret, with the largest absolute cumulative correlation of any fixed test in the class. Controlling the correlation of this sequence then controls the whole class, up to the learner's regret.

We use a learner for the averaged decision weights with a learner for the hypothesis tests $\Delta\mathcal{L}\circ\mathcal{H}$. Then, we choose predictions that make the two learners' combined correlation nonpositive, so their regret bounds control the correlations over both classes.

Definition 3.6 (Online OI learners)**.** Fix a horizon $n \geq 1$. Each learner chooses its round-t test using only feedback from preceding rounds and its own randomization. The guarantees below hold for every adaptive sequence $(x_t, q_t, y_t)_{t=1}^n$, including choices that depend on the current tests.

1. A decision learner with smoothing width $0 < r \leq 1/2$ supplies a continuous test $w_t : \Delta_K \to [-5,5]^K$ on each round t and then receives (q_t, y_t) as feedback. Over the n rounds, it guarantees
$$\sup_{\ell \in \mathcal{L}_{\mathrm{All}}} \left| \sum_{t=1}^n \langle w_{\ell,r}(q_t), e_{y_t} - q_t \rangle \right| \leq \sum_{t=1}^n \langle w_t(q_t), e_{y_t} - q_t \rangle + \mathrm{Reg}_{\mathrm{dec}}(n, r).$$
2. A hypothesis learner for $\Delta\mathcal{L} \circ \mathcal{H}$ supplies a test $g_t : \mathcal{X} \to [-5,5]^K$ on each round t and then receives (x_t, q_t, y_t) as feedback. Over the n rounds, it guarantees
$$\mathbb{E}\left[\sup_{g \in \Delta\mathcal{L} \circ \mathcal{H}} \left| \sum_{t=1}^n \langle g(x_t), e_{y_t} - q_t \rangle \right| - \sum_{t=1}^n \langle g_t(x_t), e_{y_t} - q_t \rangle \right] \leq \mathrm{Reg}_{\Delta\mathcal{L} \circ \mathcal{H}}(n),$$
where the expectation is over the learner's randomness.

Given the test w_t from the decision learner and g_t from the hypothesis learner, we choose a base prediction q so that the two tests have nonpositive combined payoff, *regardless of the* (x_t, y_t) *pair.*

To do this we choose a minimizer
$$q_t(x) \in \arg\min_{q \in \Delta_K} \max_{y \in \mathcal{Y}} \langle w_t(q) + g_t(x), e_y - q \rangle.$$
Brouwer's theorem gives a fixed point q of the continuous map that projects $q + w_t(q) + g_t(x)$ onto Δ_K. Moving from q toward any vertex e_y, the squared distance to this fixed target has initial derivative $-2\langle w_t(q) + g_t(x), e_y - q \rangle$. This must be nonnegative because q is the nearest point. Thus every label payoff is nonpositive, as required.

Lemma 3.7 (Combining the regret guarantees)**.** *Algorithm 1 satisfies*
$$\mathbb{E}\left[\begin{array}{c} \displaystyle\sup_{\ell \in \mathcal{L}_{\mathrm{All}}} \left| \frac{1}{n} \sum_{t=1}^n \langle w_{\ell,r}(q_t(x_t)), e_{y_t} - q_t(x_t) \rangle \right| \\ + \displaystyle\sup_{g \in \Delta\mathcal{L} \circ \mathcal{H}} \left| \frac{1}{n} \sum_{t=1}^n \langle g(x_t), e_{y_t} - q_t(x_t) \rangle \right| \end{array} \right] \leq \frac{\mathrm{Reg}_{\mathrm{dec}}(n, r) + \mathrm{Reg}_{\Delta\mathcal{L} \circ \mathcal{H}}(n)}{n}.$$

Proof. At the realized label y_t, the prediction rule gives
$$\langle w_t(q_t(x_t)), e_{y_t} - q_t(x_t) \rangle + \langle g_t(x_t), e_{y_t} - q_t(x_t) \rangle \leq 0.$$
Add the two expected-regret guarantees and use this pointwise inequality. Dividing by n proves the bound. □

Together with the OI reduction and the smoothing bounds, this gives expected average omniregret on any online sequence at most
$$\frac{\mathrm{Reg}_{\mathrm{dec}}(n, r) + \mathrm{Reg}_{\Delta\mathcal{L} \circ \mathcal{H}}(n)}{n} + 4r,$$
with each round's loss averaged over the prediction randomization in (3.1). In Section 3.6, we obtain the following population guarantee by feeding independent examples into the same online algorithm and applying symmetrization to both test classes.

Algorithm 1 Prediction from the decision and hypothesis learners

Input: A sample $S = ((x_i, y_i))_{i=1}^n$ with $n \geq 1$, width $0 < r \leq 1/2$, a decision learner Dec and a hypothesis learner Hyp, both with the outputs and feedback described in Definition 3.6.

Output: A randomized predictor $p : \mathcal{X} \to \Delta_K$.

Initialize Dec using Algorithm 3 with horizon n and width r.
Initialize Hyp with horizon n.
Training: one pass over the sample.
for $t = 1, \ldots, n$ **do**
    Obtain the functions w_t and g_t from Dec and Hyp using only preceding rounds, before accessing (x_t, y_t).
    Define the prediction rule $q_t : \mathcal{X} \to \Delta_K$ at *every* context $x \in \mathcal{X}$ (not only at x_t) by a fixed choice of minimizer, and store it for the output predictor:
    $q_t(x) \in \arg\min_{q \in \Delta_K} \max_{y \in \mathcal{Y}} \langle w_t(q) + g_t(x), e_y - q \rangle$.
    Receive the context x_t and predict $q_t(x_t)$, the rule evaluated at $x = x_t$.
    Receive y_t.
    Send $(q_t(x_t), y_t)$ to Dec.
    Send $(x_t, q_t(x_t), y_t)$ to Hyp.
Output: after the loop; not executed during training.
return the randomized predictor evaluated by PREDICT(x) below.
function PREDICT(x) ▷ called on test contexts; uses all stored rules $q_1, \ldots, q_n$
    Draw $\tau \sim \mathrm{Unif}(\{1, \ldots, n\})$ and, independently, $z \sim \mathrm{Unif}(\Delta_K)$.
    return $(1 - r)q_\tau(x) + rz$, using fresh draws on every call.

Theorem 3.8 (Omniprediction with a hypothesis-test learner)**.** *Fix a bounded loss class $\mathcal{L}$, a hypothesis class $\mathcal{H}$, $0 < r \leq 1/2$, and a horizon $n \geq 1$. Suppose the learner for $\Delta\mathcal{L} \circ \mathcal{H}$ has the expected-regret guarantee of Definition 3.6, with bound $\mathrm{Reg}_{\Delta\mathcal{L}\circ\mathcal{H}}(n)$.*

For any distribution $\mathcal{D}$, Algorithm 1, run with the decision learner of Algorithm 3, uses n independent examples from $\mathcal{D}$ and returns a predictor p satisfying

$$\mathbb{E}\big[\mathrm{DecOI}_{\mathcal{L}_{\mathrm{All}}}(p) + \mathrm{HypOI}_{\mathcal{L},\mathcal{H}}(p)\big] \leq C_K \, \frac{\mathrm{Reg}_{\mathrm{dec}}(n, r) + \mathrm{Reg}_{\Delta\mathcal{L}\circ\mathcal{H}}(n)}{n} + 4r.$$

Here $\mathrm{Reg}_{\mathrm{dec}}$ is the decision-learner bound in (3.4)*. By Theorem 2.2, the same bound holds for $\mathbb{E}[\mathrm{OmniReg}_{\mathcal{L},\mathcal{H}}(p)]$.*

3.4 Decision calibration learner

We want to compete with all the smoothed decision weights by learning only finitely many tests. The covering bounds below control how many tests suffice for each approximation accuracy.

3.4.1 Approximating the smoothed decision weights

Our first covering bound uses Lemma 3.3 to turn a cover of Bayes risks into a cover of smoothed decision weights. We recall the binary calculation from Section 1.2, which explains the factor $1/r$ in the approximation bound.

With two outcomes, the label-1 coordinate of $w_{\ell,r}(q)$ is the average Bayes-risk slope over an interval of length r. For a one-dimensional Bayes-risk curve $\underline{L}$, let $[a, a + r]$ denote this interval. Its average slope is

$$\frac{1}{r} \int_a^{a+r} \underline{L}'(p)\, dp = \frac{\underline{L}(a + r) - \underline{L}(a)}{r}.$$

If two curves differ in height by at most δ, each endpoint changes by at most δ. Their average slopes therefore differ by at most $2\delta/r$, as shown in Figure 2(b). Lemma 3.3 gives the corresponding multiclass bound; its proof is in Appendix A.1.

3.4.2 Covering the averaged weights

Definition 3.9. A uniform u-cover of a class $\mathcal{F}$ is a finite set $\mathcal{B}$ such that, for every $f \in \mathcal{F}$, there is $b \in \mathcal{B}$ with $\|f - b\|_\infty \le u$.

We may choose cover representatives from the class being covered. Indeed, start with a cover at half the desired radius and select one member of the class from each covering ball that meets it. The triangle inequality gives a cover at the desired radius, with no larger cardinality. This changes only the constants in the covering bounds below. In particular, representatives chosen from the smoothed decision weights are continuous and take values in $[-2, 2]^K$, by Lemma 3.3.

Lemma 3.10 (Concave-function cover)**.** *A class of concave functions on the standard d-dimensional simplex, $d \ge 1$, with uniformly bounded absolute values and a common Lipschitz constant in the ℓ_1 norm has a uniform ρ-cover of size M with*

$$\log(2M) \le C_d \rho^{-d/2}, \qquad 0 < \rho \le 1,$$

where C_d depends only on d, the common bound on the functions' absolute values, and the common Lipschitz constant.

On a cube, this is the uniform-norm entropy bound of Bronshtein [Bro76], in the form of Guntuboyina and Sen [GS13, Theorem 3.2]. The reduction from the simplex to a cube is proved in Appendix A.2.1.

We use *chaining*, as described in Section 1.2, to learn the smoothed decision weights. For each weight, choose approximations from covers at successively finer dyadic accuracies. The finest approximation is its coarsest approximation plus the successive differences. Successive approximations at accuracies $2u$ and u differ in uniform norm by at most $3u$. At each finer scale, we retain only differences between elements of these two covers satisfying this norm bound. We run a finite-test learner on the coarsest cover and on each such family of differences, then add their outputs. The norm dependence in Lemma 3.12 gives a factor of u in each difference learner's regret bound. Summing the regrets of all these learners and adding the final approximation error gives the overall bound. We therefore need covering bounds at every accuracy.

Let $M(u)$ be the size of the smallest uniform u-cover of the smoothed decision weights $\{w_{\ell,r} : \ell \in \mathcal{L}_{\mathrm{All}}\}$.

Lemma 3.11 (Two covering bounds)**.** *For every $0 < r \le 1/2$ and $0 < u \le 1$, the following bounds hold. For $K \ge 3$,*

$$\log\bigl(2M(u)\bigr) \le C_K \min\left\{(ru)^{-(K-1)/2},\ \frac{1}{u}(ru)^{-(K-2)/2}\right\}. \tag{3.3}$$

For $K = 2$,

$$\log\bigl(2M(u)\bigr) \le C \min\left\{(ru)^{-1/2},\ \frac{1}{u}\log\frac{C}{ru}\right\}.$$

The benchmark bound $C_K(ru)^{-(K-1)/2}$ in Lemma 3.11 comes from covering the $K-1$ dimensional Bayes risks at a sufficiently small constant multiple of ru using Lemma 3.10. By Lemma 3.3, such a cover gives a uniform u-cover of the smoothed decision weights.

For $K \geq 3$, we construct a uniform u-cover of the smoothed decision weights with logarithmic size $O_K\big(\frac{1}{u}(ru)^{-(K-2)/2}\big)$, proving the second bound in Lemma 3.11.

On each coordinate slice, the Bayes-risk curve can be replaced by the lower envelope of $O(1/u)$ supporting lines, with $O(u)$ error in its slopes almost everywhere. The left panels of Figure 10 show why a fixed grid of slopes with spacing of order u suffices. In the upper left panel, the gray supporting lines lie above and touch the green Bayes-risk curve, and the blue curve is their lower envelope. Between successive contact points, concavity puts the green curve's slope between the two supporting slopes, while the blue envelope uses one of them. The lower left panel compares the original coordinate decision weights with these slope approximations. The smoothed decision weights are averages of the Bayes-risk slopes over the reporting distribution. Averaging the envelope's slopes in the same way preserves the $O(u)$ approximation.

A uniform $O(ru)$ approximation to the blue envelope gives an $O_K(ru/r) = O_K(u)$ approximation to its smoothed slopes, by the earlier smoothing estimate (Figure 2(b)). Approximating each line's vertical position to accuracy $O(ru)$, with its slope fixed, gives such a function approximation and taking the lower envelope preserves this uniform error. The upper right panel shows the blue envelope and its magenta approximation; the lower right compares their averaged slopes. The integration estimate in the proof of Lemma 3.3 applies to these Lipschitz envelopes. Together, the two approximations change the smoothed weights by $O_K(u)$.

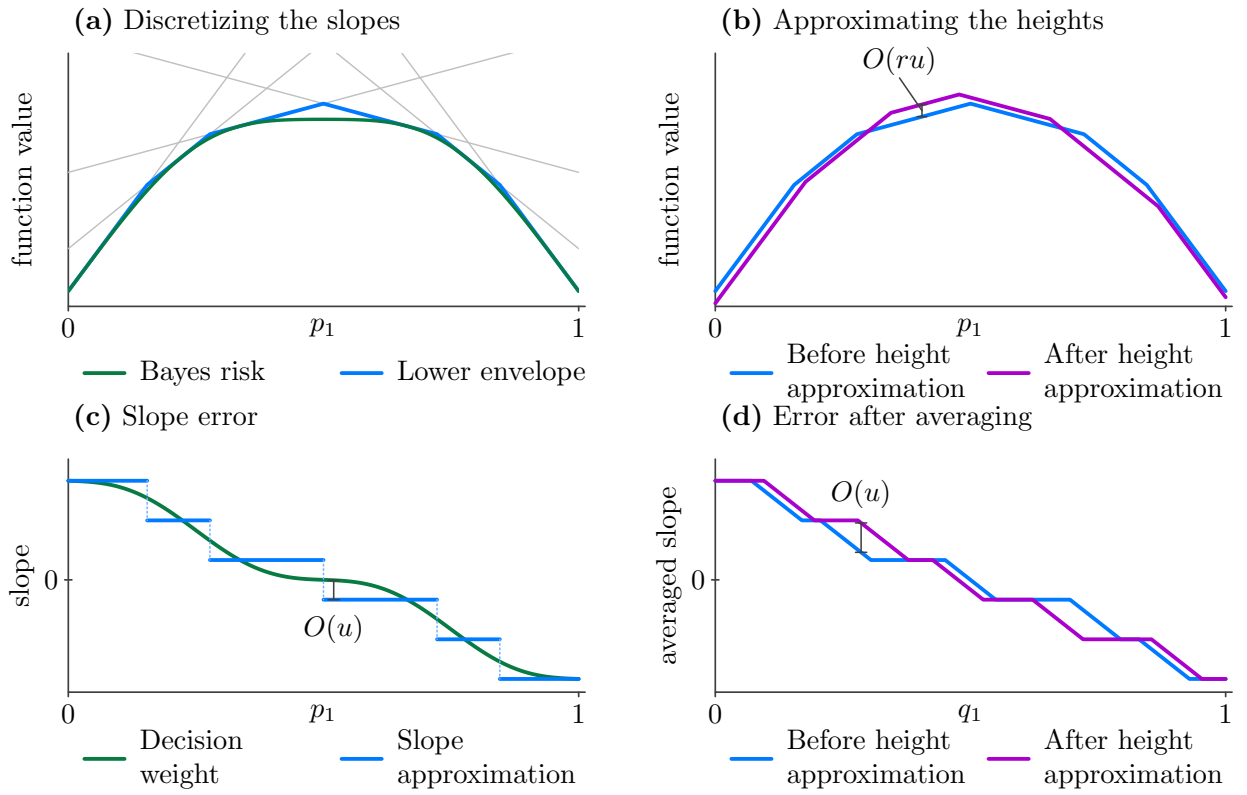


Figure 10: Two approximation steps for the smoothed decision weights, illustrated on a binary Bayes-risk curve. Here u is the desired accuracy of the smoothed decision weights and r is the smoothing width. Both upper panels show function values before smoothing. *Upper left:* the gray supporting lines have slopes spaced by order u and touch the green Bayes-risk curve; their lower envelope is blue. *Lower left:* the slopes of the green curve and blue envelope give the decision weight and its approximation, differing by $O(u)$ almost everywhere. *Upper right:* approximating the lines' vertical positions to accuracy $O(ru)$, with their slopes fixed, gives the magenta envelope. It differs uniformly by $O(ru)$ from the same blue envelope shown on the left. *Lower right:* averaging the slopes of the blue and magenta envelopes over an interval of length r gives smoothed weights differing by $O(ru)/r = O(u)$.

It therefore remains to approximate the vertical positions of the supporting lines to accuracy $O(ru)$, uniformly over slices. For a line with fixed slope, its vertical intercept specifies this position; we call this intercept the line's *height*. In Figure 11, the blue supporting lines all have the same slope and touch different parallel slices of the Bayes-risk surface. Their intercepts trace the blue curve on the back wall, which records the height of the supporting line on each slice. This is the *height function* for that slope; the orange curve gives the height function for another slope. One of the $K-1$ independent prediction coordinates varies within a slice, and the remaining $K-2$ specify the slice. Thus, for each fixed slope and coordinate direction, we need a uniform $O(ru)$-cover of the possible height functions in $K-2$ variables.

These height functions are concave, which lets us apply the same covering bound at the same accuracy in *dimension* $K-2$ *rather than* $K-1$. The same figure shows why the height functions are concave. For the blue family, concavity of the Bayes risk puts the surface above the dashed chord joining the two outer contact points, as the thin blue arc illustrates. The chord's hollow midpoint therefore lies below the surface in the middle slice, and hence below the middle blue supporting line.

Projecting back to the wall along the blue lines' common slope preserves midpoints and vertical order. The middle intercept is therefore at least the average of the two outer intercepts. The same argument applies to every convex combination of two slices, proving concavity of the height function.

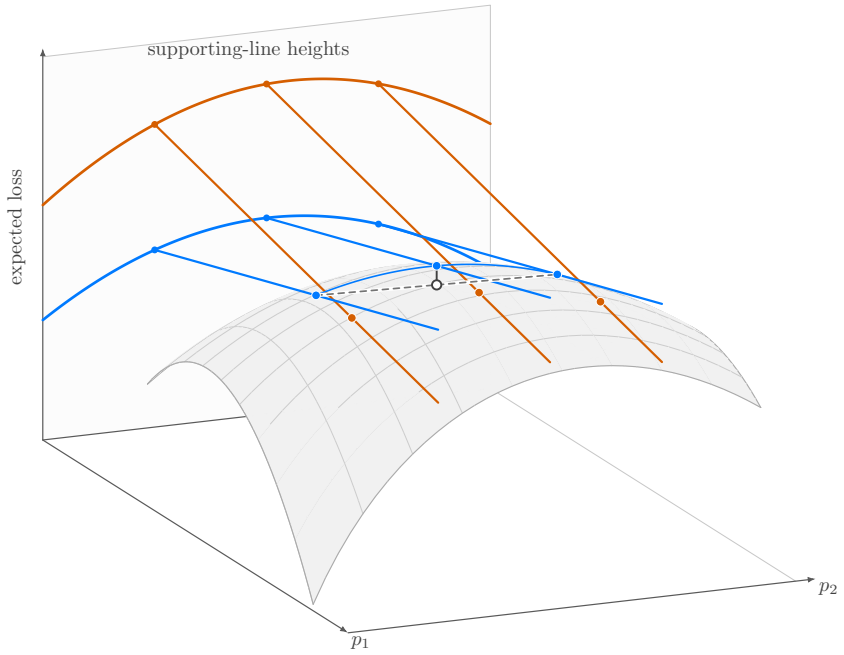


Figure 11: Concavity of the supporting-line heights. Lines of the same color have the same slope and lie above and touch different slices of the gray Bayes-risk surface. Their vertical intercepts trace a height function on the back wall. The thin blue arc lies on the surface above the dashed chord joining the outer blue contact points. The chord's hollow midpoint lies below the surface in the middle slice, and hence below the middle blue supporting line. Projection to the back wall along the blue lines' common slope preserves midpoints and vertical order, so the middle intercept is at least the average of the two outer intercepts.

The height functions also have uniformly bounded values and Lipschitz constants, as verified in Appendix A.2.2. For each fixed slope and coordinate direction, Lemma 3.10 therefore covers the possible height functions at accuracy $O(ru)$ with logarithmic size $O_K((ru)^{-(K-2)/2})$. Compared with the $O_K((ru)^{-(K-1)/2})$ cost of covering the Bayes risk itself, the lower dimension contributes a factor $\sqrt{ru}$. There are, however, $O_K(1/u)$ height functions to record for each loss, across all grid slopes and coordinate directions. Combining their covers multiplies the cover sizes, so their logarithms add. With sufficiently small constants in the approximation scales, the resulting uniform u-cover satisfies

$$\log\bigl(2M(u)\bigr) \le C_K \frac{1}{u}(ru)^{-(K-2)/2} = C_K(ru)^{-(K-1)/2}\sqrt{\frac{r}{u}}.$$

The net factor in the logarithmic cover bound is therefore $\sqrt{r/u}$. At the terminal accuracy $u \asymp r$, the two bounds are comparable. At coarser accuracies $u \gg r$, the saving from the lower dimension outweighs the cost of recording a height function for each slope. The full proof appears in Appendix A.2.2.

3.4.3 Learning from successive covers

We now turn the covers from Lemma 3.11 into an online learner for the smoothed decision weights $\{w_{\ell,r} : \ell \in \mathcal{L}_{\mathrm{All}}\}$. The learner runs exponential weights on the coarsest cover and on small differences between consecutive covers, then sums the proposed functions. Its regret is bounded by the sum of the regrets at each level and the error from the final cover.

We use the standard exponential-weights method, applied to the positive and negative copies of each test, for the decision learner and, later, the Lipschitz hypothesis learner. For a decision test b, the payoff on prediction q_t and label y_t is

$$\rho_t(b) = \langle b(q_t), e_{y_t} - q_t \rangle.$$

For a hypothesis test b at context x_t, it is

$$\rho_t(b) = \langle b(x_t), e_{y_t} - q_t \rangle.$$

Algorithm 2 Signed exponential weights: $\mathrm{EW}(\mathcal{B}, a, n)$

Input: A finite nonempty family $\mathcal{B}$ of $\mathbb{R}^K$-valued tests on a common domain, a bound $a > 0$ with $\|b\|_\infty \le a$ for every $b \in \mathcal{B}$, and a horizon $n \ge 1$.
Per-round feedback: Payoffs $\rho_t(b) \in [-2a, 2a]$ for every $b \in \mathcal{B}$, received after outputting b_t.
Output: Tests b_t with $\|b_t\|_\infty \le a$, with the regret guarantee of Lemma 3.12.

Set $\eta \leftarrow \sqrt{\log(2|\mathcal{B}|)/(2na^2)}$.
Set $\lambda_1(b, s) \leftarrow 1/(2|\mathcal{B}|)$ for every $b \in \mathcal{B}$ and $s \in \{-1, 1\}$.
for $t = 1, \dots, n$ **do**
    Output the test
    $$b_t = \sum_{b \in \mathcal{B}} \sum_{s \in \{-1,1\}} \lambda_t(b, s)\, s\, b.$$
    Receive the payoffs $\rho_t(b)$ for all $b \in \mathcal{B}$.
    For every $b \in \mathcal{B}$ and $s \in \{-1, 1\}$, set
    $$\lambda_{t+1}(b, s) \leftarrow \frac{\lambda_t(b, s) \exp\big(\eta s \rho_t(b)\big)}{\sum_{c \in \mathcal{B}} \sum_{\sigma \in \{-1,1\}} \lambda_t(c, \sigma) \exp\big(\eta \sigma \rho_t(c)\big)}.$$

Lemma 3.12 (Signed exponential weights)**.** *Let $\mathcal{B}$ consist of $M \ge 1$ tests with $\|b\|_\infty \le a$, where $a > 0$. For every horizon $n \ge 1$ and every sequence of payoffs $\rho_t(b) \in [-2a, 2a]$, Algorithm 2 satisfies*

$$\max_{b \in \mathcal{B}} \left| \sum_{t=1}^{n} \rho_t(b) \right| \le \sum_{t=1}^{n} \sum_{b \in \mathcal{B}} \sum_{s \in \{-1,1\}} \lambda_t(b, s)\, s\, \rho_t(b) + Ca\sqrt{n \log(2M)},$$

where C is an absolute constant. Its outputs satisfy $\|b_t\|_\infty \le a$.

For random payoffs, condition on the preceding rounds and all choices made before the current payoff vector is sampled. In particular, the current weights λ_t are fixed. Write $\bar{\rho}_t(b)$ for the resulting conditional mean of $\rho_t(b)$. For every $0 < \delta < 1$, with probability at least $1 - \delta$,

$$\max_{b \in \mathcal{B}} \left| \sum_{t=1}^{n} \bar{\rho}_t(b) \right| \le \sum_{t=1}^{n} \sum_{b \in \mathcal{B}} \sum_{s \in \{-1,1\}} \lambda_t(b, s) s \bar{\rho}_t(b) + Ca\sqrt{n \log(2M/\delta)}.$$

The algorithm still updates using the observed payoffs $\rho_t(b)$.

Algorithm 3 Decision learner from successive covers

Input: Horizon $n \geq 1$ and width $0 < r \leq 1/2$.
Per-round feedback: A prediction $q_t \in \Delta_K$ and label $y_t \in \mathcal{Y}$, received after outputting w_t.
Output: Continuous tests $w_t : \Delta_K \to [-5,5]^K$, satisfying the regret guarantee of Lemma 3.13.

Set $J \leftarrow \lfloor \log_2(1/r) \rfloor$ and $u_j \leftarrow 2^{-j}$ for $j = 0, \dots, J$.
Choose uniform u_j-covers $\mathcal{B}_j$ of $\{w_{\ell,r} : \ell \in \mathcal{L}_{\mathrm{All}}\}$ for $j = 0, \dots, J$, with continuous representatives taking values in $[-2,2]^K$ and the size bounds of Lemma 3.11.
Initialize $A_0 \leftarrow \mathrm{EW}(\mathcal{B}_0, 2, n)$.
for $j = 1, \dots, J$ **do**
    Initialize an EW learner A_j with horizon n and bound $3u_j$ on $\{b - b' : b \in \mathcal{B}_j,\ b' \in \mathcal{B}_{j-1},\ \|b - b'\|_\infty \leq 3u_j\}$.
for $t = 1, \dots, n$ **do**
    Output w_t as the sum of the current tests from $A_0, \dots, A_J$.
    Receive (q_t, y_t).
    Update every A_j using the payoff $\rho_t(b) = \langle b(q_t), e_{y_t} - q_t \rangle$ for each of its tests b.

The first bound is the exponential-weights potential bound applied to the $2M$ signed payoffs $s\rho_t(b)$. The second adds the Azuma–Hoeffding inequality and a union bound over the tests. The proof is in Appendix A.3.

The decision learner stops at cover accuracy of order r which will later be taken to be of order ε.

Lemma 3.13 (Online decision learner)**.** *Fix $n \geq 1$ and $0 < r \leq 1/2$, and let $J, u_j, \mathcal{B}_j$ be as in Algorithm 3. Let $M_j \geq 1$ be nondecreasing upper bounds on $|\mathcal{B}_j|$.*

For every sequence of predictions and labels, the algorithm outputs continuous tests $w_t : \Delta_K \to [-5,5]^K$ before observing (q_t, y_t) and satisfies

$$\sup_{\ell \in \mathcal{L}_{\mathrm{All}}} \left| \sum_{t=1}^{n} \langle w_{\ell,r}(q_t), e_{y_t} - q_t \rangle \right| \leq \sum_{t=1}^{n} \langle w_t(q_t), e_{y_t} - q_t \rangle + 2nu_J + C\sqrt{n} \sum_{j=0}^{J} u_j \sqrt{\log(2M_j)}.$$

The same bound holds when the residuals $e_{y_t} - q_t$ are replaced by arbitrary vectors of ℓ_1 norm at most two, in both the updates and the inequality.

Proof. Fix $\ell \in \mathcal{L}_{\mathrm{All}}$. For each j, choose $b_j \in \mathcal{B}_j$ with $\|w_{\ell,r} - b_j\|_\infty \leq u_j$. Then

$$\|b_j - b_{j-1}\|_\infty \leq u_j + u_{j-1} = 3u_j, \qquad j \geq 1,$$

so every successive difference is included in its EW learner. Moreover,

$$w_{\ell,r} = b_0 + \sum_{j=1}^{J} (b_j - b_{j-1}) + (w_{\ell,r} - b_J).$$

Since $\|e_{y_t} - q_t\|_1 \leq 2$, the last term contributes at most $2nu_J$ to the absolute cumulative payoff. Note that the proof only relies on this norm bound and so holds for more general bounded norm feedback.

For the remaining terms, use the triangle inequality and apply Lemma 3.12 to each EW learner. The weighted payoffs in these bounds sum to

$$\sum_{t=1}^{n} \langle w_t(q_t), e_{y_t} - q_t \rangle,$$

because w_t is the sum of the EW outputs.

The first learner contributes $O(\sqrt{n \log(2M_0)})$. At level $j \geq 1$, there are at most $M_j M_{j-1} \leq M_j^2$ differences, each of norm at most $3u_j$, giving regret $O(u_j \sqrt{n \log(2M_j)})$. Adding these bounds and taking the supremum over ℓ proves the inequality.

Each output is a sum of continuous functions, and

$$\|w_t\|_\infty \leq 2 + 3 \sum_{j=1}^{J} u_j \leq 5.$$

□

3.4.4 The resulting regret bound

Lemma 3.14 (Decision regret)**.** *Fix $K \geq 2$, $0 < r \leq 1/2$, and $n \geq 1$. There is an online learner satisfying the decision guarantee in Definition 3.6 with total-regret bound*

$$\mathrm{Reg}_{\mathrm{dec}}(n, r) = C_K n r + C_K \sqrt{n}\, f_K(r), \qquad f_K(r) := \begin{cases} \sqrt{\log(2/r)}, & K = 2, \\ r^{-1/4}, & K = 3, \\ r^{-1/2} \log(2/r), & K = 4, \\ r^{-(K-3)/2}, & K \geq 5. \end{cases} \tag{3.4}$$

Proof. Use continuous covers with the size bounds of Lemma 3.11 in Lemma 3.13, stopping at $r \leq u_J < 2r$. The terminal approximation term $2nu_J$ is $O(nr)$.

For $K \geq 3$, the second covering bound gives a contribution of at most

$$C_K \sqrt{n}\, r^{-(K-2)/4} u^{(4-K)/4}$$

at each accuracy u. The sum of $u_j^{(4-K)/4}$ is $O(1)$ for $K = 3$, $O(\log(2/r))$ for $K = 4$, and $O_K(r^{-(K-4)/4})$ for $K \geq 5$. Multiplying by the common factor $C_K \sqrt{n}\, r^{-(K-2)/4}$ gives the three claimed bounds.

For $K = 2$, the covering bound gives

$$u_j \sqrt{n \log(2M_j)} \leq C \sqrt{n u_j \log \frac{C}{r u_j}} \leq C' \sqrt{n \log(2/r)}\, \sqrt{u_j}.$$

The last inequality uses $u_j \geq r$. The sum $\sum_{j=0}^{J} \sqrt{u_j} = \sum_{j=0}^{J} 2^{-j/2}$ is bounded by an absolute constant, giving the binary bound. □

3.5 Hypothesis learners

3.5.1 Lipschitz losses

To prove Theorem 3.1, we now construct a hypothesis learner for Lipschitz losses and a finite benchmark class $\mathcal{H}$. We use the same chaining argument as for the decision learner, applied to covers of $\Delta\mathcal{L}_{\mathrm{Lip}} \circ \mathcal{H}$.

For $\ell \in \mathcal{L}_{\mathrm{Lip}}$, each coordinate of $\Delta\ell$ is 2-Lipschitz on the action simplex. We obtain a cover of the hypothesis tests by covering these derivatives and composing with the hypotheses in $\mathcal{H}$. Lemma 3.15 bounds the size of the resulting cover. For use in Algorithm 1, take the learner's accuracy scale r equal to the smoothing width.

Lemma 3.15 (Hypothesis tests for Lipschitz losses)**.** *Let $K \geq 2$ and let $\mathcal{H}$ be a class of $N \geq 1$ hypotheses $\mathcal{X} \to \Delta_K$. For every $0 < u \leq 1$, there is a uniform u-cover $\mathcal{B} \subseteq \Delta\mathcal{L}_{\mathrm{Lip}} \circ \mathcal{H}$ satisfying*

$$\log(2|\mathcal{B}|) \leq C_K u^{-(K-1)} + \log(2N),$$

where C_K depends only on K.

Proof. Choose a regular grid on Δ_K with $O_K(u^{-(K-1)})$ points, such that every point of the simplex is within $u/8$ of a grid point and neighboring grid points are at most $u/8$ apart. Round each coordinate of $\Delta\ell$ at the grid points to the nearest multiple of $u/2$. Two derivatives with the same rounded values differ by at most $u/2$ on the grid. Each coordinate of $\Delta\ell$ is 2-Lipschitz in the Euclidean norm, since it is the difference of two 1-Lipschitz functions. The two derivatives therefore differ by at most

$$\frac{u}{2} + 2\frac{u}{8} + 2\frac{u}{8} = u$$

at every point of the simplex. Thus choosing one derivative for each rounding pattern gives a uniform u-cover.

To count the patterns, order the grid points so that each point after the first has an earlier neighbor. At the first point there are $O_K(u^{-K})$ possible rounded values. At each later point, each coordinate differs from its earlier neighbor by at most $u/4$ before rounding and $3u/4$ after rounding. Since the rounded values are multiples of $u/2$, there are at most three choices per coordinate. There are therefore at most $\exp(C_K u^{-(K-1)})$ patterns. Composing the chosen derivatives with the N hypotheses preserves the approximation and gives the required cover of size at most $N \exp(C_K u^{-(K-1)})$. □

Algorithm 4 Hypothesis learner for Lipschitz losses

Input: Horizon $n \geq 1$, accuracy $0 < r \leq 1/2$, and a finite nonempty hypothesis class $\mathcal{H}$.
Per-round feedback: A context $x_t \in \mathcal{X}$, prediction $q_t \in \Delta_K$, and label $y_t \in \mathcal{Y}$.
Output: Tests $g_t : \mathcal{X} \to [-5, 5]^K$ with the regret bound of Lemma 3.16.

Set $J \leftarrow \lfloor \log_2(1/r) \rfloor$ and $u_j \leftarrow 2^{-j}$ for $j = 0, \ldots, J$.
For each $j = 0, \ldots, J$, choose a uniform u_j-cover $\mathcal{B}_j \subseteq \Delta\mathcal{L}_{\mathrm{Lip}} \circ \mathcal{H}$ as in Lemma 3.15.
Initialize a learner $\mathrm{EW}(\mathcal{B}_0, 2, n)$.
for $j = 1, \ldots, J$ **do**
    Initialize another EW learner with horizon n and bound $3u_j$ on $\{b - b' : b \in \mathcal{B}_j,\ b' \in \mathcal{B}_{j-1},\ \|b - b'\|_\infty \leq 3u_j\}$.
for $t = 1, \ldots, n$ **do**
    Output g_t as the sum of the tests supplied by the EW learners for this round.
    Receive (x_t, q_t, y_t).
    Update each EW learner with payoff $\langle b(x_t), e_{y_t} - q_t \rangle$ for each of its tests b.

Lemma 3.16 (Online hypothesis learner for Lipschitz losses)**.** *Fix $K \geq 2$, $n \geq 1$, $0 < r \leq 1/2$, and a hypothesis class $\mathcal{H}$ of size $N \geq 1$. For every sequence of contexts, predictions, and labels, Algorithm 4 satisfies*

$$\sup_{g \in \Delta\mathcal{L}_{\mathrm{Lip}} \circ \mathcal{H}} \left| \sum_{t=1}^{n} \langle g(x_t), e_{y_t} - q_t \rangle \right| \leq \sum_{t=1}^{n} \langle g_t(x_t), e_{y_t} - q_t \rangle + C_K n r + C_K \sqrt{n} \left(\sqrt{\log(2N)} + \begin{cases} 1, & K = 2, \\ \log(2/r), & K = 3, \\ r^{-(K-3)/2}, & K \geq 4, \end{cases} \right), \tag{3.5}$$

where C_K depends only on K.

Proof. As in the proof of Lemma 3.13, write each hypothesis test as its coarsest approximation, a sum of differences between successive approximations, and a final approximation error. Successive approximations differ in uniform norm by at most $u_j + u_{j-1} = 3u_j$, so each difference is included in the EW learner at level j. Lemma 3.12 bounds the regret for the coarse approximation and the differences by the sum of the EW regrets. The final approximation error contributes at most $2nu_J$.

At level $j \geq 1$, there are at most $|\mathcal{B}_j||\mathcal{B}_{j-1}|$ tests, each of norm at most $3u_j$. Lemma 3.15 and $u_{j-1} = 2u_j$ therefore bound the EW regret by

$$Cu_j\sqrt{n\log(2|\mathcal{B}_j||\mathcal{B}_{j-1}|)} \leq C_K\sqrt{n}\left(u_j^{(3-K)/2} + u_j\sqrt{\log(2N)}\right).$$

The initial learner has $|\mathcal{B}_0|$ tests of norm at most 2, giving regret at most $C_K\sqrt{n}(1+\sqrt{\log(2N)})$.

Since $\sum_{j=0}^{J} u_j \leq 2$, summing over the levels and taking the supremum over g gives total regret at most

$$2nu_J + C_K\sqrt{n}\left(\sqrt{\log(2N)} + \sum_{j=0}^{J} u_j^{(3-K)/2}\right).$$

The approximation error satisfies $2nu_J \leq 4nr$, because $r \leq u_J < 2r$. Using $u_j = 2^{-j}$, we obtain

$$\begin{aligned}
K = 2: \quad & \sum_{j=0}^{J} u_j^{1/2} = \sum_{j=0}^{J} 2^{-j/2} \leq \sum_{j=0}^{\infty} 2^{-j/2} = \frac{1}{1-2^{-1/2}}, \\
K = 3: \quad & \sum_{j=0}^{J} u_j^{0} = J + 1 \leq \log_2(2/r) = \frac{\log(2/r)}{\log 2}, \\
K \geq 4: \quad & \sum_{j=0}^{J} u_j^{(3-K)/2} = \sum_{j=0}^{J} 2^{j(K-3)/2} \\
& \qquad = 2^{J(K-3)/2}\sum_{j=0}^{J} 2^{-j(K-3)/2} \\
& \qquad \leq \frac{2^{J(K-3)/2}}{1-2^{-(K-3)/2}} \leq \frac{r^{-(K-3)/2}}{1-2^{-(K-3)/2}}.
\end{aligned}$$

The last inequality uses $2^J \leq 1/r$. Substituting these bounds into the preceding regret expression proves the claimed rates.

Finally, the EW output bounds give

$$\|g_t\|_\infty \leq 2 + 3\sum_{j=1}^{J} u_j \leq 5.$$

□

Theorem 3.8 uses the hypothesis learner through the expected-regret guarantee in Definition 3.6. For the finite-cover learners here, the conditional-mean bound in Lemma 3.12 gives a direct population guarantee with high probability.

Lemma A.1 in Appendix A.3 states this guarantee for the decision learner and the Lipschitz hypothesis learner: each population correlation bound holds with probability at least $1-\delta$, at the cost of an additional $C\sqrt{n\log(2/\delta)}$ term.

Proof of Theorem 3.1. Run Algorithm 1 with the decision learner from Algorithm 3 and the hypothesis learner from Algorithm 4. Take $r = c_K\varepsilon$ for a sufficiently small constant. Substituting the decision and Lipschitz hypothesis regret bounds into Theorem 3.8 gives the expected guarantee.

For the confidence guarantee, apply the two bounds in Lemma A.1 with failure probability $\delta/2$ each. The prediction rule makes their combined payoff nonpositive on every fresh example. Adding the inequalities, dividing by n, and applying the smoothing bounds gives, with probability at least $1-\delta$,

$$\mathrm{OmniReg}_{\mathcal{L}_{\mathrm{Lip}},\mathcal{H}}(p) \le C\frac{\mathrm{Reg}_{\mathrm{dec}}(n,r) + \mathrm{Reg}_{\Delta\mathcal{L}_{\mathrm{Lip}}\circ\mathcal{H}}(n)}{n} + C\sqrt{\frac{\log(2/\delta)}{n}} + 4r.$$

The displayed sample bound with $\log(2N/\delta)$ in place of $\log(2N)$ makes the right-hand side at most ε. □

3.5.2 Extending to infinite hypothesis classes

Theorem 3.1 pays $\log(2N)/\varepsilon^2$ for a class of N hypotheses. The analysis only needs a finite class that approximates $\mathcal{H}$ on average over the context distribution, and only at accuracy of order ε. We build such a class from samples, and then run Algorithm 1 on it. This replaces $\log(2N)$ by the fat-shattering dimension of $\mathcal{H}$, which is finite for many infinite classes.

Fat-shattering dimension. For $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$ and $\gamma > 0$, points $x_1, \dots, x_d$ are *γ-shattered* by $\mathcal{F}$ if there are thresholds $s_1, \dots, s_d \in \mathbb{R}$ such that for every $b \in \{\pm 1\}^d$ some $f \in \mathcal{F}$ satisfies $b_i(f(x_i) - s_i) \ge \gamma$ for all i. The *fat-shattering dimension* $\mathrm{fat}_\gamma(\mathcal{F})$ is the largest such d [KS94; Alo+97]. It is at most $\log_2|\mathcal{F}|$ for finite $\mathcal{F}$, and at most the pseudo-dimension of $\mathcal{F}$. For a class $\mathcal{H}$ of maps $\mathcal{X} \to \Delta_K$, let $\mathcal{H}_j := \{x \mapsto h(x)_j : h \in \mathcal{H}\}$ and $\mathrm{fat}_\gamma(\mathcal{H}) := \max_{j\in\mathcal{Y}} \mathrm{fat}_\gamma(\mathcal{H}_j)$.

From hypotheses to tests. Write $d_{\mathcal{D}}(h,h') := \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}\|h(x) - h'(x)\|_1$ for the average distance between two hypotheses under the context marginal $\mathcal{D}_{\mathcal{X}}$. Each coordinate of $\Delta\ell$ is 2-Lipschitz in the Euclidean norm for $\ell \in \mathcal{L}_{\mathrm{Lip}}$, and $\|e_y - p\|_1 \le 2$. Since $\mathbb{E}_{\widetilde{y}\sim p}[\ell(a,y) - \ell(a,\widetilde{y})] = \langle \Delta\ell(a), e_y - p\rangle$, for all h, h' and every predictor p,

$$\left|\mathbb{E}\left[\langle \Delta\ell(h(x)) - \Delta\ell(h'(x)), e_y - p(x)\rangle\right]\right| \le 4\, d_{\mathcal{D}}(h,h'). \tag{3.6}$$

Hence if every $h \in \mathcal{H}$ is within $d_{\mathcal{D}}$-distance γ of some $h' \in \widehat{\mathcal{H}}$, then $\mathrm{HypOI}_{\mathcal{L}_{\mathrm{Lip}},\mathcal{H}}(p) \le \mathrm{HypOI}_{\mathcal{L}_{\mathrm{Lip}},\widehat{\mathcal{H}}}(p) + 4\gamma$.

The algorithm. Draw m examples and let S be their contexts. Write d_S for the distance $d_{\mathcal{D}}$ with $\mathcal{D}_{\mathcal{X}}$ replaced by the uniform distribution on S. Let $\widehat{\mathcal{H}} \subseteq \mathcal{H}$ be a maximal subset whose elements are pairwise more than $\varepsilon/16$ apart in d_S. Then run Algorithm 1, with the decision learner of Algorithm 3 and the hypothesis learner of Algorithm 4 for the finite class $\widehat{\mathcal{H}}$, on n fresh examples. By maximality, every $h \in \mathcal{H}$ is within $\varepsilon/16$ of $\widehat{\mathcal{H}}$ in d_S. Building a cover of the class from unlabeled data in this way is classical [BI91; Hau92]; Bartlett and Long [BL98] give scale-sensitive versions. Choosing $\widehat{\mathcal{H}}$ requires searching $\mathcal{H}$ on m points; we do not address its computational cost here.

Lemma 3.17 (Empirical covers)**.** *There are absolute constants $c, C > 0$ such that the following holds for every class $\mathcal{H}$ of maps $\mathcal{X} \to \Delta_K$[5], every distribution $\mathcal{D}$, and all $0 < \varepsilon, \delta < 1$.*

[5]As usual for uniform convergence over infinite classes, we assume the coordinate classes $\mathcal{H}_j$ are permissible in the sense of Haussler [Hau92, Section 10.2, Definition 11]. Permissibility passes to the coordinate-distance classes $\{|f - f'| : f, f' \in \mathcal{H}_j\}$ through the product parameterization and the continuous map $(a,b) \mapsto |a-b|$. Hence the events $\sup_{f,f'\in\mathcal{H}_j}\left|\mathbb{E}_{\mathcal{D}_{\mathcal{X}}}|f-f'| - \mathbb{E}_S|f-f'|\right| \ge \alpha$ used in the proof of part 1 are measurable.

1. With probability at least $1-\delta$ *over the sample* S*,*

$$\sup_{h,h'\in\mathcal{H}}\bigl|d_{\mathcal{D}}(h,h')-d_S(h,h')\bigr|\le\frac{\varepsilon}{16},\qquad\textit{provided}\qquad m\ge C\frac{K^2}{\varepsilon^2}\Bigl(\mathrm{fat}_{c\varepsilon/K}(\mathcal{H})\log\frac{2K}{\varepsilon}+\log\frac{2K}{\delta}\Bigr).$$

2. For every sample, $\log(2|\widehat{\mathcal{H}}|)\le CK\,\mathrm{fat}_{c\varepsilon/K}(\mathcal{H})\log(2K/\varepsilon)+1$.

Part 1 is a uniform law of large numbers for the coordinate distances $|h_j-h'_j|$, and part 2 bounds the size of a separated set by a covering number. Both follow from the entropy bound of Mendelson and Vershynin [MV03], which controls covering numbers by the fat-shattering dimension, together with the uniform convergence bound of Anthony and Bartlett [AB99, Theorem 17.7]. The proof is in Appendix A.5.

Theorem 3.18 (Omniprediction for infinite hypothesis classes)**.** *Fix* $K\ge 2$ *and* $0<\varepsilon,\delta<1$*, and let* $\mathcal{H}$ *be any class of maps* $\mathcal{X}\to\Delta_K$*. There is a constant* $c_K>0$*, depending only on* K*, such that the algorithm above uses*

$$\widetilde{O}_K\left(\varepsilon^{-\max\{K-1,(K+2)/2\}}+\frac{\mathrm{fat}_{c_K\varepsilon}(\mathcal{H})+\log(2/\delta)}{\varepsilon^2}\right)$$

independent samples from any distribution $\mathcal{D}$ *and, with probability at least* $1-\delta$*, returns* p *with* $\mathrm{OmniReg}_{\mathcal{L}_{\mathrm{Lip}},\mathcal{H}}(p)\le\varepsilon$*.*

Proof. Apply Lemma 3.17 with failure probability $\delta/2$. On this event, every $h\in\mathcal{H}$ has $h'\in\widehat{\mathcal{H}}$ with $d_{\mathcal{D}}(h,h')\le d_S(h,h')+\varepsilon/16\le\varepsilon/8$. By (3.6) and Theorem 2.2, the returned predictor satisfies

$$\mathrm{OmniReg}_{\mathcal{L}_{\mathrm{Lip}},\mathcal{H}}(p)\le\mathrm{DecOI}_{\mathcal{L}_{\mathrm{All}}}(p)+\mathrm{HypOI}_{\mathcal{L}_{\mathrm{Lip}},\widehat{\mathcal{H}}}(p)+\frac{\varepsilon}{2}.$$

Conditional on S, $\widehat{\mathcal{H}}$ is a fixed finite class, and the remaining n examples are independent of it. The proof of Theorem 3.1, with failure probability $\delta/2$, bounds the sum of the two OI errors by $\varepsilon/2$ once n meets the bound of Theorem 3.1 for accuracy $\varepsilon/2$, with $\log(2|\widehat{\mathcal{H}}|\cdot 2/\delta)$ in place of $\log(2N/\delta)$. Part 2 of Lemma 3.17 bounds this logarithm by $O_K(\mathrm{fat}_{c_K\varepsilon}(\mathcal{H})\log(2/\varepsilon)+\log(2/\delta))$ with $c_K=c/K$, and part 1 bounds m by the same order divided by ε^2. Adding m and n proves the theorem. □

For a finite class, $\mathrm{fat}_{c_K\varepsilon}(\mathcal{H})\le\log_2 N$, so Theorem 3.18 recovers Theorem 3.1 up to logarithmic factors. Since the fat-shattering dimension is at most the pseudo-dimension, the theorem also holds with the pseudo-dimension of $\mathcal{H}$ (the maximum over j of that of $\mathcal{H}_j$) in place of $\mathrm{fat}_{c_K\varepsilon}(\mathcal{H})$.

Optimality. In the worst case, the hypothesis-dependent term cannot be improved beyond logarithmic factors. Suppose $\mathcal{H}$ contains, for some points $x_1,\dots,x_d$, all 2^d patterns of values in $\{e_0,e_1\}$ on these points. The proof of Theorem 4.6 with $M=d$ uses only these hypotheses and contexts, so it gives $n=\Omega(d/\varepsilon^2)$. For a class of hypotheses with values in $\{e_0,e_1\}$, this d can be taken to be the VC dimension of $\mathcal{H}_1$. In that case $\mathrm{fat}_\gamma(\mathcal{H})$ equals this VC dimension for every $\gamma\le 1/2$, so the upper and lower bounds match. This is a worst-case statement, not a matching lower bound for every class. For classes with real-valued coordinates, $\mathrm{fat}_{c_K\varepsilon}(\mathcal{H})$ can be much larger than the largest d for which $\mathcal{H}$ contains all $\{e_0,e_1\}$-patterns on d points. We do not know whether Theorem 3.18 is tight for each such class.

3.5.3 GLM losses

The most notable prior work on omniprediction in the multiclass setting came from Hu, Tian, and Yang [HTY26]. Their work gave a general framework for achieving omniprediction with sample complexity scaling with $\varepsilon^{-(K+1)}$. They instantiated their framework over the class of Generalized Linear Model (GLM) losses and the hypothesis class of linear models.

In this subsection we show that our algorithm improves the sample complexity for these classes to $\varepsilon^{-\max\{K-1,K/2+1\}}$. Our implementation combines the hypothesis learner of Hu, Tian, and Yang [HTY26] with our new decision learner to achieve the improved bounds. The main setup and results are stated here; see Appendix A.6 for the proof details.

GLM losses and linear classifiers. Let $\mathcal{X} \subseteq \{x \in \mathbb{R}^d : \|x\|_2 \le 1\}$, take $\mathcal{A} = [-1,1]^K$, and define

$$\mathcal{H}_{\mathrm{lin}} := \{x \mapsto Wx : W \in \mathbb{R}^{K\times d}, \|W_y\|_2 \le 1 \text{ for every } y\},$$

where W_y is row y of W. Let $\mathcal{L}_{\mathrm{GLM}}$ consist of the losses

$$\ell(a,y) = \frac{\omega(a) - \omega(0) - a_y}{2}, \qquad a \in \mathcal{A},$$

where $\omega : \mathbb{R}^K \to \mathbb{R}$ is convex and differentiable, with $\nabla\omega(a) \in \Delta_K$ for every a. The gradient condition gives $|\omega(a) - \omega(0)| \le \|a\|_\infty$, and these losses take values in $[-1,1]$ on $\mathcal{A}$. Centering and scaling preserve best responses on this cube. All best responses in this application minimize over $\mathcal{A}$. For example, $\omega(a) = \log\sum_{j=0}^{K-1} e^{a_j}$ gives multiclass logistic loss with scores in $[-1,1]^K$, after this centering and scaling.

The hypothesis tests do not depend on the potential ω. For every $\ell \in \mathcal{L}_{\mathrm{GLM}}$,

$$\Delta_i\ell(Wx) = \frac{(Wx)_0 - (Wx)_i}{2}, \qquad i = 0,\dots,K-1,$$

and, for every $q \in \Delta_K$ and $y \in \mathcal{Y}$,

$$\big\langle \Delta\ell(Wx), e_y - q\big\rangle = \frac{1}{2}\langle Wx, q - e_y\rangle. \tag{3.7}$$

Thus one learner for the linear benchmarks controls the hypothesis tests for all the losses in $\mathcal{L}_{\mathrm{GLM}}$.

We use the matrix learners of Hu, Tian, and Yang [HTY26]. Each learner chooses W_t from the preceding rounds, with $\|(W_t)_i\|_2 \le 1$ for every row i, and supplies the test

$$g_t(x)_i = \frac{(W_tx)_0 - (W_tx)_i}{2}, \qquad i = 0,\dots,K-1. \tag{3.8}$$

These tests take values in $[-1,1]^K$. For the online guarantee below, the matrix learner receives x_t and $q_t - e_{y_t}$ after observing the current prediction and label. In the population setting, it receives x_t and $q_t(x_t) - e_{y_t}$ after observing the fresh example. The following two guarantees use their online and statistical learners, respectively.

Lemma 3.19 (GLM hypothesis learners)**.** *Fix $K \ge 2$ and a horizon $n \ge 1$.*

1. *The online learner of Hu, Tian, and Yang [HTY26, Lemma 41] chooses the tests in* (3.8) *so that*

$$\begin{aligned}\sup_{g\in\Delta\mathcal{L}_{\mathrm{GLM}}\circ\mathcal{H}_{\mathrm{lin}}} \left|\sum_{t=1}^{n}\langle g(x_t), e_{y_t} - q_t\rangle\right| \\ \le \sum_{t=1}^{n}\langle g_t(x_t), e_{y_t} - q_t\rangle + \sqrt{Kn}.\end{aligned} \tag{3.9}$$

This holds for every adaptive sequence of contexts, predictions, and labels (x_t, q_t, y_t), including choices that depend on the current test g_t. In particular, it satisfies Definition 3.6 with

$$\mathrm{Reg}_{\Delta\mathcal{L}_{\mathrm{GLM}}\circ\mathcal{H}_{\mathrm{lin}}}(n) \le \sqrt{Kn}.$$

2. *Fix a distribution $\mathcal{D}$ and $0 < \delta < 1$. The statistical learner of Hu, Tian, and Yang [HTY26, Lemma 42] uses one independent example $(x_t, y_t) \sim \mathcal{D}$ per round. Suppose the whole rule $q_t : \mathcal{X} \to \Delta_K$ is chosen using preceding examples, the current tests, and independent randomness, before observing (x_t, y_t). With probability at least $1 - \delta$,*

$$\begin{aligned}\sup_{g\in\Delta\mathcal{L}_{\mathrm{GLM}}\circ\mathcal{H}_{\mathrm{lin}}} \left|\sum_{t=1}^{n} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\langle g(x), e_y - q_t(x)\rangle]\right| \\ \le \sum_{t=1}^{n} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\langle g_t(x), e_y - q_t(x)\rangle] + C\sqrt{Kn\log(2/\delta)},\end{aligned} \tag{3.10}$$

where C is an absolute constant. Each expectation holds the current prediction and test functions fixed.

For the expected guarantee below, run Algorithm 1 with the decision learner from Algorithm 3 and the online hypothesis learner of Lemma 3.19, and apply Theorem 3.8. For the confidence guarantee, use the same decision learner and the statistical hypothesis learner. For any $n \ge 1$, $0 < r \le 1/2$, and $0 < \delta < 1$, run the latter with failure probability $\delta/2$. The resulting predictor satisfies, with probability at least $1 - \delta$,

$$\mathrm{OmniReg}_{\mathcal{L}_{\mathrm{GLM}},\mathcal{H}_{\mathrm{lin}}}(p) \le C\frac{\mathrm{Reg}_{\mathrm{dec}}(n, r)}{n} + C\sqrt{\frac{K\log(4/\delta)}{n}} + 4r, \tag{3.11}$$

where $\mathrm{Reg}_{\mathrm{dec}}$ is defined in (3.4) and C is an absolute constant.

Corollary 3.20 (Multiclass GLM losses)**.** *For fixed $K \ge 2$ and $0 < \varepsilon < 1$, there is an algorithm that uses $\widetilde{O}_K(\varepsilon^{-\max\{K-1,(K+2)/2\}})$ independent examples from any distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$ and returns a predictor $p : \mathcal{X} \to \Delta_K$ satisfying*

$$\mathbb{E}\big[\mathrm{OmniReg}_{\mathcal{L}_{\mathrm{GLM}},\mathcal{H}_{\mathrm{lin}}}(p)\big] \le \varepsilon.$$

For $0 < \delta < 1$, the sample bound

$$\widetilde{O}_K\left(\varepsilon^{-\max\{K-1,(K+2)/2\}} + \frac{\log(2/\delta)}{\varepsilon^2}\right)$$

gives omniregret at most ε with probability at least $1 - \delta$ over training.

For fixed $K \ge 4$ and constant confidence, the sample bound improves from $\widetilde{O}_K(\varepsilon^{-(K+1)})$ in Hu, Tian, and Yang [HTY26, Theorem 43] to $\widetilde{O}_K(\varepsilon^{-(K-1)})$.

3.6 From online correlations to population OI

On a new context, our predictor chooses one of the saved rules $q_1, \ldots, q_n$ uniformly, applies that rule, and smooths the resulting prediction. We want to bound the omniprediction error of this predictor on fresh data. The OI reduction and smoothing lemmas reduce this task to controlling the decision and hypothesis test correlations at the base prediction, using the averaged decision weights.

The proof, in Appendix A.4, uses sequential symmetrization [RST15]. Each training example is paired with an independent copy, and a fair sign decides which member of the pair the algorithm trains on. Conditional on the training run, the unused examples are fresh draws from $\mathcal{D}$, so the expected gap between the population and training correlations, uniformly over tests, is at most a signed sum of correlations along the training sequence. An auxiliary copy of each online learner bounds this signed sum by its regret: the decision learner is run on the signed residuals, and the hypothesis learner on synthetic labels with the same conditional mean. Each auxiliary learner chooses its tests before seeing the current sign, so its own payoff has expectation zero. This bounds the two population–training gaps by $2\,\mathrm{Reg}_{\mathrm{dec}}(n,r)/n$ and $4\,\mathrm{Reg}_{\Delta\mathcal{L}\circ\mathcal{H}}(n)/n$. Adding the training bound of Lemma 3.7 and the two smoothing errors of $2r$ proves the theorem.

Proof of Theorem 3.2. Take r to be a sufficiently small multiple of ε, with the constant depending only on K and at most $1/2$. Substituting this choice into (3.4), the geometric sample requirements are, respectively, $O_K(\varepsilon^{-2}\log(2/\varepsilon))$ for $K=2$, $O_K(\varepsilon^{-5/2})$ for $K=3$, $O_K(\varepsilon^{-3}\log^2(2/\varepsilon))$ for $K=4$, and $O_K(\varepsilon^{-(K-1)})$ for $K\geq 5$. The stated lower bound on n dominates all four requirements. The population theorem, the assumed hypothesis-regret bound, and a sufficiently small choice of its constant then make the total error at most ε. □

4 Lower bounds by forcing memorization

One of the first observations about omnipredictors is that p^* is an omnipredictor [Gop+22]. However, learning a complicated p^* can be statistically infeasible. For example, assign independent fair binary labels to N equally likely contexts and keep these labels fixed. Averaged over the random labeling and a training sample of size n, every estimator of $p_1^*(x)=\Pr[y=1\mid x]$ has population squared error at least $(1-n/N)/4$, since each unobserved label remains a fair bit. Thus, when $n\ll N$, even approximating the conditional label probabilities to small squared error is impossible. This difficulty motivates the question: do *statistically efficient* omnipredictors exist?

In the binary setting the answer is strongly affirmative: omnipredictors can be learned at nearly the same statistical rate as minimizing a single loss *regardless of the complexity of the underlying Bayes rule* [OKK25]. For a fixed finite benchmark class $\mathcal{H}$, the $K=2$ case of Theorem 3.1 gives an algorithm that uses

$$\widetilde{O}\left(\frac{\log(2|\mathcal{H}|)}{\varepsilon^2}\right)$$

samples and achieves expected omniprediction error at most ε for all bounded Lipschitz losses, roughly the same statistical rate as minimizing a single loss. The bound depends on the benchmark class, but not on the complexity of $p^*(x)=\Pr[y=1\mid x]$. In particular, it holds in the agnostic learning setting where we make no assumption that p^* belongs to $\mathcal{H}$.

For multiclass prediction, minimizing a single bounded loss against a finite class $\mathcal{H}$ still takes only $O(\log(2|\mathcal{H}|)/\varepsilon^2)$ samples for expected excess loss at most ε. In contrast, we show that the sample complexity of multiclass omniprediction for Lipschitz losses grows as $(1/\varepsilon)^{K-1}$ even when the hypothesis class is small.

Our lower bounds show how broad multiclass loss classes can force an omnipredictor to memorize significant information about the Bayes rule. We formalize this notion of memorization by constructing distributions determined by independent hidden signs and show that, given any predictor with small omniprediction error, we can accurately recover most of these signs. In some sense, there are just too many losses in high dimensions; in particular, we can construct a hard distribution over choice of p^* with large support, where for each supported Bayes rule, there is a loss in the simple

class that "captures" the p^*, requiring the omnipredictor to memorize it. Much of the technical novelty of our lower bounds comes from finding these simple subclasses of several natural loss classes that force memorization of p^*.

Section 4.1 formalizes sign memorization and collects the tools shared by our lower bounds. Section 4.2 proves tight lower bounds for omniprediction with respect to $\mathcal{L}_{\mathrm{Lip}}$, even with a single hypothesis. Section 4.3 shows a barrier to Loss OI: $\mathcal{L}_{\mathrm{All}}$ Decision OI together with a single multiaccuracy condition is already statistically expensive. Section 4.4 shows that omniprediction for all proper losses is unlearnable for $K \geq 3$, even against a known class of two hypotheses, and Section 4.5 gives a lower bound for Lipschitz proper losses.

4.1 Forcing the predictor to remember hidden signs

In each construction, the distribution is determined by M independent binary choices. At each context, one choice specifies either a deterministic label or the direction of a small perturbation of the conditional label distribution. We show that small regret lets us recover at least three quarters of these choices on average, using only the learned output.

The challenge is that omniprediction controls an average over contexts. Poor performance at one context can be offset by better performance elsewhere. We choose losses for which a benchmark minimizes true expected loss at every context, so local excess losses are nonnegative. This lets us make mistakes in the hidden signs accumulate into global omniprediction regret.

Once most of the hidden signs can be reconstructed, a sample lower bound follows. Even when each observed label reveals its context's sign completely, this requires $\Omega(M)$ examples. When the two possible label distributions are only η apart, the requirement in our constructions rises to $\Omega_K(M/\eta^2)$. The following lemma gives both bounds; its proof is deferred to Appendix B.1.

Lemma 4.1 (Samples needed for sign decoding)**.** *Fix M distinct contexts $x_1, \dots, x_M$ and any distribution on them. Choose independent fair bits $\sigma_1, \dots, \sigma_M$. Under $\mathcal{D}_\sigma$, the context distribution is fixed, and σ_i selects one of two fixed label distributions at x_i.*

Let s consist of n independent examples from $\mathcal{D}_\sigma$. Suppose a decoder given s, possibly using fresh randomness, returns $\widehat{\sigma}$ satisfying

$$\mathbb{E}[d_{\mathrm{H}}(\widehat{\sigma}, \sigma)] \leq \delta M, \qquad 0 \leq \delta < \tfrac{1}{2},$$

where the expectation includes the hidden bits, the sample, and the decoder's randomness. Then

$$n \geq (1 - 2\delta)M.$$

Moreover, suppose the two possible label distributions at each context are at Euclidean distance at most $\eta > 0$, and their average assigns probability at least $1/(2K)$ to every label. Then

$$n = \Omega_{K,\delta}(M/\eta^2).$$

The constructions establish recovery for every fixed sign assignment. They use the learned predictor p, or the returned decision rules in the Lipschitz proper-loss result, together with the known construction.

Well-spaced grids in the simplex. We use well-spaced grid points as contexts so that the losses can assign separate costs to mistakes at different contexts. We want many hidden signs, but the spacing must also allow the losses to penalize treating nearby contexts alike. Some constructions use a grid in the full simplex. Others use the slice $x_0 = x_1$, which has one fewer free coordinate. Keeping these two label probabilities equal lets us encode a hidden sign by increasing one and decreasing the other, perpendicular to the grid. The following lemma gives both grids.

Lemma 4.2 (Regular grids in the simplex)**.** *Fix $K \geq 2$. For all sufficiently small $s > 0$, the region*

$$\{x \in \Delta_K : x_y \geq 1/(2K) \text{ for every } y\}$$

contains a grid with $\Theta_K(s^{-(K-1)})$ points. The intersection of this region with the slice $x_0 = x_1$ contains a grid with $\Theta_K(s^{-(K-2)})$ points. In each grid, the free coordinates vary in steps of s, and distinct points have Euclidean distance at least s.

Proof. Choose each free coordinate independently from

$$\frac{1}{K} + s\left\{0, \ldots, \left\lfloor \frac{1}{2K^2 s} \right\rfloor\right\}.$$

For the first grid, choose $x_1, \ldots, x_{K-1}$ and set $x_0 = 1 - \sum_{y=1}^{K-1} x_y$. For the second, choose $x_2, \ldots, x_{K-1}$ and set $x_0 = x_1 = (1 - \sum_{y=2}^{K-1} x_y)/2$. Each free coordinate lies between $1/K$ and $1/K + 1/(2K^2)$, so the remaining coordinates are at least $1/(2K)$. There are $\Theta_K(s^{-1})$ choices per free coordinate, giving the stated counts, and two distinct points differ by at least s in a free coordinate. □

4.2 Lipschitz losses

Theorem 4.3. *Fix $K \geq 2$. For every sufficiently small $\varepsilon > 0$, there is a finite context space $\mathcal{X} \subseteq \Delta_K$ such that the following holds. Let $h_{\mathrm{id}}(x) = x$. Consider a learner that returns a possibly randomized predictor p from n independent samples drawn from $\mathcal{D}$. If, for every distribution $\mathcal{D}$ on this context and label space,*

$$\mathbb{E}\Big[\mathrm{OmniReg}_{\mathcal{L}_{\mathrm{Lip}},\{h_{\mathrm{id}}\};\mathcal{D}}(p)\Big] \leq \varepsilon,$$

then the learner must use

$$n = \Omega_K\Big(\varepsilon^{-\max\{K-1,(K+2)/2\}}\Big) \tag{4.1}$$

samples.

For fixed K, the two constructions below together match the ε-dependence of Theorem 3.1, up to logarithmic factors, using only the identity benchmark. The constructions trade off the number of hidden labels or signs against the number of observations needed to learn each one.

The first construction places $\Theta_K(\varepsilon^{-(K-1)})$ hidden deterministic labels on a grid in the full simplex with spacing proportional to ε. One observation at a context reveals its label. Small omniprediction error forces recovery of at least three quarters of these labels on average, requiring $\Omega_K(\varepsilon^{-(K-1)})$ samples.

The second construction uses a grid on the slice $x_0 = x_1$ with spacing proportional to $\sqrt{\varepsilon}$, giving $\Theta_K(\varepsilon^{-(K-2)/2})$ contexts. At each context, a hidden sign changes the two label probabilities by order ε in the direction $e_1 - e_0$. Recovering a sign with probability at least $3/4$ takes order ε^{-2} observations at its context. Small omniprediction error again forces recovery of at least three quarters of the signs on average, now requiring

$$n = \Omega_K\Big(\varepsilon^{-(K-2)/2} \cdot \varepsilon^{-2}\Big) = \Omega_K\Big(\varepsilon^{-(K+2)/2}\Big)$$

samples.

The second construction is stronger for two or three labels, the bounds agree for four labels, and the first is stronger for at least five labels. For Theorem 4.3, choose the construction with the larger exponent.

4.2.1 Proof of the $(1/\varepsilon)^{K-1}$ lower bound

Proof sketch. We take a regular grid of $M = \Theta_K(\varepsilon^{-(K-1)})$ points $x_1, \dots, x_M$ in the simplex, separated by a sufficiently large multiple of ε, and use them as equally likely contexts. Each context x_i has an independent fair label $\sigma_i \in \{0, 1\}$. We show that small omniprediction error lets us recover most of these labels from the learned predictor.

For every hidden labeling σ, we exhibit a loss $\ell_\sigma \in \mathcal{L}_{\mathrm{Lip}}$ for which identity is optimal at every context. It might seem that the predictor could match the identity hypothesis simply by predicting $p(x) = x$. However, the predictor does not directly select the downstream action. Instead, the prediction is converted into a best response that minimizes expected loss under the predicted distribution.

We seek losses for which identity incurs $1/2 - 4\varepsilon$ under the true label at every context, while every fixed action has expected loss at least $1/2$ under an independent fair label. We also require their properizations to agree, so this gap applies at an unseen context even though the loss depends on the hidden labeling.

The main question is how ℓ_σ can depend on σ while $\ell_\sigma^{\mathrm{pr}}$ does not. As a varies over Δ_K, $\ell_\sigma(a, \cdot)$ ranges over the same set for every σ, although the actions attaining those label losses may change.

Set $\beta = 16\varepsilon$. The losses have the form

$$\ell_\sigma(a, y) = \frac{1}{2} + \frac{f_\sigma(a)^2}{4\beta} - \frac{f_\sigma(a)\tau(y)}{2},$$

where $f_\sigma : \Delta_K \to [-\beta, \beta]$, $\tau(0) = -1$, and $\tau(y) = 1$ for $y \neq 0$. For a fixed σ and any action a,

$$\mathbb{E}_{y \sim \mathrm{Ber}(1/2)}\big[\ell_\sigma(a, y)\big] = \frac{1}{2} + \frac{f_\sigma(a)^2}{4\beta} \geq \frac{1}{2},$$

because $\tau(y)$ has mean zero and the quadratic term is nonnegative. For a known outcome y, the minimum is $1/2 - \beta/4$, attained when $f_\sigma(a) = \beta\tau(y)$. Prescribing $f_\sigma(x_i) = \beta\tau(\sigma_i)$ therefore makes identity optimal.

The hidden labels determine these prescribed values, but we ensure that every f_σ takes every value in $[-\beta, \beta]$. A best response to a prediction p minimizes the same quadratic over the same interval. Its unique minimizer gives

$$f_\sigma\big(\mathrm{BR}_{\ell_\sigma}(p)\big) = \beta(1 - 2p_0).$$

Although the selected action can depend on σ, its losses for all labels depend only on p. Thus the properization is the same for every hidden labeling.

At a context x_i absent from the sample, condition on the sample and all randomness used to produce the prediction. The label σ_i remains a fair bit. Since the properization is independent of the hidden labeling, the fair-label calculation gives expected loss at least $1/2$. Identity has loss $1/2 - \beta/4$, so an unobserved context costs expected excess loss at least $\beta/4$.

Completing the square gives the common properization

$$\ell_\sigma^{\mathrm{pr}}(p, y) = \frac{1}{2} - \frac{\beta}{4} + \beta\big(p_0 - \mathbf{1}_{\{y=0\}}\big)^2.$$

For the formal recovery argument, the decoder returns label 0 when $p_0(x_i) \geq 1/2$, and label 1 otherwise. Thus every decoding error costs at least $\beta/4 = 4\varepsilon$ more than identity, while excess loss is nonnegative at every other context. Small average regret therefore forces most labels to be decoded correctly. This decoder uses only the learned predictions, without knowing the hidden labeling or its associated loss.

Appendix B gives the loss construction and the full proof of the $\Omega_K(\varepsilon^{-(K-1)})$ sample lower bound.

4.2.2 Proof of the $(1/\varepsilon)^{K/2+1}$ lower bound

We now implement the noisy-sign construction sketched in Section 1.2.2. The key is to keep the identity action optimal under the true label distribution while adding a decoy that beats identity under predictions on the slice $p_0 = p_1$ or on the wrong side of it. The grid spacing makes actions associated with other contexts more costly under the true label distribution, and the distance penalty keeps every best response among the specified actions. Together, these properties turn an incorrectly predicted sign into excess loss.

Fix $0 < \beta \le 1/(4K)$; the proof below takes $\beta = 600K\varepsilon$.

Hard distribution. By Lemma 4.2, take a grid $\mathcal{X}$ in the slice $x_0 = x_1$, with pairwise distances at least $8\sqrt{\beta}$ and every coordinate at least $1/(2K)$. Then

$$M := |\mathcal{X}| = \Theta_K\Big(\beta^{-(K-2)/2}\Big).$$

For each assignment $\sigma\colon \mathcal{X} \to \{-1, 1\}$, let $\mathcal{D}_\sigma$ sample contexts uniformly from $\mathcal{X}$, with conditional label distribution

$$p^*(x) = x + \sigma(x)\beta(e_1 - e_0).$$

We use the identity benchmark $h_{\mathrm{id}}(x) = x$ and $\mathcal{H} = \{h_{\mathrm{id}}\}$.

Hard losses. For each sign assignment σ, we construct a loss whose absolute values and Lipschitz constants are $O_K(1)$, then rescale it to belong to $\mathcal{L}_{\mathrm{Lip}}$. For each context x, add the action

$$a(x) = x + \frac{\sigma(x)}{4K}(e_1 - e_0).$$

This lies in Δ_K, since every coordinate of x is at least $1/(2K)$. On

$$A = \{x, a(x) : x \in \mathcal{X}\}$$

define

$$\ell_\sigma(x, y) = \frac{1}{2}\|x - e_y\|_2^2,$$

$$\ell_\sigma(a(x), y) = \ell_\sigma(x, y) + \sigma(x)\big(\mathbf{1}_{\{y=1\}} - \mathbf{1}_{\{y=0\}}\big) - \beta.$$

Since $\frac{1}{2}\|x - e_y\|_2^2 \in [0, 1]$ and $\beta \le 1/8$, all these assigned loss values lie in $[-3, 3]$.

For $a \in \Delta_K \setminus A$, extend the loss by adding a penalty for distance:

$$\ell_\sigma(a, y) = \min_{a' \in A}\big\{\ell_\sigma(a', y) + 72K\|a - a'\|_2\big\}.$$

The following lemma collects the properties of these losses used below and in Section 4.3. Its proof is in Appendix B.2.

Lemma 4.4 (Ridge losses)**.** *For every sign assignment σ and every context $x \in \mathcal{X}$:*

(i) $\ell_\sigma(x, y) = \frac{1}{2}\|x - e_y\|_2^2$ *for every label y;*

(ii) the action x minimizes expected loss under $p^*(x)$ over all of Δ_K;

(iii) for every prediction $p \in \Delta_K$ with $\sigma(x)(p_1 - p_0) \le 0$, every best response of ℓ_σ to p has expected loss under $p^*(x)$ at least β greater than that of x;

(iv) $\ell_\sigma/(150K) \in \mathcal{L}_{\mathrm{Lip}}$.

Let $\mathcal{L}_{\mathrm{ridge}} := \{\ell_\sigma/(150K) : \sigma\colon \mathcal{X} \to \{-1,1\}\}$. By item (iv), $\mathcal{L}_{\mathrm{ridge}} \subseteq \mathcal{L}_{\mathrm{Lip}}$, so the following proposition proves the second bound in Theorem 4.3.

Proposition 4.5. *Take $\beta = 600K\varepsilon$, with ε small enough that $\beta \le 1/(4K)$. Suppose a learner given n independent samples from $\mathcal{D}_\sigma$ returns a possibly randomized predictor p with $\mathbb{E}[\mathrm{OmniReg}_{\mathcal{L}_{\mathrm{ridge}},\{h_{\mathrm{id}}\};\mathcal{D}_\sigma}(p)] \le \varepsilon$ for every sign assignment σ. Then $n = \Omega_K(\varepsilon^{-(K+2)/2})$.*

Proof of Proposition 4.5. We first recover the hidden signs from the predictor.

Let $s \sim \mathcal{D}_\sigma^n$, and let p be the learner's output. Evaluate p once at each context, using fresh independent randomness if p is randomized. The decoder returns

$$\widehat{\sigma}(x) = \begin{cases} +1, & p_1(x) \ge p_0(x), \\ -1, & p_1(x) < p_0(x). \end{cases}$$

Each decoding error implies $\sigma(x)(p_1(x) - p_0(x)) \le 0$. By Lemma 4.4(iii), every best response then has true excess loss at least β. Positive rescaling preserves the set of best responses. For the loss $\ell_\sigma/(150K)$, each decoding error therefore incurs true excess loss at least $\beta/(150K)$, while all other conditional excess losses are nonnegative by Lemma 4.4(ii).

Averaging over the prediction randomness and the training run, and applying the learner's guarantee (the loss $\ell_\sigma/(150K)$ belongs to $\mathcal{L}_{\mathrm{ridge}}$), gives for every fixed assignment σ,

$$\frac{\beta}{150KM}\mathbb{E}[d_{\mathrm{H}}(\widehat{\sigma},\sigma)] \le \mathbb{E}\Big[\mathrm{OmniReg}_{\mathcal{L}_{\mathrm{ridge}},\mathcal{H};\mathcal{D}_\sigma}(p)\Big] \le \varepsilon.$$

Our choice $\beta = 600K\varepsilon$ gives expected Hamming error at most $M/4$.

Completing the lower bound. At each context, the two conditional label distributions are distance $2\sqrt{2}\beta$ apart, and their midpoint is x, whose coordinates are at least $1/(2K)$. Averaging over independent fair signs and applying the noisy bound in Lemma 4.1 with $\delta = 1/4$ and $\eta = 2\sqrt{2}\beta$ gives

$$n = \Omega_K\left(\frac{M}{\beta^2}\right) = \Omega_K\left(\varepsilon^{-(K+2)/2}\right),$$

using $M = \Theta_K(\beta^{-(K-2)/2})$ and $\beta = 600K\varepsilon$. □

4.2.3 Dependence on the benchmark class

The preceding constructions establish the cost of supporting all Lipschitz losses even against a single known benchmark. To complete the comparison with Theorem 3.1, we show that the $\log(2|\mathcal{H}|)/\varepsilon^2$ dependence on the benchmark class is also necessary in the worst case.

Theorem 4.6 (Lower bound for finite benchmark classes)**.** *Fix $K \ge 2$. For every sufficiently small $\varepsilon > 0$ and every integer $N \ge 1$, there are a finite context space $\mathcal{X} \subseteq \Delta_K$ and a class $\mathcal{H}$ of N hypotheses $\mathcal{X} \to \Delta_K$ with the following property. Any learner that uses n independent samples from $\mathcal{D}$ and returns a possibly randomized predictor p satisfying*

$$\mathbb{E}\big[\mathrm{OmniReg}_{\mathcal{L}_{\mathrm{Lip}},\mathcal{H};\mathcal{D}}(p)\big] \le \varepsilon$$

for every distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$ must satisfy

$$n = \Omega_K\left(\varepsilon^{-\max\{K-1,(K+2)/2\}} + \frac{\log(2N)}{\varepsilon^2}\right).$$

Proof. Theorem 4.3 gives the first term with one benchmark. Adding distinct hypotheses can only increase omniprediction error, so that lower bound also holds for a class of exactly N hypotheses.

For $N \geq 2$, set $M = \lfloor \log_2 N \rfloor$ and take M distinct contexts $x_1, \ldots, x_M \in \Delta_K$. For each $\sigma \in \{\pm 1\}^M$, let $h_\sigma(x_i) = e_1$ if $\sigma_i = 1$ and $h_\sigma(x_i) = e_0$ otherwise. Include these 2^M hypotheses in a class $\mathcal{H}$ of size N, adding distinct hypotheses if needed. Use the single loss

$$\ell(a, y) = \big|a_1 - \mathbf{1}_{\{y=1\}}\big|,$$

which is $[0,1]$-valued and 1-Lipschitz in the Euclidean norm. Let $\mathcal{D}_\sigma$ have uniform contexts and labels in $\{0,1\}$, with

$$\Pr_{\mathcal{D}_\sigma}[y = 1 \mid x_i] = \frac{1}{2} + 4\varepsilon\sigma_i,$$

taking $\varepsilon \leq 1/16$. For every action $a \in \Delta_K$,

$$\mathbb{E}_{y|x_i}\big[\ell(a, y) - \ell(h_\sigma(x_i), y)\big] = 8\varepsilon\big|a_1 - \mathbf{1}_{\{\sigma_i=1\}}\big|.$$

In particular, h_σ minimizes true expected loss at every context.

Evaluate the learned predictor once at each context and let $a_i = \mathrm{BR}_\ell(p(x_i))$. Decode $\widehat{\sigma}_i = 1$ when $(a_i)_1 \geq 1/2$, and $\widehat{\sigma}_i = -1$ otherwise. Every decoding mistake incurs excess loss at least 4ε. Thus, for every fixed σ,

$$\frac{4\varepsilon}{M}\mathbb{E}[d_{\mathrm{H}}(\widehat{\sigma}, \sigma)] \leq \mathbb{E}\big[\mathrm{OmniReg}_{\mathcal{L}_{\mathrm{Lip}}, \mathcal{H}; \mathcal{D}_\sigma}(p)\big] \leq \varepsilon.$$

Averaging over independent fair signs gives expected Hamming error at most $M/4$. The two conditional label distributions are $8\sqrt{2}\,\varepsilon$ apart and have midpoint uniform on $\{0,1\}$. The binary case of Lemma 4.1, with $\delta = 1/4$, therefore gives

$$n = \Omega(M/\varepsilon^2) = \Omega(\log N/\varepsilon^2).$$

For $N \geq 2$, $\log(2N) \leq 2\log N$, and the maximum of the two lower bounds is within a constant factor of their sum. Use the class from the construction giving the larger lower bound. For $N = 1$, the first term already dominates ε^{-2}. □

Together with Theorem 3.1, this determines the minimax sample complexity for expected omniprediction error at most ε, up to logarithmic factors for each fixed K. The worst case ranges over context spaces, benchmark classes of size N, and data distributions.

4.3 Proper calibration and identity multiaccuracy

Our upper bounds obtain omniprediction by simultaneously satisfying proper calibration and multiaccuracy for the tests determined by the losses and benchmarks, using (2.7). This is the approach used for binary omniprediction by Okoroafor, Kleinberg, and Kim [OKK25], building on the loss-OI reduction of Gopalan, Hu, Kim, Reingold, and Wieder [Gop+23].

Proper calibration alone is easy to learn by bundling all contexts into one prediction and predicting the empirical mean label vector everywhere. This uses $O(K/\varepsilon^2)$ samples for expected error ε. A single bounded multiaccuracy test also needs only $O(1/\varepsilon^2)$ samples. However, we show that requiring both guarantees from the same predictor in the multiclass setting can make the exponent of $1/\varepsilon$ grow linearly with K, even when the only multiaccuracy test is identity. For contexts in Δ_K, this error is

$$\mathrm{MA}_{\mathrm{id}}(p) = \big|\mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\langle x, e_y - p(x)\rangle\big]\big|.$$

Theorem 4.7. *Fix $K \geq 2$. Suppose that, for every distribution $\mathcal{D}$ on $\Delta_K \times \mathcal{Y}$, a learner given n independent samples outputs a possibly randomized predictor p with*

$$\mathbb{E}[\mathrm{PCal}_K(p)] \leq \varepsilon, \qquad \mathbb{E}[\mathrm{MA}_{\mathrm{id}}(p)] \leq \varepsilon.$$

For all sufficiently small $\varepsilon > 0$,

$$n = \Omega_K\left(\varepsilon^{-(K+2)/2}\right).$$

Proof. We exhibit a loss class for which all hypothesis-OI correlations are the same constant multiple of the identity correlation. The loss-OI reduction then turns proper calibration and identity multiaccuracy into omniprediction for this class, allowing us to apply the second Lipschitz lower bound.

We reuse the losses, identity hypothesis, and hard distribution from Section 4.2.2, with $\beta = 600K\varepsilon'$ for $\varepsilon' = (1 + 1/(150K))\varepsilon$. Let $\mathcal{L} = \mathcal{L}_{\mathrm{ridge}}$ consist of the normalized losses $\ell_\sigma/(150K)$ from that construction, and let $h_{\mathrm{id}}(x) = x$. By Lemma 4.4(i), for every $\ell \in \mathcal{L}$ and every context $x \in \mathcal{X}$,

$$\ell(x, y) = \frac{1}{2 \cdot 150K}\|x - e_y\|_2^2.$$

Consequently,

$$\Delta_y \ell(x) = \frac{x_0 - x_y}{150K}.$$

Since the coordinates of $e_y - p(x)$ sum to zero,

$$\langle \Delta\ell(x), e_y - p(x)\rangle = -\frac{1}{150K}\langle x, e_y - p(x)\rangle.$$

Thus every hypothesis-OI test gives the same scaled identity correlation, and for every distribution on $\mathcal{X} \times \mathcal{Y}$ and every predictor p,

$$\mathrm{HypOI}_{\mathcal{L},\{h_{\mathrm{id}}\}}(p) = \frac{1}{150K}\,\mathrm{MA}_{\mathrm{id}}(p).$$

The loss-OI reduction (2.7) therefore gives

$$\mathrm{OmniReg}_{\mathcal{L},\{h_{\mathrm{id}}\}}(p) \leq \mathrm{PCal}_K(p) + \frac{1}{150K}\,\mathrm{MA}_{\mathrm{id}}(p).$$

Taking expectations over training and applying the two assumptions of the theorem gives

$$\mathbb{E}\Big[\mathrm{OmniReg}_{\mathcal{L},\{h_{\mathrm{id}}\}}(p)\Big] \leq \left(1 + \frac{1}{150K}\right)\varepsilon.$$

Applying Proposition 4.5 at this accuracy therefore requires

$$n = \Omega_K\left(\left(\left(1 + \frac{1}{150K}\right)\varepsilon\right)^{-(K+2)/2}\right) = \Omega_K\left(\varepsilon^{-(K+2)/2}\right).$$

□

4.4 Omniprediction is unlearnable for proper losses

We now consider omniprediction for proper losses. With a single hypothesis the identity best response lets us copy a single benchmark with zero regret. However, we show that with two hypotheses, the class becomes unlearnable. We show that, already for $K = 3$, no finite sample bound guarantees small expected omniprediction error uniformly over distributions.

Theorem 4.8. *Let $K = 3$. There exist two fixed affine hypotheses $h_0, h_1 : [0,1] \to \Delta_3$ such that, writing $\mathcal{H} = \{h_0, h_1\}$, the following holds. For every sample size n and every learner returning a possibly randomized predictor p from n independent examples, there is a finitely supported distribution $\mathcal{D}$ on $[0,1] \times \{0,1\}$, with labels a deterministic function of the context, for which*

$$\mathbb{E}\Big[\mathrm{OmniReg}_{\mathcal{L}_{\mathrm{prop}},\mathcal{H};\mathcal{D}}(p)\Big] \geq \frac{1}{4}.$$

The expectation is over the training sample and training randomness. Only $[0,1]$-valued proper losses are needed.

The conclusion also holds if the learner returns a separate, possibly randomized decision rule a_ℓ for each loss ℓ, all learned from the same sample. In this case, replace $\ell(\mathrm{BR}_\ell(p(x)), y)$ by $\ell(a_\ell(x), y)$ in the regret, averaging population losses over action randomness.

Proof. We prove the stronger statement allowing a separate decision rule a_ℓ for each loss ℓ.

Fix n, choose an integer $M > 4n$, and set $x_i = i/M$ for $i = 1, \ldots, M$. For each $\sigma \in \{0,1\}^M$, let $\mathcal{D}_\sigma$ be uniform on the labeled examples (x_i, σ_i).

Use the two fixed affine hypotheses

$$h_0(x) = \left(\frac{x}{4}, \frac{x}{4}, 1 - \frac{x}{2}\right), \qquad h_1(x) = \left(\frac{x+1}{4}, \frac{x+1}{4}, \frac{1-x}{2}\right).$$

The $2M$ reports $h_0(x_i), h_1(x_i)$ are all distinct and have equal first two coordinates.

The construction uses the fact that the tie set $p_0 = p_1$ contains infinitely many reports when $K = 3$. At such a report, either loss vector $(0,1,0)$ or $(1,0,0)$ is compatible with properness. We use this freedom to encode arbitrary labels at the benchmarks' reports.

We now define the losses ℓ_σ. All losses agree at reports $p = (p_0, p_1, p_2)$ with $p_0 \neq p_1$. If $p_0 > p_1$, set $\ell_\sigma(p, 0) = 0$ and $\ell_\sigma(p, 1) = 1$. If $p_1 > p_0$, set $\ell_\sigma(p, 0) = 1$ and $\ell_\sigma(p, 1) = 0$. Thus the less likely of labels 0 and 1 incurs loss one.

At reports with $p_0 = p_1$, we choose the losses so that under $\mathcal{D}_\sigma$, hypothesis h_0 always incurs loss zero and h_1 always incurs loss one. Specifically, set

$$\ell_\sigma(h_0(x_i), 0) = \sigma_i, \qquad \ell_\sigma(h_1(x_i), 0) = 1 - \sigma_i.$$

At every other report with $p_0 = p_1$, set $\ell_\sigma(p, 0) = 0$. For all reports with $p_0 = p_1$, set $\ell_\sigma(p, 1) = 1 - \ell_\sigma(p, 0)$. Finally, set $\ell_\sigma(p, 2) = 0$ for every report p.

To verify propriety, fix a true label distribution p^* and any report a. If $p_0^* \neq p_1^*$, suppose first that $p_0^* > p_1^*$. Then

$$\begin{aligned} \mathbb{E}_{y \sim p^*}[\ell_\sigma(a, y)] &= p_0^* \ell_\sigma(a, 0) + p_1^* \ell_\sigma(a, 1) \\ &= p_1^* + (p_0^* - p_1^*)\ell_\sigma(a, 0) \\ &\geq p_1^* = \mathbb{E}_{y \sim p^*}[\ell_\sigma(p^*, y)]. \end{aligned}$$

The case $p_1^* > p_0^*$ follows by interchanging labels 0 and 1. If $p_0^* = p_1^*$, then

$$\mathbb{E}_{y \sim p^*}[\ell_\sigma(a, y)] = p_0^*\big(\ell_\sigma(a, 0) + \ell_\sigma(a, 1)\big) = p_0^* = \mathbb{E}_{y \sim p^*}[\ell_\sigma(p^*, y)].$$

Thus every ℓ_σ is proper.

Now choose σ uniformly from $\{0,1\}^M$ and draw the training sample from $\mathcal{D}_\sigma$. Condition on the sample and the learner's training randomness, fixing the returned decision rules. For each assignment $\tau \in \{0,1\}^M$, define

$$\mathrm{Reg}(\tau) := \max_{\substack{\rho\in\{0,1\}^M \\ b\in\{0,1\}}} \mathbb{E}_{(x,y)\sim\mathcal{D}_\tau}\big[\ell_\rho(a_{\ell_\rho}(x),y) - \ell_\rho(h_b(x),y)\big].$$

The expectation includes action randomness. The loss family and the returned rules are fixed; only the distribution used to evaluate them varies with τ.

Let $s \le n$ be the number of distinct observed contexts. Define σ' by keeping the observed labels and flipping all unseen labels. The unseen labels remain independent fair bits, so σ and σ' have the same conditional distribution.

Evaluate both distributions using the same loss ℓ_σ and the same returned rule a_{ℓ_σ}. The comparison under ℓ_σ is included in the regret under either distribution, since both regrets maximize over the entire constructed loss family. Under $\mathcal{D}_{\sigma'}$, hypothesis h_1 is correct at every unseen context and wrong at every observed context, giving population loss s/M.

The learner uses the same returned rule a_{ℓ_σ} under both distributions. At each unseen context the two labels are opposite, and

$$\ell_\sigma(a,0) + \ell_\sigma(a,1) = 1 \qquad \text{for every action } a,$$

also after averaging over action randomness. The learner's two population losses therefore sum to at least $(M-s)/M$. Comparing with h_0 under the first distribution and h_1 under the second gives

$$\mathrm{Reg}(\sigma) + \mathrm{Reg}(\sigma') \ge \frac{M-s}{M} - \frac{s}{M} = 1 - \frac{2s}{M}.$$

The two regrets have the same conditional expectation. Taking expectations and using $s \le n$ therefore yields

$$\mathbb{E}[\mathrm{Reg}(\sigma)] \ge \frac{1}{2} - \frac{n}{M} > \frac{1}{4}.$$

This expectation also averages over σ, so some fixed σ gives expected regret greater than $\frac{1}{4}$ over the training sample and training randomness. □

Corollary 4.9. *With $K = 3$ and the same two affine hypotheses, Theorem 4.8 also holds for the class of all $[0,1]$-valued strictly proper losses, with lower bound $\frac{1}{8}$. This includes the stronger formulation allowing a separate decision rule for each loss, all learned from the same sample.*

Proof. Fix n and use the finite family of proper losses constructed in the preceding proof. For each loss ℓ in this family, define

$$\ell'(p,y) = \frac{15}{16}\ell(p,y) + \frac{1}{32}\|p - e_y\|_2^2.$$

We first verify strict propriety. For a true label distribution p^* and any report a, expanding the squares and using $\mathbb{E}_{y\sim p^*}[e_y] = p^*$ gives

$$\begin{aligned}\mathbb{E}_{y\sim p^*}\big[\|a-e_y\|_2^2 - \|p^*-e_y\|_2^2\big] &= \|a\|_2^2 - 2\langle a, p^*\rangle + \|p^*\|_2^2 \\ &= \|a-p^*\|_2^2.\end{aligned}$$

Since ℓ is proper,

$$\mathbb{E}_{y\sim p^*}\big[\ell'(a,y) - \ell'(p^*,y)\big] \ge \frac{1}{32}\|a-p^*\|_2^2.$$

This is strictly positive whenever $a \neq p^*$, so ℓ' is strictly proper. It also takes values in $[0,1]$, since $0 \leq \ell(p,y) \leq 1$ and $\|p - e_y\|_2^2 \leq 2$.

Now fix any learner for the strictly proper losses. Construct a learner for the original finite loss family by running the given learner on the same training sample and, for each original loss ℓ, returning the rule it returns for ℓ'. The preceding proof allows arbitrary decision rules for each loss, learned from the same sample, so it applies to this learner. It gives a distribution $\mathcal{D}$ under which these rules have expected maximum regret at least $\frac{1}{4}$ when evaluated using the original losses.

Write Reg and Reg' for the maximum population regrets of these same rules under the original and perturbed loss families, respectively, on this distribution. For every report a and label y,

$$|\ell'(a,y) - \ell(a,y)| = \frac{1}{16}\left|\frac{1}{2}\|a - e_y\|_2^2 - \ell(a,y)\right| \leq \frac{1}{16},$$

because both $\frac{1}{2}\|a - e_y\|_2^2$ and $\ell(a,y)$ lie in $[0,1]$. Thus the learner's population loss and each benchmark's population loss change by at most $\frac{1}{16}$. Taking maxima over the corresponding losses and the same benchmarks gives

$$\mathrm{Reg}' \geq \mathrm{Reg} - \frac{1}{8}$$

for every training sample and realization of training randomness. Taking expectations therefore yields

$$\mathbb{E}[\mathrm{Reg}'] \geq \mathbb{E}[\mathrm{Reg}] - \frac{1}{8} \geq \frac{1}{4} - \frac{1}{8} = \frac{1}{8}.$$

All losses in the perturbed family are $[0,1]$-valued and strictly proper, proving the claim. □

4.5 Lipschitz proper losses

The preceding unlearnability construction uses discontinuous proper losses. Restricting to 1-Lipschitz losses restores finite-sample learnability for finite benchmark classes by Theorem 3.1, but the sample complexity can still have an exponent growing linearly with K. We prove this using two fixed hypotheses. As in the preceding result, the lower bound holds even when the learner returns a separate decision rule for each loss, all learned from the same sample.

Theorem 4.10 (Lipschitz proper losses)**.** *Fix $K \geq 2$, and let $\mathcal{L} = \mathcal{L}_{\mathrm{Lip}} \cap \mathcal{L}_{\mathrm{prop}}$. For every sufficiently small $\varepsilon > 0$, there are a finite context space $\mathcal{X}_{\mathrm{prop}} \subseteq \Delta_K$ and two fixed hypotheses $h_+, h_- : \mathcal{X}_{\mathrm{prop}} \to \Delta_K$ with the following property. Write $\mathcal{H}_{\mathrm{prop}} = \{h_+, h_-\}$. Any learner given n independent samples and returning a possibly randomized predictor p satisfying, for every distribution $\mathcal{D}$ on $\mathcal{X}_{\mathrm{prop}} \times \mathcal{Y}$,*

$$\mathbb{E}\Big[\mathrm{OmniReg}_{\mathcal{L},\mathcal{H}_{\mathrm{prop}};\mathcal{D}}(p)\Big] \leq \varepsilon$$

must use

$$n = \Omega_K\left(\varepsilon^{-(K+1)/2}\right)$$

samples.

The same bound holds if the learner returns a possibly randomized decision rule a_ℓ for each loss ℓ, all learned from the same sample, and the regret is defined using $\ell(a_\ell(x), y)$ in place of $\ell(\mathrm{BR}_\ell(p(x)), y)$. Population risks average over action randomness, while $\mathbb{E}_{\mathrm{train}}$ averages over the sample and training randomness.

Proof sketch. Take the full-simplex grid from Lemma 4.2 with spacing 8α, where $\alpha^2 = 192\varepsilon$. It contains $M = \Theta_K(\alpha^{-(K-1)})$ contexts $x_1, \dots, x_M$, each with every coordinate at least $1/(2K)$. Define the two hypotheses by

$$h_+(x_i) = x_i + \frac{\alpha}{2\sqrt{2}}(e_1 - e_0), \qquad h_-(x_i) = x_i - \frac{\alpha}{2\sqrt{2}}(e_1 - e_0).$$

For sufficiently small ε, both reports lie in Δ_K. They are distance α apart and have midpoint x_i. Under $\mathcal{D}_\sigma$, contexts are uniform and a hidden assignment $\sigma \in \{-1, +1\}^M$ selects the true conditional label distribution $p^*(x_i) = h_{\sigma_i}(x_i)$.

The key ingredient is a family of proper, 1-Lipschitz losses that distinguish the two benchmark reports on a chosen set of contexts and make them tie elsewhere. Write $B(a, y) = \frac{1}{2}\|a - e_y\|_2^2$. Lemma B.3 constructs a loss ℓ_J for every $J \subseteq [M]$, with $\ell_J(x_i, y) = B(x_i, y)/3$ and

$$\ell_J(h_\pm(x_i), y) = \begin{cases} B(h_\pm(x_i), y)/3, & i \in J, \\ B(x_i, y)/3, & i \notin J. \end{cases}$$

The hypotheses and this entire loss family are fixed before σ is chosen.

Fix the decision rules returned on the training sample. For each candidate assignment τ, evaluate these same rules under $\mathcal{D}_\tau$ and define

$$\mathrm{Reg}(\tau) := \max_{\substack{J \subseteq [M] \\ h \in \mathcal{H}_{\mathrm{prop}}}} \mathbb{E}_{(x,y)\sim\mathcal{D}_\tau}\big[\ell_J(a_{\ell_J}(x), y) - \ell_J(h(x), y)\big].$$

The expectation includes action randomness. For any two assignments σ, τ, use one loss for the indices where $(\sigma_i, \tau_i) = (+1, -1)$ and another for the indices with the opposite disagreement. For each loss, one fixed benchmark is optimal at every context under $\mathcal{D}_\sigma$, and the other is optimal everywhere under $\mathcal{D}_\tau$. Thus conditional excess losses cannot cancel. Properness at the midpoint x_i, together with the prescribed Brier values, gives

$$\mathrm{Reg}(\sigma) + \mathrm{Reg}(\tau) \geq \frac{\alpha^2}{24M} d_{\mathrm{H}}(\sigma, \tau),$$

as proved in Lemma B.4.

Decode the hidden assignment by choosing $\widehat{\sigma} \in \arg\min_\tau \mathrm{Reg}(\tau)$, using a fixed rule for ties. This decoder uses only the learned rules and the known candidate distributions. Since $\mathrm{Reg}(\widehat{\sigma}) \leq \mathrm{Reg}(\sigma)$,

$$\frac{\alpha^2}{24M} d_{\mathrm{H}}(\sigma, \widehat{\sigma}) \leq 2\,\mathrm{Reg}(\sigma).$$

The assumed expected regret bound therefore yields

$$\frac{1}{M}\mathbb{E}[d_{\mathrm{H}}(\sigma, \widehat{\sigma})] \leq \frac{48\varepsilon}{\alpha^2} = \frac{1}{4}.$$

This holds for every fixed σ, and hence also after averaging over independent fair signs. The two possible label distributions at each context are α apart and have midpoint x_i, so Lemma 4.1 gives

$$n = \Omega_K\left(\frac{M}{\alpha^2}\right) = \Omega_K\left(\alpha^{-(K+1)}\right) = \Omega_K\left(\varepsilon^{-(K+1)/2}\right).$$

Appendix B.3 gives the loss construction and the full proof of the separation inequality.

A Deferred proofs from Section 3

This appendix collects the proofs deferred from Section 3, in the order in which the results appear there. Appendix A.1 proves the smoothing bounds of Section 3.2. Appendix A.2 proves the covering bounds of Section 3.4.2. Appendix A.3 proves the online-learning lemmas of Sections 3.4 and 3.5.1. Appendix A.4 proves the population guarantee of Section 3.6. Appendix A.5 proves the empirical-cover lemma of Section 3.5.2. Appendix A.6 proves the results on GLM losses in Section 3.5.3.

A.1 Proof of the smoothing bounds

We prove Lemmas 3.3, 3.4 and 3.5.

Proof of Lemma 3.3. The range bound follows from $\Delta\ell^{\mathrm{pr}}(p) \in [-2,2]^K$. For $j \in \mathcal{Y}$, let

$$F_j := \{z \in \Delta_K : z_j = 0\},$$

equipped with its uniform probability measure; when $K = 2$, these measures are point masses.

View the simplex in coordinates $p_1, \dots, p_{K-1}$. The Bayes risk is Lipschitz, and its derivative in coordinate p_y agrees with $\Delta_y \ell^{\mathrm{pr}}(p)$ almost everywhere. For $y \geq 1$, integrate this derivative over slices of $(1-r)q + r\Delta_K$ parallel to $e_y - e_0$. The fundamental theorem of calculus gives

$$\big(w_{\ell,r}(q)\big)_y = \frac{K-1}{r}\left(\mathbb{E}_{z\sim\mathrm{Unif}(F_0)}\big[\underline{L}_\ell((1-r)q+rz)\big] - \mathbb{E}_{z\sim\mathrm{Unif}(F_y)}\big[\underline{L}_\ell((1-r)q+rz)\big]\right).$$

In these coordinates, the translated simplex $(1-r)q + r\Delta_K$ has volume $r^{K-1}/(K-1)!$, and its projection along coordinate p_y has volume $r^{K-2}/(K-2)!$. The factor $(K-1)/r$ is the ratio of the second volume to the first.

Replacing $\underline{L}_\ell$ by $\underline{L}_{\ell'}$ changes each facet average by at most $\|\underline{L}_\ell - \underline{L}_{\ell'}\|_\infty$. This proves the approximation bound.

Since losses take values in $[-1,1]$, the Bayes risk is 1-Lipschitz in ℓ_1: compare the expected losses of a fixed optimal action under the two distributions, and then interchange them. For $q, q' \in \Delta_K$, each facet average therefore changes by at most $(1-r)\|q-q'\|_1$. Hence

$$\|w_{\ell,r}(q) - w_{\ell,r}(q')\|_\infty \leq \frac{2(K-1)(1-r)}{r}\|q-q'\|_1 \leq \frac{2(K-1)}{r}\|q-q'\|_1.$$

The coordinate $y = 0$ is identically zero. □

Proof of Lemma 3.4. Fix $\ell \in \mathcal{L}$ and a draw of (x,y) and z, and write $q = q(x)$ and $p = (1-r)q + rz$. Applying (2.3) to the action $\mathrm{BR}_\ell(p)$ and the modeled distribution p gives

$$\ell(\mathrm{BR}_\ell(p), y) - \mathbb{E}_{\widetilde{y}\sim p}\big[\ell(\mathrm{BR}_\ell(p), \widetilde{y})\big] = \langle \Delta\ell^{\mathrm{pr}}(p), e_y - q\rangle + \langle \Delta\ell^{\mathrm{pr}}(p), q - p\rangle.$$

The coordinates of $q - p$ sum to zero, so the second inner product equals $\langle \ell^{\mathrm{pr}}(p,\cdot), q-p\rangle$. The loss vector $\ell^{\mathrm{pr}}(p,\cdot)$ lies in $[-1,1]^K$, so this term has absolute value at most $\|q-p\|_1 = r\|q-z\|_1 \leq 2r$.

The draw z is independent of (x,y), and q depends only on x. Averaging the first inner product over z therefore gives $\langle w_{\ell,r}(q), e_y - q\rangle$, by (3.2). Taking expectations over (x,y) and z, applying the triangle inequality, and taking the supremum over $\ell \in \mathcal{L}$ proves the lemma. □

Proof of Lemma 3.5. Fix $\ell \in \mathcal{L}$, $h \in \mathcal{H}$ and a context x. The modeled expected losses of the fixed action $h(x)$ under the two predictions differ by

$$\mathbb{E}_{\widetilde{y}\sim p(x)}\big[\ell(h(x),\widetilde{y})\big] - \mathbb{E}_{\widetilde{y}\sim q(x)}\big[\ell(h(x),\widetilde{y})\big] = \langle \ell(h(x),\cdot), p(x) - q(x)\rangle.$$

Since $\ell(h(x),\cdot) \in [-1,1]^K$, this has absolute value at most $\|p(x) - q(x)\|_1$. The real-world term $\mathbb{E}[\ell(h(x),y)]$ is the same for both predictors. Taking expectations, applying the triangle inequality, and taking the supremum over ℓ and h proves the first bound. For the prediction in (3.1), $\|p(x) - q(x)\|_1 = r\|q(x) - z\|_1 \le 2r$. □

A.2 Proofs of the covering bounds

We prove the reduction of Lemma 3.10 from the simplex to the cube, and the second bound in Lemma 3.11.

A.2.1 Simplex cover

Proof of Lemma 3.10. Write $\Sigma_d = \{x \in [0,1]^d : \sum_i x_i \le 1\}$. By rescaling, it suffices to consider functions bounded by 1 in absolute value and 1-Lipschitz in ℓ_1. For each such f, define

$$\widetilde{f}(x) := \max_{y\in\Sigma_d}\{f(y) - \|x - y\|_1\}, \qquad x \in [0,1]^d.$$

Lipschitzness gives $\widetilde{f} = f$ on Σ_d. The extension is also 1-Lipschitz and bounded in absolute value by $d+1$. It is concave because the expression being maximized is jointly concave in (x,y) and Σ_d is convex.

Apply the cube covering bound [GS13, Theorem 3.2] to the negatives of these extensions. Negating the cover elements and restricting them to Σ_d gives the required cover. □

A.2.2 The second covering bound

Proof of the second bounds in Lemma 3.11. Fix $0 < r \le 1/2$ and $0 < u \le 1$. For each coordinate, we keep $O(1/u)$ possible slopes and describe where they occur by recording the height of the line with each slope. These heights are concave functions of one fewer variable, and it will suffice to approximate them to accuracy of order ru.

Use the independent coordinates $p_1, \dots, p_{K-1}$, with $p_0 = 1 - \sum_{i=1}^{K-1} p_i$. By Lemma 2.6, every Bayes risk f is concave, bounded by 1 in absolute value, and 2-Lipschitz in the ℓ_1 norm on these coordinates. Its coordinate derivatives lie in $[-2,2]$ almost everywhere and average to the corresponding coordinates of the smoothed decision weight. Coordinate 0 of that weight is identically zero.

The heights are concave and Lipschitz. Fix a coordinate $j \ge 1$, write it as t, and write v for the remaining $K-2$ coordinates. The allowed interval on this slice is $0 \le t \le 1 - \|v\|_1$. For $s \in [-2,2]$, define

$$H_s^f(v) := \max_{0\le t\le 1-\|v\|_1}\{f(t,v) - st\}.$$

The line $t \mapsto st + H_s^f(v)$ lies above the slice and touches it somewhere, possibly at an endpoint. The value $H_s^f(v)$ records where to place the line of slope s.

To see that H_s^f is concave, take maximizing points on two slices. Their convex combination is feasible on the intermediate slice. Concavity of $f(t,v) - st$ then gives the concavity inequality for the maximum heights. Also, $-1 \le H_s^f \le 3$.

For the Lipschitz bound, let t maximize the expression on the slice v and put $t' = \min\{t, 1-\|v'\|_1\}$. This is feasible on the slice v', and $|t-t'| \le \|v-v'\|_1$. Therefore

$$\begin{aligned} H_s^f(v) - H_s^f(v') &\le f(t,v) - st - f(t',v') + st' \\ &\le 2\|v-v'\|_1 + (2+|s|)|t-t'| \\ &\le 6\|v-v'\|_1. \end{aligned}$$

Interchanging v and v' proves the claimed uniform Lipschitz bound. Thus the height functions satisfy the hypotheses of Lemma 3.10 in dimension $K-2$ when $K \ge 3$.

Keeping finitely many slopes changes derivatives by at most $u/4$. Choose an equally spaced grid $S \subset [-2,2]$ containing both endpoints, with spacing at most $u/4$ and $|S| = O(1/u)$. For the fixed coordinate, form the lowest of the lines with these slopes,

$$f_S(t,v) := \min_{s \in S}\{st + H_s^f(v)\}.$$

It is a minimum of finitely many lines in t with slopes in $[-2,2]$, whose heights are 6-Lipschitz in v by the bound just proved, so it is Lipschitz. We claim that

$$|\partial_t f_S(t,v) - \partial_t f(t,v)| \le u/4 \qquad \text{almost everywhere.} \tag{A.1}$$

At an interior differentiability point of the slice, observe that

$$st + H_s^f(v) = \max_{\tilde{t}}\{f(\tilde{t},v) + s(t-\tilde{t})\},$$

where $\tilde{t}$ ranges over the same slice. As a function of s, this is convex, being a maximum of affine functions. Its value is at least $f(t,v)$, with equality exactly when $s = \partial_t f(t,v)$, the unique slope of a line through $(t, f(t,v))$ lying above the slice. Its minimum over S must therefore occur at one of the two grid slopes bracketing $\partial_t f(t,v)$, or at that slope itself if it is in S. Wherever f_S is also differentiable in t, its derivative is a minimizing grid slope. This proves (A.1).

Approximating the heights gives a cover of the averaged weights. Set

$$\kappa := \frac{ru}{8(K-1)}.$$

Fix a uniform κ-cover of the height functions, and record a cover element for each coordinate and each slope in S. Group losses according to all these choices. If losses ℓ, ℓ' with Bayes risks f, g are in the same group, then, for every coordinate and slope,

$$\|H_s^f - H_s^g\|_\infty \le 2\kappa.$$

For the chosen coordinate, taking the minimum of the corresponding lines gives

$$\|f_S - g_S\|_\infty \le 2\kappa.$$

For any base prediction q, let $p = (1-r)q + rz$, with $z \sim \mathrm{Unif}(\Delta_K)$. Integrating along coordinate slices of this translated simplex gives

$$\left|\mathbb{E}_z\left[\partial_t f_S(p) - \partial_t g_S(p)\right]\right| \le \frac{2(K-1)}{r}\|f_S - g_S\|_\infty. \tag{A.2}$$

Indeed, the integral of $\partial_t(f_S - g_S)$ on each slice is the difference of its endpoint values, whose absolute value is at most $2\|f_S - g_S\|_\infty$. In the independent coordinates, the simplex has volume $r^{K-1}/(K-1)!$ and its projection onto the other coordinates has volume $r^{K-2}/(K-2)!$. Their ratio gives $(K-1)/r$. For $K = 2$, this is just the endpoint calculation on an interval of length r. This argument requires only Lipschitz continuity, not that f_S and g_S be Bayes risks.

The reporting distribution is absolutely continuous for every q, so we can also average the almost-everywhere derivative bound (A.1). Combining the two bounds gives

$$\begin{aligned}|(w_{\ell,r}(q))_j - (w_{\ell',r}(q))_j| &\le \frac{u}{4} + \frac{2(K-1)}{r}\|f_S - g_S\|_\infty + \frac{u}{4} \\ &\le \frac{u}{2} + \frac{4(K-1)\kappa}{r} = u.\end{aligned}$$

This holds for every coordinate and every q. Choosing one original loss from each nonempty group therefore gives a uniform u-cover by genuine smoothed decision weights.

Counting the height approximations. For $K \ge 3$, Lemma 3.10 supplies a height cover of size N satisfying

$$\log(2N) = O_K\left(\kappa^{-(K-2)/2}\right).$$

There are $(K-1)|S| = O_K(1/u)$ height functions to record for each loss. The number of groups is at most $N^{(K-1)|S|}$, and hence

$$\log(2M(u)) = O_K\left(\frac{1}{u}\,\kappa^{-(K-2)/2}\right) = O_K\left(\frac{1}{u}(ru)^{-(K-2)/2}\right).$$

For $K = 2$, there are no remaining coordinates, so each height is a number in $[-1, 3]$. A uniform κ-cover of this interval has $O(1/\kappa)$ elements. The same argument therefore gives

$$\log(2M(u)) = O\left(\frac{1}{u}\log\frac{2}{ru}\right).$$

□

A.3 Online learning tools

We prove Lemma 3.12 on signed exponential weights, and state and prove the high-probability guarantees for the decision learner and the Lipschitz hypothesis learner used in the proof of Theorem 3.1.

Proof of Lemma 3.12. Apply the exponential-weights potential bound to the $2M$ signed payoffs $s\rho_t(b) \in [-2a, 2a]$. The regret is at most

$$\frac{\log(2M)}{\eta} + 2\eta n a^2 = 2\sqrt{2}\,a\sqrt{n\log(2M)}.$$

Maximizing over the two signs gives the absolute value on the left. Each output b_t is a convex combination of the signed tests $s\,b$, so $\|b_t\|_\infty \le a$.

For the confidence bound, the differences $\rho_t(b) - \bar\rho_t(b)$ have conditional mean zero and absolute value at most $4a$. Since the weights are fixed under the same conditioning and sum to one, the same is true of

$$\sum_{b\in\mathcal{B}} \sum_{s\in\{-1,1\}} \lambda_t(b,s)s\big(\rho_t(b) - \bar\rho_t(b)\big).$$

Thus Azuma–Hoeffding gives, for every $u > 0$,

$$\Pr\left(\left|\sum_{t=1}^{n}\big(\rho_t(b) - \bar\rho_t(b)\big)\right| > u\right) \le 2\exp\left(-\frac{u^2}{32na^2}\right)$$

for each fixed b, and

$$\Pr\left(\sum_{t=1}^{n}\sum_{b\in\mathcal{B}}\sum_{s\in\{-1,1\}} \lambda_t(b,s)s\big(\rho_t(b) - \bar\rho_t(b)\big) > u\right) \le \exp\left(-\frac{u^2}{32na^2}\right).$$

Set $u = 4a\sqrt{2n\log((2M+1)/\delta)}$. A union bound over the M fixed tests and the learner has failure probability at most

$$(2M+1)\exp\left(-\frac{u^2}{32na^2}\right) = \delta.$$

On the complementary event, the deterministic regret bound gives

$$\begin{aligned}
\max_{b\in\mathcal{B}}\left|\sum_{t=1}^{n}\bar\rho_t(b)\right| &\le \max_{b\in\mathcal{B}}\left|\sum_{t=1}^{n}\rho_t(b)\right| + u \\
&\le \sum_{t=1}^{n}\sum_{b\in\mathcal{B}}\sum_{s\in\{-1,1\}} \lambda_t(b,s)s\rho_t(b) + 2\sqrt{2}\,a\sqrt{n\log(2M)} + u \\
&\le \sum_{t=1}^{n}\sum_{b\in\mathcal{B}}\sum_{s\in\{-1,1\}} \lambda_t(b,s)s\bar\rho_t(b) + 2\sqrt{2}\,a\sqrt{n\log(2M)} + 2u.
\end{aligned}$$

Since $\log((2M+1)/\delta) \le 2\log(2M/\delta)$, the last two terms are at most $Ca\sqrt{n\log(2M/\delta)}$ for an absolute constant C. □

Lemma A.1 (Confidence bounds for the decision and Lipschitz learners)**.** *Fix $K \ge 2$, $n \ge 1$, $0 < r \le 1/2$, and $0 < \delta < 1$. Let $(x_t, y_t)_{t=1}^n$ be independent draws from $\mathcal{D}$. On each round, choose the whole rule $q_t : \mathcal{X} \to \Delta_K$ using preceding examples, the current tests, and randomness independent of the examples, before observing (x_t, y_t). Each conclusion below holds with probability at least $1-\delta$. Each expectation over $\mathcal{D}$ holds the current prediction and test functions fixed.*

1. *If w_t is produced by Algorithm 3 with horizon n, width r, and updates $(q_t(x_t), y_t)$, then*

$$\begin{aligned}
&\sup_{\ell\in\mathcal{L}_{\mathrm{All}}}\left|\sum_{t=1}^{n}\mathbb{E}_{(x,y)\sim\mathcal{D}}[\langle w_{\ell,r}(q_t(x)), e_y - q_t(x)\rangle]\right| \\
&\quad\le \sum_{t=1}^{n}\mathbb{E}_{(x,y)\sim\mathcal{D}}[\langle w_t(q_t(x)), e_y - q_t(x)\rangle] + C\,\mathrm{Reg}_{\mathrm{dec}}(n,r) + C\sqrt{n\log(2/\delta)}.
\end{aligned} \tag{A.3}$$

2. *Let $\mathcal{H}$ be a fixed finite nonempty class of maps $\mathcal{X} \to \Delta_K$. If g_t is produced by Algorithm 4 with horizon n, accuracy r, and updates $(x_t, q_t(x_t), y_t)$, then*

$$\begin{aligned}&\sup_{g \in \Delta\mathcal{L}_{\mathrm{Lip}} \circ \mathcal{H}} \left| \sum_{t=1}^{n} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\langle g(x), e_y - q_t(x)\rangle] \right| \\ &\quad \le \sum_{t=1}^{n} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\langle g_t(x), e_y - q_t(x)\rangle] + C \operatorname{Reg}_{\Delta\mathcal{L}_{\mathrm{Lip}} \circ \mathcal{H}}(n) + C\sqrt{n \log(2/\delta)}.\end{aligned} \tag{A.4}$$

Here C is an absolute constant, $\mathrm{Reg}_{\mathrm{dec}}$ is defined in (3.4)*, and the hypothesis regret is the bound in* (3.5)*.*

Proof. Condition on the preceding examples and the randomness used to choose the current tests and q_t. This fixes the EW weights and the current functions, while (x_t, y_t) remains an independent draw from $\mathcal{D}$. For a decision test b, the conditional mean in Lemma 3.12 is therefore

$$\bar{\rho}_t(b) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\langle b(q_t(x)), e_y - q_t(x)\rangle\big].$$

For a hypothesis test b, replace $b(q_t(x))$ by $b(x)$.

For either learner, apply the confidence bound at level j with failure probability $\delta/2^{j+1}$, whose sum over all levels is at most δ. Repeat the telescoping argument in the learner's regret proof. The uniform approximation still costs $2nu_J$, and the terms involving the cover sizes are the same up to absolute constants. The additional term is

$$C\sqrt{n} \sum_{j=0}^{J} 2^{-j} \sqrt{\log(1/\delta) + (j+1)\log 2}.$$

By Cauchy–Schwarz,

$$\begin{aligned}&\left(\sum_{j=0}^{J} 2^{-j} \sqrt{\log(1/\delta) + (j+1)\log 2} \right)^2 \\ &\quad \le \left(\sum_{j=0}^{J} 2^{-j} \right) \left(\sum_{j=0}^{J} 2^{-j} \big[\log(1/\delta) + (j+1)\log 2\big] \right) \\ &\quad \le 2\big[2\log(1/\delta) + 4\log 2\big] = 4\log(4/\delta).\end{aligned}$$

Since $\log(4/\delta) \le 2\log(2/\delta)$, this adds at most $C\sqrt{n\log(2/\delta)}$ to the corresponding regret bound. This proves the asserted inequality for each learner. □

A.4 Proof of the population guarantee

We prove Theorem 3.8, which converts the learners' online regret bounds into a bound on the population OI errors of the output predictor (Section 3.6).

Proof of Theorem 3.8. The online regret guarantees already bound the average correlations on the training examples. We now need to compare these with the population correlations, simultaneously for all tests. We will show that the difference is also bounded by the learners' regrets divided by n, up to constant factors.

Comparing training and population correlations. With the training run fixed, each fixed test's population correlation is the expected correlation on a fresh example. We therefore introduce independent examples kept out of training and compare their correlations with the training correlations. We generate the training and unused examples together in pairs, randomly choosing which example trains the algorithm. This random choice will let us express the difference between unused and training correlations using a fair sign, as in sequential symmetrization [RST15].

For each round t, we draw two examples $(x_t, y_t), (x'_t, y'_t) \sim \mathcal{D}$ and a fair sign $\sigma_t \in \{-1, 1\}$, with all examples and signs drawn independently. The algorithm receives (x_t, y_t) when $\sigma_t = 1$ and (x'_t, y'_t) when $\sigma_t = -1$. The example not given to the algorithm is kept out of training and provides the fresh comparison on that round.

The selected examples are independent draws from $\mathcal{D}$. The algorithm trains on the selected examples, and q_t is its prediction rule after $t-1$ examples. We bound expected population omniregret by analyzing the performance of q_t on the unused examples.

For each fixed loss ℓ, the population average is

$$\frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\langle w_{\ell,r}(q_t(x)), e_y - q_t(x)\rangle],$$

where the saved rules are fixed in the fresh-example expectation. The training and unused averages evaluate this correlation at the selected and unused member of each pair, respectively.

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\ell} |\text{population average} - \text{training average}|\right] \\
&\overset{(1)}{=} \mathbb{E}\left[\sup_{\ell} |\mathbb{E}[\text{unused average} \mid \text{training run}] - \text{training average}|\right] \\
&\overset{(2)}{\leq} \mathbb{E}\left[\sup_{\ell} |\text{unused average} - \text{training average}|\right] \\
&\overset{(3)}{=} \frac{1}{n}\mathbb{E}\left[\sup_{\ell} \left|\sum_{t=1}^{n} \sigma_t\big(\text{primed correlation} - \text{unprimed correlation}\big)\right|\right] \\
&\overset{(4)}{\leq} \frac{1}{n}\mathbb{E}\left[\sup_{\ell} \left|\sum_{t=1}^{n} \sigma_t \langle w_{\ell,r}(q_t(x'_t)), e_{y'_t} - q_t(x'_t)\rangle\right|\right] \\
&\qquad + \frac{1}{n}\mathbb{E}\left[\sup_{\ell} \left|\sum_{t=1}^{n} \sigma_t \langle w_{\ell,r}(q_t(x_t)), e_{y_t} - q_t(x_t)\rangle\right|\right].
\end{aligned}$$

The comparison also applies to hypothesis tests, with $g(x_t)$ in place of $w_{\ell,r}(q_t(x_t))$ and the supremum over $g \in \Delta\mathcal{L} \circ \mathcal{H}$.

For (1), conditional on the training run, including the original independent randomness determining the saved rules, the unused examples are independent draws from $\mathcal{D}$ and every saved rule is fixed. For each fixed loss,

$$\begin{aligned}
&\mathbb{E}[\text{unused average} \mid \text{training run}] \\
&\quad = \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\langle w_{\ell,r}(q_t(x)), e_y - q_t(x)\rangle] = \text{population average}.
\end{aligned}$$

The training average is also fixed under this conditioning.

For (2), we keep the training run fixed and average over unused examples. For each fixed loss, the absolute value of the expected difference is at most the expectation of its absolute value. That absolute value is at most the supremum over all losses. The resulting upper bound no longer depends on the fixed loss, so it also bounds the supremum on the left. Averaging over the training run gives the corresponding bound on the expected population–training difference.

For (3), the rule q_t is determined by preceding training examples, before the current sign. Thus the sign decides which correlation is subtracted from which without changing the current prediction rule. For a fixed loss, unused correlation $-$ training correlation $= \sigma_t\big($primed correlation $-$ unprimed correlation$\big)$.

Both correlations on the right use the same rule q_t, evaluated at their respective contexts. Earlier signs can change later training examples and prediction rules.

For (4), the cumulative difference is the primed sum minus the unprimed sum. Bounding the absolute value of each sum therefore bounds the difference we need to control. Taking a separate supremum for each term gives the upper bound.

It remains to bound

$$\sup_{\ell \in \mathcal{L}_{\mathrm{All}}} \left| \sum_{t=1}^{n} \sigma_t \left\langle w_{\ell,r}(q_t(x_t)), e_{y_t} - q_t(x_t) \right\rangle \right|.$$

and correspondingly

$$\sup_{g \in \Delta\mathcal{L} \circ \mathcal{H}} \left| \sum_{t=1}^{n} \sigma_t \left\langle g(x_t), e_{y_t} - q_t(x_t) \right\rangle \right|.$$

For each fixed test, averaging over the fair signs makes its signed sum have mean zero. But this does not bound the expected supremum of the absolute signed sums: we cannot move the expectation through those two operations. The online learner lets us avoid them. Its regret guarantee compares the largest absolute correlation with the accumulated correlation of its own tests, chosen before each sign. The latter has mean zero. We carry this out directly with signed feedback for the decision learner, and in expectation using synthetic labels for the hypothesis learner.

Using the decision learner. We need to bound

$$\mathbb{E}\left[\sup_{\ell \in \mathcal{L}_{\mathrm{All}}} \left| \sum_{t=1}^{n} \sigma_t \left\langle w_{\ell,r}(q_t(x_t)), e_{y_t} - q_t(x_t) \right\rangle \right| \right] = \mathbb{E}\left[\sup_{\ell \in \mathcal{L}_{\mathrm{All}}} \left| \sum_{t=1}^{n} \left\langle w_{\ell,r}(q_t(x_t)), \sigma_t(e_{y_t} - q_t(x_t)) \right\rangle \right| \right]$$

Run an auxiliary copy of the decision learner with feedback $\sigma_t(e_{y_t} - q_t(x_t))$ in place of the residual. Lemma 3.13 gives

$$\begin{aligned} &\mathbb{E}\left[\sup_{\ell \in \mathcal{L}_{\mathrm{All}}} \left| \sum_{t=1}^{n} \sigma_t \langle w_{\ell,r}(q_t(x_t)), e_{y_t} - q_t(x_t) \rangle \right| \right] \\ &\quad \le \sum_{t=1}^{n} \mathbb{E}[\sigma_t \langle \widetilde{w}_t(q_t(x_t)), e_{y_t} - q_t(x_t) \rangle] + \mathrm{Reg}_{\mathrm{dec}}(n, r) = \mathrm{Reg}_{\mathrm{dec}}(n, r). \end{aligned}$$

For the last equality, condition on the preceding original and auxiliary rounds, the current example pair, and any randomness used to choose q_t and $\widetilde{w}_t$. The current sign is still fair and the inner product is fixed, so each expected summand is zero.

A second auxiliary copy proposes $\widetilde{w}'_t$, evaluates tests at $q_t(x'_t)$, and uses $\sigma_t(e_{y'_t} - q_t(x'_t))$ in its updates. The corresponding regret inequality applies to the primed sum. The same cancellation gives another $\mathrm{Reg}_{\mathrm{dec}}(n, r)$.

Applying the two bounds to the shared comparison gives

$$\mathbb{E}\left[\sup_{\ell\in\mathcal{L}_{\mathrm{All}}}|\text{population average} - \text{training average}|\right] \le \frac{2\,\mathrm{Reg}_{\mathrm{dec}}(n,r)}{n}.$$

Using the hypothesis learner in the same comparison. A general hypothesis learner need not accept signed residuals. Instead, run an auxiliary copy on the contexts x_t with synthetic predictions and labels

$$\widetilde{p}_t = \frac{e_{y_t} + q_t(x_t)}{2}, \qquad \widetilde{y}_t = \begin{cases} y_t, & \sigma_t = 1, \\ \text{a fresh draw from } q_t(x_t), & \sigma_t = -1. \end{cases}$$

Its proposed tests $\widetilde{g}_t$ depend only on preceding auxiliary rounds and its own independent randomness.

Fix the original example pairs, signs, and original learner's randomness, and average only over the fresh synthetic draws. These draws do not affect the original prediction rules, and

$$\mathbb{E}_{\widetilde{y}}\big[e_{\widetilde{y}_t} - \widetilde{p}_t\big] = \frac{\sigma_t}{2}\big(e_{y_t} - q_t(x_t)\big).$$

Moving the synthetic-label expectation outside the absolute value and the supremum can only increase the expression. Then applying the auxiliary learner's expected-regret guarantee gives

$$\begin{aligned} &\mathbb{E}\left[\sup_{g\in\Delta\mathcal{L}\circ\mathcal{H}}\left|\sum_{t=1}^{n}\sigma_t\langle g(x_t), e_{y_t} - q_t(x_t)\rangle\right|\right] \\ &\quad\le 2\mathbb{E}\left[\sup_{g\in\Delta\mathcal{L}\circ\mathcal{H}}\left|\sum_{t=1}^{n}\langle g(x_t), e_{\widetilde{y}_t} - \widetilde{p}_t\rangle\right|\right] \\ &\quad\le 2\sum_{t=1}^{n}\mathbb{E}\big[\langle \widetilde{g}_t(x_t), e_{\widetilde{y}_t} - \widetilde{p}_t\rangle\big] + 2\,\mathrm{Reg}_{\Delta\mathcal{L}\circ\mathcal{H}}(n) \\ &\quad= 2\,\mathrm{Reg}_{\Delta\mathcal{L}\circ\mathcal{H}}(n). \end{aligned} \tag{A.5}$$

For the last equality, condition instead on the preceding original and auxiliary rounds, the current example pair, and the randomness used to choose q_t and $\widetilde{g}_t$. The current sign remains fair. Averaging it and the fresh synthetic label gives $\mathbb{E}[e_{\widetilde{y}_t}] = (e_{y_t} + q_t(x_t))/2 = \widetilde{p}_t$ under this conditioning, so the auxiliary learner's expected payoff is zero.

The primed auxiliary copy uses the corresponding primed context, prediction, and label and gives the same bound. Averaging over the example pairs and original training randomness, adding the two bounds, and dividing by n gives

$$\mathbb{E}\left[\sup_{g\in\Delta\mathcal{L}\circ\mathcal{H}}|\text{population average} - \text{training average}|\right] \le \frac{4\,\mathrm{Reg}_{\Delta\mathcal{L}\circ\mathcal{H}}(n)}{n}.$$

Here the averages are the hypothesis correlations for the same fixed g.

Finishing the guarantee. For each test, its absolute population average is at most its absolute training average plus the absolute difference between the two. Lemma 3.7 bounds the expected sum of the two uniform training-correlation suprema by

$$\frac{\mathrm{Reg}_{\mathrm{dec}}(n,r) + \mathrm{Reg}_{\Delta\mathcal{L}\circ\mathcal{H}}(n)}{n}.$$

The decision smoothing lemma relates the averaged decision weights paired with the base residual to decision OI of this randomized predictor, with an additional error at most $2r$. The hypothesis smoothing lemma adds at most $2r$ to the hypothesis OI bound. The OI reduction bounds expected omniprediction error by the sum of the two OI errors. Adding the training bound, the two population–training bounds, and the two smoothing errors gives

$$\begin{aligned}
&\mathbb{E}\big[\mathrm{DecOI}_{\mathcal{L}_{\mathrm{All}}}(p) + \mathrm{HypOI}_{\mathcal{L},\mathcal{H}}(p)\big] \\
&\quad\le \frac{\mathrm{Reg}_{\mathrm{dec}}(n,r) + \mathrm{Reg}_{\Delta\mathcal{L}\circ\mathcal{H}}(n)}{n} + \frac{2\,\mathrm{Reg}_{\mathrm{dec}}(n,r)}{n} + \frac{4\,\mathrm{Reg}_{\Delta\mathcal{L}\circ\mathcal{H}}(n)}{n} + 2r + 2r \\
&\quad= \frac{3\,\mathrm{Reg}_{\mathrm{dec}}(n,r) + 5\,\mathrm{Reg}_{\Delta\mathcal{L}\circ\mathcal{H}}(n)}{n} + 4r \\
&\quad\le C_K \frac{\mathrm{Reg}_{\mathrm{dec}}(n,r) + \mathrm{Reg}_{\Delta\mathcal{L}\circ\mathcal{H}}(n)}{n} + 4r.
\end{aligned}$$

□

A.5 Proof of the empirical-cover lemma

Both parts combine two standard results. For $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$, a probability measure ν, and $t > 0$, let $N(\mathcal{F}, t, L_p(\nu))$ be the $L_p(\nu)$ covering number of $\mathcal{F}$ at scale t. Requiring the centers to lie in $\mathcal{F}$ at most halves the scale, and $L_2(\nu)$-covers are $L_1(\nu)$-covers.

(E) Mendelson and Vershynin [MV03, Theorem 1]: there are absolute constants $c_0, C_0 > 0$ such that $\log N(\mathcal{F}, t, L_2(\nu)) \le C_0 \,\mathrm{fat}_{c_0 t}(\mathcal{F}) \log(2/t)$ for every ν and $0 < t < 1$. (Their shattering dimension uses a gap of t rather than a margin, which only changes c_0.)

(U) Anthony and Bartlett [AB99, Theorem 17.7], applied with dummy labels and the identity loss $L(t,0) = t$, whose bound and Lipschitz constant are both 1: for permissible $\mathcal{G} \subseteq [0,1]^{\mathcal{X}}$ and m i.i.d. draws from μ, $\Pr[\sup_{g\in\mathcal{G}} |\mathbb{E}_\mu g - \mathbb{E}_S g| \ge \alpha] \le 4 \sup_\nu N(\mathcal{G}, \alpha/8, L_1(\nu))\, e^{-\alpha^2 m/32}$, where ν ranges over uniform distributions on $2m$ points.

The fat-shattering dimension is nonincreasing in the scale. Write $\gamma := \varepsilon/16$ and $\alpha := \gamma/K$.

Proof of Lemma 3.17. We take $c := c_0/512$; the constant C absorbs the numerical factors below, using $2K/\varepsilon \ge 4$ to compare logarithms.

Part 1. Fix j and let $\mathcal{G}_j := \{|f - f'| : f, f' \in \mathcal{H}_j\} \subseteq [0,1]^{\mathcal{X}}$. Since $\big||f - f'| - |b - b'|\big| \le |f - b| + |f' - b'|$, pairs from an $\alpha/16$-cover of $\mathcal{H}_j$ with centers in $\mathcal{H}_j$ form an $\alpha/8$-cover of $\mathcal{G}_j$. By (E), $\log N(\mathcal{G}_j, \alpha/8, L_1(\nu)) \le 2C_0 \,\mathrm{fat}_{c_0\alpha/32}(\mathcal{H}) \log(64/\alpha)$ for every ν, and $c_0\alpha/32 = c\varepsilon/K$. By (U) with $\mu = \mathcal{D}_{\mathcal{X}}$, the event $\sup_{g\in\mathcal{G}_j} |\mathbb{E}_{\mathcal{D}_{\mathcal{X}}} g - \mathbb{E}_S g| \ge \alpha$ has probability at most δ/K under the condition on m in the lemma. On the complement of these K events, summing over coordinates gives $|d_{\mathcal{D}}(h,h') - d_S(h,h')| < K\alpha = \varepsilon/16$ for all $h, h' \in \mathcal{H}$.

Part 2. For each j, fix a $\gamma/(2K)$-cover B_j of $\mathcal{H}_j$ in L_1 of the uniform distribution on S, of minimal size. Map each h to a tuple $(b_j)_j \in \prod_j B_j$ with b_j within $\gamma/(2K)$ of h_j. Two hypotheses with the same tuple are within γ in d_S, so the map is injective on the γ-separated set $\widehat{\mathcal{H}}$, and by (E) with $t = \gamma/(2K)$, $\log(2|\widehat{\mathcal{H}}|) \le 1 + KC_0 \,\mathrm{fat}_{c_0\varepsilon/(32K)}(\mathcal{H}) \log(64K/\varepsilon) \le 1 + CK \,\mathrm{fat}_{c\varepsilon/K}(\mathcal{H}) \log(2K/\varepsilon)$. The same bound applies to every γ-separated subset of $\mathcal{H}$, so a greedy construction stops and a maximal one exists. □

A.6 Applying the GLM hypothesis learners

We prove the statements in Section 3.5.3 using the learners of Hu, Tian, and Yang [HTY26, Lemmas 41 and 42].[6]

A.6.1 Normalization and the hypothesis tests

For every $a \in \mathbb{R}^K$, the gradient condition gives

$$\omega(a) - \omega(0) = \int_0^1 \langle \nabla\omega(ta), a\rangle \, dt, \qquad |\langle \nabla\omega(ta), a\rangle| \le \|a\|_\infty.$$

Consequently, $|\omega(a) - \omega(0)| \le \|a\|_\infty$, and for $a \in [-1,1]^K$,

$$|\ell(a,y)| \le \frac{|\omega(a) - \omega(0)| + |a_y|}{2} \le 1.$$

Subtracting $\omega(0)$ and dividing by two preserve the minimizers of expected loss over this cube. The loss is continuous and the cube is compact, so these minimizers exist. For $\omega(a) = \log \sum_{j=0}^{K-1} e^{a_j}$, the gradient has coordinates $e^{a_i}/\sum_j e^{a_j}$, which form a probability vector. The loss is therefore

$$\ell(a,y) = \frac{\log \sum_{j=0}^{K-1} e^{a_j} - a_y - \log K}{2},$$

the stated normalization of multiclass logistic loss on the cube.

Subtracting the loss at label zero removes the potential,

$$\Delta_i \ell(Wx) = \ell(Wx, i) - \ell(Wx, 0) = \frac{(Wx)_0 - (Wx)_i}{2}.$$

Since the coordinates of $e_y - q$ sum to zero,

$$\begin{aligned}\langle \Delta\ell(Wx), e_y - q\rangle &= \frac{(Wx)_0}{2} \sum_{i=0}^{K-1} (e_y - q)_i - \frac{1}{2}\langle Wx, e_y - q\rangle \\ &= \frac{1}{2}\langle Wx, q - e_y\rangle.\end{aligned}$$

This proves (3.7), including the sign of the input to the matrix learner and the factor $1/2$.

A.6.2 Matching the borrowed learners

The feature normalization in Hu, Tian, and Yang [HTY26, equation (36)] is $\|x\|_2 \le 1$, as here. Their matrix class in equation (49) also agrees with ours, since

$$\|W\|_{2\to\infty} = \max_i \|W_i\|_2.$$

Their label y is a one-hot vector, whereas ours is a label index. The remaining notation matches as follows.

[6] All references to their numbering are to the final COLT version.

Hu–Tian–Yang	This paper
$k,\ T$	$K,\ n$
$C,\ C_t$	$W,\ W_t$
y	e_y
v_t	$q_t - e_{y_t}$ in the online setting
$v_t(x, y)$	$q_t(x) - e_y$ in the population setting
$v_t \otimes x_t$	$(q_t - e_{y_t})x_t^\top$ in the online setting

Their equation (6) removes the common potential from the loss vector, leaving $-Wx$. Our discrete derivative subtracts the loss at label zero and includes the normalization by two. Equation (3.7) relates these conventions. We take all best responses over $[-1, 1]^K$ throughout.

For either learner, W_t depends only on preceding updates and has rows of Euclidean norm at most one. Thus $|(W_t x)_i| \le 1$ for every $x \in \mathcal{X}$, and the tests in (3.8) have coordinates in $[-1, 1]$. For any prediction $q \in \Delta_K$ and label y,

$$\|q - e_y\|_2^2 = \|q\|_2^2 + 1 - 2q_y \le 2.$$

Hence the vector supplied to the matrix learner has Euclidean norm at most $\sqrt{2}$, within its required bound of 2. Its update matrix satisfies

$$\|(q - e_y)x^\top\|_{\mathrm{F}} = \|q - e_y\|_2 \|x\|_2 \le \sqrt{2},$$

where this matrix norm is the square root of the sum of its squared entries.

A.6.3 Proof of Lemma 3.19

The online guarantee. With the inputs x_t and $q_t - e_{y_t}$, Hu, Tian, and Yang [HTY26, Lemma 41] gives

$$\sup_{W:\, \|W_i\|_2 \le 1 \text{ for every } i} \sum_{t=1}^{n} \langle q_t - e_{y_t}, (W - W_t)x_t \rangle \le 2\sqrt{Kn}. \tag{A.6}$$

This bound holds for every sequence of inputs. In particular, the current prediction may depend on the current matrix W_t.

Both W and $-W$ are allowed and give opposite cumulative correlations. Therefore, by (3.7),

$$\begin{aligned}
&\sup_{g \in \Delta\mathcal{L}_{\mathrm{GLM}} \circ \mathcal{H}_{\mathrm{lin}}} \left| \sum_{t=1}^{n} \langle g(x_t), e_{y_t} - q_t \rangle \right| \\
&\quad = \frac{1}{2} \sup_{W:\, \|W_i\|_2 \le 1 \text{ for every } i} \left| \sum_{t=1}^{n} \langle W x_t, q_t - e_{y_t} \rangle \right| \\
&\quad = \frac{1}{2} \sup_{W:\, \|W_i\|_2 \le 1 \text{ for every } i} \sum_{t=1}^{n} \langle W x_t, q_t - e_{y_t} \rangle.
\end{aligned}$$

The selected tests satisfy

$$\sum_{t=1}^{n} \langle g_t(x_t), e_{y_t} - q_t \rangle = \frac{1}{2} \sum_{t=1}^{n} \langle W_t x_t, q_t - e_{y_t} \rangle.$$

Subtracting this expression from the preceding supremum and using (A.6) proves (3.9). Taking expectations also gives the regret guarantee in Definition 3.6.

The population guarantee. Use the statistical learner of Hu, Tian, and Yang [HTY26, Lemma 42] with failure probability δ. Condition on the preceding examples and all randomness used to choose the current tests and prediction rule. This fixes the current matrix and the whole rule q_t, while (x_t, y_t) remains an independent draw from $\mathcal{D}$. In the omniprediction algorithm, this conditioning includes the history of both learners and the current decision test w_t. The update has conditional mean

$$\begin{aligned}&\mathbb{E}_{(x_t,y_t)\sim\mathcal{D}}\Big[(q_t(x_t)-e_{y_t})x_t^\top \;\Big|\; \text{preceding rounds and current choices}\Big]\\&\quad=\mathbb{E}_{(x,y)\sim\mathcal{D}}[(q_t(x)-e_y)x^\top].\end{aligned}$$

Together with the norm bound above, this verifies the conditional-mean assumption of their Lemma 15, used to prove Lemma 42. The latter lemma gives, with probability at least $1-\delta$,

$$\sup_{W:\,\|W_i\|_2\le 1 \text{ for every } i}\sum_{t=1}^{n}\mathbb{E}_{(x,y)\sim\mathcal{D}}[\langle q_t(x)-e_y,(W-W_t)x\rangle]\le C\sqrt{Kn\log(2/\delta)}.$$

Applying (3.7) inside each expectation divides both cumulative correlations by two. The matrices W and $-W$ again give opposite population correlations, so the same calculation as in the online case proves (3.10).

A.6.4 Combining the learners and proving the corollary

For the expected guarantee, use the online learner in Lemma 3.19. Theorem 3.8 gives

$$\mathbb{E}[\mathrm{OmniReg}_{\mathcal{L}_{\mathrm{GLM}},\mathcal{H}_{\mathrm{lin}}}(p)]\le C_K\frac{\mathrm{Reg}_{\mathrm{dec}}(n,r)+\sqrt{Kn}}{n}+4r.$$

The hypothesis learner has regret $O(\sqrt{Kn})$, so the hypothesis requirement in Theorem 3.2 is satisfied once $n\ge C_K\varepsilon^{-2}$. Since $\max\{K-1,(K+2)/2\}\ge 2$, this requirement is dominated by the expected sample bound in Corollary 3.20.

For confidence, run the prediction procedure of Algorithm 1 with the statistical learner. Apply (3.10) and the decision bound (A.3), each with failure probability $\delta/2$. Both hold with probability at least $1-\delta$. Their chosen-test terms have nonpositive sum, because the prediction rule gives, for every x, y,

$$\langle w_t(q_t(x)),e_y-q_t(x)\rangle+\langle g_t(x),e_y-q_t(x)\rangle\le 0.$$

Adding the two population bounds and dividing by n therefore gives

$$\begin{aligned}&\sup_{\ell\in\mathcal{L}_{\mathrm{All}}}\left|\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{(x,y)\sim\mathcal{D}}[\langle w_{\ell,r}(q_t(x)),e_y-q_t(x)\rangle]\right|\\&\quad+\sup_{g\in\Delta\mathcal{L}_{\mathrm{GLM}}\circ\mathcal{H}_{\mathrm{lin}}}\left|\frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_{(x,y)\sim\mathcal{D}}[\langle g(x),e_y-q_t(x)\rangle]\right|\\&\quad\le C\frac{\mathrm{Reg}_{\mathrm{dec}}(n,r)}{n}+C\sqrt{\frac{K\log(4/\delta)}{n}}.\end{aligned}$$

The two suprema are the averaged decision and hypothesis correlations when the base predictor chooses uniformly from $q_1,\dots,q_n$. Lemmas 3.4 and 3.5 each add at most $2r$ for the final randomized prediction. Theorem 2.2 then proves (3.11).

To obtain the confidence sample bound, take r to be a sufficiently small K-dependent multiple of ε. Using (3.4), the bound

$$n \geq C_K\left(\varepsilon^{-\max\{K-1,(K+2)/2\}}\log^2(2/\varepsilon) + \frac{\log(2/\delta)}{\varepsilon^2}\right)$$

makes $C\,\mathrm{Reg}_{\mathrm{dec}}(n,r)/n + 4r$ at most $\varepsilon/2$ and the remaining square-root term in (3.11) at most $\varepsilon/2$. This proves the confidence claim in Corollary 3.20.

B Deferred proofs from Section 4

This appendix proves Lemma 4.1, which converts recovery of the hidden signs into a sample lower bound for every construction in Section 4, and proves the properties of the losses used for the $(1/\varepsilon)^{(K+2)/2}$ lower bound (Lemma 4.4).

B.1 Proof of the sign decoding lemma

We prove the bounds on samples needed to recover the hidden signs. The noisy case uses the following form of the binary rate–distortion bound [CT06]. We include the elementary proof for completeness.

Lemma B.1. *Let σ be uniform on $\{0,1\}^M$, and let $\widehat{\sigma} \in \{0,1\}^M$ satisfy*

$$\mathbb{E}[d_{\mathrm{H}}(\widehat{\sigma},\sigma)] \leq \delta M, \qquad 0 \leq \delta < \tfrac{1}{2}.$$

Then

$$I(\sigma;\widehat{\sigma}) \geq M\big(\log 2 - h_2(\delta)\big),$$

where d_{H} counts sign disagreements, and $h_2(t) = -t\log t - (1-t)\log(1-t)$, with natural logarithms and $0\log 0 = 0$. If a decoder produces $\widehat{\sigma}$ from the training sample s using fresh randomness independent of the joint pair (σ, s), the same lower bound holds for $I(\sigma; s)$.

Proof. For each sign, the decoded sign and its mistake indicator determine the true sign. Therefore

$$\begin{aligned} H(\sigma \mid \widehat{\sigma}) &\leq \sum_{i=1}^{M} h_2(\Pr[\widehat{\sigma}_i \neq \sigma_i]) \\ &\leq M h_2\left(\frac{\mathbb{E}[d_{\mathrm{H}}(\widehat{\sigma},\sigma)]}{M}\right) \leq M h_2(\delta). \end{aligned}$$

The second inequality is Jensen's inequality for the concave function h_2. The third holds because the average error probability $\mathbb{E}[d_{\mathrm{H}}(\widehat{\sigma},\sigma)]/M$ is at most $\delta < 1/2$ and h_2 is increasing on $[0,1/2]$. Subtracting from $H(\sigma) = M\log 2$ proves the first claim. The sample claim follows from

$$I(\sigma;\widehat{\sigma}) \leq I(\sigma; s).$$

□

Proof of Lemma 4.1. The sample contains at most n distinct contexts. Conditional on the sample, each hidden choice at an unobserved context is still a fair bit. A decoder using independent randomness therefore makes an error with probability $1/2$ at each such context, so

$$\mathbb{E}[d_{\mathrm{H}}(\widehat{\sigma},\sigma)] \geq \frac{M-n}{2}.$$

The assumed error bound δM implies $n \geq (1-2\delta)M$.

For the noisy bound, write p_i^0 and p_i^1 for the two possible conditional label distributions at x_i, and let $q_i = (p_i^0 + p_i^1)/2$. Their distance is at most η, so $\|p_i^b - q_i\|_2 \leq \eta/2$ for $b \in \{0,1\}$. The inequality $\log t \leq t-1$ gives

$$\begin{aligned} D_{\mathrm{KL}}(p_i^b \| q_i) &\leq \sum_{y=0}^{K-1} \frac{(p_{i,y}^b - q_{i,y})^2}{q_{i,y}} \\ &\leq 2K\|p_i^b - q_i\|_2^2 \leq \frac{K\eta^2}{2}. \end{aligned}$$

The context contains no information about σ, so

$$\begin{aligned} I(\sigma; x, y) &= \sum_{i=1}^{M} \Pr[x = x_i] \, \frac{D_{\mathrm{KL}}(p_i^0 \| q_i) + D_{\mathrm{KL}}(p_i^1 \| q_i)}{2} \\ &\leq \frac{K\eta^2}{2}. \end{aligned}$$

The examples in s are independent conditional on σ, so

$$\begin{aligned} I(\sigma; s) &= H(s) - H(s \mid \sigma) \\ &\leq nH(x,y) - nH(x,y \mid \sigma) = nI(\sigma; x, y) \\ &\leq \frac{nK\eta^2}{2}. \end{aligned}$$

The decoder uses only the sample and independent randomness. Lemma B.1 therefore gives

$$M\big(\log 2 - h_2(\delta)\big) \leq I(\sigma; s) \leq \frac{nK\eta^2}{2}. \tag{B.1}$$

Rearranging proves $n = \Omega_{K,\delta}(M/\eta^2)$. □

Proof of Theorem 4.3 $1/\varepsilon^{K-1}$ **lower bound** The following lemma exhibits the required family in $\mathcal{L}_{\mathrm{Lip}}$. We state it in terms of a parameter β, which we later set to 16ε.

Lemma B.2 (Hard family of Lipschitz losses)**.** *Fix $0 < \beta \leq 2/3$, and let $x_1, \ldots, x_M, z_-, z_+ \in \Delta_K$ be pairwise 4β-separated in Euclidean distance. Then there exists a family $\{\ell_\sigma\}_{\sigma \in \{0,1\}^M} \subseteq \mathcal{L}_{\mathrm{Lip}}$ with the following properties.*

(1) For every σ and $i \in [M]$, the action x_i is a global minimizer of $\ell_\sigma(\cdot, \sigma_i)$, and

$$\ell_\sigma(x_i, \sigma_i) = \frac{1}{2} - \frac{\beta}{4}.$$

(2) For every σ, $p \in \Delta_K$, and $y \in \mathcal{Y}$,

$$\ell_\sigma^{\mathrm{pr}}(p, y) = \frac{1}{2} - \frac{\beta}{4} + \beta\big(p_0 - \mathbf{1}_{\{y=0\}}\big)^2.$$

In particular, the properization is independent of σ.

Proof. Set

$$\tau(y) := \begin{cases} -1, & y = 0, \\ 1, & y \neq 0. \end{cases}$$

For each σ, we will construct a map $f_\sigma : \Delta_K \to [-\beta, \beta]$. Given this map, define

$$\ell_\sigma(a, y) := \frac{1}{2} + \frac{f_\sigma(a)^2}{4\beta} - \frac{f_\sigma(a)\tau(y)}{2}. \tag{B.2}$$

Completing the square gives

$$\ell_\sigma(a, y) = \frac{1}{2} - \frac{\beta}{4} + \frac{(f_\sigma(a) - \beta\tau(y))^2}{4\beta}. \tag{B.3}$$

Thus, for each outcome y, the displayed expression is minimized precisely when $f_\sigma(a) = \beta\tau(y)$, and its minimum value is $1/2 - \beta/4$.

We therefore prescribe

$$f_\sigma(x_i) = \beta\tau(\sigma_i) \quad (i \in [M]).$$

Since $h_{\mathrm{id}}(x_i) = x_i$, these values make the identity hypothesis attain the minimum $1/2 - \beta/4$ at every context.

It remains to construct f_σ with these prescribed values and whose range is exactly $[-\beta, \beta]$. The auxiliary points force the two endpoint values: set

$$f_\sigma(z_-) = -\beta, \qquad f_\sigma(z_+) = \beta.$$

All prescribed values differ by at most 2β, whereas the corresponding points are pairwise 4β-separated. Thus this partial function is $1/2$-Lipschitz. The McShane extension theorem, followed by clipping to $[-\beta, \beta]$, gives a $1/2$-Lipschitz function $f_\sigma : \Delta_K \to [-\beta, \beta]$ without changing the prescribed values. Because f_σ is continuous and Δ_K is connected, $f_\sigma(\Delta_K)$ is an interval. It is contained in $[-\beta, \beta]$ and contains both endpoints, so

$$f_\sigma(\Delta_K) = [-\beta, \beta]. \tag{B.4}$$

This proves property (1).

We now prove property (2). Fix a prediction p. Since $f_\sigma(\Delta_K) = [-\beta, \beta]$, minimizing expected loss over actions is equivalent to minimizing

$$\frac{1}{2} + \frac{t^2}{4\beta} - \frac{t(1 - 2p_0)}{2}, \qquad t \in [-\beta, \beta].$$

This strictly convex quadratic is uniquely minimized at $t = \beta(1 - 2p_0)$. Every best response therefore satisfies $f_\sigma(\mathrm{BR}_{\ell_\sigma}(p)) = \beta(1 - 2p_0)$. Substituting into (B.3) gives

$$\begin{aligned} \ell_\sigma^{\mathrm{pr}}(p, y) &= \frac{1}{2} - \frac{\beta}{4} + \frac{\big(\beta(1 - 2p_0) - \beta\tau(y)\big)^2}{4\beta} \\ &= \frac{1}{2} - \frac{\beta}{4} + \beta\big(p_0 - \mathbf{1}_{\{y=0\}}\big)^2. \end{aligned}$$

This proves property (2).

It remains to verify that $\ell_\sigma \in \mathcal{L}_{\mathrm{Lip}}$. For fixed y, regard (B.2) as a function of $t = f_\sigma(a)$. Its derivative is

$$\frac{t}{2\beta} - \frac{\tau(y)}{2},$$

whose absolute value is at most 1 on $[-\beta, \beta]$. Since f_σ is 1/2-Lipschitz, $\ell_\sigma(\cdot, y)$ is 1/2-Lipschitz. Moreover, both $f_\sigma(a)$ and $\beta\tau(y)$ lie in $[-\beta, \beta]$, so (B.3) gives

$$\ell_\sigma(a, y) \in \left[\frac{1}{2} - \frac{\beta}{4}, \frac{1}{2} + \frac{3\beta}{4}\right] \subseteq [-1, 1],$$

where the last inclusion uses $\beta \le 2/3$. Thus $\ell_\sigma \in \mathcal{L}_{\mathrm{Lip}}$. □

Completing the proof.

Proof. Set $\beta = 16\varepsilon$. For sufficiently small ε, we have $\beta \le 1/16$. By Lemma 4.2, take a grid $\mathcal{X} = \{x_1, \dots, x_M\}$ whose distinct points are at least 4β apart, with $M = \Theta_K(\beta^{-(K-1)})$ and every coordinate at least $1/(2K)$. Set $z_- = e_0$ and $z_+ = e_1$. For $x \in \mathcal{X}$ and $j \in \{0, 1\}$,

$$\|x - e_j\|_2 \ge 1 - x_j \ge \frac{K-1}{2K} \ge \frac{1}{4} \ge 4\beta.$$

The two vertices are distance $\sqrt{2}$ apart, so all $M + 2$ points are pairwise 4β-separated. Apply Lemma B.2 to obtain the family $\{\ell_\sigma\}_{\sigma \in \{0,1\}^M}$.

For each fixed assignment $\sigma \in \{0, 1\}^M$, define $\mathcal{D}_\sigma$ by sampling contexts uniformly and setting $y = \sigma_i$ at $x = x_i$. Run the learner on n independent examples, and evaluate its output p once at every context, using fresh prediction randomness if needed. Define

$$\widehat{\sigma}_i = \begin{cases} 0, & p_0(x_i) \ge 1/2, \\ 1, & p_0(x_i) < 1/2. \end{cases}$$

The true probability of label 0 is either 1, when $\sigma_i = 0$, or 0, when $\sigma_i = 1$. The decoder rounds $p_0(x_i)$ to the nearer of these two values and returns the corresponding hidden label. By property (2) of Lemma B.2,

$$\ell^{\mathrm{pr}}_\sigma(p(x_i), \sigma_i) - \ell_\sigma(x_i, \sigma_i) = \beta\big(p_0(x_i) - \mathbf{1}_{\{\sigma_i = 0\}}\big)^2 \ge \frac{\beta}{4}\mathbf{1}_{\{\widehat{\sigma}_i \ne \sigma_i\}}.$$

Average first over the fresh prediction draws and then over the training run. For every fixed assignment σ, uniformity of the context distribution gives

$$\frac{\beta}{4M}\mathbb{E}[d_{\mathrm{H}}(\widehat{\sigma}, \sigma)] \le \mathbb{E}\Big[\mathrm{OmniReg}_{\mathcal{L}_{\mathrm{Lip}}, \{h_{\mathrm{id}}\}; \mathcal{D}_\sigma}(p)\Big] \le \varepsilon.$$

Thus $\beta = 16\varepsilon$ gives expected Hamming error at most $M/4$.

Averaging over independent fair labels and applying the first bound in Lemma 4.1 with $\delta = 1/4$ yields

$$n \ge \frac{M}{2} = \Omega_K\left(\varepsilon^{-(K-1)}\right).$$

□

B.2 Proof of the ridge-loss lemma

Proof of Lemma 4.4. We use the notation of Section 4.2.2. Item (i) holds by definition. The remaining items follow from three claims. These losses satisfy the following properties.

(1) Under $p^*(x)$, every action in $A \setminus \{x\}$ has expected loss at least β greater than that of x. For any prediction p, if label 1 is more likely than label 0 under $p^*(x)$ but $p_1 \le p_0$, or label 0 is more likely than label 1 under $p^*(x)$ but $p_0 \le p_1$, then $a(x)$ has expected loss under p at least β smaller than that of x.

(2) For each $a \notin A$, some $a' \in A$ satisfies $\ell_\sigma(a, y) > \ell_\sigma(a', y)$ for every label y.

(3) The scaled loss satisfies

$$\frac{\ell_\sigma}{150K} \in \mathcal{L}_{\mathrm{Lip}}.$$

Proof of Claim 1. Fix a context x. We first compare all actions in A under the true distribution $p^*(x)$. For every $x' \in \mathcal{X}$,

$$\begin{aligned}\mathbb{E}_{y\sim p^*(x)}\big[\ell_\sigma(x', y) - \ell_\sigma(x, y)\big] &= \frac{1}{2}\|x' - x\|_2^2 - \langle x' - x, p^*(x) - x\rangle \\ &= \frac{1}{2}\|x' - x\|_2^2.\end{aligned}$$

The inner product vanishes because $x' - x$ changes coordinates 0 and 1 equally, whereas $p^*(x) - x$ changes them by opposite amounts. Since $p_1^*(x) - p_0^*(x) = 2\sigma(x)\beta$, the added action at x' satisfies

$$\mathbb{E}_{y\sim p^*(x)}\big[\ell_\sigma(a(x'), y) - \ell_\sigma(x, y)\big] = \frac{1}{2}\|x' - x\|_2^2 + \big(2\sigma(x')\sigma(x) - 1\big)\beta.$$

The action $a(x)$ has true excess loss β. For $x' \neq x$, the grid spacing gives $\frac{1}{2}\|x' - x\|_2^2 \ge 32\beta$, so the true excess losses of x' and $a(x')$ are at least 32β and 29β, respectively. Thus x minimizes true expected loss on A, and every other action in A has true excess loss at least β.

For any prediction p with $\sigma(x)(p_1 - p_0) \le 0$,

$$\mathbb{E}_{y\sim p}\big[\ell_\sigma(a(x), y) - \ell_\sigma(x, y)\big] = \sigma(x)(p_1 - p_0) - \beta \le -\beta.$$

Hence x cannot minimize expected loss under p on A. Any minimizer under p in A therefore has expected loss under $p^*(x)$ at least β greater than that of x.

Proof of Claims 2 and 3. For each label y, half the squared loss is 2-Lipschitz on the simplex. If $\sigma(x) = \sigma(x')$, then

$$\begin{aligned}|\ell_\sigma(a(x), y) - \ell_\sigma(a(x'), y)| &= |\ell_\sigma(x, y) - \ell_\sigma(x', y)| \\ &\le 2\|x - x'\|_2 = 2\|a(x) - a(x')\|_2.\end{aligned}$$

Every action $a \in A$ satisfies

$$a_1 - a_0 \in \left\{0, \frac{1}{2K}, -\frac{1}{2K}\right\}.$$

Actions with different values of $a_1 - a_0$ are at least $1/(4K)$ apart, and their losses differ by at most 6. Thus $\ell_\sigma(\cdot, y)$ is $24K$-Lipschitz on A.

For $a \in A$, this bound gives

$$\min_{a'\in A}\big\{\ell_\sigma(a', y) + 72K\|a - a'\|_2\big\} = \ell_\sigma(a, y),$$

with the minimum attained at $a' = a$. Hence, on the whole simplex, $\ell_\sigma(\cdot, y)$ is a minimum of $72K$-Lipschitz functions and is itself $72K$-Lipschitz.

For $a \notin A$, choose $a_\star \in A$ nearest to a. For every $a' \in A$, the Lipschitz bound and the triangle inequality give

$$\begin{aligned}\ell_\sigma(a', y) + 72K\|a - a'\|_2 &\geq \ell_\sigma(a_\star, y) - 24K\|a' - a_\star\|_2 + 72K\|a - a'\|_2 \\ &\geq \ell_\sigma(a_\star, y) + 48K\|a - a'\|_2 - 24K\|a - a_\star\|_2 \\ &\geq \ell_\sigma(a_\star, y) + 24K\|a - a_\star\|_2.\end{aligned}$$

Taking the minimum over a' yields

$$\ell_\sigma(a, y) \geq \ell_\sigma(a_\star, y) + 24K\|a - a_\star\|_2 > \ell_\sigma(a_\star, y).$$

Thus every best response belongs to A. Since x minimizes true expected loss on A, it also minimizes true expected loss over the full simplex.

The losses on A lie in $[-3, 3]$, and any two points in the simplex are at distance at most 2. Therefore

$$-3 \leq \ell_\sigma(a, y) \leq 3 + 144K \leq 150K.$$

Dividing by $150K$ puts the loss values in $[-1, 1]$ and makes the Lipschitz constant at most $72/150 < 1$. Thus $\ell_\sigma/(150K) \in \mathcal{L}_{\text{Lip}}$.

Claims 1 and 2 give items (ii) and (iii): every best response lies in A, the action x minimizes true expected loss on A, and when $\sigma(x)(p_1 - p_0) \leq 0$ every minimizer under p in A is at least β worse than x under $p^*(x)$. Claim 3 is item (iv). □

B.3 Lower bound for Lipschitz proper losses

Section 1.2.2 sketches a lower bound for proper losses by evaluating the Brier loss at a reference prediction nearest to the report. The resulting losses are discontinuous. We extend this construction to 1-Lipschitz proper losses, proving Theorem 4.10. The extension assigns weights to the references and adds a quadratic penalty on the distance between their weighted mean and the report.

We prove the stronger statement in which the learner returns a decision rule a_ℓ for each loss ℓ, with all these rules learned from the same sample. The rules may be randomized. A learned omnipredictor p induces such rules through $a_\ell(x) = \text{BR}_\ell(p(x))$.

B.3.1 Hard distributions and the required loss values

Let $\alpha > 0$ be sufficiently small, depending only on K. We will eventually choose α^2 to be a constant multiple of ε. Take $\mathcal{X}_{\text{prop}} = \{x_1, \ldots, x_M\}$ to be the grid in the full simplex from Lemma 4.2 with spacing 8α. Then $M = \Theta_K(\alpha^{-(K-1)})$, distinct points are at least 8α apart, and every coordinate is at least $1/(2K)$. Each x_i is also a probability vector, which we use as the midpoint of the two possible label distributions at that context. Define the two hypotheses by

$$h_+(x_i) = x_i + \frac{\alpha}{2\sqrt{2}}(e_1 - e_0), \qquad h_-(x_i) = x_i - \frac{\alpha}{2\sqrt{2}}(e_1 - e_0), \tag{B.5}$$

and write $\mathcal{H}_{\text{prop}} = \{h_+, h_-\}$. The coordinate lower bound ensures that both predictions belong to Δ_K for sufficiently small α. Their midpoint is x_i, their distance from each other is α, and each is at distance $\alpha/2$ from x_i. The 8α separation between contexts will let us prescribe the loss values at the two hypotheses' reports separately for each context.

For each assignment $\sigma \in \{-1,+1\}^M$, define $\mathcal{D}_\sigma$ by sampling contexts uniformly and using the true conditional label distribution

$$p^*(x_i) = \begin{cases} h_+(x_i), & \sigma_i = +1, \\ h_-(x_i), & \sigma_i = -1. \end{cases}$$

The contexts and the hypotheses are the same for every assignment; only the choice of the true label distribution at each context changes.

When the signs are independent and uniform, recovering at least three quarters of them on average requires $\Omega_K(M/\alpha^2)$ samples by Lemma 4.1. We will show that expected regret at most ε forces this accuracy when α^2 is a sufficiently large constant multiple of ε.

Each regret compares the learner with a single hypothesis over the whole context space. We therefore need losses for which one fixed hypothesis is optimal at every context, so that conditional excess losses cannot cancel. For each $J \subseteq [M]$, we construct a loss ℓ_J that gives $h_+(x_i)$ and $h_-(x_i)$ their respective scaled Brier losses when $i \in J$, and gives both the same loss when $i \notin J$. For two assignments, choose J to consist of the indices where the first selects h_+ and the second selects h_-. Then h_+ will be optimal everywhere under the first assignment, and h_- under the second. A second loss handles the contexts with the opposite choices. The following lemma states the required properties. We give its construction and proof after completing the sample lower bound. Write $B(a,y) = \frac{1}{2}\|a - e_y\|_2^2$ for the Brier loss.

Lemma B.3 (Losses for the hidden-sign construction)**.** *For every $J \subseteq [M]$, there is a loss ℓ_J, depending only on the contexts, the two hypotheses, and J, with the following properties.*

(1) Properness and boundedness. *The loss ℓ_J is proper and takes values in* $[0,1]$.

(2) Lipschitz continuity. *For every label y and reports $a, a' \in \Delta_K$, $|\ell_J(a,y) - \ell_J(a',y)| \le \|a - a'\|_2$.*

(3) Loss values at x_i, $h_+(x_i)$, and $h_-(x_i)$. *For every $i \in [M]$ and label y, $\ell_J(x_i,y) = B(x_i,y)/3$. If $i \in J$, then*

$$\ell_J(h_+(x_i),y) = \frac{1}{3}B(h_+(x_i),y), \qquad \ell_J(h_-(x_i),y) = \frac{1}{3}B(h_-(x_i),y).$$

If $i \notin J$, then

$$\ell_J(h_+(x_i),y) = \ell_J(h_-(x_i),y) = \frac{1}{3}B(x_i,y).$$

The value at x_i will let us use properness under the average of the two possible true distributions to lower-bound the learner's combined loss under those distributions.

B.3.2 Recovering the hidden signs from small regret

Fix the learner's returned collection of decision rules. For each candidate assignment $\tau \in \{-1,+1\}^M$, evaluate these same rules under $\mathcal{D}_\tau$. Define their regret against $\mathcal{H}_{\mathrm{prop}}$ for the loss ℓ_J by

$$\mathrm{Reg}_J(\tau) := \mathbb{E}_{(x,y)\sim\mathcal{D}_\tau}[\ell_J(a_{\ell_J}(x),y)] - \min_{h\in\mathcal{H}_{\mathrm{prop}}} \mathbb{E}_{(x,y)\sim\mathcal{D}_\tau}[\ell_J(h(x),y)],$$

where the first expectation includes the randomness of the decision rule a_{ℓ_J}. We hold the learned rules fixed while varying τ. Take the largest regret over the constructed losses and choose the assignment for which it is smallest:

$$\mathrm{Reg}(\tau) := \max_{J\subseteq[M]} \mathrm{Reg}_J(\tau), \qquad \widehat{\sigma} \in \operatorname*{arg\,min}_{\tau\in\{-1,+1\}^M} \mathrm{Reg}(\tau),$$

using a fixed rule for ties. This choice of $\widehat{\sigma}$ uses only the learned rules and the known candidate distributions. Because the true assignment σ is one of the candidates, $\mathrm{Reg}(\widehat{\sigma}) \le \mathrm{Reg}(\sigma)$. Thus small regret under the true assignment also gives small regret under the assignment chosen by the decoder.

The next lemma shows why this forces the decoder to recover most signs: the sum of the regrets under two assignments is at least $\alpha^2/24$ times the fraction of signs on which they disagree. Applied to σ and $\widehat{\sigma}$, it will give

$$\frac{\alpha^2}{24M} d_{\mathrm{H}}(\sigma, \widehat{\sigma}) \le \mathrm{Reg}(\sigma) + \mathrm{Reg}(\widehat{\sigma}) \le 2\,\mathrm{Reg}(\sigma). \tag{B.6}$$

Thus $\mathbb{E}_{\mathrm{train}}[\mathrm{Reg}(\sigma)] \le \varepsilon$ will imply an expected fraction of incorrectly recovered signs at most $48\varepsilon/\alpha^2$. With $\alpha^2 = 192\varepsilon$, this fraction is at most $1/4$.

Lemma B.4 (Separating hidden assignments)**.** *For every fixed collection of possibly randomized decision rules and every two assignments $\sigma, \tau \in \{-1, +1\}^M$,*

$$\mathrm{Reg}(\sigma) + \mathrm{Reg}(\tau) \ge \frac{\alpha^2}{24M}\, d_{\mathrm{H}}(\sigma, \tau). \tag{B.7}$$

Proof. We use a separate loss for each direction of disagreement:

$$J_+ = \{i : \sigma_i = +1,\ \tau_i = -1\}, \qquad J_- = \{i : \sigma_i = -1,\ \tau_i = +1\}.$$

Consider first the loss ℓ_{J_+}. At a context x_i with $i \in J_+$, the true label distribution is $h_+(x_i)$ under $\mathcal{D}_\sigma$ and $h_-(x_i)$ under $\mathcal{D}_\tau$. Properness makes $h_+(x_i)$ optimal in the first case and $h_-(x_i)$ optimal in the second. If $i \notin J_+$, property (3) of Lemma B.3 gives the two reports the same loss for every label. Under either assignment, one report is the true distribution, so it is optimal by properness. The equality makes the other report optimal as well. Consequently, h_+ is optimal at every context under $\mathcal{D}_\sigma$ for the loss ℓ_{J_+}, and h_- is optimal at every context under $\mathcal{D}_\tau$ for that same loss.

For ℓ_{J_-}, the true distribution at each context indexed by J_- is instead $h_-(x_i)$ under $\mathcal{D}_\sigma$ and $h_+(x_i)$ under $\mathcal{D}_\tau$. At every context x_i with $i \notin J_-$, both reports are again optimal because their losses are equal. Thus h_- is optimal at every context under $\mathcal{D}_\sigma$ for ℓ_{J_-}, and h_+ is optimal at every context under $\mathcal{D}_\tau$ for that loss. For each loss and assignment, we have identified a fixed hypothesis attaining the minimum in the regret definition. Since it is optimal at every context, the regret is the uniform average of nonnegative conditional excess losses against that hypothesis.

We next lower-bound the sum of the conditional excess losses at a context where the two assignments disagree. Let J be either J_+ or J_- and fix $i \in J$. For any fixed action a, the identity $h_+(x_i) + h_-(x_i) = 2x_i$ gives

$$\mathbb{E}_{y\sim h_+(x_i)}[\ell_J(a, y)] + \mathbb{E}_{y\sim h_-(x_i)}[\ell_J(a, y)] = 2\mathbb{E}_{y\sim x_i}[\ell_J(a, y)] \ge 2\mathbb{E}_{y\sim x_i}[\ell_J(x_i, y)].$$

The inequality uses properness with true label distribution $p^* = x_i$. At each of the reports x_i, $h_+(x_i)$, and $h_-(x_i)$, property (3) identifies ℓ_J with one third of the Brier loss. Using $\mathbb{E}_{y\sim p^*}[B(p^*, y)] = (1 - \|p^*\|_2^2)/2$ and subtracting the expected loss of the truthful report under each of the two possible distributions, we obtain

$$\begin{aligned}
&\mathbb{E}_{y\sim h_+(x_i)}[\ell_J(a, y) - \ell_J(h_+(x_i), y)] + \mathbb{E}_{y\sim h_-(x_i)}[\ell_J(a, y) - \ell_J(h_-(x_i), y)] \\
&\ge \frac{1}{6}\left(\|h_+(x_i)\|_2^2 + \|h_-(x_i)\|_2^2 - 2\|x_i\|_2^2\right) \\
&= \frac{1}{6}\left(\|h_+(x_i) - x_i\|_2^2 + \|h_-(x_i) - x_i\|_2^2\right) = \frac{\alpha^2}{12}.
\end{aligned}$$

The first equality uses $h_+(x_i) + h_-(x_i) = 2x_i$. The last uses that both reports are at distance $\alpha/2$ from x_i.

For each fixed loss ℓ_J, its returned rule $a_{\ell_J}(x_i)$ is evaluated under both assignments. Even if that rule is randomized, its distribution over actions at x_i is the same in the two evaluations. We can therefore average the displayed inequality over that action distribution. The rule for ℓ_{J_+} and the rule for ℓ_{J_-} may be different; the argument applies to each rule separately.

Every index where σ and τ disagree belongs to exactly one of J_+ and J_-. For the corresponding loss, the sum of the two conditional excess losses is at least $\alpha^2/12$. At all other contexts, conditional excess losses are nonnegative because the fixed comparator for that loss and assignment is optimal there. Summing over $i \in J_+$ for ℓ_{J_+} and over $i \in J_-$ for ℓ_{J_-}, and averaging over the uniform context distribution, gives

$$\begin{aligned}\frac{\alpha^2}{12M}\, d_{\mathrm{H}}(\sigma,\tau) &\le \mathrm{Reg}_{J_+}(\sigma) + \mathrm{Reg}_{J_+}(\tau) + \mathrm{Reg}_{J_-}(\sigma) + \mathrm{Reg}_{J_-}(\tau) \\ &\le 2\,\mathrm{Reg}(\sigma) + 2\,\mathrm{Reg}(\tau).\end{aligned}$$

There are two regrets under σ, each at most $\mathrm{Reg}(\sigma)$, and two under τ, each at most $\mathrm{Reg}(\tau)$. Dividing by two proves the lemma. □

Proof of Theorem 4.10. Use the contexts, hypotheses, and losses above with $\alpha^2 = 192\varepsilon$, taking ε sufficiently small that the grid and the two hypotheses are well defined. By properties (1) and (2) of Lemma B.3, all these losses belong to $\mathcal{L}_{\mathrm{Lip}} \cap \mathcal{L}_{\mathrm{prop}}$. Since $\mathrm{Reg}(\sigma)$ maximizes only over this family, the learner's guarantee gives $\mathbb{E}_{\mathrm{train}}[\mathrm{Reg}(\sigma)] \le \varepsilon$ for every fixed assignment σ. Taking expectations in (B.6) therefore gives

$$\mathbb{E}[d_{\mathrm{H}}(\sigma,\widehat{\sigma})] \le \frac{48M\varepsilon}{\alpha^2} = \frac{M}{4}.$$

Now draw the signs independently and uniformly from $\{-1,+1\}$. The same Hamming bound holds after averaging over these signs. The decoder is a function of the training sample and independent training randomness. At each context, the two possible conditional label distributions $h_+(x_i)$ and $h_-(x_i)$ are distance α apart, and their midpoint x_i has every coordinate at least $1/(2K)$. Lemma 4.1 therefore gives

$$n = \Omega_K\left(\frac{M}{\alpha^2}\right) = \Omega_K\left(\alpha^{-(K+1)}\right) = \Omega_K\left(\varepsilon^{-(K+1)/2}\right).$$

□

B.3.3 Constructing the loss family

Proof of Lemma B.3. Fix $J \subseteq [M]$. Include every x_i as a reference prediction, and also include $h_+(x_i)$ and $h_-(x_i)$ whenever $i \in J$:

$$A_J = \{x_1,\dots,x_M\} \cup \{h_+(x_i), h_-(x_i) : i \in J\}.$$

In place of choosing a single nearest reference, we allow each $b \in A_J$ to receive a nonnegative weight $\lambda(b)$, with $\sum_{b\in A_J} \lambda(b) = 1$. These weights specify how much each reference contributes to the loss. We combine the weighted average of the reference Brier losses with the Brier loss at their weighted mean. The Brier loss at the weighted mean supplies the squared-distance bound used below to prove Lipschitz continuity.

For each report a, choose weights λ_a minimizing

$$\sum_{b\in A_J} \lambda(b)\|b-a\|_2^2 + \left\|\sum_{b\in A_J} \lambda(b)b - a\right\|_2^2. \tag{B.8}$$

The first term is a weighted average of squared distances from the individual references to a; the second is the squared distance from their weighted mean to a. A minimizer exists by continuity on the compact simplex of weights. Write $m_a = \sum_{b\in A_J} \lambda_a(b)b$ and define

$$\ell_J(a,y) = \frac{1}{6}\left(\sum_{b\in A_J} \lambda_a(b)B(b,y) + B(m_a,y)\right). \tag{B.9}$$

Property (1): properness and boundedness. Fix a true distribution $p^* \in \Delta_K$. For any weights λ with mean $m = \sum_{b\in A_J} \lambda(b)b$, the Brier identity $\mathbb{E}_{y\sim p^*}[B(b,y)] = \frac{1}{2}\|b-p^*\|_2^2 + \frac{1}{2}(1-\|p^*\|_2^2)$ gives

$$\begin{aligned}&\frac{1}{6}\mathbb{E}_{y\sim p^*}\left[\sum_{b\in A_J} \lambda(b)B(b,y) + B(m,y)\right]\\ &\quad= \frac{1}{12}\left(\sum_{b\in A_J} \lambda(b)\|b-p^*\|_2^2 + \|m-p^*\|_2^2\right) + \frac{1}{6}(1-\|p^*\|_2^2).\end{aligned}$$

The part depending on the weights is precisely (B.8) with $a = p^*$, scaled by $1/12$. Thus λ_{p^*} minimizes this expected value over all choices of weights. In particular, reporting p^* has expected loss no greater than reporting any a, which selects the weights λ_a. This proves properness. Both Brier terms in (B.9) take values in $[0,1]$, so $\ell_J(a,y) \in [0,1/3] \subseteq [0,1]$.

Property (2): Lipschitz continuity. Fix two reports a, a'. For $0 < t < 1$, the weights $(1-t)\lambda_a(b) + t\lambda_{a'}(b)$ are nonnegative, sum to one, and have mean $(1-t)m_a + tm_{a'}$. Substituting these weights into the right-hand side of (B.9) and expanding the Brier term at their mean gives

$$(1-t)\ell_J(a,y) + t\ell_J(a',y) - \frac{t(1-t)}{12}\|m_a - m_{a'}\|_2^2.$$

When the true distribution is a, the weights λ_a minimize the expected value over all weights, as just shown. Thus $\mathbb{E}_{y\sim a}[\ell_J(a,y)]$ cannot exceed the expectation of the displayed expression. Subtract this expected loss from both sides, divide by t, and let t decrease to zero. We obtain

$$\mathbb{E}_{y\sim a}[\ell_J(a',y) - \ell_J(a,y)] \geq \frac{1}{12}\|m_a - m_{a'}\|_2^2.$$

To combine this inequality with the one obtained by interchanging a and a', expand the loss as

$$\ell_J(a,y) = \frac{1}{12}\left(\sum_{b\in A_J} \lambda_a(b)\|b\|_2^2 + \|m_a\|_2^2 + 2\right) - \frac{1}{3}(m_a)_y.$$

Only the last term depends on y. When we add the two inequalities, the other terms cancel, giving

$$\begin{aligned}\frac{1}{6}\|m_a - m_{a'}\|_2^2 &\leq \mathbb{E}_{y\sim a}[\ell_J(a',y) - \ell_J(a,y)] + \mathbb{E}_{y\sim a'}[\ell_J(a,y) - \ell_J(a',y)]\\ &= \frac{1}{3}\langle a - a', m_a - m_{a'}\rangle.\end{aligned}$$

Cauchy–Schwarz therefore gives $\|m_a - m_{a'}\|_2 \le 2\|a - a'\|_2$.

We now bound the loss difference for a fixed label y. The same expansion gives

$$\ell_J(a, y) - \ell_J(a', y) = \mathbb{E}_{y' \sim a}[\ell_J(a, y') - \ell_J(a', y')] + \frac{1}{3}\langle a - e_y, m_a - m_{a'}\rangle.$$

The expected difference is nonpositive by properness under the true distribution a. Since $\|a - e_y\|_2 \le \sqrt{2}$, this bounds the loss difference above by $(\sqrt{2}/3)\|m_a - m_{a'}\|_2$. Interchanging a and a' bounds its negative by the same quantity. Consequently,

$$|\ell_J(a, y) - \ell_J(a', y)| \le \frac{\sqrt{2}}{3}\|m_a - m_{a'}\|_2 \le \frac{2\sqrt{2}}{3}\|a - a'\|_2 \le \|a - a'\|_2.$$

Property (3): loss values at x_i, $h_+(x_i)$, and $h_-(x_i)$. For any report $a \in A_J$, putting all weight at a makes (B.8) zero. Both terms are nonnegative, so every minimizing choice must satisfy $\sum_{b \in A_J} \lambda_a(b)\|b - a\|_2^2 = 0$. Hence all its weight is at a, and (B.9) gives $\ell_J(a, y) = B(a, y)/3$. This proves the required value at every x_i, and at both $h_+(x_i)$ and $h_-(x_i)$ when $i \in J$.

Fix $i \notin J$ and let a be either $h_+(x_i)$ or $h_-(x_i)$. We show that every minimizing choice puts all its weight at x_i. Putting all weight there gives the value $2\|x_i - a\|_2^2 = \alpha^2/2$ in (B.8). For any weights λ with mean m, the difference from this value is

$$\begin{aligned}&\sum_{b \in A_J} \lambda(b)\|b - a\|_2^2 + \|m - a\|_2^2 - 2\|x_i - a\|_2^2 \\ &\quad = \sum_{b \in A_J} \lambda(b)\left(\|b - x_i\|_2^2 - 4\langle a - x_i, b - x_i\rangle\right) + \|m - x_i\|_2^2.\end{aligned}$$

Every other reference $b \ne x_i$ is either x_j, $h_+(x_j)$, or $h_-(x_j)$ for some $j \ne i$, and is therefore within distance $\alpha/2$ of x_j. Thus $\|b - x_i\|_2 \ge 8\alpha - \alpha/2 = 15\alpha/2$. Since $\|a - x_i\|_2 = \alpha/2$, Cauchy–Schwarz gives

$$\|b - x_i\|_2^2 - 4\langle a - x_i, b - x_i\rangle \ge \|b - x_i\|_2(\|b - x_i\|_2 - 2\alpha) > 0.$$

The expression in parentheses in the sum is zero at $b = x_i$, and the remaining squared norm is nonnegative. Thus any positive weight outside x_i strictly increases (B.8). Every minimizing choice therefore puts all its weight at x_i. Equation (B.9) gives $\ell_J(h_+(x_i), y) = \ell_J(h_-(x_i), y) = B(x_i, y)/3$, as required. □

References

[AB99] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[ABH11] J. Abernethy, P. L. Bartlett, and E. Hazan. "Blackwell approachability and no-regret learning are equivalent". In: *Proceedings of the 24th Annual Conference on Learning Theory*. JMLR Workshop and Conference Proceedings. 2011, pp. 27–46.

[Alo+97] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. "Scale-Sensitive Dimensions, Uniform Convergence, and Learnability". In: *Journal of the ACM* 44.4 (1997), pp. 615–631.

[BI91] G. M. Benedek and A. Itai. "Learnability with respect to fixed distributions". In: *Theoretical Computer Science* 86.2 (1991), pp. 377–389.

[BL98] P. L. Bartlett and P. M. Long. "Prediction, learning, uniform convergence, and scale-sensitive dimensions". In: *Journal of Computer and System Sciences* 56.2 (1998), pp. 174–190.

[Bla56] D. Blackwell. "An analog of the minimax theorem for vector payoffs." In: *Pacific Journal of Mathematics* 6.1 (1956), pp. 1–8. URL: https://doi.org/10.2140/pjm.1956.6.1.

[BLR25] Y. Bechavod, J. Lu, and A. Roth. "Online omniprediction with long-term constraints". In: *arXiv preprint arXiv:2509.11357* (2025).

[Bro76] E. M. Bronshtein. "ε-Entropy of Convex Sets and Functions". In: *Siberian Mathematical Journal* 17.3 (1976), pp. 393–398.

[CK25] S. Casacuberta and V. Kanade. "Selective omniprediction and fair abstention". In: *Advances in Neural Information Processing Systems*. Vol. 38. 2025, pp. 39435–39477.

[CT06] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. 2nd ed. Wiley-Interscience, 2006.

[Dud67] R. M. Dudley. "The sizes of compact subsets of Hilbert space and continuity of Gaussian processes". In: *Journal of Functional Analysis* 1.3 (1967), pp. 290–330.

[Dwo+21] C. Dwork, M. P. Kim, O. Reingold, G. N. Rothblum, and G. Yona. "Outcome indistinguishability". In: *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*. 2021, pp. 1095–1108.

[Dwo+22] C. Dwork, M. P. Kim, O. Reingold, G. N. Rothblum, and G. Yona. "Beyond bernoulli: Generating random outcomes that cannot be distinguished from nature". In: *International Conference on Algorithmic Learning Theory*. PMLR. 2022, pp. 342–380.

[Dwo+25] C. Dwork, C. Hays, N. Immorlica, J. C. Perdomo, and P. Tankala. "From Fairness to Infinity: Outcome-Indistinguishable (Omni)Prediction in Evolving Graphs". In: *Proceedings of Thirty Eighth Conference on Learning Theory*. Ed. by N. Haghtalab and A. Moitra. Vol. 291. Proceedings of Machine Learning Research. PMLR, 30 Jun–04 Jul 2025, pp. 1564–1637.

[FP26] G. Farina and J. C. Perdomo. "An Efficient Black-Box Reduction from Online Learning to Multicalibration, and a New Route to Φ-Regret Minimization". In: *CoRR* abs/2604.19592 (2026). arXiv: 2604.19592 [cs.LG].

[Gar+24] S. Garg, C. Jung, O. Reingold, and A. Roth. "Oracle efficient online multicalibration and omniprediction". In: *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM. 2024, pp. 2725–2792.

[GHR24] P. Gopalan, L. Hu, and G. N. Rothblum. "On computationally efficient multi-class calibration". In: *The Thirty Seventh Annual Conference on Learning Theory*. PMLR. 2024, pp. 1983–2026.

[GKR23] P. Gopalan, M. P. Kim, and O. Reingold. "Swap Agnostic Learning, or Characterizing Omniprediction via Multicalibration". In: *Advances in Neural Information Processing Systems*. Vol. 36. 2023. URL: https://proceedings.neurips.cc/paper_files/paper/2023/hash/7d693203215325902ff9dbdd067a50ac-Abstract-Conference.html.

[Gop+22] P. Gopalan, A. T. Kalai, O. Reingold, V. Sharan, and U. Wieder. "Omnipredictors". In: *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*. Vol. 215. LIPIcs. 2022, 79:1–79:21.

[Gop+23] P. Gopalan, L. Hu, M. P. Kim, O. Reingold, and U. Wieder. "Loss Minimization Through the Lens of Outcome Indistinguishability". In: *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*. Vol. 251. LIPIcs. 2023, 60:1–60:20.

[Gop+24] P. Gopalan, P. Okoroafor, P. Raghavendra, A. Shetty, and M. Singhal. "Omnipredictors for regression and the approximate rank of convex functions". In: *The Thirty Seventh Annual Conference on Learning Theory*. PMLR. 2024, pp. 2027–2070.

[Gop+26] P. Gopalan, K. Stavropoulos, K. Talwar, and P. Tankala. "The Importance of Being Smoothly Calibrated". In: *7th Symposium on Foundations of Responsible Computing (FORC 2026)*. Vol. 368. Leibniz International Proceedings in Informatics (LIPIcs). Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2026, 21:1–21:22.

[GR07] T. Gneiting and A. E. Raftery. "Strictly proper scoring rules, prediction, and estimation". In: *Journal of the American statistical Association* 102.477 (2007), pp. 359–378.

[GS13] A. Guntuboyina and B. Sen. "Covering Numbers for Convex Functions". In: *IEEE Transactions on Information Theory* 59.4 (2013), pp. 1957–1965.

[GT26] I. Gibbs and R. J. Tibshirani. "Sample-Efficient Omniprediction for Proper Losses". In: *Proceedings of the Thirty-Ninth Conference on Learning Theory*. Vol. 336. Proceedings of Machine Learning Research. PMLR, 2026, pp. 2679–2719. URL: https://proceedings.mlr.press/v336/gibbs26a.html.

[Hau92] D. Haussler. "Decision theoretic generalizations of the PAC model for neural net and other learning applications". In: *Information and Computation* 100.1 (1992), pp. 78–150.

[Héb+18] U. Hébert-Johnson, M. Kim, O. Reingold, and G. Rothblum. "Multicalibration: Calibration for the (Computationally-Identifiable) Masses". In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. Proceedings of Machine Learning Research. 2018, pp. 1939–1948.

[HTY25] L. Hu, K. Tian, and C. Yang. "Omnipredicting single-index models with multi-index models". In: *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*. 2025, pp. 1762–1773.

[HTY26] L. Hu, K. Tian, and C. Yang. "Simultaneous Blackwell Approachability and Applications to Multiclass Omniprediction". In: *Proceedings of Thirty Ninth Conference on Learning Theory* (2026), pp. 3593–3634.

[Hu+23] L. Hu, I. R. Livni Navon, O. Reingold, and C. Yang. "Omnipredictors for constrained optimization". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 13497–13527.

[HV25] L. Hu and S. Vadhan. "Generalized and unified equivalences between hardness and pseudoentropy". In: *Theory of Cryptography Conference*. Springer. 2025, pp. 258–288.

[HWY25] J. Hartline, Y. Wu, and Y. Yang. "Smooth Calibration and Decision Making". In: *6th Symposium on Foundations of Responsible Computing (FORC 2025)*. Vol. 329. Leibniz International Proceedings in Informatics (LIPIcs). Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2025, 16:1–16:26.

[KGZ19] M. P. Kim, A. Ghorbani, and J. Zou. "Multiaccuracy: Black-box post-processing for fairness in classification". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019, pp. 247–254.

[Kle+23] B. Kleinberg, R. P. Leme, J. Schneider, and Y. Teng. "U-calibration: Forecasting for an unknown agent". In: *The Thirty Sixth Annual Conference on Learning Theory*. PMLR. 2023, pp. 5143–5145.

[KP23] M. P. Kim and J. C. Perdomo. "Making Decisions Under Outcome Performativity". In: *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*. Vol. 251. Schloss Dagstuhl–Leibniz-Zentrum für Informatik. 2023, p. 79.

[KS94] M. J. Kearns and R. E. Schapire. "Efficient Distribution-Free Learning of Probabilistic Concepts". In: *Journal of Computer and System Sciences* 48.3 (1994), pp. 464–497.

[Li+22] Y. Li, J. D. Hartline, L. Shan, and Y. Wu. "Optimization of scoring rules". In: *Proceedings of the 23rd ACM Conference on Economics and Computation*. 2022, pp. 988–989.

[LRS25] J. Lu, A. Roth, and M. Shi. "Sample efficient omniprediction and downstream swap regret for non-linear losses". In: *arXiv preprint arXiv:2502.12564* (2025).

[LSS24] H. Luo, S. Senapati, and V. Sharan. "Optimal multiclass u-calibration error and beyond". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 7521–7551.

[LSS25] H. Luo, S. Senapati, and V. Sharan. "Improved Bounds for Swap Multicalibration and Swap Omniprediction". In: *Advances in Neural Information Processing Systems*. Vol. 38. 2025. URL: https://proceedings.neurips.cc/paper_files/paper/2025/hash/1d6a7d85fc445891ef9caea6876875aa-Abstract-Conference.html.

[MV03] S. Mendelson and R. Vershynin. "Entropy and the Combinatorial Dimension". In: *Inventiones Mathematicae* 152.1 (2003), pp. 37–55.

[OKK25] P. Okoroafor, R. Kleinberg, and M. P. Kim. "Near-optimal algorithms for omniprediction". In: *2025 IEEE 66th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2025, pp. 1595–1609.

[QZ25] M. Qiao and E. Zhao. "Truthfulness of Decision-Theoretic Calibration Measures". In: *Proceedings of Thirty Eighth Conference on Learning Theory*. Ed. by N. Haghtalab and A. Moitra. Vol. 291. Proceedings of Machine Learning Research. PMLR, 30 Jun–04 Jul 2025, pp. 4686–4739.

[Ros+25] R. Rossellini, J. A. Soloff, R. F. Barber, Z. Ren, and R. Willett. "Can a calibration metric be both testable and actionable?" In: *Proceedings of Thirty Eighth Conference on Learning Theory*. Ed. by N. Haghtalab and A. Moitra. Vol. 291. Proceedings of Machine Learning Research. PMLR, 30 Jun–04 Jul 2025, pp. 4937–4972.

[RST15] A. Rakhlin, K. Sridharan, and A. Tewari. "Sequential Complexities and Uniform Martingale Laws of Large Numbers". In: *Probability Theory and Related Fields* 161.1–2 (2015), pp. 111–153.

[Zha+21] S. Zhao, M. Kim, R. Sahoo, T. Ma, and S. Ermon. "Calibrating predictions to decisions: A novel approach to multi-class calibration". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 22313–22324.

C Acknowledgments and AI Usage

The project documented in this manuscript was conceived through conversations amongst the three authors, as well as Robert Kleinberg. We are grateful for Bobby's initial ideas and contributions to the directions studied.

The research process involved AI usage at various levels. The authors verified the correctness

and originality of all content, including references.

- The algorithmic framework and chaining arguments in Section 3 were developed primarily by the authors. GPT 6 suggested the symmetrization argument used to prove Theorem 3.8, presented in Section 3.6 and Appendix A.4. The second covering bound in Lemma 3.11, which yields the sharp sample-complexity exponents for $K = 3, 4$, was also suggested in an interaction with GPT.

- AI played a larger role in developing and completing the lower-bound proofs. The authors provided GPT 5.6 and GPT 6 Astra with proof sketches for the $\Omega_K(\varepsilon^{-(K-1)})$ Lipschitz-loss lower bound in Theorem 4.3 and the lower bound for Lipschitz proper losses in Theorem 4.10. The models supplied the remaining proof details, which the authors verified and revised for exposition.

  The sharp Lipschitz-loss lower bound for $K = 3$, obtained from Proposition 4.5, was developed entirely through an interaction with GPT 5.6. For proper calibration together with identity multiaccuracy, the authors had an alternative construction giving a weaker lower bound. GPT 5.6 suggested how to strengthen that construction by a factor of $1/\sqrt{\varepsilon}$, yielding the bound in Theorem 4.7. The unlearnability result for bounded proper losses (Theorem 4.8) was also suggested by GPT, through a modification of an earlier construction giving a lower bound exponential in K.

- Throughout the project, AI models assisted with literature searches, proofreading, and comparisons with prior work, particularly the results on omniprediction for GLM losses of Hu, Tian, and Yang [HTY26].

- The authors used GPT 5 and GPT 6 to help edit paragraphs, draft proofs from sketches, and suggest alternative formulations. These interactions typically involved several rounds of revision, with the authors supplying their own edits and feedback before settling on the final text.

- GPT 6 Astra and Sol were used to convert handwritten figure sketches into TikZ. The authors supplied the sketches for all figures.