Panprediction: Optimal Predictions for Any Downstream Task and Loss

Sivaraman Balakrishnan[1], Nika Haghtalab[2], Daniel Hsu[3], Brian Lee[2], and Eric Zhao[2]

[1]Carnegie Mellon University
[2]University of California, Berkeley
[3]Columbia University

Abstract

Supervised learning is classically formulated as training a model to minimize a fixed loss function over a fixed distribution, or task. However, an emerging paradigm instead views model training as extracting enough information from data so that the model can be used to minimize many losses on many downstream tasks. We formalize a mathematical framework for this paradigm, which we call *panprediction*, and study its statistical complexity. Formally, panprediction generalizes omniprediction and sits upstream from multi-group learning, which respectively focus on predictions that generalize to many downstream losses or many downstream tasks, but not both. Concretely, we design algorithms that learn deterministic and randomized panpredictors with $\widetilde{O}(1/\varepsilon^3)$ and $\widetilde{O}(1/\varepsilon^2)$ samples, respectively. Our results demonstrate that under mild assumptions, simultaneously minimizing infinitely many losses on infinitely many tasks can be as statistically easy as minimizing one loss on one task. Along the way, we improve the best known sample complexity guarantee of deterministic omniprediction by a factor of $1/\varepsilon$, and match all other known sample complexity guarantees of omniprediction and multi-group learning. Our key technical ingredient is a nearly lossless reduction from panprediction to a statistically efficient notion of calibration, called *step calibration*.

1 Introduction

Consider the problem of predicting the probability of an adverse medical event for a patient based on their health records—be it in-hospital mortality over 12 hours, re-admission over 30 days, or a cardiovascular event over 10 years. A range of decision makers across the healthcare system—ICU doctors, discharge planners, actuaries, and more—wish to incorporate these probabilities into their decisions. However, this problem is complicated by the fact that each decision maker defines prediction quality differently: an ICU doctor who must rapidly take action may care most about the zero-one loss, while an actuary who seeks precise probability estimates may care most about the square loss. Further complicating the problem is how each decision maker focuses on different, possibly overlapping subgroups of patients: those with specific pre-existing conditions, those in a certain age group, and more.

> *What does it mean for a single predictor to be "good" for such a heterogeneous population of decision makers? How can such a predictor be learned?*

From one perspective, we have simply described a collection of binary prediction problems, each defined by a loss function and subgroup, or task, of interest. Classic supervised learning techniques can readily produce a good model for each problem. However, this approach requires that the number of samples needed to produce good predictions for all decision makers scales linearly in the number of losses and subgroups of interest, which can be prohibitively large.

In this work, we introduce *panprediction*, an alternative framework for constructing a single probability predictor that any decision maker—focused on any loss and any subgroup—can easily post-process to solve their specialized problem. Importantly, panpredictors guarantee that decision makers could not have made better predictions by training a bespoke model for their problem. Moreover, the sample complexity of panprediction scales only logarithmically in the number of *basis functions* needed to linearly approximate losses of interest, and also in the number of subgroups.

Formally, we say a probability predictor $p^\star : \mathcal{X} \to [0,1]$ is a $(\mathcal{L}, \mathcal{G}, \mathcal{H}, \varepsilon)$-panpredictor if for every loss $\ell \in \mathcal{L}$ and group $g \in \mathcal{G}$, the predictor $p^\star$ can be post-processed into a predictor $h^\star$ that is ε-optimal, meaning

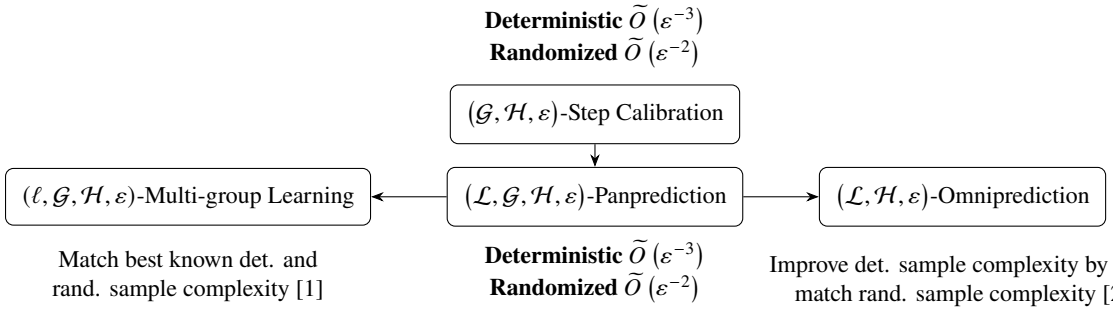

Figure 1: Panprediction is a nearly lossless generalization of omniprediction and multi-group learning in the sense that sample complexity guarantees are preserved (up to logarithmic factors), or improved. Similarly, the reduction from panprediction to step calibration preserves sample complexity guarantees.

$$\mathbb{E}_{(x,y)\sim D}\left[\ell\big(h^*(x), y\big) \;\middle|\; g(x) = 1\right] \approx_\varepsilon \min_{h\in\mathcal{H}} \mathbb{E}_{(x,y)\sim D}\left[\ell\big(h(x), y\big) \;\middle|\; g(x) = 1\right],$$

where $\mathcal{H}$ is a competitor hypothesis class. Supposing for simplicity that $\mathcal{G}$ and $\mathcal{H}$ have finite cardinality, our proposed algorithms learn deterministic panpredictors with $\widetilde{O}(\varepsilon^{-3}\log(|\mathcal{G}||\mathcal{H}|))$ samples, and randomized panpredictors with $\widetilde{O}(\varepsilon^{-2}\log(|\mathcal{G}||\mathcal{H}|))$ samples. Notably, our sample complexity guarantees account for $\mathcal{L}$ through a $\log(1/\varepsilon)$ factor, with no dependence on the cardinality or covering number of $\mathcal{L}$.

Our results are enabled by the observation that panprediction can be reduced to *step calibration* [3], a statistically efficient notion of calibration also known as proper calibration [2]. This answers our first question: a single predictor is "good" for a heterogeneous array of losses and subgroups if it is step calibrated. Algorithmically, we construct step calibrated predictors using tools from multi-objective learning, a powerful framework recently developed to tighten the sample complexity of various calibration notions [4, 5]. This answers our second question on how such "good" predictors can be learned.

Panprediction builds on recent advances in machine learning theory that seek models with greater adaptability to many downstream losses or tasks. Omniprediction [6] studies predictors whose outputs can be post-processed to perform well according to many loss functions, but for a single fixed distribution, or task. Conversely, multi-group learning [7, 1] ensures robust performance across many downstream subgroups, or tasks, but for a single fixed loss function. We generalize both lines of work and show that simultaneous adaptability to many downstream losses *and* tasks is readily attainable. Figure 1 illustrates the relationship between these concepts.

Our approach yields several benefits over prior works. Quantitatively, we improve the sample complexity of deterministic omniprediction by a factor of $1/\varepsilon$, and match all other known sample complexities of omniprediction and multi-group learning. Qualitatively, we provide a unified algorithmic path to omniprediction and multi-group learning—which prior works have treated with bespoke approaches that had difficulty learning, e.g., deterministic predictors—and significantly simplify proofs.

1.1 Related Works

Omniprediction Gopalan et al. [6] initiated the study of omniprediction by showing that an appropriately calibrated predictor can be post-processed to minimize a range of convex and Lipschitz losses. Subsequent work deepened the conceptual connection between omniprediction and calibration by formalizing the notion of a predictor being indistinguishable from the Bayes predictor, as measured by a set of loss functions [8]. Using these insights, recent work improved the sample complexity and oracle efficiency of omniprediction by sidestepping full calibration and relying on statistically efficient notions of calibration [2]. We carefully extend this line of work to account for the more challenging setting where loss functions are evaluated over overlapping subgroups of the domain.

Multi-group learning Motivated by subgroup fairness, Blum and Lykouris [9] and Rothblum and Yona [7] respectively initiated the study of multi-group learning in the sequential and batch settings. A rich line of work has since

refined sample complexity guarantees [1] and studied variants with oracle-efficiency [10], hierarchical group structure [11], and robustness concerns [12]. Typical multi-group algorithms leverage reductions to the sleeping experts problem. In contrast, we leverage a fundamental connection between multi-group learning and calibration.

Calibration Calibration originates from the online forecasting literature [13, 14], where it was proposed as a basic sanity check for predictions. More recently, Hebert-Johnson et al. [15] initiated the study of multicalibration, which connects calibration to subgroup robustness. Subsequent work developed interpretations of calibration as requiring that a learned predictor be indistinguishable from the Bayes predictor according to a class of tests [16], and variants of calibration that provide downstream decision-theoretic guarantees [17, 3]. Multi-objective learning is a flexible algorithmic framework for multicalibration [4] and related multi-distribution learning problems [5]. Our techniques build on these works, especially ideas from indistinguishability and algorithms from multi-objective learning.

2 Models and Preliminaries

We study batch prediction with binary labels. Let D be a joint distribution over context space $\mathcal{X}$ and binary labels $\mathcal{Y} = \{0, 1\}$. Let $\mathcal{H} \subset \{h : \mathcal{X} \to \widehat{\mathcal{Y}}\}$ be a class of hypothesis functions, where $\widehat{\mathcal{Y}}$ is the prediction space. We study both the binary prediction setting, where $\widehat{\mathcal{Y}} = \{0, 1\}$, and the probabilistic prediction setting, where $\widehat{\mathcal{Y}} = [0, 1]$. Let $\mathcal{P} = \{h : \mathcal{X} \to [0, 1]\}$ be the class of all real-valued hypothesis functions. Denote the simplex over $\mathcal{H}$ and $\mathcal{P}$ as $\Delta(\mathcal{H})$ and $\Delta(\mathcal{P})$, respectively. The simplex in $\mathbb{R}^d$ is denoted by Δ^{d-1}. For tractability, each real-valued hypothesis is quantized to return predictions on the λ-net of the unit interval, $I_\lambda = \{0, \lambda, 2\lambda, \dots, 1\}$, where the quantization parameter $\lambda \in (0, 1)$ is chosen by the learner with knowledge of the target error tolerance ε. Let $\mathcal{L} \subset \{\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to [-1, 1]\}$ be a class of loss functions that only take the prediction and label as inputs. Let $\mathcal{G} \subseteq 2^{\mathcal{X}}$ be a set of (possibly overlapping) groups on $\mathcal{X}$, with each group identified by a group membership function $g : \mathcal{X} \to \{0, 1\}$. Overloading notation, we say $x \in \mathcal{X}$ belongs to group $g \in \mathcal{G}$ if $g(x) = 1$. We denote group size under D by $P_g = \Pr_{(x,y)\sim D}(g(x) = 1)$.

2.1 Panprediction

We formalize the problem of constructing a flexible predictor that downstream decision makers can adapt to minimize many different loss functions on many different tasks. To obtain this flexibility, we require that the universe of all losses, tasks (represented by groups over the domain), and competitor hypotheses of interest to downstream decision makers be pre-specified and collected into the classes $(\mathcal{L}, \mathcal{G}, \mathcal{H})$. Given this triplet, the goal of panprediction is to use samples from the distribution D to construct a predictor that, for any post-hoc choice of loss $\ell \in \mathcal{L}$ and group $g \in \mathcal{G}$, can be post-processed to compete with the best hypothesis $h \in \mathcal{H}$ on the group g, as measured by the loss ℓ. We make the following assumptions about $(\mathcal{L}, \mathcal{G}, \mathcal{H})$.

First, we place a regularity condition on $\mathcal{L}$. The total variation of a function $f : [0, 1] \to [-1, 1]$ is defined as

$$V(f) = \sup_{m\in\mathbb{N}} \sup_{0=z_0<\cdots<z_m=1} \sum_{j=1}^{m} \left|f(z_j) - f(z_{j-1})\right|.$$

Assumption 1. *All losses $\ell \in \mathcal{L}$ have bounded variation in the first argument, meaning*

$$\mathcal{L} \subseteq \mathcal{L}_{BV} \triangleq \left\{\ell : \sup_{y\in\mathcal{Y}} V(\ell(\cdot, y)) \le 1\right\}.$$

This is a very mild restriction: $\mathcal{L}_{\text{BV}}$ has infinite cardinality and subsumes all standard loss functions—zero-one loss, the hinge loss, square loss, and pinball loss—as well as 1-Lipschitz functions and proper scoring rules (up to re-scaling). Losses having bounded variation is the weakest assumption made in omniprediction [2], a related problem setting discussed in Section 5.

Next, we mildly constrain the expressivity of $\mathcal{H}$ and $\mathcal{G}$.

Assumption 2. *$\mathcal{H}$ has finite combinatorial dimension.*

If $\mathcal{H}$ is a class of binary predictors, Assumption 2 requires that $\mathcal{H}$ has bounded VC dimension. If $\mathcal{H}$ is real-valued, then Assumption 2 requires that $\mathcal{H}$ has bounded pseudo-dimension. Both are standard assumptions in statistical learning theory.

Assumption 3. *$\mathcal{G}$ has finite VC dimension.*

Intuitively, Assumption 3 allows groups to have infinite cardinality and encode rich structure, but requires that they are at least *learnable* from data.

Given Assumptions 1, 2, and 3, we can formally define the desiderata of panprediction.

Definition 2.1 (Deterministic Panprediction)**.** *Given loss class $\mathcal{L}$, hypothesis class $\mathcal{H}$, and set of groups $\mathcal{G}$, a deterministic predictor $p^* : \mathcal{X} \to [0,1]$ is a $(\mathcal{L}, \mathcal{G}, \mathcal{H}, \varepsilon)$-panpredictor if for all $\ell \in \mathcal{L}$ and $g \in \mathcal{G}$,*

$$\mathop{\mathbb{E}}_{(x,y)\sim D}\left[\ell(k_\ell(p^*(x)), y) \mid g(x) = 1\right] \leq \min_{h\in\mathcal{H}} \mathop{\mathbb{E}}_{(x,y)\sim D}\left[\ell(h(x), y) \mid g(x) = 1\right] + \varepsilon \cdot \sqrt{P_g^{-1}},$$

where the post-processing function $k_\ell : [0,1] \to \widehat{\mathcal{Y}}$ is

$$k_\ell(p) = \arg\min_{\widehat{y}\in\widehat{\mathcal{Y}}} \mathop{\mathbb{E}}_{y\sim\mathrm{Ber}(p)}\left[\ell(\widehat{y}, y)\right] = \arg\min_{\widehat{y}\in\widehat{\mathcal{Y}}} \left(p \cdot \ell(\widehat{y}, 1) + (1-p)\cdot \ell(\widehat{y}, 0)\right). \tag{1}$$

The post-processing step involves solving a simple optimization problem, which downstream decision makers can do with no access to D, and often in closed form.

We can also define a less restrictive notion of panprediction that allows the predictor p^* to randomize.

Definition 2.2 (Randomized Panprediction)**.** *Given loss class $\mathcal{L}$, hypothesis class $\mathcal{H}$, and set of groups $\mathcal{G}$, a randomized predictor $\boldsymbol{p}^* \in \Delta(\mathcal{P})$ is a $(\mathcal{L}, \mathcal{G}, \mathcal{H}, \varepsilon)$-panpredictor if for all $\ell \in \mathcal{L}$ and $g \in \mathcal{G}$,*

$$\mathop{\mathbb{E}}_{(x,y)\sim D,\, p^*\sim\boldsymbol{p}^*}\left[\ell(k_\ell(p^*(x)), y) \mid g(x) = 1\right] \leq \min_{h\in\mathcal{H}} \mathop{\mathbb{E}}_{(x,y)\sim D}\left[\ell(h(x), y) \mid g(x) = 1\right] + \varepsilon \cdot \sqrt{P_g^{-1}}.$$

Remark 2.3 (Deterministic vs Randomized)**.** *All things equal, we prefer deterministic predictors to randomized ones—though they can be more sample-intensive to learn. In the sequel, we state most definitions and results for deterministic predictors, and defer a formal statement of the randomized variants to Appendix A.*

Remark 2.4 (Conditional Guarantees)**.** *We define panprediction risk as a group-conditional expectation, and error tolerance as scaling inversely with the square root of each group's size under D. Both are strong guarantees sought in multi-group learning and multicalibration [1, 4]. The choice of error tolerance scaling, in particular, reflects the optimal sample complexity of learning each group using samples from D.*

All our definitions of panprediction are non-vacuous, since the Bayes predictor $\mathbb{E}[y \mid x]$ is a panpredictor for *all* $\mathcal{L}$, $\mathcal{G}$, and $\mathcal{H}$. Consequently, a panpredictor can be interpreted as a coarsening of the Bayes predictor with respect to a particular set of losses, groups, and hypotheses. We make this notion of "coarsening" precise in the following discussion of step calibration.

2.2 Step Calibration

Calibration—which requires that a probabilistic predictor is unbiased on its own level set—is the key property of the Bayes predictor that makes it useful for a wide range of downstream decision makers. Calibration is a sufficient condition for decision making, in the sense that decision makers that treat *any* calibrated predictor as the Bayes predictor enjoy provable loss minimization guarantees [18, 6]. However, full calibration can be more statistically expensive than direct loss minimization, and recent work has explored efficiently attainable notions of calibration that retain decision-theoretic guarantees [17].

One such notion is *step calibration* [3], also known as proper calibration [2]. In contrast to full calibration, step calibration only requires that a predictor is unbiased on its *sublevel* set. That is, on the set of all instances whose predicted probability is at most a given threshold. Below, we introduce a multi-objective variant of step calibration that not only holds marginally, but also on carefully defined subsets of the domain.

Definition 2.5 ($(\mathcal{G},\mathcal{H},\varepsilon)$-Step Calibration)**.** *A deterministic predictor $p^*:\mathcal{X}\to[0,1]$ is $(\mathcal{G},\mathcal{H},\varepsilon)$-step calibrated if for all $v,w\in[0,1]$, $h\in\mathcal{H}$, and $g\in\mathcal{G}$,*

$$\left|\mathop{\mathbb{E}}_{D}\left[(y-p^*(x))\cdot\mathbb{1}[p^*(x)\le v,h(x)\le w]\mid g(x)=1\right]\right|\le\varepsilon\cdot\sqrt{P_g^{-1}}.$$

In addition to serving as the indicator for sublevel sets of p^*, the step function $\mathbb{1}[p^*(x)\le v]$ serves an approximation-theoretic purpose, as a natural basis for bounded variation functions. This connection is further discussed in Section 3. Also notice that in Definition 2.5, we condition on the event $\{g(x)=1\}$, as in the formulation of panprediction risk, and take an indicator on the events $\{p^*(x)\le v\}$ and $\{h(x)\le w\}$.

We sometimes invoke a weaker notion of unbiasedness, called multiaccuracy, that does not take an indicator on the (sub)level sets of the predictor.

Definition 2.6 ($(\mathcal{G},\mathcal{H},\varepsilon)$-Multiaccuracy)**.** *A deterministic predictor $p^*:\mathcal{X}\to[0,1]$ is $(\mathcal{G},\mathcal{H},\varepsilon)$-multiaccurate if for all $w\in[0,1]$, $h\in\mathcal{H}$, and $g\in\mathcal{G}$,*

$$\left|\mathop{\mathbb{E}}_{(x,y)\sim D}\left[(y-p^*(x))\cdot\mathbb{1}[h(x)\le w]\mid g(x)=1\right]\right|\le\varepsilon\cdot\sqrt{P_g^{-1}}.$$

Randomized variants of Definition 2.5 and 2.6 are deferred to Appendix A.

3 Reducing Panprediction to Step Calibration

We now show that panprediction admits a clear reduction to step calibration by carefully extending the language of outcome indistinguishability, which was originally developed for omniprediction [8, 16]. In this section, we assume $\widehat{\mathcal{Y}}=[0,1]$. The setting with $\widehat{\mathcal{Y}}=\{0,1\}$ is strictly easier and follows readily from our arguments.

3.1 Loss Outcome Indistinguishability

The Loss OI framework [8] is useful for deducing sufficient conditions for panprediction guarantees to hold. For *any* deterministic predictor $p^*:\mathcal{X}\to[0,1]$, by the definition of the post-processing function k_ℓ, the following inequality holds: for all $\ell\in\mathcal{L}$, $g\in\mathcal{G}$, $h\in\mathcal{H}$,

$$\mathop{\mathbb{E}}_{\substack{x\sim D\\ \widetilde{y}\sim\mathrm{Ber}(p^*(x))}}\left[\ell(k_\ell(p^*(x)),\widetilde{y})\mid g(x)=1\right]\le\mathop{\mathbb{E}}_{\substack{x\sim D\\ \widetilde{y}\sim\mathrm{Ber}(p^*(x))}}\left[\ell(h(x),\widetilde{y})\mid g(x)=1\right]. \tag{2}$$

To derive the deterministic $(\mathcal{L},\mathcal{G},\mathcal{H},O(\varepsilon))$-panpredictor guarantee from here, it suffices to show that p^* approximates the true Bayes predictor well in some sense. Concretely, for all $\ell\in\mathcal{L}$, $g\in\mathcal{G}$, and $h\in\mathcal{H}$, we want the following inequalities to hold:

$$\left|\mathop{\mathbb{E}}_{(x,y)\sim D}\left[\ell(k_\ell(p^*(x)),y)\mid g(x)=1\right]-\mathop{\mathbb{E}}_{\substack{x\sim D\\ \widetilde{y}\sim\mathrm{Ber}(p^*(x))}}\left[\ell(k_\ell(p^*(x)),\widetilde{y})\mid g(x)=1\right]\right|\le\frac{O(\varepsilon)}{\sqrt{P_g}},$$

$$\left|\mathop{\mathbb{E}}_{(x,y)\sim D}\left[\ell(h(x),y)\mid g(x)=1\right]-\mathop{\mathbb{E}}_{\substack{x\sim D\\ \widetilde{y}\sim\mathrm{Ber}(p^*(x))}}\left[\ell(h(x),\widetilde{y})\mid g(x)=1\right]\right|\le\frac{O(\varepsilon)}{\sqrt{P_g}}.$$

The first condition is called Decision OI, and the second condition, Hypothesis OI. To characterize each in terms of familiar terms, take the following definition.

Definition 3.1 (Discrete Derivative)**.** *Given a loss $\ell\in\mathcal{L}$, define the discrete derivative*

$$\Delta\ell(p)=\ell(p,1)-\ell(p,0).$$

Given a loss class $\mathcal{L}$, let $\Delta\mathcal{L}=\{\Delta\ell\mid\ell\in\mathcal{L}\}$ be the class of corresponding discrete derivatives.

Intuitively, $\Delta\ell(p)$ quantifies how sharply ℓ distinguishes between $y = 1$ and $y = 0$ given a prediction $p \in [0, 1]$. Then Decision and Hypothesis OI can be respectively written as

$$\left|\mathop{\mathbb{E}}_{D}\left[(y - p^*(x)) \cdot \Delta\ell(k_\ell(p^*(x))) \mid g(x) = 1\right]\right| \leq \frac{O(\varepsilon)}{\sqrt{P_g}},$$
$$\left|\mathop{\mathbb{E}}_{D}\left[(y - p^*(x)) \cdot \Delta\ell(h(x)) \mid g(x) = 1\right]\right| \leq \frac{O(\varepsilon)}{\sqrt{P_g}}.$$

This reformulation suggests that Decision OI can be deduced from a certain calibration condition, and Hypothesis OI, from a certain multiaccuracy condition.

3.2 Reducing Deterministic Panprediction to Determnistic Step Calibration

Now we show that deterministic $(\mathcal{G}, \mathcal{H}, \varepsilon)$-step calibration implies both sufficient conditions found above.

Theorem 3.2. *If a deterministic predictor $p^* : \mathcal{X} \to [0, 1]$ is $(\mathcal{G}, \mathcal{H}, \varepsilon)$-step calibrated, then p^* is a deterministic $(\mathcal{L}, \mathcal{G}, \mathcal{H}, O(\varepsilon))$-panpredictor.*

We prove Theorem 3.2 by decomposing the step calibration guarantee into a calibration and a multiaccuracy guarantee (Lemma 3.3), then showing that the calibration guarantee implies Decision OI (Lemma 3.5), and that the multiaccuracy guarantee implies Hypothesis OI (Lemma 3.6).

Lemma 3.3. *If a deterministic predictor $p^* : \mathcal{X} \to [0, 1]$ is $(\mathcal{G}, \mathcal{H}, \varepsilon)$-step calibrated, then p^* is $(\mathcal{G}, \emptyset, \varepsilon)$-step calibrated and $(\mathcal{G}, \mathcal{H}, \varepsilon)$-multiaccurate.*

The proof of Lemma 3.3 is straightforward and is deferred to Appendix B. To prove the remaining two lemmas, it is useful to invoke the following technical result on approximating (the discrete derivative of) bounded variation functions with step functions.

Lemma 3.4. *Suppose $\ell \in \mathcal{L}_{\mathrm{BV}}$. Then for any predictor $f : \mathcal{X} \to [0, 1]$ and $g \in \mathcal{G}$,*

$$\left|\mathop{\mathbb{E}}_{(x,y)\sim D}\left[(y - p^*(x)) \cdot \Delta\ell(f(x)) \mid g(x) = 1\right]\right| \leq 9 \sup_{v\in[0,1]}\left|\mathop{\mathbb{E}}_{D}\left[(y - p^*(x)) \cdot \mathbb{1}[f(x) \leq v] \mid g(x) = 1\right]\right| + \varepsilon \cdot \sqrt{P_g^{-1}}.$$

Lemma 3.4 can be invoked with $f = p^*$ or $f = h$, for any $h \in \mathcal{H}$, and helps upper bound expressions arising from Decision and Hypothesis OI in terms of step calibration error. Its proof is deferred to Appendix B.

Lemma 3.5. *If a deterministic predictor p^* is $(\mathcal{G}, \emptyset, \varepsilon)$-step calibrated, then for all $g \in \mathcal{G}$,*

$$\left|\mathop{\mathbb{E}}_{D}\left[(y - p^*(x)) \cdot \Delta\ell(k_\ell(p^*(x))) \mid g(x) = 1\right]\right| \leq \frac{O(\varepsilon)}{\sqrt{P_g}}.$$

Proof of Lemma 3.5. For any loss $\ell : [0, 1] \times \mathcal{Y} \to [-1, 1]$ and constants $p, q \in [0, 1]$, by definition of k_ℓ,

$$\mathop{\mathbb{E}}_{y\sim\mathrm{Ber}(p)}\left[\ell(k_\ell(p), y)\right] \leq \mathop{\mathbb{E}}_{y\sim\mathrm{Ber}(p)}\left[\ell(k_\ell(q), y)\right].$$

Hence, $\ell(k_\ell(\cdot), y) : [0, 1] \to [-1, 1]$ is a proper scoring rule and is contained in $\mathcal{L}_{\mathrm{BV}}$. Then by invoking Lemma 3.4 with $f = p^*$, we have that

$$\begin{aligned}\left|\mathop{\mathbb{E}}_{D}\left[(y - p^*(x)) \cdot \Delta\ell(k_\ell(p^*(x))) \mid g(x) = 1\right]\right| &\leq 9 \sup_{v\in[0,1]}\left|\mathop{\mathbb{E}}_{D}\left[(y - p^*(x)) \cdot \mathbb{1}[p^*(x) \leq v] \mid g(x) = 1\right]\right| + \varepsilon \cdot \sqrt{P_g^{-1}} \\ &\leq O(\varepsilon) \cdot \sqrt{P_g^{-1}},\end{aligned}$$

where the last inequality follows from the $(\mathcal{G}, \emptyset, \varepsilon)$-step calibration guarantee. □

Lemma 3.6. *If a deterministic predictor p^* is $(\mathcal{G}, \mathcal{H}, \varepsilon)$-multiaccurate, then for all $h \in \mathcal{H}$, $g \in \mathcal{G}$,*

$$\left|\mathop{\mathbb{E}}_{D}\left[(y - p^*(x)) \cdot \Delta\ell(h(x)) \mid g(x) = 1\right]\right| \leq \frac{O(\varepsilon)}{\sqrt{P_g}}.$$

Proof of Lemma 3.6. By assumption, $\ell \in \mathcal{L}_{\mathrm{BV}}$. Then by applying Lemma 3.4 with $f = h$ for any $\in \mathcal{H}$,

$$\begin{aligned}\left|\mathop{\mathbb{E}}_{D}\left[(y - p^*(x)) \cdot \Delta\ell(h(x)) \mid g(x) = 1\right]\right| &\leq 9 \sup_{w \in [0,1]} \left|\mathop{\mathbb{E}}_{D}\left[(y - p^*(x)) \cdot \mathbb{1}[h(x) \leq w] \mid g(x) = 1\right]\right| + \varepsilon \cdot \sqrt{P_g^{-1}} \\ &\leq O(\varepsilon) \cdot \sqrt{P_g^{-1}},\end{aligned}$$

where the last inequality follows from the $(\mathcal{G}, \mathcal{H}, \varepsilon)$-multiaccuracy guarantee. □

Proof of Theorem 3.2. Starting with Equation 2 and invoking Lemmas 3.3, 3.5, and 3.6 to swap terms on both sides of the inequality, at the cost of additive error $O(\varepsilon \cdot \sqrt{P_g^{-1}})$, immediately yields the result. □

3.3 Extensions for Randomized Panprediction

By extending the Loss OI machinery to work with the definitions randomized step calibration and randomized multiaccuracy, we can also show that randomized step calibration implies randomized panprediction. The statement and proof of the randomized variants of results are deferred to Appendix B.

4 Step Calibration Algorithms via Multi-objective Learning

In this section, we derive algorithms for deterministic and randomized step calibration using the multi-objective learning framework. Our deterministic step calibration algorithm improves on the best known sample complexity guarantee of deterministic omniprediction by a factor of $1/\varepsilon$. Moreover, the sample complexity of our algorithms match the best known sample complexity guarantees of deterministic and randomized multi-group learning, up to logarithmic factors, demonstrating that panprediction guarantees can be attained "for free" whenever multi-group guarantees are sought.

4.1 Multi-objective Learning

Fix a distribution D, a hypothesis class $\mathcal{F}$, and a set of bounded objectives $\mathcal{O} = \{\ell : \mathcal{F} \times \mathcal{X} \times \mathcal{Y} \to [a, b]\}$. The goal of the $(D, \mathcal{O}, \mathcal{H})$-multi-objective learning framework, formalized by Haghtalab, Jordan, and Zhao [4], is finding a deterministic predictor $f^* \in \mathcal{F}$ such that

$$\max_{\ell \in \mathcal{O}} \mathop{\mathbb{E}}_{(x,y)\sim D}\left[\ell(f^*(x), y)\right] \leq \min_{f \in \mathcal{F}} \max_{\ell \in \mathcal{O}} \mathop{\mathbb{E}}_{(x,y)\sim D}\left[\ell(f(x), y)\right] + \varepsilon. \tag{3}$$

A randomized predictor $\mathbf{h}^* \in \Delta(\mathcal{H})$ can also be found, with the expectation on the left hand side of Equation 3 holding on average over the draw of $h^* \sim \mathbf{h}^*$. Below, we show that the $(\mathcal{G}, \mathcal{H}, \varepsilon)$-step calibration problem can be formulated as a multi-objective problem. In the rest of the paper, we assume that the group probabilities $\{P_g\}_{g \in \mathcal{G}}$ are known constants, and denote $\gamma = \min_{g \in \mathcal{G}} P_g$. This is a standard assumption in, e.g., multi-group learning [1].

Theorem 4.1. *Fix a distribution D, the competitor hypothesis class $\mathcal{H}$, and a set of groups $\mathcal{G}$. For each $\sigma \in \{\pm 1\}$, $v, w \in [0, 1]$, $h \in \mathcal{H}$, and $g \in \mathcal{G}$, define the objective $\ell_{\sigma,v,w,h,g} : \mathcal{P} \times \mathcal{X} \times \mathcal{Y} \to [-1/\sqrt{\gamma}, 1/\sqrt{\gamma}]$ as*

$$\ell_{\sigma,v,w,h,g}(p, (x, y)) = \frac{\sigma \cdot (y - p(x))}{\sqrt{P_g}} \cdot \mathbb{1}[p(x) \leq v, h(x) \leq w, g(x) = 1].$$

Let $\mathcal{O}_{\mathrm{sc}} = \{\ell_{\sigma,v,w,h,g}\}$ be the set of all such objectives. If $p^* \in \mathcal{P}$ is an ε-optimal solution to the $(D, \mathcal{O}_{\mathrm{sc}}, \mathcal{P})$-multi-objective learning problem, then p^* is $(\mathcal{G}, \mathcal{H}, O(\varepsilon))$-step calibrated.

Algorithm 1 Deterministic Step Calibration

Require: $\varepsilon, \delta \in (0,1)$, $T \in \mathbb{N}$, $c \in [0,1]$, $C \in \mathbb{N}$, sampling access to D, best response oracle $\mathcal{A}$

Initialize Hedge iterate $p^{(1)} \leftarrow [\frac{1}{2}, \frac{1}{2}]^{\mathcal{X}}$

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

Update

$$\ell^{(t)} \leftarrow \mathcal{A}\big(p^{(t)}, \widehat{O}_{\mathrm{sc}}^{\gamma}, c\varepsilon\sqrt{\gamma}\big),$$

where $\widehat{O}_{\mathrm{sc}}^{\gamma}$ is a modification of O_{sc} specified in Appendix C

For each $x \in \mathcal{X}$, update

$$p^{(t+1)}(x) \leftarrow \mathrm{Hedge}\big(c_x^{(1)}, \ldots, c_x^{(t)}\big),$$

where $c_x^{(t)}$ is a modification of $\ell^{(t)}$ specified in Appendix C

end for

Draw $m \leftarrow \frac{C\log(T/\delta)}{\varepsilon^2}$ samples $z^{(1)}, \ldots, z^{(m)} \sim D$

$t^* \leftarrow \arg\min_{t\in[T]} \sum_{(x,y)\in z^{(1:m)}} \ell^{(t)}\big(p^{(t)}, (x,y)\big)$

return $p^{(t^*)}$

Proof of Theorem 4.1. By assumption that p^* is an ε-optimal solution to the $(\mathcal{D}, O_{\mathrm{sc}}, \mathcal{P})$-multi-objective learning problem, we have that

$$\begin{aligned}
\max_{\sigma,v,w,h,g} \mathop{\mathbb{E}}_{(x,y)\sim D}\left[\ell_{\sigma,v,w,h,g}(p^*, (x,y))\right] &= \max_{v,w,h,g} \frac{\left|\mathbb{E}\left[(y - p^*(x)) \cdot \mathbb{1}[p^*(x) \le v, h(x) \le w, g(x) = 1]\right]\right|}{\sqrt{P_g}} \\
&\le \min_{p\in\mathcal{P}} \max_{v,w,h,g} \mathop{\mathbb{E}}_{(x,y)\sim D}\left[\ell_{\sigma,v,w,h,g}(p, (x,y))\right] + \varepsilon \\
&= \varepsilon,
\end{aligned}$$

where the last equality holds because the expectation is zero when p is the Bayes predictor. Since

$$\left|\mathop{\mathbb{E}}_{D}\left[(y - p^*(x)) \cdot \mathbb{1}[p^*(x) \le v, h(x) \le w, g(x) = 1]\right]\right| = P_g \left|\mathop{\mathbb{E}}_{D}\left[(y - p^*(x))\mathbb{1}[p^*(x) \le v, h(x) \le w] \mid g(x) = 1\right]\right|$$

it follows that for all $v, w \in [0,1]$, $h \in \mathcal{H}$, and $g \in \mathcal{G}$,

$$\left|\mathop{\mathbb{E}}_{D}\left[(y - p^*(x)) \cdot \mathbb{1}[p^*(x) \le v, h(x) \le w] \mid g(x) = 1\right]\right| \le \varepsilon \cdot \sqrt{P_g^{-1}}. \qquad \square$$

Theorem 4.1 establishes that step calibrated predictors correspond to the approximate equilibria of a particular two-player, zero-sum game. Following Haghtalab, Jordan, and Zhao [4], we construct step calibration algorithms that use tools from online learning to solve this game.

A technical detail that must be handled before designing algorithms is the fact that O_{sc} can have infinite cardinality. In order to make algorithms tractable, we leverage Assumptions 1, 2, and 3 to construct a finite cover, denoted by $\widehat{O}_{\mathrm{sc}}$. This covering argument is standard in statistical learning theory, and details are deferred to Appendix C.

4.2 Deterministic Step Calibration

Below, we build intuition about our algorithms and state guarantees. Technical details are deferred to Appendix C.

Per Haghtalab, Jordan, and Zhao [4], deterministic solutions to the multiobjective problem can be found via "no-regret vs best-response" dynamics between a fictitious player, who constructs a sequence of predictors using a no-regret algorithm, and a fictitious adversary, who constructs a sequence of objectives using a best-response oracle. In our setting, the player is instantiated with a copy of the Hedge algorithm [19] per point $x \in \mathcal{X}$. The player's prediction at $x \in \mathcal{X}$ at iteration t is denoted by $p^{(t)}(x)$ and evolves according to the Hedge update rule, which takes (a light modification of) the historical loss functions $\ell^{(1)}, \ldots, \ell^{(t-1)}$ as input. In turn, at each iteration, the adversary selects

Algorithm 2 Randomized Step Calibration

Require: $\varepsilon, \delta \in (0,1)$, $T \in \mathbb{N}$, $c \in [0,1]$, $C \in \mathbb{N}$, sampling access to D

1: Initialize Hedge iterates $p^{(1)} \leftarrow [\frac{1}{2}, \frac{1}{2}]^{\mathcal{X}}$ and $q^{(1)} = \text{Uniform}(\widehat{O}_{\text{sc}}^{\gamma})$.
2: **for** $t = 1, \ldots, T$ **do**
3: Sample objective $\ell^{(t)} \sim q^{(t)}$ and data point $(x^{(t)}, y^{(t)}) \sim D$
4: For each $x \in \mathcal{X}$, update

$$p^{(t+1)}(x) \leftarrow \text{Hedge}\big(c_x^{(1)}, \ldots, c_x^{(t)}\big),$$

where $c_x^{(t)}(\widehat{y})$ is a modification of $\ell^{(t)}$ as specified in Appendix C
5: Let $q^{(t+1)} = \text{Hedge}(c_{\text{adv}}^{(1)}, \ldots, c_{\text{adv}}^{(t)})$ where $c_{\text{adv}}^{(t)} = 1 - \ell_{(\cdot)}(p^{(t)}, (x^{(t)}, y^{(t)}))$
6: **end for**
7: **return** $\boldsymbol{p}^*$, a uniform distribution over $p^{(1)}, \ldots, p^{(T)}$

a loss $\ell^{(t)} \in O_{\text{sc}}$ that approximately maximizes penalty against the player's Hedge algorithm. That is, given predictor $p^{(t)}$, objectives O_{sc}, and error tolerance ε', $\ell^{(t)} \in O_{\text{sc}}$ is chosen such that

$$\mathop{\mathbb{E}}_{(x,y)\sim D}\left[\ell(p^{(t)}, (x,y))\right] \geq \max_{\ell^* \in O_{\text{sc}}} \mathop{\mathbb{E}}_{(x,y)\sim D}\left[\ell^*(p^{(t)}, (x,y))\right] - \varepsilon'.$$

The above computation is abstracted into a best-response oracle $\mathcal{A}$. This oracle requires samples from D to approximate the expectation, and uses techniques from adaptive data analysis [20] in order to reduce sample complexity overhead. Putting everything together in Algorithm 1, we have the following result.

Theorem 4.2. *Fix $\varepsilon > 0$ and $\delta \in (0,1)$. Let d_H be the appropriate combinatorial dimension of H and d_G the VC dimension of $\mathcal{G}$. Then with probability at least $1-\delta$ and at most $\widetilde{O}\left(\varepsilon^{-3} \cdot (d_H + d_G) \cdot \log(1/\delta)\right)$ samples from D, Algorithm 1 returns a deterministic predictor p^* that is $(\mathcal{G}, \mathcal{H}, \varepsilon)$-step calibrated.*

The proof of Theorem 4.2 is deferred to Appendix C. Notably, the $\widetilde{O}\left(\varepsilon^{-3} \cdot d_H\right)$ sample complexity guarantee improves the best known sample complexity guarantee of $\widetilde{O}\left(\varepsilon^{-4} \cdot d_H\right)$ for deterministic omniprediction [2] by a factor of ε^{-1}. Moreover, it matches the best known sample complexity guarantee for deterministic multi-group learning [1] up to logarithmic factors.

4.3 Randomized Step Calibration

Per Haghtalab, Jordan, and Zhao [4], randomized solutions to the multi-objective problem can be found via "no-regret vs no-regret" dynamics. In our setting, this corresponds to using instantiations of Hedge to construct the predictor $p^{(t)}$ and another instantiation of Hedge to select the losses $\ell^{(t)}$ on which the predictor is updated. An advantage of this dynamic, over the no-regret vs best-response dynamics considered for deterministic multi-objective learning, is that it only incurs a sample complexity of $\widetilde{O}(1/\varepsilon^2)$.

Theorem 4.3. *Fix $\varepsilon > 0$ and $\delta \in (0,1)$. Let d_H be the appropriate combinatorial dimension of H and d_G the VC dimension of $\mathcal{G}$. Then with probability at least $1-\delta$ and at most $\widetilde{O}\left(\varepsilon^{-2} \cdot (d_H + d_G) \cdot \log(1/\delta)\right)$ samples from D, Algorithm 2 returns a randomized predictor p^* that is $(\mathcal{G}, \mathcal{H}, \varepsilon)$-step calibrated.*

The statement of the algorithm and the proof of Theorem 4.3 are deferred to Appendix C. Importantly, the above result matches the best known sample complexity guarantees of randomized omniprediction [2] and multi-group learning [1] up to logarithmic factors. Furthermore, the $\widetilde{O}(1/\varepsilon^2)$ rate is minimax optimal for minimizing one loss over one task. This establishes that under mild assumptions—such as allowing for randomized predictors—simultaneously minimizing infinitely many losses on infinitely many tasks can be as statistically easy as minimizing one loss on one task. This strengthens the conclusion of Okoroafor, Kleinberg, and Kim [2], who showed that simultaneously minimizing infinitely many losses can be as statistically easy as minimizing one loss.

5 Connections to Omniprediction and Multi-group Learning

In this section, we make precise how panprediction generalizes the problems of *omniprediction* [6, 2] and *multi-group learning* [7, 1], which respectively study predictions that generalize to many losses or many tasks, but not both.

5.1 Omniprediction

The goal of omniprediction is to use samples from D to construct a predictor that, for any post-hoc choice of loss $\ell \in \mathcal{L}$, can be post-processed to compete with the best hypothesis $h \in \mathcal{H}$, as measured by the loss ℓ.

Definition 5.1 (Deterministic Omniprediction)**.** *Given loss class $\mathcal{L}$ and hypothesis class $\mathcal{H}$, a deterministic predictor $p^* : \mathcal{X} \to [0,1]$ is a $(\mathcal{L}, \mathcal{H}, \varepsilon)$-omnipredictor if for each $\ell \in \mathcal{L}$,*

$$\mathop{\mathbb{E}}_{(x,y)\sim D}[\ell(k_\ell(p^*(x)), y)] \le \min_{h\in\mathcal{H}} \mathop{\mathbb{E}}_{(x,y)\sim D}[\ell(h(x), y)] + \varepsilon.$$

The post-processing k_ℓ is defined as in Equation 1.

By symmetry in the definitions, a deterministic panpredictor with the trivial group over the entire domain, $\mathcal{G} = \{\mathcal{X}\}$, is a deterministic omnipredictor.

Proposition 5.2. *Let $\mathcal{L}$ be a loss class and $\mathcal{H}$ a hypothesis class. If the deterministic predictor $p^* : \mathcal{X} \to [0,1]$ is a $(\mathcal{L}, \{\mathcal{X}\}, \mathcal{H}, \varepsilon)$-panpredictor, then p^* is a deterministic $(\mathcal{L}, \mathcal{H}, \varepsilon)$-omnipredictor.*

The definition of randomized omniprediction, and a corresponding result that a randomized panpredictor with $\mathcal{G} = \{\mathcal{X}\}$ is a randomized omnipredictor, are deferred to Appendix D.

5.2 Multi-group Learning

The goal of multi-group learning is to use samples from D to construct a predictor that, on each group $g \in \mathcal{G}$, competes with best hypothesis $h \in \mathcal{H}$ for that group, as measured by the zero-one loss.

Definition 5.3 (Deterministic MG Learning)**.** *Given the zero-one loss ℓ, set of groups $\mathcal{G}$, and binary hypothesis class $\mathcal{H}$, a deterministic hypothesis $h^* : \mathcal{X} \to \{0,1\}$ is a $(\ell, \mathcal{G}, \mathcal{H}, \varepsilon)$-multi-group learner if for each group $g \in \mathcal{G}$,*

$$\mathop{\mathbb{E}}_{(x,y)\sim D}[\ell(h^*(x), y) \mid g(x) = 1] \le \min_{h\in\mathcal{H}} \mathop{\mathbb{E}}_{(x,y)\sim D}[\ell(h(x), y) \mid g(x) = 1] + \varepsilon \cdot \sqrt{P_g^{-1}}.$$

Panprediction sits upstream from multi-group learning, in the sense that the post-processing of an appropriate panpredictor is a multi-group learning solution.

Proposition 5.4. *Fix the zero-one loss ℓ, groups $\mathcal{G}$, and binary hypothesis class $\mathcal{H}$. If a deterministic predictor $p^* : \mathcal{X} \to [0,1]$ is a $(\{\ell\}, \mathcal{G}, \mathcal{H}, \varepsilon)$-panpredictor, then the binary classifier $h(x) = \mathbb{1}[p^*(x) \ge 0.5]$ is a deterministic $(\ell, \mathcal{G}, \mathcal{H}, \varepsilon)$-multi-group learner.*

The definition of randomized multi-group learning, and a corresponding result that the post-processing of an appropriate randomized panpredictor is a randomized multi-group learning solution, are deferred to Appendix D.

6 Discussion

We formalize the problem of panprediction in the binary prediction setting and design statistically efficient panpredictors via a reduction to step calibration. Important questions for future work include:

- Is the ε^{-1} sample complexity gap between deterministic and randomized panprediction fundamental? This question also remains open in both omniprediction [2] and multi-group learning [1].
- What is the sample complexity of oracle-efficient panprediction, and is there a gap compared to the information-theoretic bounds we prove? Similar questions on statistical-computational gaps are only just being explored in omniprediction [2] and multi-group learning [10].
- Can multi-class panprediction be achieved with the same sample complexity as direct multi-class loss minimization? This question also remains open in omniprediction.

7 Acknowledgements

NH was supported in part by the National Science Foundation under grant CCF-2145898, by the Office of Naval Research under grant N00014-24-1-2159, a Google Research Scholar Award, an Alfred P. Sloan fellowship, and a Schmidt Science AI2050 fellowship. DH was supported by the Office of Navel Research under grant N00014-24-1-2700. This work was partially done while the authors were visitors at the Simons Institute for the Theory of Computing.

References

[1] Christopher J. Tosh and Daniel Hsu. Simple and near-optimal algorithms for hidden stratification and multi-group learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 21633–21657. PMLR, 2022.

[2] Princewill Okoroafor, Robert Kleinberg, and Michael P. Kim. Near-Optimal Algorithms for Omniprediction, January 2025. arXiv:2501.17205 [stat].

[3] Mingda Qiao and Eric Zhao. Truthfulness of Decision-Theoretic Calibration Measures, March 2025.

[4] Nika Haghtalab, Michael I. Jordan, and Eric Zhao. A unifying perspective on multi-calibration: Game dynamics for multi-objective learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[5] Zihan Zhang, Wenhao Zhan, Yuxin Chen, Simon S. Du, and Jason D. Lee. Optimal multi-distribution learning, 2023.

[6] Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In Mark Braverman, editor, *13th Innovations in Theoretical Computer Science Conference, ITCS 2022, January 31 - February 3, 2022, Berkeley, CA, USA*, volume 215 of *LIPIcs*, pages 79:1–79:21. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022.

[7] Guy N. Rothblum and Gal Yona. Multi-group agnostic PAC learnability. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 9107–9115. PMLR, 2021.

[8] Parikshit Gopalan, Lunjia Hu, Michael P. Kim, Omer Reingold, and Udi Wieder. Loss Minimization through the Lens of Outcome Indistinguishability, December 2022. arXiv:2210.08649 [cs].

[9] Avrim Blum and Thodoris Lykouris. Advancing Subgroup Fairness via Sleeping Experts. *LIPIcs, Volume 151, ITCS 2020*, 151:55:1–55:24, 2020. Artwork Size: 24 pages, 580511 bytes ISBN: 9783959771344 Medium: application/pdf Publisher: Schloss Dagstuhl – Leibniz-Zentrum für Informatik Version Number: 1.0.

[10] Samuel Deng, Daniel Hsu, and Jingwen Liu. Group-wise oracle-efficient algorithms for online multi-group learning. *Advances in Neural Information Processing Systems*, 37:39462–39500, December 2024.

[11] Samuel Deng and Daniel Hsu. Multi-group Learning for Hierarchical Groups. In *Proceedings of the 41st International Conference on Machine Learning*, pages 10440–10487. PMLR, July 2024. ISSN: 2640-3498.

[12] Saba Ahmadi, Avrim Blum, Omar Montasser, and Kevin M. Stangl. Agnostic Multi-Robust Learning using ERM. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pages 2242–2250. PMLR, April 2024. ISSN: 2640-3498.

[13] A Philip Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982. Publisher: Taylor & Francis.

[14] Dean P. Foster and Rakesh V. Vohra. Asymptotic Calibration. *Biometrika*, 85(2):379–390, 1998. Publisher: [Oxford University Press, Biometrika Trust].

[15] Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-Identifiable) Masses. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1939–1948. PMLR, July 2018. ISSN: 2640-3498.

[16] Cynthia Dwork, Michael P. Kim, Omer Reingold, Guy N. Rothblum, and Gal Yona. Outcome indistinguishability. In Samir Khuller and Virginia Vassilevska Williams, editors, *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021*, pages 1095–1108. ACM, 2021.

[17] Robert Kleinberg, Renato Paes Leme, Jon Schneider, and Yifeng Teng. U-calibration: Forecasting for an unknown agent. In *Conference on Learning Theory (COLT)*, pages 5143–5145, 2023.

[18] Dean P. Foster and Rakesh V. Vohra. Calibrated Learning and Correlated Equilibrium. *Games and Economic Behavior*, 21(1):40–55, October 1997.

[19] Yoav Freund and Robert E Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1):119–139, August 1997.

[20] Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic Stability for Adaptive Data Analysis, November 2015. arXiv:1511.02513 [cs].

[21] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust Stochastic Approximation Approach to Stochastic Programming. *SIAM Journal on Optimization*, 19(4):1574–1609, January 2009. Publisher: Society for Industrial and Applied Mathematics.

[22] Frank McSherry and Kunal Talwar. Mechanism Design via Differential Privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103, October 2007. ISSN: 0272-5428.

[23] Gyora M. Benedek and Alon Itai. Learnability with respect to fixed distributions. *Theoretical Computer Science*, 86(2):377–389, September 1991.

[24] Jessica Dai, Nika Haghtalab, and Eric Zhao. Learning With Multi-Group Guarantees For Clusterable Subpopulations, December 2024.

A Deferred Definitions from Section 2

A.1 Quantizing Real-valued Hypotheses

We say a hypothesis is real-valued if $h : \mathcal{X} \to [0,1]$. For tractability, real-valued hypotheses are quantized so that they return predictions that are on the λ-net of the unit interval $[0,1]$, denoted by I_λ. Importantly, λ is a quantization parameter that the learner selects with full knowledge of the target error tolerance ε. Formally, given a class of real-valued hypotheses $\mathcal{H}$, we can define the quantized class

$$\mathcal{H}_\lambda = \{h_\lambda : \mathcal{X} \to I_\lambda \mid h_\lambda(x) = Q_\lambda(h(x)),\ h \in \mathcal{H}\},$$

where the "nearest neighbor" quantization map $Q_\lambda : [0,1] \to I_\lambda$ is defined as

$$Q_\lambda(p) = \arg\min_{p' \in I_\lambda} |p - p'|\,.$$

In all that follows, it suffices to set $\lambda = \Theta(\varepsilon)$.

A.2 Combinatorial Dimensions

We recall the definitions of three combinatorial dimensions we consider for the classes $\mathcal{H}$ and $\mathcal{G}$. For binary hypotheses, we consider the VC dimension. For real-valued hypotheses, we consider the pseudo-dimension. Roughly, hypothesis classes with infinite cardinality but bounded combinatorial dimension admit finite covers with cardinality that is exponential in the combinatorial dimension.

Definition A.1 (VC Dimension)**.** *Suppose $\mathcal{H}$ is a class of binary hypotheses, i.e., $h : \mathcal{X} \to \{0,1\}$ for all $h \in \mathcal{H}$. A finite set $S = \{x_1, \ldots, x_m\} \subset \mathcal{X}$ is shattered by $\mathcal{H}$ if for every labeling $y \in \{0,1\}^m$, there exists $h \in \mathcal{H}$ with $h(x_i) = y_i$ for all $i \in [m]$. The VC dimension of $\mathcal{H}$, denoted by $\mathrm{VC}(\mathcal{H})$, is the size of the largest set S that is shattered by $\mathcal{H}$.*

Definition A.2 (Fat-Shattering Dimension)**.** *Suppose $\mathcal{H}$ is a class of real-valued hypotheses, i.e., $h : \mathcal{X} \to [0,1]$ for all $h \in \mathcal{H}$. Fix $\varepsilon' > 0$. A finite set $S = \{x_1, \ldots, x_m\} \subset \mathcal{X}$ is ε'-shattered by $\mathcal{H}$ if there exist threshold values $w_1, \ldots, w_m \in [0,1]$ such that for every subset $S' \subseteq S$, there exists $h_{S'} \in \mathcal{H}$ with*

$$\begin{aligned} h_{S'}(x_i) &\geq w_i + \varepsilon' \quad \text{for} \quad x_i \in S', \\ h_{S'}(x_i) &\leq w_i - \varepsilon' \quad \text{for} \quad x_i \notin S'. \end{aligned}$$

The fat-shattering dimension of $\mathcal{H}$ at scale ε', denoted by $\mathrm{fat}_{\varepsilon'}(\mathcal{H})$, is the size of the largest set S that is ε'-shattered by $\mathcal{H}$.

Definition A.3 (Pseudo-dimension)**.** *Suppose $\mathcal{H}$ is a class of real-valued hypotheses. The pseudo-dimension of $\mathcal{H}$, denoted by $\mathrm{Pdim}(\mathcal{H})$, is defined as $\lim_{\varepsilon' \downarrow 0} \mathrm{fat}_{\varepsilon'}(\mathcal{H})$.*

A.3 Definitions with Randomized Predictors

We define the variants of step calibration and multi-accuracy that allow for randomized predictors $\mathbf{p}^* \in \Delta(\mathcal{P})$. These definitions follow straightforwardly from Definitions2.5 and 2.6, by taking an additional expectation over the draw of $p^* \sim \mathbf{p}^*$.

Definition A.4 (Randomized $(\mathcal{G}, \mathcal{H}, \varepsilon)$-Step Calibration)**.** *A randomized predictor $\boldsymbol{p}^* \in \Delta(\mathcal{P})$ is $(\mathcal{G}, \mathcal{H}, \varepsilon)$-step calibrated if for all $v \in [0,1]$, $w \in [0,1]$, $h \in \mathcal{H}$, and $g \in \mathcal{G}$,*

$$\left| \mathop{\mathbb{E}}_{(x,y)\sim D,\, p^* \sim \boldsymbol{p}^*} [(y - p^*(x)) \cdot \mathbb{1}[p^*(x) \leq v, h(x) \leq w] \mid g(x) = 1] \right| \leq \varepsilon \cdot \sqrt{P_g^{-1}}.$$

Definition A.5 (Randomized $(\mathcal{G}, \mathcal{H}, \varepsilon)$-Multiaccuracy)**.** *A randomized predictor $\boldsymbol{p}^* \in \Delta(\mathcal{P})$ is $(\mathcal{G}, \mathcal{H}, \varepsilon)$-multiaccurate if for all $w \in [0,1]$, $h \in \mathcal{H}$, and $g \in \mathcal{G}$,*

$$\left| \mathop{\mathbb{E}}_{(x,y)\sim D,\, p^* \sim \boldsymbol{p}^*} [(y - p^*(x)) \cdot \mathbb{1}[h(x) \leq w] \mid g(x) = 1] \right| \leq \varepsilon \cdot \sqrt{P_g^{-1}}.$$

B Deferred Results and Proofs from Section 3

B.1 Loss Outcome Indistinguishability

For completeness, we show that the two characterizations of Decision and Hypothesis OI that we gave in Section 3 are equivalent. This derivation is a straightforward extension of the machinery that Gopalan et al. [8] developed for the omniprediction setting.

Suppose the predictor p^* is deterministic. We can deconstruct the Decision OI condition as follows.

$$\begin{aligned}
&\left| \mathop{\mathbb{E}}_{(x,y)\sim D} [\ell(k_\ell(p^*(x)), y) \mid g(x) = 1] - \mathop{\mathbb{E}}_{\substack{x\sim D \\ \widetilde{y}\sim \mathrm{Ber}(p^*(x))}} [\ell(k_\ell(p^*(x)), \widetilde{y}) \mid g(x) = 1] \right| \\
&= \left| \mathop{\mathbb{E}}_{(x,y)\sim D} [y \cdot \ell(k_\ell(p^*(x)), 1) + (1-y) \cdot \ell(k_\ell(p^*(x)), 0) \mid g(x) = 1] \right. \\
&\qquad \left. - \mathop{\mathbb{E}}_{\substack{x\sim D \\ \widetilde{y}\sim \mathrm{Ber}(p^*(x))}} [\widetilde{y} \cdot \ell(k_\ell(p^*(x)), 1) + (1-\widetilde{y}) \cdot \ell(k_\ell(p^*(x)), 0) \mid g(x) = 1] \right| \\
&= \left| \mathop{\mathbb{E}}_{(x,y)\sim D} [y \cdot \Delta\ell(k_\ell(p^*(x))) + \ell(k_\ell(p^*(x)), 0) \mid g(x) = 1] \right. \\
&\qquad \left. - \mathop{\mathbb{E}}_{x\sim D} \left[\mathop{\mathbb{E}}_{\widetilde{y}\sim \mathrm{Ber}(p^*(x))} [\widetilde{y} \mid x] \cdot \Delta\ell(k_\ell(p^*(x))) + \ell(k_\ell(p^*(x)), 0) \mid g(x) = 1 \right] \right| \\
&= \left| \mathop{\mathbb{E}}_{(x,y)\sim D} [(y - p^*(x)) \cdot \Delta\ell(k_\ell(p^*(x))) \mid g(x) = 1] \right|.
\end{aligned}$$

An identical argument applies to Hypothesis OI.

$$\begin{aligned}
&\left| \mathop{\mathbb{E}}_{(x,y)\sim D} [\ell(h(x), y) \mid g(x) = 1] - \mathop{\mathbb{E}}_{\substack{x\sim D \\ \widetilde{y}\sim \mathrm{Ber}(p^*(x))}} [\ell(h(x), \widetilde{y}) \mid g(x) = 1] \right| \\
&= \left| \mathop{\mathbb{E}}_{(x,y)\sim D} [y \cdot \ell(h(x), 1) + (1-y) \cdot \ell(h(x), 0) \mid g(x) = 1] \right. \\
&\qquad \left. - \mathop{\mathbb{E}}_{\substack{x\sim D \\ \widetilde{y}\sim \mathrm{Ber}(p^*(x))}} [\widetilde{y} \cdot \ell(h(x), 1) + (1-\widetilde{y}) \cdot \ell(h(x), 0) \mid g(x) = 1] \right| \\
&= \left| \mathop{\mathbb{E}}_{(x,y)\sim D} [y \cdot \Delta\ell(h(x)) + \ell(h(x), 0) \mid g(x) = 1] \right. \\
&\qquad \left. - \mathop{\mathbb{E}}_{x\sim D} \left[\mathop{\mathbb{E}}_{\widetilde{y}\sim \mathrm{Ber}(p^*(x))} [\widetilde{y} \mid x] \cdot \Delta\ell(h(x)) + \ell(h(x), 0) \mid g(x) = 1 \right] \right| \\
&= \left| \mathop{\mathbb{E}}_{(x,y)\sim D} [(y - p^*(x)) \cdot \Delta\ell(h(x)) \mid g(x) = 1] \right|.
\end{aligned}$$

The same steps apply to randomized $\boldsymbol{p}^*$, with an additional expectation taken over the draw of $p^* \sim \boldsymbol{p}^*$. For

Decision OI, this means

$$\begin{aligned}
&\left|\mathop{\mathbb{E}}_{(x,y)\sim D,\, p^*\sim \boldsymbol{p}^*}[\ell(k_\ell(p^*(x)), y) \mid g(x)=1] - \mathop{\mathbb{E}}_{\substack{x\sim D,\, p^*\sim \boldsymbol{p}^* \\ \widetilde{y}\sim \mathrm{Ber}(p^*(x))}}[\ell(k_\ell(p^*(x)), \widetilde{y}) \mid g(x)=1]\right| \\
&= \left| \mathop{\mathbb{E}}_{(x,y)\sim D,\, p^*\sim \boldsymbol{p}^*}[y\cdot \ell(k_\ell(p^*(x)),1) + (1-y)\cdot \ell(k_\ell(p^*(x)),0) \mid g(x)=1] \right. \\
&\qquad \left. - \mathop{\mathbb{E}}_{\substack{x\sim D,\, p^*\sim \boldsymbol{p}^* \\ \widetilde{y}\sim \mathrm{Ber}(p^*(x))}}[\widetilde{y}\cdot \ell(k_\ell(p^*(x)),1) + (1-\widetilde{y})\cdot \ell(k_\ell(p^*(x)),0) \mid g(x)=1] \right| \\
&= \left| \mathop{\mathbb{E}}_{(x,y)\sim D,\, p^*\sim \boldsymbol{p}^*}[y\cdot \Delta\ell(k_\ell(p^*(x))) + \ell(k_\ell(p^*(x)),0) \mid g(x)=1] \right. \\
&\qquad \left. - \mathop{\mathbb{E}}_{x\sim D,\, p^*\sim \boldsymbol{p}^*}\left[\mathop{\mathbb{E}}_{\widetilde{y}\sim \mathrm{Ber}(p^*(x))}[\widetilde{y} \mid x,\, p^*]\cdot \Delta\ell(k_\ell(p^*(x))) + \ell(k_\ell(p^*(x)),0) \mid g(x)=1\right] \right| \\
&= \left| \mathop{\mathbb{E}}_{(x,y)\sim D,\, p^*\sim \boldsymbol{p}^*}[(y-p^*(x))\cdot \Delta\ell(k_\ell(p^*(x))) \mid g(x)=1] \right|.
\end{aligned}$$

For Hypothesis OI, this means

$$\begin{aligned}
&\left|\mathop{\mathbb{E}}_{(x,y)\sim D,\, p^*\sim \boldsymbol{p}^*}[\ell(h(x), y) \mid g(x)=1] - \mathop{\mathbb{E}}_{\substack{x\sim D,\, p^*\sim \boldsymbol{p}^* \\ \widetilde{y}\sim \mathrm{Ber}(p^*(x))}}[\ell(h(x), \widetilde{y}) \mid g(x)=1]\right| \\
&= \left| \mathop{\mathbb{E}}_{(x,y)\sim D,\, p^*\sim \boldsymbol{p}^*}[y\cdot \ell(h(x),1) + (1-y)\cdot \ell(h(x),0) \mid g(x)=1] \right. \\
&\qquad \left. - \mathop{\mathbb{E}}_{\substack{x\sim D,\, p^*\sim \boldsymbol{p}^* \\ \widetilde{y}\sim \mathrm{Ber}(p^*(x))}}[\widetilde{y}\cdot \ell(h(x),1) + (1-\widetilde{y})\cdot \ell(h(x)0) \mid g(x)=1] \right| \\
&= \left| \mathop{\mathbb{E}}_{(x,y)\sim D,\, p^*\sim \boldsymbol{p}^*}[y\cdot \Delta\ell(h(x)) + \ell(h(x),0) \mid g(x)=1] \right. \\
&\qquad \left. - \mathop{\mathbb{E}}_{x\sim D,\, p^*\sim \boldsymbol{p}^*}\left[\mathop{\mathbb{E}}_{\widetilde{y}\sim \mathrm{Ber}(p^*(x))}[\widetilde{y} \mid x,\, p^*]\cdot \Delta\ell(h(x)) + \ell(h(x),0) \mid g(x)=1\right] \right| \\
&= \left| \mathop{\mathbb{E}}_{(x,y)\sim D,\, p^*\sim \boldsymbol{p}^*}[(y-p^*(x))\cdot \Delta\ell(h(x)) \mid g(x)=1] \right|.
\end{aligned}$$

B.2 From Deterministic Step Calibration to Deterministic Panprediction

First, we restate and prove Lemma 3.3, the decomposition helper lemma used to prove Theorem 3.2.

Lemma 3.3. *If a deterministic predictor $p^* : \mathcal{X} \to [0,1]$ is $(\mathcal{G}, \mathcal{H}, \varepsilon)$-step calibrated, then p^* is $(\mathcal{G}, \emptyset, \varepsilon)$-step calibrated and $(\mathcal{G}, \mathcal{H}, \varepsilon)$-multiaccurate.*

Proof of Lemma 3.3. For the first statement, fix any $h \in \mathcal{H}$ and $w = 1$. Recall that by definition of $(\mathcal{G}, \mathcal{H}, \varepsilon)$-step calibration, we have that for all $v \in [0,1]$ and $g \in \mathcal{G}$,

$$\begin{aligned}
&\left|\mathop{\mathbb{E}}_{(x,y)\sim D}[(y-p^*(x))\cdot \mathbb{1}[p^*(x)\le v, h(x)\le 1] \mid g(x)=1]\right| \\
&\qquad = \left|\mathop{\mathbb{E}}_{(x,y)\sim D}[(y-p^*(x))\cdot \mathbb{1}[p^*(x)\le v] \mid g(x)=1]\right| \le \varepsilon\cdot\sqrt{P_g^{-1}}.
\end{aligned}$$

Hence, p^* is $(\mathcal{G}, \emptyset, \varepsilon)$-step calibrated.

For the second statement, fix $v = 1$. $(\mathcal{G}, \mathcal{H}, \varepsilon)$-step calibration also guarantees that for all $h \in \mathcal{H}$, $w \in [0,1]$, and

$g \in \mathcal{G}$,

$$\left| \mathop{\mathbb{E}}_{(x,y)\sim D} \left[(y - p^*(x)) \cdot \mathbb{1}[p^*(x) \le 1, h(x) \le w] \mid g(x) = 1 \right] \right|$$
$$= \left| \mathop{\mathbb{E}}_{(x,y)\sim D} \left[(y - p^*(x)) \cdot \mathbb{1}[h(x) \le w] \mid g(x) = 1 \right] \right| \le \varepsilon \cdot \sqrt{P_g^{-1}}.$$

Hence, p^* is $(\mathcal{G}, \mathcal{H}, \varepsilon)$-multiaccurate. □

Next, we prove Lemma 3.4, the approximation-theoretic helper lemma used to prove Theorem 3.2. The following exposition closely follows that of Okoroafor, Kleinberg, and Kim [2].

We begin by recalling Assumption 1.

Definition B.1 (Bounded variation loss)**.** *The total variation of a loss function $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to [-1, 1]$ in its first argument is*

$$\sup_{y \in \mathcal{Y}} V(\ell(\cdot, y)) = \sup_{y \in \mathcal{Y}} \sup_{m \in \mathbb{N}} \sup_{0 = z_0 < \cdots < z_m = 1} \sum_{j=1}^{m} \left| \ell(z_j, y) - \ell(z_{j-1}, y) \right|.$$

The class of all loss functions whose total variation in the first argument is bounded is denoted by $\mathcal{L}_{\mathrm{BV}}$.

For simplicity, we can say that $\mathcal{L}_{\mathrm{BV}} = \{\ell : [0, 1] \to [-1, 1] \mid \sup_y V(\ell(\cdot, y)) \le 1\}$. This is without loss of generality, since ℓ with total variation $v < \infty$ can be rescaled to $v^{-1} \cdot \ell$ to fall into this simpler class. Rescaling does not change relevant properties of the loss, e.g., what predictor minimizes the loss. Observe that $\Delta\mathcal{L}_{\mathrm{BV}} = \{\Delta\ell \mid \ell \in \mathcal{L}_{\mathrm{BV}}, V(\Delta\ell) \le 2\}$ is also a class of bounded variation functions.

An important subset of bounded variation losses is proper losses.

Definition B.2 (Proper loss)**.** *A bounded loss function $\ell : [0, 1] \times \mathcal{Y} \to [-1, 1]$ is a proper loss if for all $p, q \in [0, 1]$,*

$$\mathop{\mathbb{E}}_{y\sim \mathrm{Ber}(p)} [\ell(p, y)] \le \mathop{\mathbb{E}}_{y\sim \mathrm{Ber}(p)} [\ell(q, y)].$$

A proper loss ℓ has bounded variation, as $\Delta\ell(p) = \ell(p, 1) - \ell(p, 0)$ decreases monotonically as $p \to 1$. Additionally, when ℓ is bounded on $[-1, 1]$, it follows that $V(\Delta\ell) \le 2$.

Given a loss class $\mathcal{L} \subseteq \mathcal{L}_{\mathrm{BV}}$, our goal is to approximate functions in $\Delta\mathcal{L} \subseteq \Delta\mathcal{L}_{\mathrm{BV}}$ with a finite basis of simple functions. To begin, we rigorously define what we consider an approximate basis, then give two examples. At a technical level, our approach combines those of Okoroafor, Kleinberg, and Kim [2] and Qiao and Zhao [3].

Definition B.3 (ε-approximate basis)**.** *Let $\mathcal{F} \subseteq \{f : [0, 1] \to [-1, 1]\}$ be a function class. A set of functions $\mathcal{B} = \{\beta : [0, 1] \to [-1, 1]\}$ forms a ε-approximate basis with sparsity $s \in \mathbb{N}$ and coefficient norm $C > 0$ if for all $f \in \mathcal{F}$, there exists a finite subset $\beta_1, \ldots, \beta_s \in \mathcal{B}$ and coefficients $c_1, \ldots, c_s \in [-1, 1]$ such that for all $p \in [0, 1]$,*

$$\left| f(x) - \sum_{j=1}^{s} c_j \cdot \beta_j(x) \right| \le \varepsilon \quad \text{and} \quad \sum_{j=1}^{s} |c_j| \le C.$$

Definition B.4 (Threshold functions)**.** *For $v \in [0, 1]$, define $\mathrm{Th}_v : [0, 1] \to \{-1, 1\}$,*

$$\mathrm{Th}_v(p) = \mathrm{sign}(v - p) = \begin{cases} +1 & \text{if} \quad p \le v \\ -1 & \text{if} \quad p > v. \end{cases}$$

Definition B.5 (Step functions)**.** *For $v \in [0, 1]$, define Step$_v : [0, 1] \to \{0, 1\}$, $\mathrm{Step}_v(p) = \mathbb{1}[p \le v]$.*

As observed in Lemma A.1 of Qiao and Zhao [3], threshold functions and step functions are related by the fact that

$$\mathrm{Th}_v(p) = \mathbb{1}[p \le v] - \mathbb{1}[p > v] = 2 \cdot \mathbb{1}[p \le v] - \mathbb{1}[p \le 1]. \tag{4}$$

Proposition B.6 (Lemma 5.6 in [2])**.** *For any $\varepsilon > 0$, the uncountably infinite set of threshold functions*

$$\mathbf{Th} \coloneqq \{\mathrm{Th}_v \mid v \in [0,1]\}$$

forms a ε-basis for $\Delta\mathcal{L}_{\mathrm{BV}} = \{\Delta\ell \mid \ell \in \mathcal{L}_{\mathrm{BV}}\}$ with sparsity $\lceil \frac{2}{\varepsilon} + 1 \rceil$ and coefficient norm 3.

The fact that the approximate basis of $\Delta\mathcal{L}_{\mathrm{BV}}$ is uncountably infinite is not a problem for the proof of Lemma 3.4, but can be problematic for the design of algorithms that take the approximate basis functions as input. Fortunately, an approximate basis of finite size can always be found when the inputs to $\Delta\ell$, i.e., the predictions made by $p^* \in \mathcal{P}$ and $h \in \mathcal{H}$, are quantized to the λ-net of the unit interval, I_λ. This observation is formalized below.

Proposition B.7 (Lemma 5.7 in [2])**.** *Suppose the inputs to each $\Delta\ell \in \Delta\mathcal{L}_{\mathrm{BV}}$ is restricted to I_λ. Then for all $\varepsilon > 0$, the finite set of threshold functions*

$$\mathbf{Th}_\lambda \coloneqq \{\mathrm{Th}_v \mid v \in I_\lambda\}$$

forms an ε-basis for $\Delta\mathcal{L}_{\mathrm{BV}}$ with sparsity $\lceil \frac{2}{\varepsilon} + 1 \rceil$ and coefficient norm 3.

With the above results in hand, we can prove Lemma 3.4.

Lemma 3.4. *Suppose $\ell \in \mathcal{L}_{\mathrm{BV}}$. Then for any predictor $f : \mathcal{X} \to [0,1]$ and $g \in \mathcal{G}$,*

$$\left| \mathop{\mathbb{E}}_{(x,y)\sim D} \left[(y - p^*(x)) \cdot \Delta\ell(f(x)) \mid g(x) = 1 \right] \right| \leq 9 \sup_{v \in [0,1]} \left| \mathop{\mathbb{E}}_{D} \left[(y - p^*(x)) \cdot \mathbb{1}[f(x) \leq v] \mid g(x) = 1 \right] \right| + \varepsilon \cdot \sqrt{P_g^{-1}}.$$

Proof of Lemma 3.4. By Definition B.3, Proposition B.6, and the triangle inequality, we have that

$$\begin{aligned}
&\left| \mathop{\mathbb{E}}_{(x,y)\sim D} \left[(y - p^*(x)) \cdot \Delta\ell(f(x)) \mid g(x) = 1 \right] \right| \\
&\quad \leq \left| \mathop{\mathbb{E}}_{(x,y)\sim D} \left[(y - p^*(x)) \cdot \sum_{j=1}^{s} c_j \cdot \mathrm{Th}_{v_j}(f(x)) \mid g(x) = 1 \right] \right| \\
&\qquad + \left| \mathop{\mathbb{E}}_{(x,y)\sim D} \left[(y - p^*(x)) \cdot \left(\Delta\ell(f(x)) - \sum_{j=1}^{s} c_j \cdot \mathrm{Th}_{v_j}(f(x)) \right) \mid g(x) = 1 \right] \right| \\
&\quad \leq \left| \sum_{j=1}^{s} c_j \cdot \mathop{\mathbb{E}}_{(x,y)\sim D} \left[(y - p^*(x)) \cdot \mathrm{Th}_{v_j}(f(x)) \mid g(x) = 1 \right] \right| + \varepsilon.
\end{aligned}$$

The above line can be further upper bounded as

$$\begin{aligned}
&\sum_{j=1}^{s} |c_j| \cdot \left| \mathop{\mathbb{E}}_{(x,y)\sim D} \left[(y - p^*(x)) \cdot \mathrm{Th}_{v_j}(f(x)) \mid g(x) = 1 \right] \right| + \varepsilon \\
&\quad \leq \left(\sum_{j=1}^{s} |c_j| \right) \cdot \max_{j \in [s]} \left| \mathop{\mathbb{E}}_{(x,y)\sim D} \left[(y - p^*(x)) \cdot \mathrm{Th}_{v_j}(f(x)) \mid g(x) = 1 \right] \right| + \varepsilon \\
&\quad \leq 3 \sup_{v \in [0,1]} \left| \mathop{\mathbb{E}}_{(x,y)\sim D} \left[(y - p^*(x)) \cdot \mathrm{Th}_{v}(f(x)) \mid g(x) = 1 \right] \right| + \varepsilon,
\end{aligned}$$

where the second inequality follows from recalling that threshold functions form an approximate basis with coefficient norm 3. Recalling Equation 4 and invoking the triangle inequality, we can rewrite the above line as

$$\begin{aligned}
&3 \sup_{v \in [0,1]} \left| \mathop{\mathbb{E}}_{(x,y)\sim D} \left[(y - p^*(x)) \cdot (2 \cdot \mathbb{1}[f(x) \leq v] - \mathbb{1}[f(x) \leq 1]) \mid g(x) = 1 \right] \right| + \varepsilon \\
&\quad \leq 9 \sup_{v \in [0,1]} \left| \mathop{\mathbb{E}}_{(x,y)\sim D} \left[(y - p^*(x)) \cdot \mathbb{1}[f(x) \leq v] \mid g(x) = 1 \right] \right| + \varepsilon \\
&\quad \leq 9 \sup_{v \in [0,1]} \left| \mathop{\mathbb{E}}_{(x,y)\sim D} \left[(y - p^*(x)) \cdot \mathbb{1}[f(x) \leq v] \mid g(x) = 1 \right] \right| + \varepsilon \cdot \sqrt{P_g^{-1}}.
\end{aligned}$$

The final inequality is trivial, since P_g is positive and at most 1. □

B.3 From Randomized Step Calibration to Randomized Panprediction

We show that the same loss outcome indistinguishability approach used above can be used to prove that randomized step calibration implies randomized panprediction.

Theorem B.8. *If a randomized predictor $\boldsymbol{p}^* \in \Delta(\mathcal{P})$ is $(\mathcal{G}, \mathcal{H}, \varepsilon)$-step calibrated, then $\boldsymbol{p}^*$ is a randomized $(\mathcal{L}, \mathcal{G}, \mathcal{H}, O(\varepsilon))$-panpredictor.*

As seen before, we prove Theorem B.8 by decomposing the step calibration guarantee into a calibration and a multiaccuracy guarantee (Lemma B.9), then showing that the calibration guarantee implies Decision OI (Lemma B.10) and that the multiaccuracy guarantee implies Hypothesis OI (Lemma B.11).

Lemma B.9. *If a randomized predictor $\boldsymbol{p}^* \in \Delta(\mathcal{P})$ is $(\mathcal{G}, \mathcal{H}, \varepsilon)$-step calibrated, then $\boldsymbol{p}^*$ is $(\mathcal{G}, \emptyset, \varepsilon)$-step calibrated and $(\mathcal{G}, \mathcal{H}, \varepsilon)$-multiaccurate.*

Proof of Lemma B.9. For the first statement, fix any $h \in \mathcal{H}$ and $w = 1$. Recall that from the definition of $(\mathcal{G}, \mathcal{H}, \varepsilon)$-step calibration, we have that for all $v \in [0, 1]$ and $g \in \mathcal{G}$,

$$\begin{aligned}&\left|\mathop{\mathbb{E}}_{(x,y)\sim D,\, p^*\sim \boldsymbol{p}^*}\left[(y - p^*(x)) \cdot \mathbb{1}[p^*(x) \le v, h(x) \le 1] \mid g(x) = 1\right]\right| \\ &\quad = \left|\mathop{\mathbb{E}}_{(x,y)\sim D,\, p^*\sim \boldsymbol{p}^*}\left[(y - p^*(x)) \cdot \mathbb{1}[p^*(x) \le v] \mid g(x) = 1\right]\right| \le \varepsilon \cdot \sqrt{P_g^{-1}}.\end{aligned}$$

Hence, p^* is $(\mathcal{G}, \emptyset, \varepsilon)$-step calibrated.

For the second statement, fix $v = 1$. $(\mathcal{G}, \mathcal{H}, \varepsilon)$-step calibration also guarantees that for all $h \in \mathcal{H}$, $w \in [0, 1]$, and $g \in \mathcal{G}$,

$$\begin{aligned}&\left|\mathop{\mathbb{E}}_{(x,y)\sim D,\, p^*\sim \boldsymbol{p}^*}\left[(y - p^*(x)) \cdot \mathbb{1}[p^*(x) \le 1, h(x) \le w] \mid g(x) = 1\right]\right| \\ &\quad = \left|\mathop{\mathbb{E}}_{(x,y)\sim D,\, p^*\sim \boldsymbol{p}^*}\left[(y - p^*(x)) \cdot \mathbb{1}[h(x) \le w] \mid g(x) = 1\right]\right| \le \varepsilon \cdot \sqrt{P_g^{-1}}.\end{aligned}$$

Hence, p^* is $(\mathcal{G}, \mathcal{H}, \varepsilon)$-multiaccurate. □

Lemma B.10. *If a randomized predictor $\boldsymbol{p}^* \in \Delta(\mathcal{P})$ is $(\mathcal{G}, \emptyset, \varepsilon)$-step calibrated, then for all $g \in \mathcal{G}$,*

$$\left|\mathop{\mathbb{E}}_{(x,y)\sim D,\, p^*\sim \boldsymbol{p}^*}\left[(y - p^*(x)) \cdot \Delta\ell(k_\ell(p^*(x))) \mid g(x) = 1\right]\right| \le O(\varepsilon) \cdot \sqrt{P_g^{-1}}.$$

Proof of Lemma B.10. Since for any $\ell : [0, 1] \times \mathcal{Y} \to [-1, 1]$ and $p, q \in [0, 1]$, by the definition of the function k_ℓ,

$$\mathop{\mathbb{E}}_{y\sim \mathrm{Ber}(p)}\left[\ell(k_\ell(p), y)\right] \le \mathop{\mathbb{E}}_{y\sim \mathrm{Ber}(p)}\left[\ell(k_\ell(q), y)\right],$$

the map $\ell(k_\ell(\cdot), y) : [0, 1] \to [-1, 1]$ is a proper scoring rule. Hence this map is contained in the bounded variation loss class $\mathcal{L}_{\mathrm{BV}}$. Then by Lemma 3.4,

$$\begin{aligned}&\left|\mathop{\mathbb{E}}_{(x,y)\sim D}\left[(y - p^*(x)) \cdot \Delta\ell(k_\ell(p^*(x))) \mid g(x) = 1\right]\right| \\ &\quad \le 9 \sup_{v\in[0,1]}\left|\mathop{\mathbb{E}}_{(x,y)\sim D,\, p^*\sim \boldsymbol{p}^*}\left[(y - p^*(x)) \cdot \mathbb{1}[p^*(x) \le v] \mid g(x) = 1\right]\right| + \varepsilon \cdot \sqrt{P_g^{-1}} \\ &\quad \le O(\varepsilon) \cdot \sqrt{P_g^{-1}},\end{aligned}$$

where the last inequality follows from the $(\mathcal{G}, \emptyset, \varepsilon)$-step calibration guarantee. □

Lemma B.11. *If a randomized predictor $\boldsymbol{p}^* \in \Delta(\mathcal{P})$ is $(\mathcal{G}, \mathcal{H}, \varepsilon)$-multiaccurate, then for all $h \in \mathcal{H}$ and $g \in \mathcal{G}$,*

$$\left|\mathop{\mathbb{E}}_{(x,y)\sim D,\, p^*\sim \boldsymbol{p}^*}\left[(y - p^*(x)) \cdot \Delta\ell(h(x)) \mid g(x) = 1\right]\right| \le O(\varepsilon) \cdot \sqrt{P_g^{-1}}.$$

Proof of Lemma B.11. By assumption, $\ell \in \mathcal{L}_{\mathrm{BV}}$. Then by applying Lemma 3.4 with any $h \in \mathcal{H}$, $h : \mathcal{X} \to [0,1]$,

$$\begin{aligned}
&\left| \mathop{\mathbb{E}}_{(x,y)\sim D,\, p^*\sim \boldsymbol{p}^*} \left[(y - p^*(x)) \cdot \Delta\ell(h(x)) \mid g(x) = 1 \right] \right| \\
&\quad \le 9 \sup_{w\in[0,1]} \left| \mathop{\mathbb{E}}_{(x,y)\sim D,\, p^*\sim \boldsymbol{p}^*} \left[(y - p^*(x)) \cdot \mathbb{1}[h(x) \le w] \mid g(x) = 1 \right] \right| + \varepsilon \cdot \sqrt{P_g^{-1}} \\
&\quad \le O(\varepsilon) \cdot \sqrt{P_g^{-1}},
\end{aligned}$$

where the last inequality follows from the $(\mathcal{G}, \mathcal{H}, \varepsilon)$-multiaccuracy guarantee. □

C Deferred Results and Proofs from Section 4

In this section, we state and prove guarantees about our deterministic and randomized step calibration algorithms.

C.1 Background

We begin by collecting notation and useful results from online learning, adaptive data analysis, and multi-objective learning. This exposition closely follows that of Haghtalab, Jordan, and Zhao [4].

Online learning Online learning models a T-round interaction between a learner and an adversary. On each round $t \in [T]$, the learner chooses an action $a^{(t)}$ from an action set $\mathcal{A}$ and the adversary chooses a bounded cost function $c^{(t)} : \mathcal{A} \to [0,1]$. In the sequel, we focus on stochastic cost functions $c^{(t)} : \mathcal{A} \times (\mathcal{X} \times \mathcal{Y}) \to [0,1]$ that take both the learner's action and a datapoint $(x, y) \sim D$ as inputs. It is natural to minimize expected cost $\mathcal{R}_{c^{(t)}}(\cdot) = \mathbb{E}_{(x,y)\sim D}\left[c^{(t)}(\cdot, (x,y))\right] : \mathcal{A} \to [0,1]$. When the choices of action and cost function are non-deterministic, that is, $p \in \Delta(\mathcal{A})$ and $q \in \Delta(C)$, we can also consider $\mathcal{R}_q(p) = \mathbb{E}_{a\sim p,\, c\sim q}\left[\mathcal{R}_c(a)\right] = \mathbb{E}_{a\sim p,\, c\sim q}\left[\mathbb{E}_{(x,y)\sim D}\left[c(a,(x,y)\right]\right]$. Importantly, the stochastic cost functions we consider have linear structure in the action, that is, for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$, $c(\cdot, (x,y)) : \mathcal{A} \to [0,1]$ is linear.

The learner's regret (to the best fixed action) is defined as

$$\mathrm{Reg}(a^{(1:T)}, c^{(1:T)}) = \sum_{t=1}^{T} c^{(t)}(a^{(t)}) - \min_{a^*\in\mathcal{A}} \sum_{t=1}^{T} c^{(t)}(a^*).$$

When working with stochastic cost functions, we have that

$$\begin{aligned}
&\mathrm{Reg}(a^{(1:T)}, \{\mathcal{R}_{c^{(t)}}(\cdot)\}^{(1:T)}) \\
&\qquad = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{(x,y)\sim D}\left[c^{(t)}(a^{(t)}, (x,y))\right] - \min_{a^*\in\mathcal{A}} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{(x,y)\sim D}\left[c^{(t)}(a^*, (x,y))\right].
\end{aligned}$$

The learner's weak regret to a minimax benchmark $T \cdot \min_{a^*\in\mathcal{A}} \max_{c^*\in C} c^*(a^*)$ is given by

$$\mathrm{Reg}_{\mathrm{weak}}(a^{(1:T)}, c^{(1:T)}) = \sum_{t=1}^{T} c^{(t)}(a^{(t)}) - T \cdot \min_{a^*\in\mathcal{A}} \max_{c^*\in C} c^*(a^*).$$

Identically, when working with stochastic cost functions, we have that

$$\begin{aligned}
&\mathrm{Reg}_{\mathrm{weak}}(a^{(1:T)}, \{\mathcal{R}_{c^{(t)}}(\cdot)\}^{(1:T)}) \\
&\qquad = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{(x,y)\sim D}\left[c^{(t)}(a^{(t)}, (x,y))\right] - T \cdot \min_{a^*\in\mathcal{A}} \max_{c^*\in C} \mathop{\mathbb{E}}_{(x,y)\sim D}\left[c^*(a^*, (x,y))\right].
\end{aligned}$$

The standard Hedge algorithm [19] obtains a sublinear regret bound.

Proposition C.1 (Hedge [19])**.** *If the action sequence in the k-simplex $a^{(1:T)} \in \Delta_k$ is selected by the Hedge algorithm, then for any adversarial choice of cost functions $c^{(1:T)}$,*

$$\mathrm{Reg}(a^{(1:T)}, c^{1:T)}) \le C\sqrt{\log(k)T},$$

where $C > 0$ is a universal constant.

Given sampling access to distribution D and an action $a \in \mathcal{A}$, a natural way to estimate the expected cost $\mathcal{R}_c(a)$ is sampling $(x, y) \sim D$ and evaluating $c(a, (x, y))$. When c has linear structure, this approach incurs sublinear estimation error with high probability.

Proposition C.2 (Stochastic approximation [21])**.** *Let $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(T)}, y^{(T)}) \overset{\text{iid}}{\sim} D$. Suppose that all $c \in \mathcal{C}$ are linear and that at each round $t \in [T]$, after picking action $a^{(t)}$ with an online learning algorithm, the expected cost $\mathcal{R}_{c^{(t)}}(a^{(t)})$ is estimated with $\widehat{c}^{(t)}(a) \coloneqq c^{(t)}(a, (x^{(t)}, y^{(t)}))$. Then with probability at least $1 - \delta$,*

$$\left|\mathrm{Reg}(a^{(1:T)}, \{\mathcal{R}_{c^{(t)}}(\cdot)\}^{(1:T)}) - \mathrm{Reg}(a^{(1:T)}, \widehat{c}^{(1:T)})\right| \le C\sqrt{T\log(1/\delta)},$$

where $C > 0$ is a universal constant.

We sometimes consider best-response oracles of the following form.

Definition C.3 (Best-response oracle)**.** *A best response oracle takes an action $a \in \mathcal{A}$, a set of stochastic cost functions $\mathcal{C}$, and error tolerance ε, then returns a cost function $c \in \mathcal{C}$ such that*

$$\mathop{\mathbb{E}}_{(x,y)\sim D}[c(a, (x, y))] \le \min_{c^* \in \mathcal{C}} \mathop{\mathbb{E}}_{(x,y)\sim D}[c^*(a, (x, y))] + \varepsilon.$$

Such a choice of $c \in \mathcal{C}$ is referred to as a ε-best response to the action a.

Adaptive data analysis Adaptive data analysis models a T-round interaction between a data analyst and a mechanism. At each round $t \in [T]$, the analyst poses a query about a fixed distribution D. Using only samples $z^{(1:n)} = \{(x_1, y_1), \ldots, (x_n, y_n)\} \overset{\text{iid}}{\sim} D$, the mechanism must sequentially T queries (where each query may adversarially depend on all previous queries and answers) up to small additive error with high probability.

We focus on minimization queries over finite sets, which take the following form. Suppose a family of loss functions $\{\ell\}$ is parameterized by a finite set of parameters Θ. That is, for each loss function $\ell \in \mathcal{L}$, $\ell : (\mathcal{X} \times \mathcal{Y})^n \times \Theta \to [0, 1]$. Further suppose that each loss has Δ-sensitivity. That is, for all "datasets" $z^{(1:n)}$ and $\overline{z}^{(1:n)}$ that differ in just one data point,

$$\sup_{\theta\in\Theta} \left|\ell(z^{(1:n)}, \theta) - \ell(\overline{z}^{(1:n)}, \theta)\right| \le \Delta.$$

Fix $\alpha, \beta \in (0, 1)$. The (α, β)-minimization query associated with loss ℓ seeks $\theta \in \Theta$ such that with probability $1 - \beta$,

$$\mathop{\mathbb{E}}_{z^{(1:n)}\sim D^n}\left[\ell(z^{(1:n)}, \theta)\right] \le \min_{\theta^* \in \Theta} \mathop{\mathbb{E}}_{z^{(1:n)}\sim D^n}\left[\ell(z^{(1:n)}, \theta^*)\right] + \alpha.$$

Minimization queries can be naturally used to best respond to stochastic cost functions discussed above, as long as the cost functions are parameterized by a finite set Θ. Importantly, adaptive data analysis allows cost functions to be chosen adaptively.

Proposition C.4 (Adaptive minimization queries [20])**.** *There exists an algorithm that, with probability at least $1 - \delta$, can ε-best respond to an adaptive sequence of T stochastic costs that each have sensitivity $1/n$ and are bounded in $[0, 1]$, using at most*

$$n = O\left(\frac{\sqrt{T} \cdot \log\left(|\Theta|/\varepsilon\right) \cdot \log^{3/2}(1/\varepsilon\delta)}{\varepsilon^2}\right)$$

samples from $\mathcal{D}$.

The algorithm that witnesses this sample complexity is an instantiation of the exponential mechanism, and is detailed in prior works [22, 20].

C.2 Multi-objective Learning

To begin, recall the definition of multi-objective learning.

Definition C.5 ([4]). *A multi-objective learning problem is defined by a distribution D, a hypothesis class $\mathcal{F} \subseteq \{f : \mathcal{X} \to [0,1]\}$, and a set of objectives $\mathcal{O} = \{\ell : \mathcal{F} \times (\mathcal{X} \times \mathcal{Y}) \to [a,b]\}$. The goal of multi-objective learning is to find a hypothesis $\mathbf{f}^* \in \Delta(\mathcal{F})$ such that*

$$\max_{\ell \in \mathcal{O}} \mathop{\mathbb{E}}_{(x,y)\sim D,\, f^*\sim \mathbf{f}^*} [\ell(f^*(x), y)] \le \min_{f \in \mathcal{F}} \max_{\ell \in \mathcal{O}} \mathop{\mathbb{E}}_{(x,y)\sim D} [\ell(f(x), y)] + \varepsilon.$$

If $\mathbf{f}^*$ is a point mass on a hypothesis $f^* \in \mathcal{H}$, then f^* is called a deterministic ε-optimal solution to $(D, \mathcal{O}, \mathcal{F})$. Otherwise, $\mathbf{f}^*$ is called a randomized ε-optimal solution to $(D, \mathcal{O}, \mathcal{F})$.

Deterministic solutions to multi-objective problems can be found via "no-regret vs best-response" dynamics that select a sequence of hypotheses with no-regret online learning algorithms and a sequence of objectives using best-response oracles. The following propositions formalize this result.

Proposition C.6 (No-Regret vs Best-Response [4]). *Consider a multi-objective learning problem $(D, \mathcal{O}, \mathcal{H})$ where the learner chooses $p^{(1:T)} \in \Delta(\mathcal{H})$ and the adversary chooses $q^{(1:T)} \in \Delta(\mathcal{O})$. If the learner is no-regret, such that $\mathrm{Reg}_{\mathrm{weak}}(p^{(1:T)}, \{\mathcal{R}_{q^{(t)}}(\cdot)\}^{(1:T)})$, and the adversary ε-best responds to the sequence of costs $\{1 - \mathcal{R}_{(\cdot)}(p^{(t)})\}^{(1:T)}$ using $q^{(1:T)}$, then there is a timestep $t \in [T]$ where $p^{(t)}$ is a deterministic 2ε-optimal solution to $(D, \mathcal{O}, \mathcal{H})$.*

Proposition C.7 (Deterministic solution [4]). *Suppose a sequence of hypotheses $p^{(1:T)}$ contains a deterministic ε-optimal solution to $(D, \mathcal{O}, \mathcal{H})$. Then a deterministic 2ε-optimal solution $p^{(t)} \in p^{(1:T)}$ can be identified using $O(\varepsilon^{-2} \log(|\mathcal{O}|T/\delta))$ samples.*

Randomized solutions to multi-objective problems can be found via "no-regret vs no-regret" dynamics that uses no-regret online learning algorithms to select both the sequence of hypotheses and the sequence of objectives. The following proposition formalizes this result.

Proposition C.8 (No-Regret vs No-Regret [4]). *Consider a multi-objective learning problem $(D, \mathcal{O}, \mathcal{H})$ where the learner chooses $p^{(1:T)} \in \Delta(\mathcal{H})$ and the adversary chooses $q^{(1:T)} \in \Delta(\mathcal{O})$. If both players are no-regret, such that $\mathrm{Reg}_{\mathrm{weak}}(p^{(1:T)}, \{\mathcal{R}_{q^{(t)}}(\cdot)\}^{(1:T)}) \le T\varepsilon$ and $\mathrm{Reg}(q^{(1:T)}, \{1 - \mathcal{R}_{(\cdot)}(p^{(t)})\}^{(1:T)}) \le T\varepsilon$, then the non-deterministic hypothesis $\overline{p} = \mathrm{Uniform}(p^{(1:T)})$ is a 2ε-optimal solution to $(D, \mathcal{O}, \mathcal{H})$.*

Application to Step Calibration Theorem 4.1 shows that $(\mathcal{G}, \mathcal{H}, \varepsilon)$-step calibration can be formulated as a $(D, \mathcal{O}_{\mathrm{sc}}, \mathcal{P})$-multi-objective learning problem. It remains to address the fact that $\mathcal{O}_{\mathrm{sc}}$ is an uncountably infinite set of losses parameterized by

$$(\sigma, v, w, h, g) \in \{\pm 1\} \times [0,1] \times [0,1] \times \mathcal{H} \times \mathcal{G}.$$

For our algorithms to work, a finite cover of $\mathcal{O}_{\mathrm{sc}}$, denoted by $\widehat{\mathcal{O}}_{\mathrm{sc}}$, must be found. If $\mathcal{H}$ and $\mathcal{G}$ have finite cardinality, and real-valued hypotheses are quantized to I_λ for $\lambda = \Theta(\varepsilon)$, a (lossless) finite cover of size $O(\varepsilon^{-2}|\mathcal{H}||\mathcal{G}|)$ can be constructed by choosing parameters

$$(\sigma, v, w, h, g) \in \{\pm 1\} \times I_{\Theta(\varepsilon)} \times I_{\Theta(\varepsilon)} \times \mathcal{H} \times \mathcal{G}.$$

In the more general case, where $\mathcal{H}$ and $\mathcal{G}$ have infinite cardinality but bounded combinatorial dimensions, a finite cover can still be constructed. Namely, assume that $\mathrm{Pdim}(\mathcal{H}) \le d_H$ and $\mathrm{VC}(\mathcal{G}) \le d_G$. Then the composed class $\mathrm{Step} \circ \mathcal{H}$ has VC dimension at most d_H. Furthermore, the class of pointwise conjunctions between $\mathrm{Step} \circ \mathcal{H}$ and $\mathcal{G}$, denoted by $(\mathrm{Step} \circ \mathcal{H}) \wedge \mathcal{G}$, has VC dimension at most $d_H + d_G$. It is well-known that given $\widetilde{O}(\varepsilon^{-1}(d_H + d_G + \log(1/\delta'))$ samples from D, an ε-cover of $(\mathrm{Step} \circ \mathcal{H}) \wedge \mathcal{G}$ in the $L_1(D)$ metric can be algorithmically constructed, with failure probability at most δ'. In fact, this cover is constructed by combining an ε-cover of $\mathrm{Step} \circ \mathcal{H}$ (denoted by $(\mathrm{Step} \circ \mathcal{H})'$) and an ε-cover of $\mathcal{G}$ (denoted by $\mathcal{G}'$). By standard learning-theoretic arguments, the combined cover has size at most $1/\varepsilon^{O(d_H + d_G)}$ [23, 24]. In the sequel, we assume that such a cover, denoted by $(\mathrm{Step} \circ \mathcal{H})' \times \mathcal{G}'$, has been constructed, with failure probability $\delta' = \delta/3$. All in all, $\widehat{\mathcal{O}}_{\mathrm{sc}}$ is the finite set of losses parameterized by

$$(\sigma, v, f, g) \in \{\pm 1\} \times I_{\Theta(\varepsilon)} \times (\mathrm{Step} \circ \mathcal{H})' \times \mathcal{G}'.$$

Algorithm 3 Deterministic Step Calibration

Require: $\varepsilon, \delta \in (0,1)$, $T \in \mathbb{N}$, $c \in [0,1]$, $C \in \mathbb{N}$, sampling access to D, best response oracle $\mathcal{A}$

Initialize Hedge iterate $p^{(1)} \leftarrow [\frac{1}{2}, \frac{1}{2}]^{\mathcal{X}}$

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

  Update $\ell^{(t)} \leftarrow \mathcal{A}\big(p^{(t)}, \widehat{O}_{\text{sc}}^{\gamma}, c\varepsilon\sqrt{\gamma}\big)$

  For each $x \in \mathcal{X}$, update $p^{(t+1)}(x) \leftarrow \text{Hedge}\big(c_x^{(1:t)}\big)$, where $c_x^{(t)}(\widehat{y})$ is defined as

$$\frac{1}{2} \cdot \left(1 + \sqrt{\frac{\gamma}{\Pr(g^{(t)}(x) = 1)}} \cdot \sigma^{(t)} \cdot \widehat{y} \cdot \mathbb{1}[p^{(t)}(x) \le v^{(t)}] \cdot f^{(t)}(x) \cdot \mathbb{1}[g^{(t)}(x) = 1]\right)$$

end for

Draw $m \leftarrow \frac{C \log(T/\delta)}{\varepsilon^2}$ samples $z^{(1:m)} \sim D$

$t^* \leftarrow \arg\min_{t \in [T]} \sum_{(x,y) \in z^{(1:m)}} \ell^{(t)}\big(p^{(t)}, (x,y)\big)$

return $p^{(t^*)}$

Given that $\widehat{O}_{\text{sc}}$ is a finite set, there exist straightforward ways to use best-response oracles or no-regret learning algorithms to select a sequence of objectives. The learner's task of selecting a hypothesis $p \in \mathcal{P}$ is trickier, given that $\mathcal{P}$ is a rich class containing all real-valued predictors.

Proposition C.9 ([4])**.** *Consider $\mathcal{P}$ the set of all real-valued predictors and any adversarial sequence of stochastic costs $q^{(1:T)} \in \Delta(O_{\text{sc}})$. There exists a no-regret algorithm that outputs deterministic predictors $p^{(1:T)}$ such that $\text{Reg}_{\text{weak}}(p^{(1:T)}, \{\mathcal{R}_{q^{(t)}(\cdot)}\}^{(1:T)}) \le C\sqrt{\log(2)T}$. This algorithm requires no samples from D.*

The proposed algorithm involves running Hedge at each $x \in \mathcal{X}$, then aggregating by taking $p^{(t)}(x)$ to be the prediction made by the instantiation of Hedge associated with x at time t. By the linearity of expectation, the aggregated algorithm inherits the Hedge regret bound. We defer further details to Theorems 3.7 and 5.2 in Haghtalab, Jordan, and Zhao [4].

C.3 Deterministic Step Calibration

Remark C.10 (Re-scaling and re-centering $\widehat{O}_{\text{sc}}$)**.** *To facilitate the use of no-regret algorithms and best-response oracles that require losses that are bounded in $[0,1]$, we define the re-scaled and re-centered class of objectives*

$$\widehat{O}_{\text{sc}}^{\gamma} = \left\{\frac{1}{2}\left(1 + \sqrt{\gamma} \cdot \ell\right) \mid \ell \in \widehat{O}_{\text{sc}}\right\},$$

where $\gamma = \min_{g \in \mathcal{G}} P_g = \min_{g \in \mathcal{G}} \Pr_{(x,y) \sim D}(g(x) = 1)$ is a known constant, by assumption.

Given a set of parameters $\theta \coloneqq (\sigma, v, f, g) \in \{\pm 1\} \times I_{\Theta(\varepsilon)} \times (\text{Step} \circ \mathcal{H})' \times \mathcal{G}'$, $\ell_\theta \in \widehat{O}_{\text{sc}}^{\gamma}$ takes the form

$$\ell_\theta(p, (x,y)) = \frac{1}{2} \cdot \left(1 + \sqrt{\frac{\gamma}{P_g}} \cdot \sigma \cdot (y - p(x)) \cdot \mathbb{1}[p(x) \le v] \cdot f(x) \cdot \mathbb{1}[g(x) = 1]\right) \in [0,1].$$

The following corollary of Theorem 4.1 is immediate.

Corollary C.11. *A $\varepsilon\sqrt{\gamma}$-solution to the $(D, \widehat{O}_{\text{sc}}^{\gamma}, \mathcal{P})$-multi-objective learning problem corresponds to a $O(\varepsilon)$-solution to the $(D, \widehat{O}_{\text{sc}}, \mathcal{P})$-multi-objective learning problem, and hence to a $(\mathcal{G}, \mathcal{H}, O(\varepsilon))$-step calibrated predictor.*

Though we treat γ as a constant, we track dependence on it in the rest of our derivations to facilitate future work that treats group membership probabilities as variables.

The following is a more complete restatement of Theorem 4.2.

Theorem C.12 (Deterministic Step Calibration)**.** *Fix $\varepsilon > 0$ and $\delta \in (0,1)$. Let d_H be the appropriate combinatorial dimension of H and d_G the VC dimension of $\mathcal{G}$. Then with probability at least $1 - \delta$ and at most*

$$n = O\left(\frac{1}{\varepsilon^3 \gamma^{3/2}} \cdot (d_H + d_G) \cdot \log\left(\frac{1}{\varepsilon \gamma^{1/2} \delta}\right)\right) \tag{5}$$

samples from D, Algorithm 3 returns a deterministic predictor p^* that is $(\mathcal{G}, \mathcal{H}, \varepsilon)$-step calibrated.

Algorithm 4 Randomized Step Calibration

Require: $\varepsilon, \delta \in (0,1)$, $T \in \mathbb{N}$, $c \in [0,1]$, $C \in \mathbb{N}$, sampling access to D

1: Initialize Hedge iterates $p^{(1)} \leftarrow [\frac{1}{2}, \frac{1}{2}]^{\mathcal{X}}$ and $q^{(1)} = \text{Uniform}(\widehat{O}^{\gamma}_{\text{sc}})$.
2: **for** $t = 1, \ldots, T$ **do**
3: Sample objective $\ell^{(t)} \sim q^{(t)}$ and data point $(x^{(t)}, y^{(t)}) \sim D$
4: For each $x \in \mathcal{X}$, update $p^{(t+1)}(x) \leftarrow \text{Hedge}\big(c_x^{(1:t)}\big)$, where $c_x^{(t)}(\widehat{y})$ is defined as

$$\frac{1}{2} \cdot \left(1 + \sqrt{\frac{\gamma}{\Pr(g^{(t)}(x) = 1)}} \cdot \sigma^{(t)} \cdot \widehat{y} \cdot \mathbb{1}[p^{(t)}(x) \le v^{(t)}] \cdot f^{(t)}(x) \cdot \mathbb{1}[g^{(t)}(x) = 1]\right)$$

Let $q^{(t+1)} = \text{Hedge}(c_{\text{adv}}^{(1:t)})$ where $c_{\text{adv}}^{(t)} = 1 - \ell_{(\cdot)}(p^{(t)}, (x^{(t)}, y^{(t)}))$
5: **end for**
6: **return** $\boldsymbol{p}^*$, a uniform distribution over $p^{(1)}, \ldots, p^{(T)}$

Proof. We prove this result in two steps.

By Proposition C.9, there exists a universal constant $C > 0$ such that for $T = C\varepsilon^{-2}\gamma^{-1}$, the learner's regret is $\text{Reg}_{\text{weak}}(p^{(1:T)}, \{\mathcal{R}_{\ell^{(t)}}(\cdot)\}^{(1:T)}) \le T\varepsilon\sqrt{\gamma}/8$. Suppose that at each iteration, the adversary $\varepsilon\sqrt{\gamma}/8$-best responds. By Proposition C.6, there exists a timestep $t \in [T]$ such that the predictor $p^{(t)}$ is a $\varepsilon\sqrt{\gamma}/4$-optimal solution to the $(D, \widehat{O}^{\gamma}_{\text{sc}}, \mathcal{P})$-multi-objective learning problem. By Proposition C.7, with probability $1 - \delta/3$, the predictor p^* returned by the algorithm is a deterministic $\varepsilon\sqrt{\gamma}/2$-optimal solution. Then by Corollary C.11, p^* is $(\mathcal{G}, \mathcal{H}, \varepsilon)$-step calibrated.

By Proposition C.4, the claimed sample complexity is sufficient to answer $T = C\varepsilon^{-2}\gamma^{-1}$ adaptive $\varepsilon\sqrt{\gamma}/8$-best response queries with probability $1 - \delta/3$. □

C.4 Randomized Step Calibration

The following is a more complete statement of Theorem 4.3.

Theorem C.13 (Randomized Step Calibration)**.** *Fix $\varepsilon > 0$ and $\delta \in (0,1)$. Let d_H be the appropriate combinatorial dimension of H and d_G the VC dimension of $\mathcal{G}$. Then with probability at least $1 - \delta$ and at most*

$$n = \widetilde{O}\left(\frac{1}{\varepsilon^2\gamma} \cdot (d_H + d_G) \cdot \log\left(\frac{1}{\varepsilon\delta}\right)\right) \tag{6}$$

samples from D, Algorithm 4 returns a randomized predictor $\boldsymbol{p}^*$ that is $(\mathcal{G}, \mathcal{H}, \varepsilon)$-step calibrated.

Proof. By Proposition C.9, there exists a universal constant $C > 0$ such that for $T \ge C \cdot \varepsilon^{-2}\gamma^{-1}$, the learner's regret is $\text{Reg}_{\text{weak}}(p^{(1:T)}, \{\mathcal{R}_{\ell^{(t)}}(\cdot)\}^{(1:T)}) \le T\varepsilon\sqrt{\gamma}/12$. By Proposition C.1, there exists a universal constant $C' > 0$ such that for $T \ge C' \cdot \varepsilon^{-2}\gamma^{-1} \cdot (d_H + d_G) \cdot \log\left(\varepsilon^{-1}\right)$, the adversary's regret is $\text{Reg}(q^{(1:T)}, \{1 - \ell_{(\cdot)}(p^{(t)}, (x^{(t)}, y^{(t)}))\}^{(1:T)}) \le T\varepsilon\sqrt{\gamma}/12$. By $\ell_{(\cdot)}(p^{(t)}, (x^{(t)}, y^{(t)}))$, we mean the adversary chooses the loss $\ell \in \widehat{O}^{\gamma}_{\text{sc}}$ to plug into the cost function of the given form.

It remains to be shown that the adversary's regret to cost functions of the form $1 - \ell_{(\cdot)}(p^{(t)}, (x^{(t)}, y^{(t)})$ closely approximates their regret to cost functions of the form $1 - \mathcal{R}_{(\cdot)}(p^{(t)})$. By Proposition C.2, there exists a universal constant $C'' > 0$ such that for $T \ge C'' \cdot \varepsilon^{-2}\gamma^{-1} \cdot (d_H + d_G) \cdot \log\left(\varepsilon^{-1}\delta^{-1}\right)$,

$$\left|\text{Reg}(q^{(1:T)}, \{1 - \ell_{(\cdot)}(p^{(t)}, (x^{(t)}, y^{(t)}))\}^{(1:T)}) - \text{Reg}(q^{(1:T)}, \{1 - \mathcal{L}_{\ell^{(t)}}(\cdot)\}^{(1:T)})\right| \le T\varepsilon\sqrt{\gamma}/12$$

with probability $1 - \delta/3$. Invoking Proposition C.2 once again, there exists a universal constant $C''' > 0$ such that for $T \ge C''' \cdot \varepsilon^{-2}\gamma^{-1} \cdot (d_H + d_G) \cdot \log\left(\varepsilon^{-1}\delta^{-1}\right)$,

$$\left|\text{Reg}(q^{(1:T)}, \{1 - \mathcal{L}_{\ell^{(t)}}(\cdot)\}^{(1:T)}) - \text{Reg}(\ell^{(1:T)}, \{1 - \mathcal{L}_{\ell^{(t)}}(\cdot)\}^{(1:T)})\right| \le T\varepsilon\sqrt{\gamma}/12$$

with probability $1 - \delta/3$.

By the triangle inequality, the adversary's regret $\text{Reg}(\ell^{(1:T)}, \{1 - \mathcal{L}_{\ell^{(t)}}(\cdot)\}^{(1:T)}) \le T\varepsilon\sqrt{\gamma}/4$. By Proposition C.8, the predictor $\boldsymbol{p}^*$ is a randomized $\varepsilon\sqrt{\gamma}/2$-solution to the $(D, \widehat{O}^{\gamma}_{\text{sc}}, \mathcal{P})$-multi-objective learning problem. By Corollary C.11, the randomized predictor $\boldsymbol{p}^*$ is $(\mathcal{G}, \mathcal{H}, O(\varepsilon))$-step calibrated.

Since the algorithm requires one sample per iteration, the sample complexity is exactly T. □

D Deferred Results and Proofs from Section 5

In this section, we prove that deterministic panprediction implies deterministic multi-group learning, and that randomized panprediction implies randomized multi-group learning.

Proposition 5.4. *Fix the zero-one loss ℓ, groups $\mathcal{G}$, and binary hypothesis class $\mathcal{H}$. If a deterministic predictor $p^* : \mathcal{X} \to [0,1]$ is a $(\{\ell\}, \mathcal{G}, \mathcal{H}, \varepsilon)$-panpredictor, then the binary classifier $h(x) = \mathbb{1}[p^*(x) \geq 0.5]$ is a deterministic $(\ell, \mathcal{G}, \mathcal{H}, \varepsilon)$-multi-group learner.*

Proof of Proposition 5.4. Using the panprediction guarantee, the loss of any competitor hypothesis $h \in \mathcal{H}$ on any group $g \in \mathcal{G}$ can be bounded as

$$\mathop{\mathbb{E}}_{(x,y)\sim D}[\ell(h(x), y) \mid g(x) = 1] \geq \mathop{\mathbb{E}}_{(x,y)\sim D}[\ell(k_\ell(p^*(x)), y) \mid g(x) = 1] - \varepsilon \cdot \sqrt{P_g^{-1}}.$$

It suffices to observe that $k_\ell(p^*(x)) = \mathbb{1}[p^*(x) \geq 0.5]$. □

Proposition D.1. *Fix the zero-one loss ℓ, groups $\mathcal{G}$, and binary hypothesis class $\mathcal{H}$. If a randomized predictor $\boldsymbol{p}^* \in \Delta(\mathcal{P})$ is a $(\{\ell\}, \mathcal{G}, \mathcal{H}, \varepsilon)$-panpredictor, then the binary classifier $h(x) = \mathbb{1}[\boldsymbol{p}^*(x) \geq 0.5]$ is a randomized $(\ell, \mathcal{G}, \mathcal{H}, \varepsilon)$-multi-group learner.*

The proof mirrors the previous argument and is omitted.

E Additional Connections

In this section, we show that multi-group learning cannot be reduced to multiaccuracy.

For a counterexample, consider the setting where $\mathcal{X} = \{x, x'\}$ and and $\mathcal{G} = \{\mathcal{X}\}$. Suppose that under distribution D, $\Pr(x) = \Pr(x') = 1/2$ and $\Pr(y = 0 \mid x) = \Pr(y = 1 \mid x') = 1$. Now consider the hypotheses $\mathcal{H} = \{h, h'\}$ where $h(x) = 0$ and $h(x') = 1$, and $h'(x) = 1$ and $h'(x') = 0$. Both hypotheses are 0-multiaccurate since

$$\left|\mathop{\mathbb{E}}_{(x,y)\sim D}[(y - h(x))]\right| = \left|\frac{1}{2} \cdot (1-1) + \frac{1}{2} \cdot (0-0)\right| = 0,$$
$$\left|\mathop{\mathbb{E}}_{(x,y)\sim D}[(y - h'(x))]\right| = \left|\frac{1}{2} \cdot (0-1) + \frac{1}{2} \cdot (1-0)\right| = 0.$$

But h' cannot be a multi-group learner for any $\varepsilon \in (0,1)$.

This counterexample can be eliminated by further conditioning on the level sets of the hypotheses, which naturally motivates notions of calibration. This shows that calibration can be useful for multi-group learning, even when we do not seek omniprediction-style guarantees.