Proceedings of Machine Learning Research vol 247:1–44, 2024 37th Annual Conference on Learning Theory

Omnipredictors for regression and the approximate rank of convex functions

Parikshit Gopalan PARIKG@APPLE.COM
Apple

Princewill Okoroafor PCO9@CORNELL.EDU
Cornell University

Prasad Raghavendra RAGHAVENDRA@BERKELEY.EDU
UC Berkeley

Abhishek Shetty SHETTY@BERKELEY.EDU
UC Berkeley

Mihir Singhal MIHIRS@BERKELEY.EDU
UC Berkeley

Editors: Shipra Agrawal and Aaron Roth

Abstract

Consider the supervised learning setting where the goal is to learn to predict labels $\mathbf{y}$ given points $\mathbf{x}$ from a distribution. An *omnipredictor* for a class $\mathcal{L}$ of loss functions and a class $\mathcal{C}$ of hypotheses is a predictor whose predictions incur less expected loss than the best hypothesis in $\mathcal{C}$ for every loss in $\mathcal{L}$. Since the work of Gopalan et al. (2021) that introduced the notion, there has been a large body of work in the setting of binary labels where $\mathbf{y} \in \{0,1\}$, but much less is known about the regression setting where $\mathbf{y} \in [0,1]$ can be continuous. The naive generalization of the previous approaches to regression is to predict the probability distribution of y, discretized to ε-width intervals. The running time would be exponential in the size of the output of the omnipredictor, which is $1/\varepsilon$.

Our main conceptual contribution is the notion of *sufficient statistics* for loss minimization over a family of loss functions: these are a set of statistics about a distribution such that knowing them allows one to take actions that minimize the expected loss for any loss in the family. The notion of sufficient statistics relates directly to the approximate rank of the family of loss functions. Thus, improved bounds on the latter yield improved runtimes for learning omnipredictors.

Our key technical contribution is a bound of $O(1/\varepsilon^{2/3})$ on the ϵ-approximate rank of convex, Lipschitz functions on the interval $[0,1]$, which we show is tight up to a factor of $\mathrm{polylog}(1/\epsilon)$. This yields improved runtimes for learning omnipredictors for the class of all convex, Lipschitz loss functions under weak learnability assumptions about the class $\mathcal{C}$. We also give efficient omnipredictors when the loss families have low-degree polynomial approximations, or arise from generalized linear models (GLMs). This translation from sufficient statistics to faster omnipredictors is made possible by lifting the technique of loss outcome indistinguishability introduced by Gopalan et al. (2023a) for Boolean labels to the regression setting.

Keywords: omniprediction, regression, approximate rank, loss outcome indistinguishability

1. Introduction

Loss minimization is the dominant paradigm for training machine learning models. In the supervised learning setting, given a distribution $\mathcal{D}^*$ on point-label pairs (which we refer to as nature's distribution), we pick a family of hypotheses $\mathcal{C}$, a loss function ℓ, and find the hypothesis from $\mathcal{C}$ that minimizes the expected loss over the distribution. This reduces the task of learning to an optimization problem over a parameter space. While this recipe has proven extremely successful, one can ask whether it adequately models a process as complex as learning.

A weakness of this paradigm is that learning is not robust to the choice of loss function. Different losses result in different optimization problems (which must be solved afresh), and hence typically different optimal hypotheses. One would imagine that each time we minimize a different loss, we learn something new about nature. Is there a universal and rigorous way to synthesize all that we learn into a single model that describes our complete understanding of nature, and does well on all of these losses? Standard loss minimization does not provide a solution for this goal.

Quite often, the exact loss function is not known *a priori*. To illustrate this, we present a simple scenario here. Suppose that a retailer is building a model to forecast demand for an item in each of its stores.

- The retailer has a feature vector $\mathbf{x}$ associated with each store, such as geographical location, foot traffic, which they use to forecast the demand $p(\mathbf{x})$ for the item in the store. Based on the forecast $p(\mathbf{x})$, they decide how much of the item to stock up, which is a number $t \in [0, 1]$.
- The realized demand is given by $\mathbf{y}^* \in [0, 1]$ which represents how much demand for the item there actually was. We assume a joint distribution on $(\mathbf{x}, \mathbf{y}^*)$, but for each $\mathbf{x}$ we only see a single draw $\mathbf{y}^*$ from the joint distribution.
- Assume that the retailer sells the item at a fixed retail price of \$1 per unit. If the retailer procures the item at a wholesale price per unit c that is determined by the market, and can fluctuate day to day. The loss incurred by the retailer is given by $\ell^c(\mathbf{y}^*, t) = c \cdot t - 1 \cdot \min(t, \mathbf{y}^*)$

The key observation is that the exact loss function ℓ^c depends on the wholesale price per unit c which may be unknown *a priori* and probably fluctuates over time. At the time of training the model, the forecaster knows the general shape of the loss function family, but not the exact loss they need to minimize.

The stylized scenario described here is just one example of a recurring theme in applications of forecasting, where the true loss functions are not known *a priori*. This can occur because the loss functions depend on parameters that are not fixed yet. Alternatively, the same forecasts may be used in many different settings, each of which requires its own distinct loss function. This raises the question, can we have forecasts which are guaranteed to do well, as measured by any loss drawn from a broad family?

Omniprediction. This motivated the study of omniprediction, initiated in the work of Gopalan, Kalai, Sharan, Reingold and Wieder Gopalan et al. (2021). We will now formally describe the notion of omnipredictors.

- *Point and label distribution:* As in supervised learning, the central object being learnt is specified by the *nature's distribution*, which is a joint distribution $\mathcal{D}^*$ over points $\mathbf{x} \in \mathcal{X}$ and corresponding labels $y^* \in \mathcal{Y}$.

In the demand forecasting example, $\mathbf{x}$ will be the vector of features for each store. The label $\mathbf{y}^*$, a real number in $[0, 1]$, is the demand for the item in the store $\mathbf{x}$, hence $\mathcal{Y} = [0, 1]$.

- *Actions and Loss families:* The agent, who was the retailer in our running example, intends to use the output of a prediction algorithm towards selecting from a set of actions $\mathcal{A}$.

  The loss incurred is a function of the true label and the action chosen, i.e., the loss function family is specified as $\mathcal{L} = \{\ell : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}\}$ a family of loss functions. In this work, the labels y^* would be real valued and therefore $\mathcal{Y} := [0, 1]$.

  In our running example, the action space consisted of how much of the item to stock up (denoted by $t \in [0, 1]$), and the retailer wishes to choose an optimal value for it. The family of loss functions $\mathcal{L}$ were given by $\ell^c(y, t) = ct - \min(t, y)$ for $c \in [0, 1]$.

- *Predictions and optimal actions:* An omnipredictor consists of two efficiently computable functions.

  - *Prediction function:* $p : \mathcal{X} \to \mathcal{P}$
    Given an input label $\mathbf{x}$, an omnipredictor outputs a prediction $p(\mathbf{x})$ in some range $\mathcal{P}$, which we will specify shortly. The output $p(\mathbf{x})$ of the omnipredictor should be thought of as its prediction of the conditional distribution of the label $\mathbf{y}^*|\mathbf{x}$.
  - *Post-processing function:* $k : \mathcal{P} \times \mathcal{L} \to \mathcal{A}$
    Given the prediction $p(\mathbf{x})$ and a loss function $\ell \in \mathcal{L}$ in the family, outputs a predicted action $k(p(\mathbf{x}), \ell)$ for the agent.

  For example, an omnipredictor could simply output a distribution $\mathcal{D}$ over $[0, 1]$, which is its prediction for the conditional distribution for $\mathbf{y}|\mathbf{x}$. In our running example, $\mathcal{D}$ would be its prediction of the probability distribution of demand in the store $\mathbf{x}$. In this case, the range $\mathcal{P}$ is the space of all probability distributions over $[0, 1]$. Then, the optimal action on a given loss function $\ell \in \mathcal{L}$ is given by,

  $$k(\mathcal{D}, \ell) = \arg \min_{\theta \in [0,1]} \mathbb{E}_{y \in \mathcal{D}}[\ell(y, \theta)]$$

  Predicting the entire distribution is not succinct for real-valued labels. [1] One of the major thrusts of our work will be to find more succinct descriptions of the distribution.

  Crucially, the omnipredictor is trained once and for all without knowing a specific loss. Although the post-processing depends on the loss, it does not require further learning or access to the data set. For instance, take the Boolean setting when $\mathbf{y}^* \in \{0, 1\}$, and the predictor $p(\mathbf{x}) \in [0, 1]$ is an estimate for $\mathbb{E}[\mathbf{y}^*|\mathbf{x}]$. For the ℓ_1 loss $\ell_1(y, t) = |y - t|$, we take the action 1 if $p(x) > 1/2$ and 0 otherwise, while for the squared loss, it is the identity function.

- *Performance guarantee:* For both computational and information-theoretic reasons, it is often infeasible to even estimate how far the recommended actions of an omnipredictor are from optimal. This motivates defining a guarantee for the performance of an omnipredictor relative to the best hypothesis from a concept class.

  Fix a concept class of hypotheses, $\mathcal{C} = \{c : \mathcal{X} \to \mathcal{A}\}$ that given the features outputs a recommended action. An $(\mathcal{L}, \mathcal{C})$-omnipredictor is one whose expected losses under the distribution $\mathcal{D}$

1. We do not make parametric assumptions about the distribution of $\mathbf{y}^*|\mathbf{x}$.

compete with the best hypothesis in $\mathcal{C}$ for any loss $\ell \in \mathcal{L}$. The power of this guarantee comes from the fact that prediction algorithm p makes predictions without knowing the loss function ℓ. Yet, these predictions (with the right post-processing function k) can compete against the benchmark

$$\min_{c \in \mathcal{C}} \mathbb{E}_{(\mathbf{x},\mathbf{y}^*)\sim\mathcal{D}}[\ell(y^*, c(\mathbf{x}))]$$

which is very much dependent on the choice of ℓ.

Omniprediction in the Boolean setting The work of Gopalan et al. (2021) which introduced the notion of omniprediction, studied the Boolean setting where $\mathbf{y}^* \in \{0,1\}$. Their starting point is the observation that if we could learn the conditional distribution $\mathbf{y}^*|\mathbf{x}$, then subsequently we could take actions that optimize any loss function, without any further learning or access to the data. Since the labels are Boolean, the conditional distribution $\mathbf{y}^*|\mathbf{x}$ is fully described by a single number, namely $p^*(\mathbf{x}) = \mathbb{E}[\mathbf{y}^*|\mathbf{x}]$, and this is what our predictor attempts to predict.

Learning p^* is not feasible in general, for computational and information-theoretic reasons. Yet, Gopalan et al. (2021) showed that one can efficiently learn $(\mathcal{L}_{\text{cvx}}, \mathcal{C})$ omnipredictors where $\mathcal{L}_{\text{cvx}}$ is the family of convex, Lipschitz loss functions for all $\mathcal{C}$ that satisfy a basic learnability condition called weak agnostic learnability. They show this via a surprising connection to a multigroup fairness notion known as multicalibration, introduced by Hebert-Johnson, Kim, Reingold and Rothblum Hébert-Johnson et al. (2018). There has since been a large body of work on this topic, giving omnipredictors for other classes of loss functions Gopalan et al. (2023a); Dwork et al. (2023), for constrained optimization Hu et al. (2023); Globus-Harris et al. (2023a), for other prediction scenarios Garg et al. (2023); Kim and Perdomo (2023) and proposing stronger notions Gopalan et al. (2023b). Most of this work considers either the Boolean setting or the multiclass setting (see the discussion of related work).

1.1. Omnipredictors for regression

In this work, we present a comprehensive theory of omniprediction for the challenging regression setting where the labels $\mathbf{y}^* \in [0,1]$ are allowed to be continuous. We briefly describe our main contributions. [2]

The natural motivation for considering continuous labels comes from the fact that many real-life forecasting tasks involve predicting real-valued attributes: the amount of rain tomorrow, the price of a stock next week, the temperature of a city in ten years. However, extending prior results on omniprediction to the regression setting has proved challenging (the only prior result we are aware of comes from (Gopalan et al., 2021, Section 8)). Indeed, several techniques used in prior work, do not generalize to the more complex continuous setting, see the related work section for a detailed discussion.

Sufficient statistics: To develop a theory of omnipredictors for regression, the first question to answer is: *what information should the prediction $p(\mathbf{x})$ convey about the distribution of the label?* This question has been studied recently in the literature, most relevant to us are the works of Jung et al. (2021) and Dwork et al. (2022). The answer naturally depends on what the predictions are

2. There is ambiguity in the use of the term regression in the literature. In our paper, we will use the term regression to mean that the variable $\mathbf{y}$ is continuous, the loss can arbitrary. In the literature, regression can sometimes refer to (certain) loss minimization problems where $\mathbf{y} \in \{0,1\}$, as in logistic regression.

meant to achieve. For omniprediction, the predictions should enable expected loss minimization for any loss function drawn from the family $\mathcal{L}$ of loss functions.

The naive solution would require that $p(\mathbf{x})$ reveals the conditional probability distribution $\mathbf{y}^*|\mathbf{x}$. But since $\mathbf{y}^*$ is a continuous random variable, the distribution $\mathbf{y}^*|\mathbf{x}$ may not have a finite description.

For an arbitrary distribution $\mathcal{D}$ over $[0,1]$, we use the term *"statistic"* to refer to the expectation $\mathbb{E}_{\mathbf{y}\sim\mathcal{D}}[s(\mathbf{y})]$ of a bounded function $s:[0,1]\to[-1,1]$. For reasons that will be clear later, we limit ourselves to this class of statistics called *linearizing statistics* by Dwork et al. (2022). A natural alternative (considered in Jung et al. (2021); Dwork et al. (2022)) would be to have the predictions be the statistics associated with the distribution $\mathbf{y}^*|\mathbf{x}$. This raises the question of when a family of statistics is *sufficient* for omniprediction for a loss family $\mathcal{L}$. We abstract the requirement in the following definition (the notation ignores the conditioning on $\mathbf{x}$):

A family of statistics $\mathcal{S}$ are said to be sufficient for the loss family $\mathcal{L}$ if for any loss $\ell\in\mathcal{L}$ and distribution $\mathcal{D}$ over $[0,1]$, the value of the statistics $\mathcal{S}$ for distribution $\mathcal{D}$ determine the (near) optimal action that (approximately) minimizes the loss. In other words

$$k(\mathcal{D},\ell)=\arg\min_{t\in\mathcal{A}}\mathbb{E}_{\mathbf{y}\sim\mathcal{D}}[\ell(\mathbf{y},t)],$$

is determined by the statistics $\mathcal{S}$ of distribution $\mathcal{D}$.

If there is a (small) set of *sufficient statistics* $\mathcal{S}$ for the family $\mathcal{L}$, then our omnipredictor would try and predict these statistics for every $\mathbf{x}$. For Lipschitz loss functions, there is a simple set of $1/\varepsilon$ sufficient statistics: the probabilities of the events $\mathbf{y}\in[i\varepsilon,(i+1)\varepsilon)$ for every i. We call these the CDF statistics, since they tell us the CDF of $\mathbf{y}$ to within accuracy ε. This gives sufficient information to minimize the expected loss over actions for any Lipschitz loss function within an additive ε. This was the approach taken in (Gopalan et al., 2021, Section 8), which treats the problem of predicting these probabilities as a multiclass labeling problem with $1/\varepsilon$ labels.

For specific families of loss functions, one could hope to get more succinct statistics. For instance, consider the family of ℓ_p loss functions for even p, given by $\{\ell_p(y,t)=(y-t)^p\}$. Since this space is spanned by the monomials $\mathcal{S}=\{s_i(y)=y^i\mid i=0,\ldots,p\}$, the first p moments are sufficient statistics for this family.

This naturally raises the question: what is the smallest set of sufficient statistics for a family $\mathcal{L}$ of losses? Let us see why it holds the key to more efficient algorithms for omniprediciton.

Omniprediction from indistinguishability: The work of Gopalan et al. (2023a) gives a template for establishing omniprediction by establishing a stronger condition they call loss outcome indistinguishability, which is inspired by the notion of outcome indistinguishability introduced by Dwork et al. (2021). Specifically, they show that in the Boolean case, $(\mathcal{L},\mathcal{C})$-omniprediction against a class of loss functions $\mathcal{L}$ and concept class $\mathcal{C}$ is implied the following properties of the prediction function p.

1. Calibration: Conditioned on a prediction $p(\mathbf{x})\in[0,1]$, the expectation is close to the predicted values, i.e., $\mathbb{E}[\mathbf{y}^*|p(\mathbf{x})]\approx_\epsilon p(\mathbf{x})$.

2. Multiaccuracy: the error in prediction $p(\mathbf{x})-\mathbf{y}^*$ is uncorrelated with a class of tests derived from $\mathcal{L}$ and $\mathcal{C}$.

The proof is via an indistinguishability argument. They show that one can replace nature's labels $\mathbf{y}^*$ with labels $\tilde{\mathbf{y}}$ from a simulation that corresponds to the predictor's predictions, without

much change in the loss suffered by either the omnipredictor or a hypothesis from $\mathcal{C}$. The predictor p is Bayes optimal for the simulation, and hence it is an omnipredictor. Indistinguishability lets us conclude that it is also an omnipredictor for nature's distribution.

Lifting the above result to the regression poses several challenges, which are detailed in Section C, primarily that there is ambiguity in defining what the simulation being predicted by our predictor is, unlike in the Boolean case. Yet we are able to prove a qualitatively similar statement. If a predictor p for a family of *sufficient* statistics $\mathcal{S}$ is calibrated, and is multiaccurate with respect to an associated class of tests derived from $\mathcal{L}, \mathcal{S}$ and $\mathcal{C}$, then it is an $(\mathcal{L}, \mathcal{C})$-omnipredictor. We defer the formal statement of this theorem to Theorem 28.

Motivated by this sufficient condition for $(\mathcal{L}, \mathcal{C})$-omniprediction, we generalize the calibrated multiaccuracy algorithm of Gopalan et al. (2023a) to work in the setting of regression in Theorem 32. The running time is exponential in the size d of the family of sufficient statistics $\mathcal{S}$, arising from the need to ensure that our predictions, which take values in $[-1, 1]^d$ are calibrated. Consequently, shrinking the size of the family of sufficient statistics results in drastic reductions in running time.

Approximate rank & sufficient statistics: What is the smallest family of sufficient statistics for a given family of loss functions $\mathcal{L}$? The answer is directly related to the so-called "ε-approximate dimension" of a family of functions derived from the loss family $\mathcal{L}$.

Definition 1 *Given a family of functions $\mathcal{F} = \{\ell : [0, 1] \to [-1, 1]\}$, their ε-approximate dimension denoted by $\dim_\varepsilon(\mathcal{F})$, is the smallest dimension of a subspace of functions $\mathcal{V}$ such that for every $f \in \mathcal{F}$, there exists $\hat{f} \in V$ which is an ϵ-approximation to f in the ℓ_∞ norm, i.e., $\|f - \hat{f}\| \leq \epsilon$.*

The notion of ε-approximate rank has been studied in the literature, with motivations ranging from communication complexity to approximate Nash equilibria Alon (2009); Lee and Shraibman (2009b,a); Alon et al. (2013).

Given a family of loss functions $\mathcal{L} = \{\ell : [0, 1] \times \mathcal{A} \to \mathbb{R}\}$, consider the function family $\mathcal{L}_t$ obtained by fixing the actions, i.e.,

$$\mathcal{L}_t = \{\ell_t := \ell(\cdot, t) \mid \ell \in \mathcal{L}\}$$

Suppose there is a basis of functions $\mathcal{S}$ that uniformly approximates the function family $\mathcal{L}_t$. Then for any distribution $\mathcal{D}$ over $[0, 1]$, loss function $\ell \in \mathcal{L}$, and action t, the expected loss $\mathbb{E}_{\mathbf{y}\sim\mathcal{D}}[\ell(t, \mathbf{y})]$ can be approximately estimated from the expectations of statistics $\{\mathbb{E}_{\mathbf{y}\sim\mathcal{D}}[s(\mathbf{y})]\}$. This allows us to choose the best action for each ℓ (at least information-theoretically), as required by the definition of sufficient statistics. Therefore there is a tight connection between *sufficient statistics for loss family* $\mathcal{L}$ and the ϵ-approximate dimension of the corresponding function family $\mathcal{L}_t$: an ε-approximate basis for the latter gives us functions whose expectations are sufficient statistics for the former.

Thus upper bounds on approximate dimension of loss families lead to upper bounds on the complexity of learning omnipredictors. Our next contribution is to show that many natural loss families admit non-trivial uniform approximations.

Approximate rank of convex Lipschitz functions: A recurring property of loss functions that arise in a myriad of contexts is *convexity*. For example, the loss family in the example of demand

forecasting for a retailer were convex functions over $[0,1]$. This makes it especially important to understand the approximate dimension of the space of convex Lipschitz functions over $[0,1]$ [3].

In the absence of convexity, if one considers the family of all Lipschitz functions denoted $\mathcal{F}_{\text{lip}}$, it is easy to show that $\dim_\varepsilon(\mathcal{F}_{\text{lip}}) = \Theta(1/\epsilon)$. The upper bound follows by a straightforward basis consisting of indicators of intervals $1[y \geq i\epsilon]$ for $i \in \{0, 1, \ldots, 1/\varepsilon\}$. Using a linear algebraic argument, one can show that $1/\varepsilon$ statistics are indeed necessary.

It is natural to ask if convexity leads to better approximations to the functions. Our main technical result shows that the answer is yes, and in fact, we exhibit construction of a set of $\tilde{O}(1/\varepsilon^{2/3})$ statistics that suffice for every function in $\mathcal{F}_{\text{cvx}}$, the family of bounded, Lipschitz, convex functions on $[0,1]$. Moreover this bound is essentially tight: we show a lower bound of $\Omega(1/\varepsilon^{2/3})$ which holds even for the family $\{\text{ReLU}_{i\varepsilon}(y)\}_{i=1}^{1/\varepsilon}$ which is a subset of $\mathcal{F}_{\text{cvx}}$.

An interesting implication is that the number of statistics required to approximate (the expectations of) ℓ_1 losses of the form $|y-t|$ is very different from the number required for ℓ_2 losses $(y-t)^2$. We show that the former require $\Omega(1/\varepsilon^{2/3})$ statistics, whereas for the latter we only need a constant number of statistics: $\mathbb{E}[y]$, and $\mathbb{E}[y^2]$ suffice.

Omnipredictors for loss families Using the technical above, we give the first efficient omnipredictors for several important families of loss functions.

- A main application of our result on the approximate rank of convex Lipschitz functions is an omnipredictor for the family of all Lipschitz, convex loss functions in $\mathbf{y}$ which we refer to as $\mathcal{L}_{\text{cvx}}$. We show in Theorem 34 that a predictor that is calibrated for a family of statistics that arise from our approximation theorems and is multiaccurate with respect to bounded postprocessings of $\mathcal{C}$ is a $(\mathcal{L}_{\text{cvx}}, \mathcal{C}, \epsilon)$-omnipredictor and can be computed in time $\exp(\tilde{O}(\epsilon^{-2/3}))$ time. This is a significant improvement over the $\exp(\tilde{O}(\epsilon^{-1}))$ time algorithm one would arrive at by predictor that is calibrated with respect to the CDF statistics. The result of (Gopalan et al., 2021, Section 8) give a running time of $\exp(O(1/\varepsilon))$ again using CDF statistics, but with the requirement that the loss is convex in t (it need not be convex in y). Their result requires multicalibration for $\mathcal{C}$, whereas we require calibrated multiaccuracy for postprocessings of $\mathcal{C}$.

- For the class of ℓ_p for even $p \leq d$, we show, in Theorem 38, that calibration with respect to the moment statistics of degree $\tilde{O}(\sqrt{d})$ and multiaccuracy with respect to polynomials postprocessings of $\mathcal{C}$ leads to an omnipredictor. This leads to an omnipredictor that runs in time $(1/\epsilon)^{\tilde{O}(\sqrt{d})}$ which improves upon the naive $(1/\epsilon)^d$ algorithm that predicts the first d moments.

- For the class of losses corresponding to generalized linear models with respect to a family of statistics $\mathcal{S}$, we show, in Theorem 41, that a predictor p that is calibrated with respect to $\mathcal{S}$ and multiaccurate with respect to $\mathcal{C}$ is an omnipredictor. This leads to a $(1/\epsilon)^d$ time algorithm for producing an omnipredictor. This result should be contrasted with the earlier results due to the fact we need only access to a weak learner for the original class $\mathcal{C}$ as opposed to postprocessings of it. This generalizes a result of Gopalan et al. (2023a) in the Boolean setting.

3. Lipschitzness is a natural constraint here to make the question of ϵ-approximations invariant to scaling.

1.2. Overview of technical contributions

In this section, we highlight the main new technical contributions of this work. We first discuss the approximate rank of convex functions, and then our generalization of loss outcome indistinguishability to real valued labels.

1.2.1. APPROXIMATING UNIVARIATE CONVEX FUNCTIONS.

Recall that $\mathcal{F}_{\text{cvx}}$ denotes the space of convex 1-Lipschitz functions on the interval $[0, 1]$. We prove our bound on the approximate rank by a sequence of reductions. Let $g \in \mathcal{F}_{\text{cvx}}$ be a convex function that we wish to approximate uniformly.

Reduction to discrete functions. We place a δ-grid on $[0, 1]$ and use the piecewise linear approximation to g to reduce the problem to a discrete problem of approximating functions $f : [m] \to \mathbb{R}$ where $m = 1/\delta$. The Lipschitzness translates to the fact that the first finite differences of these functions are bounded, and convexity corresponds to positivity of the second finite difference.

Reduction to the ReLU functions. The ReLU family of functions mapping $[m]$ to $\mathbb{Z}$ is defined as $\mathrm{ReLU}_i(x) = 0$ for $x < i$ and $x - i$ for $x \geq i$. We prove a discrete Taylor theorem to show that η-uniform approximations to these functions imply $O(\eta)$ uniform approximations to all functions from $\mathcal{F}_{\text{cvx}}$. This step is similar in spirit to (Kleinberg et al., 2023, Theorem 8), which implies that the family of functions $|x - i|$ is a basis for all convex functions with small ℓ_1 norm.[4]

From ReLU to intervals. If we were to form a basis which simply contained all these ReLU_i, then we would end up with a basis of size m, which has the same size as the trivial basis of size $1/\delta$. It turns out, though, that it is possible to approximate the ReLU functions using a smaller basis. We first observe that

$$\mathrm{ReLU}_a(x) - \mathrm{ReLU}_{a+1}(x) = 1[x \geq a],$$

where $1[x \geq a]$ is the indicator of $\mathbf{x} \geq a$, or equivalently the indicator function of the interval $[a, m]$. More generally, the differences between ReLU functions can be expressed as sums of indicators of intervals. It turns out that it is possible to effectively approximate the interval functions. Therefore, we use the following natural strategy to construct a basis approximating ReLU functions.

Our final basis will combine evenly spaced ReLU functions with a basis for the interval functions. Specifically, for $t = m^{1/3}$, we add $\mathrm{ReLU}_{i \cdot t}$ to the basis for each i. We take the union of this with a $(1/m^{1/3})$-approximate basis for all interval functions. Then, we can approximate any given ReLU function by starting with the ReLU function at the nearest multiple of t, and then adding approximations to interval functions. For the appropriate basis of interval functions, this will give the desired basis of size $\tilde{O}(m^{2/3}) = \tilde{O}(1/\delta^{2/3})$ for all convex functions.

Approximating intervals. The final step is to approximate all interval functions. By the dyadic decomposition of intervals, it suffices to consider only dyadic intervals. For simplicity, consider all intervals containing a single point. A low rank approximations to all such functions is equivalent to a low rank approximation to the $m \times m$ identity matrix. Approximate low-rank factorizations of the identity matrix arise in the context of Johnson-Lindenstrauss lemma. They can be explicitly constructed using codewords from a binary code of distance $1/2 - \mu$ and rate $\Omega(1/\mu^2)$. It is known that an ε-approximation can be obtained using rank $\log(m)/\varepsilon^2$. Matching lower bounds on the rank are proved by Alon Alon (2009).

4. Their motivation, which is quite different form ours, is from the new notion of U-calibration that they define.

1.2.2. LOSS OUTCOME INDISTINGUISHABILITY FOR PREDICTING STATISTICS

Outcome indistinguishability (OI) was introduce in Dwork et al. (2021) for the Boolean setting, and generalized to regression in Dwork et al. (2022). The work of Gopalan et al. (2023a) connected it to omniprediction in the Boolean seeting, introducing the notion of loss OI. The generalization of loss OI Gopalan et al. (2023a) to real values is not straightforward. For the sake of concreteness, assume that the statistics we predict are the first d moments $\{y^i\}_{i\in[d]}$. In the Boolean setting, when we predict $p(\mathbf{x}) = 0.7$, it is clear that we mean $\mathbf{y}^*|\mathbf{x}$ is drawn from the Bernoulli distribution with parameter 0.7. When we predict the first d moments of a distribution, there might be many distributions matching those moments. Or there might not be any! The first d moments have to satisfy various moment inequalities, which our predictor would need to satisfy in order to make predictions that a realizable via some distribution. For the moments, there are indeed efficient (SDP-based) methods to ensure feasibility of predictions (See e.g. (Schmüdgen, 2020, Theorem 3.1)). For other families of statistics $\mathcal{S}$, such characterizations might not exist or might be computationally infeasible.

We require our predictors to satisfy two conditions: calibration and multiaccuracy, as in Gopalan et al. (2023a). Calibration requires that conditioned on a prediction, the expectations are close to the predicted values, whereas multiaccuracy requires that the errors in prediction $p_i(x) - s_i(\mathbf{y}^*)$ are uncorrelated with a class of tests derived from $\mathcal{L}$ and $\mathcal{C}$.

For the analysis, we define a *simulation* distribution $(\mathbf{x}, \tilde{\mathbf{y}})$ where $\tilde{\mathbf{y}}|\mathbf{x} \sim \mathbf{y}^*|p(\mathbf{x})$. In a sense, this is the random variable whose statistics our predictor p predicts (with some error). This is different from Boolean setting Dwork et al. (2021); Gopalan et al. (2023a), where the simulation is based on the predictor alone, and is independent of the distribution $\mathcal{D}^*$ that is being learnt. It is more reminiscent of the view of Gopalan et al. (2021) for the Boolean case, who view predictors as partitions of the space into level sets, with the canonical prediction which is the expectation over the level set. The simulation shows that calibration approximately solves the feasibility issue above, since if a predictor is α calibrated, then on expectation over $\mathcal{D}^*$, it holds that

$$|p_i(\mathbf{x}) - \mathbb{E}[s_i(\tilde{\mathbf{y}})]| = |p_i(\mathbf{x}) - \mathbb{E}[s_i(\mathbf{y}^*)|p(\mathbf{x})]| \leq \alpha.$$

With this definition in place, we can deduce omniprediction using a similar high-level strategy to the one used in Gopalan et al. (2023a): for any loss $\ell \in \mathcal{L}$, we show that the expected loss of the omnipredictor, where we make decisions based on $p(\mathbf{x})$ does not change if the labels are drawn from $\mathbf{y}^*$ or $\tilde{\mathbf{y}}$, nor does the expected loss suffered by any hypothesis in $\mathcal{C}$. The implementation departs from the Boolean case. There the first condition (called decision OI) is guaranteed by calibration alone, and the second (called hypothesis OI) by multiaccuracy. In our setting, we do not have similar decomposition. Showing that the expected loss for $c \in \mathcal{C}$ does not change much when we switch between $\tilde{\mathbf{y}}$ and $\mathbf{y}^*$ requires both calibration and multiaccuracy, this stems from the fact that our simulation is dependent on both p and the distribution $\mathcal{D}^*$.

Algorithm for calibrated multiaccuracy. We present an algorithm that achieves calibrated multiaccuracy for $\mathcal{S}$-predictors, assuming access to a suitable weak agnostic learner. This generalizes the algorithm from Gopalan et al. (2023a). Note that calibrated multiaccuracy is much weaker than multicalibration, and hence is more efficient to achieve. The running time for calibration is exponential in d, the number of statistics, since verifying if a d-dimensional predictor is calibrated requires $\exp(d)$ samples. Thus efficient algorithms crucially rely on the cardinality d of the sufficient statistics being small.

The multiaccuracy is for a family of tests $\mathcal{B} = \{f_\ell \circ c : c \in \mathcal{C}\}$ where $f_\ell : [-1, 1] \to [-1, 1]$ is a family of bounded post-processing functions derived from the loss family $\mathcal{L}$. The family of such tests also shows up in the work of Gopalan et al. (2023a), who refer to it as $\mathrm{level}(\mathcal{C})$. The reason is that it is a family of post-processing functions, so its level sets are (unions of) the level sets of $\mathcal{C}$. One can equivalently think of $\mathrm{level}(\mathcal{C})$ as the closure of $\mathcal{C}$ under post-processing functions. As observed by Gopalan et al. (2023a), when $\mathcal{C}$ is the family of decision trees of bounded size or a family of Boolean functions, $\mathcal{B}$ and $\mathcal{C}$ are essentially the same. Similarly, when the action space $\mathcal{A}$ is discrete, $\mathcal{B}$ is just a mapping of actions into real space. But for other hypotheses classes like low-degree polynomials, $\mathcal{B}$ might be richer than $\mathcal{C}$, so the problem of weak agnostic learning for it is harder.

It was already observed in the work of Gopalan et al. (2023a), Calibrated multiaccuracy is computationally much more efficient than multicalibration in. This difference is even more pronounced for statistic predictors. The exponential dependence on d in the running time for calibrated multiaccuracy arises from the need for calibration. The number of calls to the weak learner is (only) polynomial in d. By contrast, achieving multicalibration for a statistic predictor (as in Jung et al. (2021); Gopalan et al. (2021)) requires an exponential number of calls to the weak learner using the best known algorithms for multicalibration. Thus, this presents an improvement over using multicalibration (assuming weak learning is approximately equally difficult over $\mathcal{C}$ and $\mathcal{B}$).

1.3. Organization

We present our results on the approximate rank of convex functions in Appendix A. The construction is self-contained and does not use any machinery beyond JL matrices and does not require any knowledge of multigroup fairness or omniprediction. In Appendix B, we formally introduce the notion of sufficient statistics for families of loss functions and define calibration and multiaccuracy for statistic predictors. In Appendix C, we show how to obtain omniprediction from loss outcome indistinguishability of statistic predictors. In Appendix D, we obtain omniprediction guarantees for convex lipschitz losses, low-degree polynomials, and generalized linear models. In Appendix E, we present our algorithm for achieving calibrated multiaccuracy. Further discussion of related works can be found in Section 2.

2. Further Related work

Multi-group fairness. The fairness notions of multiaccuracy and multicalibration were introduced in the influential work of Hebert-Johnson, Kim, Reingold and Rothblum Hébert-Johnson et al. (2018), see also the work of Kleinberg et al. (2017); Kearns et al. (2018). There has been a large body of followup work, extending it to the regression setting Jung et al. (2021), to other notions of calibration Gopalan et al. (2022) and much more. Connections between multicalibration and boosting are established in the works of Gopalan et al. (2021); Globus-Harris et al. (2023b). The recent work of Blasiok et al. (2024) shows that multicalibration for neural networks can be obtained from squared loss minimization. The elegant work of Dwork et al. (2021) introduced the notion of outcome indistinguishability and related it to multigroup fairness notions of varying strength in the Boolean setting. This work was extended beyond the setting of Bernoulli labels by Dwork et al. (2022). Further connections between pseudorandomness and multigroup fairness were discovered in the work of Dwork et al. (2023), who also prove some new omniprediction results. Outcome in-

distinguishability was used to construct multigroup agnostic learners in the work of Rothblum and Yona (2021).

Omniprediction. The work of Gopalan et al. (2021) introduces the notion of omniprediction. The subsequent work of Gopalan et al. (2023a) brings the outcome indistinguishability perspective to omniprediction, introducing a general technique based on a simulated distribution that we generalize to the regression setting in this work. Reverse connections between generalizations of omniprediction and multigroup fairness notions were established in the work of Gopalan et al. (2023b). Omniprediction in a constrained setting where the predictor is required to satisfy other constraints which might be motivated for instance by fairness was considered in the works of Hu et al. (2023); Globus-Harris et al. (2023a). Omnipredictors for performative prediction were studied in Kim and Perdomo (2023). The problem of omniprediction in the online rather than the batch setting was recently studied in the work of Garg et al. (2023). The recent work of Gollakota et al. (2023) uses calibrated multiaccuracy to give the first agnostic learning guarantees for Single Index Models (SIMs), with respect to the squared loss.

Approximate rank. The notion of approximate rank of matrices arises naturally in communication complexity. Both the ϵ-error randomized and quantum communication complexities of a function are lower bounded by constant times the log of the approximate rank of the communication matrices (see (Lee and Shraibman, 2009a, Chapters 4 and 5)). It further turns out that this notion is closely connected to notions such as factorization norms Linial and Shraibman (2007); Lee and Shraibman (2009b), hereditary discrepancy Matousek et al. (2020) and sign rank Alon et al. (2013). In addition, the notion turns up in fundamental algorithmic problems such as density subgraph and approximate Nash equilibria. See Alon et al. (2013) and references therein for further discussion. Most of the mentioned works above focus on the case of sign matrices. Another area that is generally related to the notions considered in our work is the area of approximation theory (see Carothers (1998); Szego (1939)) but most work in the area focused on approximations in terms of functions families such as polynomials and rational functions. To our knowledge, our work is the first to study the notion of approximate rank (and thus approximation in terms of an arbitrary basis of functions) for bounded convex functions on $[0, 1]$.

References

Noga Alon. Perturbed identity matrices have high rank: Proof and applications. *Combinatorics, Probability and Computing*, 18(1-2):3–15, 2009. doi: 10.1017/S0963548307008917.

Noga Alon, Troy Lee, Adi Shraibman, and Santosh S. Vempala. The approximate rank of a matrix and its algorithmic applications: approximate rank. In Dan Boneh, Tim Roughgarden, and Joan Feigenbaum, editors, *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, pages 675–684. ACM, 2013.

Keith Ball. An elementary introduction to modern convex geometry. *Flavors of geometry*, 31(1–58): 26, 1997.

Jaroslaw Blasiok, Parikshit Gopalan, Lunjia Hu, Adam Tauman Kalai, and Preetum Nakkiran. Loss minimization yields multicalibration for large neural networks. In *Innovations in Theoretical Computer Science (ITCS) (to appear)*, 2024.

Neal L Carothers. A short course on approximation theory. *Bowling Green State University, Bowling Green, OH*, 1998.

W. J. Cody. A survey of practical rational and polynomial approximation of functions. *SIAM Review*, 12(3):400–423, 1970. ISSN 00361445. URL http://www.jstor.org/stable/2028556.

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

Cynthia Dwork, Michael P. Kim, Omer Reingold, Guy N. Rothblum, and Gal Yona. Beyond bernoulli: Generating random outcomes that cannot be distinguished from nature. In *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*, page 342–380. PMLR, Mar 2022. URL https://proceedings.mlr.press/v167/dwork22a.html.

Cynthia Dwork, Daniel Lee, Huijia Lin, and Pranay Tankala. From pseudorandomness to multigroup fairness and back. In *The Thirty Sixth Annual Conference on Learning Theory, COLT 2023, 12-15 July 2023, Bangalore, India*, volume 195 of *Proceedings of Machine Learning Research*, pages 3566–3614. PMLR, 2023.

Sumegha Garg, Christopher Jung, Omer Reingold, and Aaron Roth. Oracle efficient online multicalibration and omniprediction. volume abs/2307.08999, 2023. doi: 10.48550/ARXIV.2307.08999. URL https://doi.org/10.48550/arXiv.2307.08999.

Ira Globus-Harris, Varun Gupta, Christopher Jung, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Multicalibrated regression for downstream fairness. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2023*, pages 259–286. ACM, 2023a.

Ira Globus-Harris, Declan Harrison, Michael Kearns, Aaron Roth, and Jessica Sorrell. Multicalibration as boosting for regression. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 11459–11492. PMLR, 2023b.

Aravind Gollakota, Parikshit Gopalan, Adam R. Klivans, and Konstantinos Stavropoulos. Agnostically learning single-index models using omnipredictors. In *NeurIPS*, 2023.

Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors, 2021.

Parikshit Gopalan, Michael P. Kim, Mihir Singhal, and Shengjia Zhao. Low-degree multicalibration. In *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 3193–3234. PMLR, 2022.

Parikshit Gopalan, Lunjia Hu, Michael P. Kim, Omer Reingold, and Udi Wieder. Loss minimization through the lens of outcome indistinguishability, 2023a.

Parikshit Gopalan, Michael P. Kim, and Omer Reingold. Characterizing notions of omniprediction via multicalibration. *CoRR*, abs/2302.06726, 2023b. URL https://doi.org/10.48550/arXiv.2302.06726.

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.

Lunjia Hu, Inbal Rachel Livni Navon, Omer Reingold, and Chutong Yang. Omnipredictors for constrained optimization. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 13497–13527. PMLR, 2023.

Michael I Jordan. An introduction to probabilistic graphical models, 2003.

Christopher Jung, Changhwa Lee, Mallesh M. Pai, Aaron Roth, and Rakesh Vohra. Moment multicalibration for uncertainty estimation. In *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pages 2634–2678. PMLR, 2021.

Michael J. Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2569–2577. PMLR, 2018.

Michael P. Kim and Juan C. Perdomo. Making decisions under outcome performativity. In *14th Innovations in Theoretical Computer Science Conference, ITCS 2023*, volume 251 of *LIPIcs*, pages 79:1–79:15, 2023.

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

Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In Christos H. Papadimitriou, editor, *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*, volume 67 of *LIPIcs*, pages 43:1–43:23. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017.

Troy Lee and Adi Shraibman. Lower bounds in communication complexity. *Found. Trends Theor. Comput. Sci.*, 3(4):263–398, 2009a.

Troy Lee and Adi Shraibman. An approximation algorithm for approximation rank. In *Proceedings of the 24th Annual IEEE Conference on Computational Complexity, CCC 2009, Paris, France, 15-18 July 2009*, pages 351–357. IEEE Computer Society, 2009b.

Nati Linial and Adi Shraibman. Lower bounds in communication complexity based on factorization norms. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 699–708, 2007.

Jiri Matousek, Aleksandar Nikolov, and Kunal Talwar. Factorization norms and hereditary discrepancy. *International Mathematics Research Notices*, 2020(3):751–780, 2020.

Peter McCullagh. *Generalized linear models*. Routledge, 2019.

Georgy Noarov, Ramya Ramalingam, Aaron Roth, and Stephan Xie. High-dimensional prediction for sequential decision making, 2023.

Guy N. Rothblum and Gal Yona. Multi-group agnostic PAC learnability. In *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.

Konrad Schmüdgen. Ten lectures on the moment problem, 2020.

Gabor Szego. *Orthogonal polynomials*, volume 23. American Mathematical Soc., 1939.

Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Appendix A. Uniform approximations to convex functions

We will be interested in uniform approximations of functions from a (possibly infinite) family $\mathcal{F}$ using linear combinations of functions from a (small) finite basis $\mathcal{S} = \{s_i\}_{i=1}^{d}$. We will use the following definition.

Definition 2 (ε-approximate basis, ε-approximate dimension) *Fix a family of functions $\mathcal{F} = \{f : \mathcal{D} \to \mathbb{R}\}$ of functions over a domain $\mathcal{D}$. A basis $\mathcal{S} = \{s_i : \mathcal{D} \to [-1, 1]\}$ is said to ε-approximately span, or be an ε-approximate basis for, the family $\mathcal{F}$ if for every $f \in \mathcal{F}$, there exists $\{r_i \in \mathbb{R}\}_{i=0}^{d}$ such that*

$$\left\| r_0 + \sum_{i=1}^{|\mathcal{S}|} r_i s_i - f \right\|_\infty \leq \varepsilon.$$

Moreover, we define the ε-approximate dimension of $\mathcal{F}$, denoted $\dim_\varepsilon(\mathcal{F})$, to be the smallest size of any ε-approximate basis of $\mathcal{F}$.

A function family of key interest is the space of convex 1-Lipschitz functions over an interval, say $[0, 1]$. We will use $\mathcal{F}_{\text{cvx}}$ to denote this family of functions.

In this work, we obtain tight upper and lower bounds (up to logarithmic factors) on the size of basis that uniformly approximates $\mathcal{F}_{\text{cvx}}$. Specifically, we will show the following result.

Theorem 3 *For every $\delta > 0$, we have*

$$\Omega\left(\frac{1}{\delta^{2/3}}\right) \leq \dim_\delta(\mathcal{F}_{\text{cvx}}) \leq O\left(\frac{1}{\delta^{2/3}} \cdot \log^3(1/\delta)\right).$$

We will prove this result in the coming subsections.

A.1. Constructing a basis

We will explicitly construct a basis $\mathcal{S}$ using a series of reductions starting with the function family $\mathcal{F}_{\text{cvx}}$ to progressively simpler families. The first step in this series of reductions is discretization.

Discretization and scaling Since the functions in $\mathcal{F}_{\text{cvx}}$ are 1-Lipschitz, they can be approximated by piecewise constant functions by a straightforward discretization.

For notational convenience, we will assume that $\delta = 1/m$ where $m \in \mathbb{Z}^+$ is a power of 2. Consider the δ-grid on the interval $[0,1]$ consisting of the points $G_\delta = \{i\delta\}_{i=0}^{m}$. Given a function $g : [0,1] \to \mathbb{R}$, one can construct a piece-wise constant function $\hat{g}$, by setting $\hat{g}(x) = g\left(\delta \lfloor \frac{x}{\delta} \rfloor\right)$.

Lemma 4 (Discretization) *Given a 1-Lipschitz convex function $g : [0,1] \to \mathbb{R}$, construct a piecewise constant function $\hat{g}$ by fixing $\hat{g}(y) = g\left(\delta \cdot \lfloor \frac{y}{\delta} \rfloor\right)$ for all $y \in [0,1]$. Then, $|\hat{g}(y) - g(y)| \leq \delta$ for all $y \in [0,1]$.*

Proof For any $i \in \{0, 1, \ldots, 1/\delta\}$ and any $y \in [i\delta, (i+1)\delta]$,

$$|g(y) - \hat{g}(y)| \leq |g(i\delta) - \hat{g}(i\delta)| + |g(y) - g(i\delta)| + |\hat{g}(y) - \hat{g}(i\delta)|.$$

Since g is 1-Lipschitz, we have $|g(y) - g(i\delta)| \leq |y - i\delta| \leq \delta$. By definition, $\hat{g}(y) - \hat{g}(i\delta) = 0$ and $g(i\delta) - \hat{g}(i\delta) = 0$ and thus we have the result. ■

Clearly, approximating these piecewise constant functions $\hat{g}$ reduces to approximating their values on the δ-grid G_δ. For notational simplicity, we will scale the domain by a factor of $\frac{1}{\delta}$, and consider the related approximation problem for vectors, i.e., functions over $[m] = \{0, 1, \ldots, m\}$.

The functions $\hat{g}$ already exist in a vector space of size $m+1 = 1/\delta+1$, so it already follows that there is a trivial $(\delta+1)$-approximate basis for $\mathcal{F}_{\text{cvx}}$. However, we will show that we can actually construct an $\tilde{O}(1/\delta^{2/3})$-basis.

To this end, let us begin by defining the difference operators for functions on $[m]$. Define the first difference operator Δ on functions over $[m]$ by setting

$$\Delta f(y) = f(y+1) - f(y),$$

for all $f : [m] \to \mathbb{R}$. (Note that the domain of Δf is then $[m-1]$.) Similarly, define the second difference operator Δ^2 as

$$\Delta^2 f(y) = (\Delta \circ \Delta) f(y) = f(y+2) + f(y) - 2f(y+1).$$

Let $\mathbb{L}_{\text{cvx}}$ denote the set of vectors in $\mathbb{R}^m$ that satisfy a discrete version of convexity and 1-Lipschitzness.

$$\mathbb{L}_{\text{cvx}} = \left\{ f : [m] \to \mathbb{R} \;\middle|\; \begin{array}{ll} (\textbf{Convexity}) \;\; \Delta^2 f(i) = f(i+2) - 2f(i+1) + f(i) \geq 0 & \forall i \in [m-2] \\ (1\textbf{-Lipschitz}) \;\; |\Delta f(j)| = |f(j+1) - f(j)| \leq 1 & \forall j \in [m-1] \end{array} \right\}$$

For any convex function $f : [0, m] \to \mathbb{R}$, its restriction to integers satisfies the discrete convexity property above. Also, observe that uniform approximations the set of vectors $\mathbb{L}_{\text{cvx}}$ yields corresponding approximations to the space of convex functions $\mathcal{F}_{\text{cvx}}$. Formally, we have:

Lemma 5 *For all $\eta, \delta > 0$, suppose that there is a basis $\hat{\mathcal{S}}$ of functions over $[m]$ which η-approximately spans the space $\mathbb{L}_{\text{cvx}}$. Define the corresponding basis $\mathcal{S}$ of functions over $[0,1]$ as follows:*

$$\mathcal{S} = \left\{ g : [0,1] \to \mathbb{R} \;\middle|\; g(x) = \delta \cdot \hat{g}\left(\lfloor \tfrac{x}{\delta} \rfloor\right), \hat{g} \in \hat{\mathcal{S}} \right\}.$$

Then, $\mathcal{S}$ $(\delta(1+\eta))$-approximately spans $\mathcal{F}_{\text{cvx}}$.

The above claim follows immediately from Theorem 4 and the correspondence between piecewise constant functions and vectors. The rest of the section will be devoted to constructing a basis $\hat{\mathcal{S}}$ which $\Theta(1)$-approximately spans $\mathbb{L}_{\text{cvx}}$.

From piece-wise linear convex functions to ReLU In the next step, we show that all functions in $\mathbb{L}_{\text{cvx}}$ are well-approximated by the class of ReLU functions. We begin by recalling the definition of the ReLU function family. For each $i \in [m]$, define $\text{ReLU}_i(y)$ as

$$\text{ReLU}_i(y) = \begin{cases} 0 & \text{for } y < i \\ y - i & \text{for } y \geq i \end{cases}.$$

Next we show that $\mathbb{L}_{\text{cvx}}$ lies in the span of the $\{\text{ReLU}_i\}_{i\in[m]}$ via the following expansion for functions, which is essentially the discrete version of a Taylor series.

Lemma 6 (Discrete Taylor series expansion) *Every function $f : [m] \to \mathbb{R}$ can be written as*

$$f(y) = f(0) + (\Delta f(0)) \cdot y + \sum_{i=0}^{m-2} \left(\Delta^2 f(i)\right) \cdot \text{ReLU}_{i+1}(y).$$

Proof For $y = 0$, it is easy to see that both sides equal $f(0)$. Now suppose that $y \geq 1$. Then, expanding the right-hand side of the above equation,

$$\begin{aligned} & f(0) + (\Delta f(0)) \cdot y + \sum_{i=0}^{m-2} \left(\Delta^2 f(i)\right) \text{ReLU}_{i+1}(y) \\ &= f(0) + (\Delta f(0)) \cdot y + \sum_{i=0}^{y-2} \left(\Delta f(i+1) - \Delta f(i)\right)(y - i - 1) \\ &= f(0) + \sum_{i=0}^{y-1} \Delta f(i) \\ &= f(0) + \sum_{i=0}^{y-1} (f(i+1) - f(i)) \\ &= f(y). \end{aligned}$$

■

Thus, every function on $[m]$ can be approximated by a linear combination of ReLU functions. Moreover, it turns out that for functions in $\mathbb{L}_{\text{cvx}}$, the sum of the coefficients of the ReLU functions is actually $O(1)$. This allows us to conclude that approximately spanning $\mathbb{L}_{\text{cvx}}$ is actually equivalent to just approximately spanning the ReLU functions:

Corollary 7 *Suppose that a $\mathcal{S}$ is an η-approximate basis for the family $\{\text{ReLU}_i \mid i \in [m-1]\}$. Then, $\mathcal{S}$ also (3η)-approximately spans all of $\mathbb{L}_{\text{cvx}}$.*

Proof Notice that $\text{ReLU}_0(y) = y$, so Theorem 6 lets us write f as a linear combination over ReLU_i for $i \in [m-1]$.

For each i, let $\hat{\text{ReLU}}_i$ denote the η-approximation to ReLU_i given by the basis $\mathcal{S}$, i.e., $\|\text{ReLU}_i - \hat{\text{ReLU}}_i\|_\infty \leq \eta$ and $\hat{\text{ReLU}}_i \in \text{Span}\{\mathcal{S}\}$. For any convex function f, define its approximation $\hat{f}$ by

$$\hat{f}(y) = f(0) + (\Delta f(0)) \cdot y + \sum_{i=0}^{m-2} \left(\Delta^2 f(i)\right) \cdot \hat{\text{ReLU}}_{i+1}(y).$$

Thenm

$$\begin{aligned}\|f-\hat{f}\|_\infty &= \max_y \left|(\Delta f(0))\cdot(\mathrm{ReLU}_0(y)-\hat{\mathrm{ReLU}}_0(y)) + \sum_{i=0}^{m-2}\left(\Delta^2 f(i)\right)\cdot(\mathrm{ReLU}_{i+1}(y)-\hat{\mathrm{ReLU}}_{i+1}(y))\right| \\ &\le |\Delta f(0)|\cdot\|\mathrm{ReLU}_0-\hat{\mathrm{ReLU}}_0\|_\infty + \sum_{i=0}^{m-2}|\Delta^2 f(i)|\cdot\|\mathrm{ReLU}_{i+1}-\hat{\mathrm{ReLU}}_{i+1}\|_\infty \\ &\le \eta\cdot\left(|\Delta f(0)| + \sum_{i=0}^{m-2}|\Delta^2 f(i)|\right)\end{aligned}$$

Since f is 1-Lipschitz, $|\Delta f(0)| \le 1$. Moreover, since f is convex, $\Delta^2 f$ is positive everywhere, so we can bound

$$\begin{aligned}\sum_{i=0}^{m-2}|\Delta^2 f(i)| &= \sum_{i=0}^{m-2}\Delta f(i+1)-\Delta f(i) \\ &= \Delta f(m-1)-\Delta f(0) \\ &\le 2.\end{aligned}$$

Therefore, $\|f-\hat{f}\|_\infty \le 3\eta$, so f is approximated with an error of 3η by the basis $\mathcal{S}$. ■

From ReLU to Interval functions Via Theorem 7 and Theorem 5, our problem is reduced to finding uniform approximations to the class of ReLU functions $\{\mathrm{ReLU}_i \mid i \in [m-1]\}$.

If we were to form a basis which simply contained all these ReLU_i, then we would end up with a basis of size m, which has the same size as the trivial basis of size $1/\delta$. It turns out, though, that it is possible to approximate the ReLU functions using a smaller basis. We first observe that

$$\mathrm{ReLU}_a(y) - \mathrm{ReLU}_{a+1}(y) = 1[y \ge a],$$

where $1[y \ge a]$ is the indicator of $y \ge a$, or equivalently the indicator function of the interval $[a, m]$.

More generally, the differences between ReLU functions can be expressed as sums of indicators of intervals. Formally, let $\mathbb{I}_{a,b}(y) = 1\,[y \in [a,b]]$ denote the indicator function for the interval $[a,b]$. Then, we have the following proposition.

Proposition 8 *For $0 \le j \le k \le m$ and $y \in [m]$*

$$\mathrm{ReLU}_j(y) - \mathrm{ReLU}_k(y) = \sum_{i=j+1}^{k}\mathbb{I}_{i,m}(y). \tag{1}$$

Arguably, the class of interval functions are a simpler class than ReLU functions, and conceivably, they admit better uniform approximations than ReLU functions. Therefore, we use the following natural strategy to construct a basis approximating ReLU functions.

- Pick a subset of ReLU functions, specifically, ReLU_i for offsets i at regular spacing in $[m]$, namely,
  $$\{\mathrm{ReLU}_{i\cdot t} \mid i = 0, 1, \ldots, m/t-1\},$$
  for some t (we again assume for convenience that t divides m). Include all these functions in the basis.

- Include uniform approximations of interval functions $\mathbb{I}_{a,m}(y) = 1[y \in [a,m]]$ for all $a \in [m-1]$.
- For each $k \in [m-1]$, that is not a multiple of t, reconstruct ReLU_k from the previous multiple of t by adding interval functions, using (1).

The following proposition follows immediately from (1), and we include the proof here for the sake of completeness.

Proposition 9 *Suppose that $\mathcal{S}$ is an η-approximate basis for the class of interval functions $\mathbb{I}^m = \{\mathbb{I}_{a,b} \mid a,b \in [m]\}$. Then, $\mathcal{S} \cup \{\mathrm{ReLU}_{it} \mid i \in [m/t-1]\}$ is an (ηt)-approximate basis for the class of all ReLU functions $\{\mathrm{ReLU}_i \mid i \in [m-1]\}$.*

Proof For $k \in [m]$, let $j = t \cdot \lfloor k/t \rfloor$ denote the largest multiple of t less than or equal to k. We can express ReLU_k as $\mathrm{ReLU}_k = \mathrm{ReLU}_j - \sum_{i=j+1}^{k} \mathbb{I}_{i,m}$. Suppose $\sum_{s\in\mathcal{S}} r_s^{(i)} \cdot s$ is a η-approximation to $\mathbb{I}_{i,m}$ in ℓ_∞ norm, for each i. Then,

$$\left\| \sum_{i=j+1}^{k} \mathbb{I}_{i,m} - \sum_{s\in\mathcal{S}} \left(\sum_{i=j+1}^{k} r_s^{(i)} \right) \cdot s \right\|_\infty \leq \sum_{i=j+1}^{k} \left\| \mathbb{I}_{i,m} - \left(\sum_{s\in\mathcal{S}} r_s^{(i)} \right) \cdot s \right\|_\infty \leq |k-j| \cdot \eta \leq t\eta \,.$$

■

Approximately spanning intervals via JL matrices Let $\mathbb{I}^m$ denote the set of all interval functions — that is, let $\mathbb{I}^m = \{\mathbb{I}_{a,b} | a,b \in [m]\}$. In this section, we will construct a small basis that uniformly approximates all interval functions.

The set of dyadic intervals will serve as a stepping stone towards approximating all intervals. The subset of dyadic intervals consists of all intervals $\mathbb{I}^m_{j,k}$ where $j = i2^h$ and $k = (i+1)2^h - 1$ for integers i, h, we denote it by $\mathbb{D}^{(m)}$. Note that every interval in $\mathbb{I}^m$ can be written as the disjoint union of $2\log(m)$ intervals from $\mathbb{D}^{(m)}$:

Proposition 10 *Every interval function in $\mathbb{I}^m$ can be expressed as a sum of at most $2\log(m)$ dyadic intervals from $\mathbb{D}^m$.*

We will use low-rank approximate factorizations of the identity matrix as described in the following lemma.

Lemma 11 (Low-rank factorization of the identity matrix) *There exists $c > 0$ such that the following holds for all $\mu > 0$ and n. There exists an $n \times k$ matrix V where $k = c\log(n)/\mu^2$ such that*

$$(VV^T)_{ij} = \begin{cases} 1 & \text{if } i = j \\ \leq \mu & \text{if } i \neq j \end{cases}$$

Low-rank factorizations of the identity matrix arise in the context of Johnson-Lindenstrauss lemma (see Alon (2009)). They can be explicitly constructed using codewords from a binary code of distance $1/2 - \mu$ and rate $\Omega(1/\mu^2)$.

Given a $n \times k$ matrix V with the above properties, it is clear that the columns of the matrix V approximately span the rows of the identity matrix. In fact, the columns of V form a basis of size $k = O(\log n/\mu^2)$ that yield a μ-approximation to each of the n rows of the identity matrix. We will use these vectors to construct a low rank approximation to the interval functions.

Definition 12 (The basis $\hat{W}(\mu, m)$) *Fix $\mu \in [0,1]$ and $m = 2^r$ for some $r \in \mathbb{N}$. For each $h \in \{0, \dots, \log m\}$, let $V^{(h)}$ denote $\frac{m}{2^h} \times k_h$ the matrix given by Theorem 11 with $n = m/2^h$ and μ. Let $v_1^{(h)}, \dots, v_{k_h}^{(h)} \in \mathbb{R}^{m/2^h}$ denote the columns of the matrix $V^{(h)}$. For $k \in [k_h]$, define the functions $\hat{w}_k^h : [m] \to [-1,1]$, by setting*

$$\hat{w}_k^h(y) = v_k^h(a) \quad \forall a \in [m/2^h] \text{ and } y \in [2^h \cdot a, 2^h \cdot (a+1)).$$

In other words, the vector $\hat{w}_k^h$ is obtained from v_k^h by repeating each element 2^h times consecutively.

Let

$$\hat{W}^{(h)}(\mu, m) = \{\hat{w}_k^h\}_{k \in [k_h]}$$

and let

$$\hat{W}(\mu, m) = \bigcup_{h=0}^{\log m} W^{(h)}(\mu, m)$$

Lemma 13 (Approximating intervals) *For every $h = \{0, \dots, \log m\}$, the vectors $\hat{W}^{(h)}(\mu, m)$ μ-approximately span the dyadic intervals $\mathbb{I}_{a \cdot 2^h, (a+1)2^h - 1}$ for all $a \in [m/2^h]$. Therefore, their union $\hat{W}(\mu, m)$ μ-approximately spans all dyadic intervals in $[m]$.*

Proof First, consider the case $h = 0$. In this case, the entries of the matrix $V^{(h)}(V^{(h)})^T$ approximate the entries of the identity matrix within an error μ. Hence the columns of the matrix $V^{(h)}$ μ approximately span the rows of the identity matrix. Note that the rows of the identity matrix are the dyadic interval functions $\mathbb{I}_{a,a}$ for $a \in [m]$.

By the same argument, for any $h \in \{1, \dots, \log m\}$, the columns of $V^{(h)}$ μ-approximately span the rows of the identity matrix of dimension $m/2^h$. Note that the entries of vectors $\hat{w}_k^h$ are obtained by repeating each entry of v_k^h 2^h times consecutively. For any row of the identity matrix of dimension $m/2^h$, repeating its entries 2^h times consecutively will yield the indicator function of a dyadic interval of length 2^h. Hence it follows that the dyadic interval functions of length 2^h are μ-approximately spanned by $\hat{W}^{(h)}(\mu, m)$. ■

Putting things together We now have all the ingredients to prove the upper bound of Theorem 3.
Proof [Proof of upper bound from Theorem 3] Fix $m = \frac{1}{\delta}$, $t = m^{1/3}$ and $\mu = \frac{1}{12m^{1/3}\log m}$. Using Theorem 13, we get that $\hat{W}(\mu, m)$ is a μ-approximate basis for all dyadic interval functions $\mathbb{D}^m$.

By Theorem 10, every interval in $\mathbb{I}^m$ is a union of at most $2\log m$ dyadic intervals, and thus, by the triangle inequality, $\hat{W}(\mu, m)$ is a $(2\mu \log m)$-approximately basis for all intervals $\mathbb{I}^{(m)}$.

Now we appeal to Theorem 9 with $t = m^{1/3}$. Thereby, we conclude that the set of functions $\hat{W}(\mu, m) \cup \{\mathrm{ReLU}_{it} \mid i \in [m^{2/3}]\}$ is an approximate basis for all ReLU functions with an error of $t \cdot (2\mu \log m) \leq 1/6$.

By Theorem 7, the same basis approximates all functions in $\mathbb{L}_{\mathrm{cvx}}$ with an error of $3 \cdot 1/6 = 1/2$. Finally, using Theorem 5, this yields a corresponding basis of functions over $[0,1]$ that is a $\delta \cdot \frac{1}{2} + \delta = 3\delta/2$-approximation for $\mathcal{F}_{\mathrm{cvx}}$.

The size of the family $\hat{W}(\mu, m)$ is given by,

$$|\hat{W}(\mu, m)| = \sum_{h=0}^{\log m} |\hat{W}^{(h)}(\mu, m)| = \sum_{h=0}^{\log m} k_h \leq O\left(\log m \cdot \frac{\log m}{\mu^2}\right) = O(m^{2/3} \log^3 m)$$

and thus the total size of the basis is $O(m^{2/3} \log^3 m) + O(m/t) = O(m^{2/3} \log^3 m)$.

This completes the proof of the upper bound Theorem 3 (after substituting δ for $2\delta/3$, so that we have a δ-approximation in the end). ■

A.2. Lower bounds for δ-approximate dimension

The main goal of this section will be to prove the lower bound of Theorem 3, i.e., that any δ-approximate basis for $\mathcal{F}_{\text{cvx}}$ must have size $\Omega(1/\delta^{2/3})$.

First, as an aside, we show that approximating all Lipschitz functions on $[0, 1]$ actually requires a basis of size $\Omega(1/\delta)$. (This is tight by the remarks after Theorem 4.) Thus, restricting ourselves to convex functions actually gives an advantage in the ε-approximate dimension, reducing it from $\Theta(1/\delta)$ to $\tilde{\Theta}(1/\delta^{2/3})$. Specifically, let $\mathcal{F}_{\text{lip}}$ denote the set of 1-Lipschitz functions on $[0, 1]$. Then, we have the following result.

Theorem 14 *For all $\delta \geq 0$, $\dim_\delta(\mathcal{F}_{\text{lip}}) \geq \lfloor 1/4\delta \rfloor$.*

Proof Let $G_{4\delta} = \{4\delta t \mid t = 0, \ldots, \lfloor 1/4\delta \rfloor\}$ denote a grid over $[0, 1]$ of separation 4δ. Observe that any function $f : G_{4\delta} \to \{-2\delta, 2\delta\}$ can be extended to a 1-Lipschitz function over $[0, 1]$. Indeed, one can pick a piecewise linear function that coincides with f on $G_{4\delta}$, and is linear in all intermediate intervals.

Now assume for contradiction that there exists $\mathcal{S}$ such that $\dim(\mathcal{S}) < \lfloor 1/4\delta \rfloor$ which gives δ-uniform approximations to $\mathcal{F}_{\text{lip}}$. Since $|G_{4\delta}| = \lfloor 1/4\delta \rfloor + 1$, by dimension counting, there exists some function on $G_{4\delta}$ which is orthogonal the restriction of every function in $\mathcal{S}$ to $G_{4\delta}$. Specifically, there exists nonzero $g : G_{4\delta} \to \mathbb{R}$ which is non-zero, such that for every $s \in \mathcal{S}$ (and thus for every $s \in \text{Span}(\mathcal{S})$),

$$\sum_{y \in G_{4\delta}} g(y)s(y) = 0$$

Consider the function $f \in \mathcal{F}$ where

$$f(y) = \text{sign}(g(y)) \cdot 2\delta \;\; \forall y \in G_{4\delta}.$$

The function f admits a 1-Lipschitz extension to $[0, 1]$, and therefore can be δ-approximated by functions in $\text{Span}(\mathcal{S})$. That is, there exists $s \in \text{Span}(\mathcal{S})$ which is a δ-approximation to f. But this means that

$$\text{sign}(s(y)) = \text{sign}(f(y)) = \text{sign}(g(y)),$$

and $|s(y)| \geq \delta$ for $y \in G_{4\delta}$. But then g cannot be orthogonal to s, giving a contradiction. ■

To prove a lower bound on the ε-approximate dimension of ReLU, we use Alon's lower bound on the approximate rank of the identity matrix.

Theorem 15 (Alon (2009)) *Let I_n be the $n \times n$ identity matrix, and let $1/(2\sqrt{n}) \leq \mu \leq 1/4$. Then,*

$$\mathrm{rank}_\mu(I_n) \geq \frac{d \log(n)}{\mu^2 \log(1/\mu)},$$

for some absolute constant d. (Here, $\mathrm{rank}_\mu(I_n)$, the ε-approximate rank of I_n, denotes the ε-approximate dimension of its rows.)

Theorem 16 *For any absolute constant $c > 0$ and all $m \in \mathbb{Z}^+$,*

$$\dim_c(\mathrm{ReLU}_{[m]}) \geq \Omega(m^{2/3}),$$

where $\mathrm{ReLU}_{[m]}$ denotes the family $\mathrm{ReLU}_{[m]} = \{\mathrm{ReLU}_i \mid i \in [m]\}$ of functions on $[m]$.

Proof Suppose that $\mathcal{S}$ is a c-approximate basis for $\mathrm{ReLU}_{[m]}$.

Fix $t = m^{1/3}$, and let $A = \{1, 2, \ldots, m/t - 1\}$. Let $\mathcal{S}'$ denote the functions in $\mathcal{S}$ restricted to the domain $t \cdot A = \{t \cdot i | i \in A\}$. Clearly, $|\mathcal{S}'| \leq |\mathcal{S}|$.

For any $i \in A$, consider the function $f_i : A \to \mathbb{R}$ defined as

$$f_i(y) = \frac{1}{t}\big(\mathrm{ReLU}_{(i+1)t}(yt) + \mathrm{ReLU}_{(i-1)t}(yt) - 2\mathrm{ReLU}_{it}(yt)\big).$$

By substituting the values of y, it is easy to check that

$$f_i(y) = \begin{cases} 1 & \text{if } y = i \\ 0 & \text{if } y \neq i \end{cases}$$

In other words, the functions f_i are the rows of the identity matrix of dimension $|A|$.

However, if the basis $\mathcal{S}$ yields an c-approximation for each of the three functions $\mathrm{ReLU}_{(i+1)t}, \mathrm{ReLU}_{it}$ and $\mathrm{ReLU}_{(i-1)t}$, then $\mathcal{S}'$ yields a $4c/t$-approximation for the functions f_i. By appealing to Alon's lower bound on the approximate rank of the identity matrix, we get that

$$|\mathcal{S}'| \geq d\frac{\log|A|}{(4c/t)^2 \log(t/4c)} \geq \Omega(m^{2/3}),$$

as desired. ■

From the above, we immediately conclude a lower bound on the δ-approximate dimension of $\mathcal{F}_{\mathrm{cvx}}$:

Corollary 17 *For all $\delta > 0$,*

$$\dim_\delta(\mathcal{F}_{\mathrm{cvx}}) \geq \Omega\left(\frac{1}{\delta^{2/3}}\right),$$

where Ω hides an absolute constant factor.

Proof Fix $m = \lfloor 1/\delta \rfloor$. Given a δ-approximation to $\mathcal{F}_{\mathrm{cvx}}$, we get a $\Omega(1)$-approximation to $\mathrm{ReLU}_{[m]}$ by considering the ReLU functions over $[0, 1]$, and restricting to the evaluation points $\{i/m \mid i \in [m]\}$. ■

Since ReLU functions are just linear transformations of L_1 loss functions, this also similarly implies that the δ-approximate dimension of L_1 loss functions is large:

Corollary 18 *Let $\mathcal{L}_1$ denote the set of L_1 loss functions of the form $|y-t|$ for $t \in [0,1]$. Then, for all $\delta > 0$,*

$$\dim_\delta(\mathcal{L}_1) \geq \Omega\left(\frac{1}{\delta^{2/3}}\right).$$

Appendix B. Loss Minimization

B.1. Loss Functions, Sufficient Statistics and Uniform Approximations

A loss function is a function $\ell : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$ which assigns a real value $\ell(y,t)$ to a pair of inputs where $y \in \mathcal{Y}$ is a label and $t \in \mathcal{A}$ is an action. We will focus on the case where $\mathcal{Y} \subseteq [0,1]$. We will allow for arbitrary actions sets $\mathcal{A}$. They can be discrete (e.g buy or not buy), or continuous (e.g in some bounded interval [-B, B]).[5] Let $\mathcal{L}_{\text{lip}}$ denote the set of all ℓ that are 1-Lipschitz in y for every $t \in \mathcal{A}$. Let $\mathcal{L}_{\text{cvx}} \subseteq \mathcal{L}_{\text{lip}}$ denote the subset of functions where $\ell(y,t)$ is convex in y for every $t \in \mathcal{A}$.

We define a family of statistics to be a set of functions $\mathcal{S} = \{s_i : \mathcal{Y} \to [-1,1]\}_{i=0}^d$, with the convention that $s_0 = 1$ is always the constant function.

Definition 19 ((d,λ,δ)-uniform approximations, sufficient statistics) *Let $\mathcal{S}$ be family of statistics and $\mathcal{L}$ be a family of loss functions. We say that $\mathcal{S}$ gives (d,λ,δ)-uniform approximations to $\mathcal{L}$ where $d = |\mathcal{S}|$ if for every $\ell \in \mathcal{L}$, there exist $\{r_i^\ell : \mathcal{A} \to \mathbb{R}\}_{i=0}^d$ such that*

$$\left|\sum_{s_i \in \mathcal{S}} r_i^\ell(t) s_i(y) - \ell(y,t)\right| \leq \delta$$

and

$$\sum_{i=0}^d |r_i^\ell(t)| \leq \lambda$$

for all $y \in \mathcal{Y}$ and for all $t \in \mathcal{A}$. Equivalently, $\mathcal{S}$ ϵ-approximately spans the family $\{\ell_t(y) = \ell(y,t)\}$ for every $t \in \mathcal{A}$ with coefficients of total magnitude at most λ. We refer to $\mathcal{S}$ as a set of sufficient statistics for $\mathcal{L}$, and to $\mathcal{R} = \{r_i^\ell\}_{i \in [d], \ell \in \mathcal{L}}$ as the coefficient family.

In light of the above definition, it is useful to define the function $\hat{\ell} : [-1,1]^d \times \mathcal{A} \to \mathbb{R}$ as

$$\hat{\ell}(v,t) = r_0^\ell(t) + \sum_{i=1}^d r_i^\ell(t) v_i$$

which acts on the statistics from $\mathcal{S}$ directly instead of on y. Note that for a distribution $\mathcal{D}$ on $\mathcal{Y}$, for $\ell \in \mathcal{L}$ and $t \in \mathcal{A}$, Theorem 19 implies

$$\hat{\ell}\left(\mathbb{E}_{\mathcal{D}}[s_i(\mathbf{y})], t\right) = r_0^\ell(t) + \sum_{i=1}^d r_i^\ell(t) \mathbb{E}_{\mathcal{D}}[s_i(\mathbf{y})] = \mathbb{E}_{\mathcal{D}}[\ell(\mathbf{y},t)] \pm \delta.$$

5. Our notion of an abstract space of actions departs from some prior work on omniprediction Gopalan et al. (2021, 2023a) which required the set of actions to be a bounded subset of $\mathbb{R}$, bringing it in line with the calibration literature (for instance Kleinberg et al. (2023)).

Hence the expectations of functions in $\mathcal{S}$ gives a set of statistics that lets us approximate the loss associated with each action in $\mathcal{A}$ for every loss in $\mathcal{L}$, thus justifying the term *sufficient statistics*.

For sake of intuition, we present the example $\mathcal{L}^C_{\{0,1\}}$ of all bounded loss functions $\ell : \{0,1\} \times \mathcal{A} \to [-C, C]$ in the case of $\mathcal{Y} = \{0,1\}$ i.e. binary classification.

Proposition 20 *For all $C > 0$, $\mathcal{S} = \{1, y\}$ gives $(1, 2C, 0)$ uniform approximations to $\mathcal{L}^C_{\{0,1\}}$.*

Proof We can write $\ell(y,t)$ using its multilinear expansion in y as

$$\ell(y,t) = \ell(0,t) + y(\ell(1,t) - \ell(0,t)).$$

We take

$$s_1(y) = y,\ r^\ell_0(t) = \ell(0,t),\ r^\ell_1(t) = \ell(1,t) - \ell(0,t).$$

It follows that $\lambda = \max_t |r^\ell_1(t)| \le 2C$. ■

We record the following corollary that allows us to assume that $\lambda = O(d)$ in a family of sufficient statistics. The fact is standard in convex geometry and follows from a simple application of John's theorem; we include a proof for completeness in Appendix F.

Corollary 21 *Let $\mathcal{L}$ be a family of loss functions bounded by C, with a family of sufficient statistics $\mathcal{S}$ that gives a (d, λ, δ)-approximation to $\mathcal{L}$ for some λ. Then, there exists a family of statistics $\mathcal{S}'$ consisting of functions also bounded by C which gives a $(d, (1 + \delta/C)d, \delta)$-approximation to $\mathcal{L}$.*

B.2. Statistics, predictors, and calibration

Next, we will define the notion of a predictor corresponding to a family of statistics. Let $\mathcal{D}^*$ denote a distribution on $\mathcal{X} \times \mathcal{Y}$. We denote samples from $\mathcal{D}^*$ by $(\mathbf{x}, \mathbf{y}^*)$. Given a family of statistics $\mathcal{S} = \{s_i : \mathcal{Y} \to [-1,1]\}_{i \in [d]}$, let $s(y) = (s_i(y))_{i \in [d]}$. An $\mathcal{S}$-predictor is a function $p : \mathcal{X} \to [-1,1]^d$ with the interpretation that $p(x)$ is an estimate for $\mathbb{E}[s(\mathbf{y})|\mathbf{x} = x]$. As an example, consider $s_i(y) = y^i$. Predictors for this family would predict the first d moments of $\mathbf{y}|x$ for each x.

We now define the notion of a calibrated predictor of statistics.

Definition 22 *Let $\mathcal{S}$ be a family of statistics. We say the predictor p is β-calibrated for $\mathcal{S}$ under $\mathcal{D}^*$ if*

$$\mathbb{E}_{\mathcal{D}^*}\left[\|\mathbb{E}[s(\mathbf{y}^*)|p(\mathbf{x})] - p(\mathbf{x})\|_\infty\right] \le \beta.$$

Perfect calibration is said to holds when $\mathbb{E}_{\mathcal{D}^*}[s(\mathbf{y}^*)|p(\mathbf{x})] = p(\mathbf{x})$ i.e. $\beta = 0$. [6]

Let $\mathrm{Im}(\mathcal{S}) \subseteq [-1,1]^d$ denote the set of values $\mathbb{E}[s(\mathbf{y})]$ can take over all distributions on $\mathbf{y}$. This set is generally a proper subset of $[-1,1]^d$, due to relationships between functions in $\mathcal{S}$. For instance when $\mathcal{S}$ is the set of the first d moments, it needs to satisfy various moment inequalities. Perfect calibration ensures that every prediction lies in $\mathrm{Im}(\mathcal{S})$. We will next show a robust analogue of this, which shows that β-calibration implies our predictions are close to the expectations of s_is for a suitably defined distribution on labels. In order to do this, we define the following *simulated distribution* corresponding to a predictor.

6. We could use ℓ_1 or other ℓ_p norms in place of ℓ_∞, all such definitions are equivalent up to polynomials in d.

Definition 23 (Simulated distribution) *Let $\mathcal{D}^*$ be a distribution on $\mathcal{X} \times \mathcal{Y}$ and let p be a $\mathcal{S}$-predictor for a statistics family $\mathcal{S}$. We will associate a distribution $\tilde{\mathcal{D}} = \tilde{\mathcal{D}}(p)$ on points and labels to p sampled as $(\mathbf{x}, \tilde{\mathbf{y}}) \sim \tilde{\mathcal{D}}$, we first sample $\mathbf{x} \sim \mathcal{D}^*$ and let $\tilde{\mathbf{y}}|\mathbf{x} \sim \mathbf{y}^*|p(\mathbf{x})$.*

From above definition, the marginal distribution of $\mathbf{x}$ matches $\mathcal{D}^*$, whereas $\tilde{\mathbf{y}}$ is identically distributed over each level set of p. This lets us couple the distributions $\mathcal{D}^*$ and $\tilde{\mathcal{D}}$. We sample $\mathbf{x}$ according to the common marginal, and then sample $\mathbf{y}^*|\mathbf{x} \sim \mathcal{D}^*$ and $\tilde{\mathbf{y}}|\mathbf{x} \sim \tilde{\mathcal{D}}$ independently.

We note that our definiton of the simulation is different from Boolean setting Dwork et al. (2021); Gopalan et al. (2023a); Dwork et al. (2022), where the simulation is based on the predictor alone, and is independent of the distribution $\mathcal{D}^*$ that is being learnt. It is reminiscent of the view of Gopalan et al. (2021), who view predictors as partitions of the space into level sets, and define a *canonical prediction* which is the expectation over the level set. Our next lemma may be viewed as showing the closeness of a calibrated predictor to precisely such a canonical predictor.

Lemma 24 *If p is β-calibrated under $\mathcal{D}^*$,*

$$\mathbb{E}_{\tilde{\mathcal{D}}}\left[\|p(\mathbf{x}) - \mathbb{E}[s(\tilde{\mathbf{y}})|\mathbf{x}]\|_\infty\right] \leq \beta.$$

Proof From the definition of $\tilde{\mathbf{y}}$, it follows that for $s_i \in \mathcal{S}$, $\mathbb{E}[s_i(\tilde{\mathbf{y}})|\mathbf{x}] = \mathbb{E}[s_i(\mathbf{y}^*)|p(\mathbf{x})]$. Hence,

$$\mathbb{E}[\|\mathbb{E}[s(\tilde{\mathbf{y}})|\mathbf{x}] - p(\mathbf{x})\|_\infty] = \mathbb{E}[\|\mathbb{E}[s(\mathbf{y}^*)|p(\mathbf{x})] - p(\mathbf{x})\|_\infty] \leq \beta.$$

where the inequality is by Definition 22. ■

B.3. Optimal Actions under Loss Functions

Given a distribution $\mathcal{D}$ and a loss function ℓ, we can define the *optimal* action as

$$k_\ell(\mathcal{D}) = \arg\min_{t \in \mathcal{A}} \mathbb{E}_{\mathbf{y} \sim \mathcal{D}}[\ell(\mathbf{y}, t)].$$

As defined, the function k_ℓ requires full knowledge of the distribution $\mathcal{D}$. We will see that if if $\mathcal{S}$ is a set of sufficient statistics for ℓ, then one can approximate k_ℓ with just the knowledge of $\mathbb{E}[s(\mathbf{y})]$.

Assume that $\mathcal{S}$ gives $(d, \lambda, \varepsilon)$-uniform approximations to ℓ, so that ℓ is ε-approximated by $\hat{\ell}$. Selecting action via $k_{\hat{\ell}}$ results in actions that are at most $O(\varepsilon)$ far from optimal for ℓ. But

$$k_{\hat{\ell}}(\mathcal{D}) = \arg\min_{t \in \mathcal{A}} \mathbb{E}_{\mathcal{D}}[\hat{\ell}(s(\mathbf{y}), t)] = \arg\min_{t \in \mathcal{A}} \left[r_0^\ell(t) + \sum_{i=1}^{d} r_i^\ell(t) \mathbb{E}_{\mathcal{D}}[s_i(\mathbf{y})] \right],$$

so $k_{\hat{\ell}}$ only depends on $\mathbb{E}_{\mathcal{D}}[s(\mathbf{y})]$ rather than the entire distribution $\mathcal{D}$.

Abusing notation, we extend the definition of $k_{\hat{\ell}}$ to take predictions in $[-1, 1]^d$ as its argument. That is, define $k_{\hat{\ell}} : [-1, 1]^d \to \mathcal{A}$ as

$$k_{\hat{\ell}}(v) = \arg\min_{t \in \mathcal{A}} r_0^\ell(t) + \sum_{i=1}^{d} r_i^\ell(t) v_i.$$

Note that, for $v = \mathbb{E}[s(\mathbf{y})] \in \mathrm{Im}(\mathcal{S})$, this matches our prior definition, since $k_{\hat{\ell}}(\mathbb{E}[s(\mathbf{y})]) = k_{\hat{\ell}}(\mathcal{D})$. But it also allows for general $v \notin \mathrm{Im}(\mathcal{S})$. This will be important since our $\mathcal{S}$-predictors are not guaranteed to make predictions in $\mathrm{Im}(\mathcal{S})$.

We do not impose any constraints on the action space $\mathcal{A}$, or how the loss family $\mathcal{L}$ depends on it. We only require the existence of an oracle for $\mathcal{A}$ that solves the minimization problem required for computing $k_{\hat{\ell}}$. In the case of a discrete set of actions, this can simply be done by enumeration. In the case when $\mathcal{A}$ is a compact set such as a bounded interval, and the loss functions ℓ to be Lipschitz in the actions, we could discretize the action space and compute the value at each choice of the discretization, to find an approximate minimum. Our reason for abstracting away the complexity of computing $k_{\hat{\ell}}$ is that even if we learnt the Bayes optimal $\mathcal{S}$-predictor, we would still need to compute $k_{\hat{\ell}}$, so the complexity of this function is extraneous to the task of learning a good predictor.

B.4. Multiaccuracy

Finally, we define the notion of multiaccuracy with respect to a class of tests $\mathcal{B} = \{b : \mathcal{X} \to \mathbb{R}\}$. The notion was defined in the context of binary classification by Hébert-Johnson et al. (2018), though similar notions have appeared previously in the literature on boosting and learning. [7]

We extend the definition of multiaccuracy to the setting of statistics prediction (similar to Jung et al. (2021); Dwork et al. (2022)). Intuitively, multiaccuracy for a predictor p for a family of statistics $\mathcal{S}$ requires that no test b in the class $\mathcal{B}$ can distinguish the true value of a statistic $s_i \in \mathcal{S}$ from the predicted value p_i.

Definition 25 (Multiaccuracy) *Let $\mathcal{S}$ be a family of statistics, $\mathcal{B} = \{b : \mathcal{X} \to \mathbb{R}\}$ be a class os tests and $\alpha > 0$. We say that an $\mathcal{S}$-predictor $p : \mathcal{X} \to [-1, 1]^d$ is $(\mathcal{B}, \alpha)$-multiaccurate if for every $i \in [d]$ and $b \in \mathcal{B}$, it holds that*

$$|\mathbb{E}_{\mathcal{D}^*}[(s_i(\mathbf{y}^*) - p_i(\mathbf{x}))b(\mathbf{x})]| \leq \alpha.$$

Appendix C. Omniprediction via outcome indisinguishability

An omnipredictor, introduced in the work of Gopalan et al. (2021), is a predictor can be postprocessed to get an action that suffers lesser loss than any hypothesis in the class $\mathcal{C}$. The original definition was in the setting of binary or multiclass classification, where the predictor returns a probability distribution on labels. The following definition generalizes this notion to $\mathcal{S}$-predictors.

Definition 26 (Omnipredictor Gopalan et al. (2021)) *Let $\mathcal{L}$ be a family of loss functions and $\mathcal{C}$ be family of hypotheses, and $\varepsilon > 0$. Let $\mathcal{S}$ be a set of sufficient statistics for $\mathcal{L}$. An $\mathcal{S}$-predictor $p : \mathcal{X} \to [-1, 1]^d$ is an $(\mathcal{L}, \mathcal{C}, \varepsilon)$-omnipredictor if for every $\ell \in \mathcal{L}$, there exists $k : [-1, 1]^d \to \mathcal{A}$ such that*

$$\mathbb{E}_{\mathcal{D}^*}[\ell(\mathbf{y}^*, k(p(\mathbf{x})))] \leq \min_{c \in \mathcal{C}} \mathbb{E}_{\mathcal{D}^*}[\ell(\mathbf{y}^*, c(\mathbf{x})] + \varepsilon.$$

Note that the set of sufficient statistics $\mathcal{S}$ needs to give a (d, λ, δ)-uniform approximations to $\mathcal{L}$ with $\delta \leq \varepsilon$. Recall that $\mathcal{S}$ is associated with a coefficient family $\mathcal{R} = \{r_i^\ell\}_{i \in [d], \ell \in \mathcal{L}}$. We let $\mathcal{R} \circ \mathcal{C}$ denote all functions of the form $r_i^\ell \circ c$ where $r_i^\ell \in \mathcal{R}$ and $c \in \mathcal{C}$.

7. In previous work, multiaccuracy was defined with respect to a hypothesis class $\mathcal{C}$, which mapped $\mathcal{X}$ to $\mathbb{R}$. Since we define hypotheses classes to map to $\mathcal{A}$, we use the term tests for functions mapping to $\mathbb{R}$. The specific tests we use will compose $c \in \mathcal{C}$ with a function $r_\ell : \mathcal{A} \to \mathbb{R}$.

Remark 27 *Note that $\mathcal{R} \circ \mathcal{C}$ only considers composition of c with r_i for $i > 0$. In particular, it does not consider compositions of c with r_0. For our main result, it will suffice to not consider compositions of c with r_0.*

Our main result in this section establishes sufficient conditions for omniprediction. It shows that for any family of loss functions that can be well-approximated by a family of statistics, we can get an omnipredictor through calibration and multiaccuracy.

Theorem 28 *Let $\mathcal{S}$ be family of statistics and $\mathcal{L}$ be a family of loss functions such that that $\mathcal{S}$ gives (d, λ, δ)-uniform approximations to $\mathcal{L}$ with coefficient family $\mathcal{R}$. If the $\mathcal{S}$-predictor p is $(\mathcal{R} \circ \mathcal{C}, \alpha)$-multiaccurate and β-calibrated then it is an $(\mathcal{L}, \mathcal{C}, \varepsilon)$-omnipredictor for*

$$\varepsilon = 3(d\alpha + \lambda\beta + \delta)$$

using the functions $k_{\hat{\ell}}$ for choosing actions.

Following the loss outcome indistinguishability paradigm of Gopalan et al. (2023a), we will prove this result by showing:

- The predictor $p(\mathbf{x})$ is an omnipredictor for the distribution $\tilde{\mathcal{D}}$ Theorem 23 and the family of losses $\hat{\mathcal{L}} = \{\hat{\ell}\}_{\ell \in \mathcal{L}}$ where we $k_{\hat{\ell}}(p(\mathbf{x}))$ to choose actions.
- One can switch the label distribution from $\tilde{\mathbf{y}}$ to $\mathbf{y}^*$ and the losses from $\hat{\ell}$ to ℓ without much change in the expected loss.

To implement this, we show a sequence of lemmas showing various forms of indistinguishability for labels and loss functions. The first shows indistinguishability for expected loss $\hat{\ell}$ when the actions are functions of the prediction $p(\mathbf{x})$.

Lemma 29 *For all functions $k : [-1, 1]^d \to \mathcal{A}$*

$$\mathbb{E}_{\mathcal{D}^*}[\hat{\ell}(s(\mathbf{y}^*), k(p(\mathbf{x})))] = \mathbb{E}_{\tilde{\mathcal{D}}}[\hat{\ell}(s(\tilde{\mathbf{y}}), k(p(\mathbf{x})))].$$

Proof We can write

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}^*}[\hat{\ell}(s(\mathbf{y}^*), k(p(\mathbf{x})))] &= \mathbb{E}\left[\sum_{i=0}^{d} r_i^{\ell}(k(p(\mathbf{x})))s_i(\mathbf{y}^*)\right] \\
&= \mathbb{E}\left[\sum_{i=0}^{d} \mathbb{E}[r_i^{\ell}(k(p(\mathbf{x})))s_i(\mathbf{y}^*)|p(\mathbf{x})]\right] \\
&= \mathbb{E}\left[\sum_{i=0}^{d} \mathbb{E}[r_i^{\ell}(k(p(\mathbf{x})))s_i(\tilde{\mathbf{y}})|\mathbf{x}]\right] \\
&= \mathbb{E}\left[\sum_{i=0}^{d} r_i^{\ell}(k(p(\mathbf{x})))s_i(\tilde{\mathbf{y}})\right] \\
&= \mathbb{E}_{\tilde{\mathcal{D}}}[\hat{\ell}(s(\tilde{\mathbf{y}}), k(p(\mathbf{x}))].
\end{aligned}$$

■

The next lemma shows that if p is well-calibrated, then distinguishing the predictions $p(\mathbf{x})$ from $\mathbb{E}[s(\tilde{\mathbf{y}})|x]$ using $\hat{\ell}$ is hard, even allowing for arbitrary actions.

Lemma 30 *If p is β-calibrated, then for all functions $b : \mathcal{X} \to \mathcal{A}$*

$$\left|\mathbb{E}[\hat{\ell}(s(\tilde{\mathbf{y}}), b(\mathbf{x}))] - \mathbb{E}[\hat{\ell}(p(\mathbf{x}), b(\mathbf{x}))]\right| \leq \lambda\beta.$$

Proof For any function $b : \mathcal{X} \to \mathcal{A}$, we have

$$\begin{aligned}
\left|\mathbb{E}_{\tilde{\mathcal{D}}}[\hat{\ell}(s(\tilde{\mathbf{y}}), b(\mathbf{x}))] - \hat{\ell}(p(\mathbf{x}), b(\mathbf{x}))\right| &= \left|\mathbb{E}\left[\sum_{i=1}^{d} r_i^{\ell}(b(\mathbf{x}))(p_i(\mathbf{x}) - s_i(\tilde{\mathbf{y}}))\right]\right| \\
&= \left|\mathbb{E}\left[\sum_{i=1}^{d} r_i^{\ell}(b(\mathbf{x}))(p_i(\mathbf{x}) - \mathbb{E}[s_i(\tilde{\mathbf{y}})|\mathbf{x}])\right]\right| \\
&\leq \mathbb{E}\left[\left(\sum_{i=1}^{d}\left|r_i^{\ell}(b(\mathbf{x}))\right|\right)\max_{i\in[d]}|p_i(\mathbf{x}) - \mathbb{E}[s_i(\tilde{\mathbf{y}})|\mathbf{x}]|\right] \\
&\qquad \text{(Holder's inequality)} \\
&\leq \mathbb{E}[\lambda \left\|p(\mathbf{x}) - \mathbb{E}[s(\tilde{\mathbf{y}})|\mathbf{x}]\right\|_{\infty}] \\
&\leq \lambda\beta.
\end{aligned}$$

■

Next we show more general conditions under which ℓ does cannot distinguish $\tilde{\mathbf{y}}$ from $\mathbf{y}^*$. Theorem 29 gave such a result but for limited actions. Here we allow more general action functions, but we also make more assumptions about the predictor.

Corollary 31 *If p is $(\mathcal{R} \circ \mathcal{C}, \alpha)$-multiaccurate and β-calibrated then*

$$\left|\mathbb{E}_{\tilde{\mathcal{D}}}[\hat{\ell}(s(\tilde{\mathbf{y}}), c(\mathbf{x}))] - \mathbb{E}_{\mathcal{D}^*}[\hat{\ell}(s(\mathbf{y}^*), c(\mathbf{x}))]\right| \leq d\alpha + \lambda\beta.$$

Proof We can write

$$\begin{aligned}
\left|\mathbb{E}_{\tilde{\mathcal{D}}}[\hat{\ell}(s(\tilde{\mathbf{y}}), c(\mathbf{x}))] - \mathbb{E}_{\mathcal{D}^*}[\hat{\ell}(s(\mathbf{y}^*), c(\mathbf{x}))]\right| \leq &\left|\mathbb{E}[\hat{\ell}(s(\tilde{\mathbf{y}}), c(\mathbf{x}))] - \mathbb{E}_{\mathcal{D}}[\hat{\ell}(p(\mathbf{x}), c(\mathbf{x}))]\right| \\
&+ \left|\mathbb{E}[\hat{\ell}(s(\mathbf{y}^*), c(\mathbf{x}))] - \mathbb{E}[\hat{\ell}(p(\mathbf{x}), c(\mathbf{x}))]\right|.
\end{aligned}$$

The first term can be bounded by $\lambda\beta$ using β-calibration together with Theorem 30. To bound the second term, we note that

$$\begin{aligned}
\left|\mathbb{E}[\hat{\ell}(s(\mathbf{y}^*), c(\mathbf{x}))] - \mathbb{E}_{\mathcal{D}}[\hat{\ell}(p(\mathbf{x}), c(\mathbf{x}))]\right| &= \left|\mathbb{E}\left[\sum_{i=1}^{d} r_i^{\ell}(c(\mathbf{x}))(s_i(\mathbf{y}^*) - p_i(\mathbf{x})\right]\right| \\
&\leq \sum_{i=1}^{d}\left|\mathbb{E}[r_i^{\ell}(c(\mathbf{x}))(s_i(\mathbf{y}^*) - p_i(\mathbf{x}))]\right| \\
&\leq d\alpha
\end{aligned}$$

where we use multiaccuracy for each i. ■

We now complete the proof of Theorem 28, our main result on omniprediction.

Proof [Proof of Theorem 28] We have the following chain of inequalities

$$\begin{aligned}
\mathbb{E}[\hat{\ell}(s(\tilde{\mathbf{y}}), k_{\hat{\ell}}(p(\mathbf{x})))] &\leq \mathbb{E}[\hat{\ell}(p(\mathbf{x}), k_{\hat{\ell}}(p(\mathbf{x})))] + \lambda\beta && \text{(By Theorem 30)} \\
&\leq \mathbb{E}[\hat{\ell}(p(\mathbf{x}), c(\mathbf{x}))] + \lambda\beta && \text{(by definition of } k_{\hat{\ell}}\text{)} \\
&\leq \mathbb{E}[\hat{\ell}(s(\tilde{\mathbf{y}}), c(\mathbf{x})) + 2\lambda\beta. && (2)
\end{aligned}$$

To switch the label distribution from from $\tilde{\mathbf{y}}$ to $\mathbf{y}^*$ we use

$$\begin{aligned}
\mathbb{E}[\hat{\ell}(s(\mathbf{y}^*), k_{\hat{\ell}}(p(\mathbf{x})))] &= \mathbb{E}[\hat{\ell}(s(\tilde{\mathbf{y}}), k_{\hat{\ell}}(p(\mathbf{x})))] && \text{(By Theorem 29)} \\
\mathbb{E}[\hat{\ell}(s(\tilde{\mathbf{y}}), c(\mathbf{x})) &\leq \mathbb{E}[\hat{\ell}(s(\mathbf{y}^*), c(x))] + (d\alpha + \lambda\beta) && \text{(Theorem 31)}
\end{aligned}$$

Plugging these into Equation (2) gives

$$\mathbb{E}[\hat{\ell}(s(\mathbf{y}^*), k_{\hat{\ell}}(p(\mathbf{x})))] \leq \mathbb{E}[\hat{\ell}(s(\mathbf{y}^*), c(\mathbf{x}))] + d\alpha + 3\lambda\beta. \tag{3}$$

Finally we can switch each loss from $\hat{\ell}$ to ℓ by incurring an additional δ. We use the uniform approximation property with $k_{\hat{\ell}}(p(\mathbf{x}))$ and $c(\mathbf{x})$, which gives

$$\begin{aligned}
\left|\mathbb{E}[\ell(\mathbf{y}^*, k_{\hat{\ell}}(p(\mathbf{x})))] - \mathbb{E}[\hat{\ell}(s(\mathbf{y}^*), k_{\hat{\ell}}(p(\mathbf{x})))]\right| &\leq \delta, \\
\left|\mathbb{E}[\ell(\mathbf{y}^*, c(\mathbf{x}))] - \mathbb{E}[\hat{\ell}(s(\mathbf{y}^*), c(\mathbf{x}))]\right| &\leq \delta
\end{aligned}$$

Plugging these into Equation (3) gives the desired bound. ■

Appendix D. Main Applications

In this section, we will derive omnipredictors for various classes of loss functions. First, in Appendix D.1, we will present an omnipredictor for the class of convex, Lipschitz loss functions. In Appendix D.2, we will present an omnipredictor for the class of functions approximated by low degree polynomials, in particular, for the class of ℓ_p losses. In Appendix D.3, we present an omnipredictor for the class of losses corresponding to generalized linear models.

As mentioned earlier, the main idea is to find a family of sufficient statistics that approximates a family of loss functions. Given the approximations, the main algorithmic result driving the omnipredictors for various classes is the following theorem below. This theorem bounds the sample complexity of achieving a predictor that satisfies calibrated multiaccuracy with respect to family of statistics and a family of test functions (corresponding to the composition of the coefficient family in the approximation of the losses and the comparison class of hypotheses). We state the theorem here and defer the proof and further discussion to Appendix E.

Theorem 32 *Let $\mathcal{S}$ be family of statistics and $\mathcal{L}$ be a family of loss functions such that that $\mathcal{S}$ gives $(d, \lambda, \varepsilon)$-uniform approximations to $\mathcal{L}$ with coefficient family $\mathcal{R}$, there exists an algorithm that returns an $(\mathcal{L}, \mathcal{C}, 4\varepsilon)$-omnipredictor[8], satisfying the following properties.*

- *The algorithm makes $O(d/\sigma^2)$ calls to a (ρ, σ)-weak learner for $\mathcal{R} \circ \mathcal{C}$ where $\sigma \leq \rho \leq \varepsilon/12\lambda$.*
- *The algorithm has time and sample complexity $\tilde{O}\left(d\left(\frac{\lambda}{\varepsilon}\right)^{d+5} + \frac{d}{\sigma^2}Z\right)$ where Z is the runtime/sample complexity of the weak learner.*

8. Note that once $\mathcal{S}$ is fixed, the 3ε factor in the omniprediction slack is unavoidable.

D.1. Omniprediction for Convex Lipschitz Losses

Recall that $\mathcal{L}_{\text{cvx}}$ denotes the family of convex, Lipschitz loss functions i.e. loss functions $\ell(y,t)$ that are Lipschitz and convex in y. For this class, we can derive approximations in terms of a small sized family of statistics as required for Theorem 28 based on our univariate approximation Theorem 3.

Recall that the set of functions

$$\mathcal{S}_{\mathcal{L}_{\text{cvx}},\delta} = \hat{W}(\mu, m) \cup \{\text{ReLU}_{it} | i \in [m/t]\},$$

for $m = \frac{1}{\delta}$, $t = m^{1/3}$ and $\mu = 1/(12m^{1/3}\log m)$, δ-approximately spans $\mathcal{F}_{\text{cvx}}$, the family of convex, Lipschitz functions on $[0,1]$.

Corollary 33 *For $\delta > 0$ sufficiently small, the set of statistics $\mathcal{S}_{\mathcal{L}_{\text{cvx}},\delta}$ gives*

$$(O(\log(1/\delta)^{4/3}/\delta^{2/3}), O(\log(1/\delta)^{4/3}/\delta^{2/3}), \delta)$$

approximation to $\mathcal{L}_{\text{cvx}}$.

In fact, in the above theorem, carefully keeping track of the coefficients in Theorem 3, we can bound λ by $O(1)$ but we do not need this strengthening. Using the theorem above, we get the following result for learning an omnipredictor for the family of Lipschitz losses. Given a class of functions $\mathcal{C}$ denote by $\mathcal{B}_{\text{post}}$ the class of tests obtained by postprocessing the functions in $\mathcal{C}$ with an arbitrary bounded functions that is

$$\mathcal{B}_{\text{post},\delta} = \left\{ f \circ c : c \in \mathcal{C} \quad f : \mathbb{R} \to \mathbb{R} \quad |f(x)| \leq O(\log(1/\delta)^{4/3}/\delta^{2/3}) \right\}.$$

The above theorem when combined with Theorem 28 gives the following result that states that calibration with respect to the family of statistics $\mathcal{S}_{\mathcal{L}_{\text{cvx}},\delta}$ and multiaccuracy with respect to the class $\mathcal{B}_{\text{post},\delta}$ gives an omnipredictor for the family of convex, Lipschitz losses.

Corollary 34 *Let $\mathcal{C}$ be a hypothesis class. Let $\epsilon \in (0,1)$ and $\rho, \sigma \in (0,1)$. Given access to a (ρ,σ)-weak learner for $\mathcal{B}_{\text{post},\epsilon/4}$ with $\sigma \leq \rho \leq \varepsilon^{4/3}/6$ and sample complexity Z, there is an algorithm that runs in time and sample complexity*

$$\tilde{O}\left(2^{\tilde{O}(\epsilon^{-2/3})} + \frac{Z}{\epsilon^{2/3}\sigma^2}\right)$$

that produces a $\mathcal{S}_{\mathcal{L}_{\text{cvx}},\epsilon/4}$ predictor p that is a $(\mathcal{L}_{\text{cvx}}, \mathcal{C}, \epsilon)$ omnipredictor. [9]

Proof From Theorem 33, we have that the family of statistics $\mathcal{S}_{\mathcal{L}_{\text{cvx}},\delta}$ gives an approximation to $\mathcal{L}_{\text{cvx}}$ with parameters $(O(\log(1/\delta)^{4/3}/\delta^{2/3}), O(\log(1/\delta)^{4/3}/\delta^{2/3}), \delta)$. Setting $\delta = \epsilon/4$ and plugging into Theorem 32, we get the desired result. ■

9. Here $\tilde{O}$ hides polylogarithmic factors in $1/\epsilon$.

D.2. Omniprediction via moments for low-degree loss functions

Here we show how to obtain omniprediction for low-degree polynomial loss functions via sufficient statistics. We first define the family of low-degree loss functions. Let $\mathcal{M}_d = \left\{x^i : i \leq d\right\}$ denote the family of monomials of degree at most d in x.

Definition 35 *Let $\mathcal{L}^{\text{poly}}_{d,\lambda,\delta}$ be the family of loss functions for which $\mathcal{M}_d$ forms a set of sufficient statistics i.e. $\mathcal{M}_d$ gives (d,λ,δ)-uniform approximations to $\mathcal{L}^{\text{poly}}_{d,\lambda,\delta}$. Explicitly, for each $\ell \in \mathcal{L}^{\text{poly}}_{d,\lambda,\delta}$, there exits $r^\ell_i : \mathcal{A} \to \mathbb{R}$ for $i \leq d$ such that*

$$\left|\ell(y,t) - \sum_{i=0}^{d} r^\ell_i(t) y^i\right| \leq \delta$$

and

$$\sum_{i=0}^{d} \left|r^\ell_i(t)\right| \leq \lambda.$$

Furthermore, if r^ℓ_i is a polynomial of degree at most k, we say that $(\mathcal{M}_d, \mathcal{R})$ gives (d,λ,δ,k)-uniform polynomial approximations to ℓ. We denote this subclass of loss functions as $\mathcal{L}^{\text{poly}}_{d,\lambda,\delta,k}$.

The main example of losses in this class are the ℓ_p losses for even p. That is,

$$\ell_p(y,t) = (y-t)^p\,.$$

First, we will look at the basic representation of this family of loss functions. Note that

$$(y-t)^p = \sum_{i=0}^{p} \binom{p}{i} (-t)^i y^{p-i}.$$

Thus, we have

$$r^{\ell_p}_i(t) = \binom{p}{i}(-t)^i$$

and

$$\sum_{i=0}^{p} \left|r^{\ell_p}_i(t)\right| = \sum_{i=0}^{p} \binom{p}{i} |t|^i \leq \sum_{i=0}^{p} \binom{p}{i} = 2^p.$$

Clearly, for $p \leq k$, we have that $\mathcal{R}$ is a family of polynomials of degree at most k. Thus, instantiating Theorem 32, in this setting we have a $\exp(k^2)\epsilon^{-k}$ time algorithm to get omnipredictors for a class that includes all ℓ_p losses for $p < k$. Our main result in this section is an improved bound using better uniform approximations.

We present an improved approximation with degree $\sqrt{p}$ and coefficients of size $2^{\sqrt{p}}$. The following theorem is a standard application of Chebyshev approximations, but we include a proof for completeness. First, recall that the Chebyshev polynomial of degree j is defined as

$$T_j(x) = j \sum_{k=0}^{j} (-2)^k \frac{(j+k-1)!}{(j-k)!(2k)!}(1-x)^k$$

and is a key tool in approximation theory. The following is a important result from approximation theory. We include a proof of the following lemma in Appendix G for completeness.

Lemma 36 *For any $n \in \mathbb{N}$ and $\epsilon \in (0,1)$, there exists a polynomial q of degree $d = \sqrt{n \log(1/\epsilon)}$ such that*

$$|x^n - q(x)| \leq \epsilon$$

for all $x \in [0,1]$. Further, q can be represented as

$$q(x) = 2^{1-n}\left[\sum_{j \leq d} \binom{n}{\frac{n-j}{2}} \cdot T_j(x) + \mathbb{I}\left[n \equiv 0 \pmod 2\right] \cdot \frac{T_0(x)}{2}\binom{n}{\frac{n}{2}}\right]$$

where T_j is the degree j Chebyshev polynomial.

Lemma 37 *For $p \leq n$, we have that $\ell_p \in \mathcal{L}^{\text{poly}}_{d,\lambda,\delta,k}$ for $k = d = \sqrt{n \log(1/\delta)}$ and $\lambda = d^3 2^{4d}$.*

Proof Recall that from Theorem 36, we have for $d = \sqrt{n \log(1/\delta)}$ that

$$\begin{aligned}
(x-t)^n &\approx_\delta 2^{1-n} \sum_{j \leq d} \binom{n}{\frac{n-j}{2}} \cdot T_j(x-t) \\
&= 2^{1-n} \sum_{j \leq d} \binom{n}{\frac{n-j}{2}} j \sum_{k=0}^{j} (-2)^k \frac{(j+k-1)!}{(j-k)!(2k)!}(1-x+t)^k \\
&= 2^{1-n} \sum_{j \leq d} \sum_{k=0}^{j} \binom{n}{\frac{n-j}{2}} j (-2)^k \frac{(j+k-1)!}{(j-k)!(2k)!}(1-x+t)^k \\
&= 2^{1-n} \sum_{j \leq d} \sum_{k=0}^{j} \binom{n}{\frac{n-j}{2}} j (-2)^k \frac{(j+k-1)!}{(j-k)!(2k)!} \sum_{h=0}^{k} \binom{k}{h} (-x)^k (1+t)^{k-h} \\
&= 2^{1-n} \sum_{j \leq d} \sum_{k=0}^{j} \sum_{h=0}^{k} \binom{n}{\frac{n-j}{2}} j (-2)^k \frac{(j+k-1)!}{(j-k)!(2k)!} \binom{k}{h} (-x)^k (1+t)^{k-h}
\end{aligned}$$

Note that each of the terms in the above is bounded by 2^{4d}, To see this note that $2^{-n}\binom{n}{\frac{n-j}{2}} \leq 1$, $\frac{(j+k-1)!}{(j-k)!(2k)!} = \frac{1}{j+k}\binom{j+k}{2k} \leq 2^{j+k} \leq 2^{2d}$ and $\binom{k}{h} \leq 2^k \leq 2^d$ Therefore, we get the sum of the coefficients to be bounded by $d^3 2^{4d}$ and the degree is bounded by $O\left(\sqrt{n \log(1/\delta)}\right)$. ■

Given a class of functions $\mathcal{C}$ define the set of tests obtained by composing the functions with monomials of degree at most k as $\mathcal{C}^k$ i.e.

$$\mathcal{C}^k = \left\{c^j : j \leq k \quad c \in \mathcal{C}\right\}.$$

This class of tests allows us to state the omniprediction result from Theorem 32 for the particular case of low degree losses.

Theorem 38 (Omniprediction for low degree losses) *Let $\mathcal{C}$ be a hypothesis class. Let $\epsilon \in (0,1)$, $\rho, \sigma \in (0,1)$, $\lambda \in \mathbb{R}$ and $d, k \in \mathbb{N}$. Given access to a (ρ, σ)-weak learner for $\mathcal{C}^k$ with $\sigma \leq \rho \leq \varepsilon/6\lambda$ and sample complexity Z, there is an algorithm that runs in time*

$$\tilde{O}\left(d\left(\frac{\lambda}{\varepsilon}\right)^{d+5} + \frac{d}{\sigma^2}Z\right)$$

and outputs an $(\mathcal{L}^{\text{poly}}_{d,\lambda,\delta,k}, \mathcal{C}, \epsilon)$ omnipredictor.

In light of Theorem 37, this gives an improved omnipredictor for the class of ℓ_p losses for $p \leq n$ and p even. We formally state this in the following corollary.

Corollary 39 (Omniprediction for ℓ_p losses) *Let $\mathcal{C}$ be a hypothesis class. Let $\epsilon \in (0,1)$ and $\rho, \sigma \in (0,1)$. For all $n \in \mathbb{N}$ let $d = \sqrt{n \log(1/\epsilon)}$. Given access to a (ρ, σ)-weak learner for $\mathcal{C}^d$ with $\sigma \leq \rho \leq \varepsilon d^{-3} 2^{-4d}/6$ and sample complexity Z, there is an algorithm that runs in time*

$$O\left(2^{O(n \log^2(1/\epsilon))} + \frac{\sqrt{n \log(1/\epsilon)}}{\sigma}Z\right)$$

and produces an $\left(\{\ell_p\}_{p \leq n; p \text{ even}}, \mathcal{C}, \epsilon\right)$-omnipredictor.

The dependence on $\lambda = 2^n$ in the above can be improved by switching to a different basis of polynomials instead of the moment basis to get $\lambda = d$ as in Theorem 21 but we state it the above theorem in terms of the monomial basis due to the natural interpretation in terms of moments.

D.3. GLM Loss Minimization

In this section, we consider the problem of minimizing losses corresponding to generalized linear models (GLM). This family of losses arises in many natural machine learning applications due to their intimate connections to exponential families, graphical models and Bregman divergences. In particular, regression using GLM losses corresponds to maximum likelihood estimation when the data generating model is assumed to be an exponential family. Given these connections, GLM losses have been extensively studied in statistics and machine learning McCullagh (2019); Jordan (2003); Wainwright and Jordan (2008).

In the setup of GLM loss minimization, the action space is $\mathcal{A} = [-1,1]^d$. Note that this deviates from the results in the previous subsections where the actions were one dimensional. Let g be a convex loss function and let $\mathcal{S} = \{s_i\}_{i \leq d}$ be a family of functions, which we will refer to as the family of statistics. Define the loss function

$$\ell_{g,\mathcal{S}}(y,t) = g(t) - \sum_{i=1}^{d} s_i(y) t_i.$$

Here $t \in \mathcal{A}$ and $g : \mathbb{R}^d \to \mathbb{R}$ is a convex function. $\ell_{g,\mathcal{S}}$ is referred to as the generalized linear model loss corresponding to g and $\mathcal{S}$. Define the class of loss functions

$$\mathcal{L}_{\mathcal{S},\text{GLM}} = \{\ell_{g,\mathcal{S}} : g \text{ is a convex function and } g(t) \text{ is bounded in } [-1,1]\}$$

as the set of GLM losses with statistics $\mathcal{S}$. Note that the bound on t and the boundedness of g holds for many loss functions of interest. In situations that the boundedness does not hold, we can approximate the loss function by a function such that it holds. See (Gopalan et al., 2023a, Section 5) for an extended discussion on GLM losses.

First, we note that the family of statistics $\mathcal{S}$ forms a $(d, d+1, 0)$-uniform approximation for the loss function $\mathcal{L}_{\mathcal{S},\mathrm{GLM}}$.

Theorem 40 *Let $\mathcal{S}$ be a family of statistics and let $d = |\mathcal{S}|$. Then, $\mathcal{S}$ forms a $(d, d+1, 0)$-uniform approximation for the class of losses $\mathcal{L}_{\mathcal{S},\mathrm{GLM}}$ with coefficient functions $r_i(t) = -t_i$ and $r_0(t) = g(t)$.*

Proof From the definition of $\mathcal{L}_{\mathcal{S},\mathrm{GLM}}$, we have that

$$\ell_{g,\mathcal{S}}(y,t) = g(t) - \sum_{i=1}^{d} s_i(y)t_i$$

Also, recall that we use the convention $s_0 \in \mathcal{S}$ is the constant function. Thus, we have that $\ell_{g,\mathcal{S}}$ is approximated by the statistics $\mathcal{S}$ with error 0 and coefficient functions $r_i(t) = -t_i$ and $r_0(t) = g(t)$. Further, we have for all t

$$|g(t)| + \sum_{i=1}^{d} |t_i| \leq 1 + d$$

as required. ■

The main theorem that we show in this section is that for the class of GLM losses, we can compute an omnipredictor using calibration and multiaccuracy. The key aspect of this result that distinguishes it from the results in previous subsections is that the set of tests for which multiaccuracy is required is the original class of hypothesis $\mathcal{C}$.

Note that in the setting of GLM loss minimization it is natural to consider a class of function $\mathcal{C}$ consisting of functions c whose output is in d dimensions. For such a class of functions it is natural to define multiaccuracy coordinatewise i.e.

$$|\mathbb{E}_{\mathcal{D}^*}[(s_i(\mathbf{y}^*) - p_i(\mathbf{x}))c_i(\mathbf{x})]| \leq \alpha.$$

where c_i is the ith coordinate of c. The algorithmic result in Theorem 32 can be extended to this setting by assuming a weak learner for each coordinate and in the below theorem we will refer to this simply as a weak learner for the class $\mathcal{C}$.

Theorem 41 *Let $\mathcal{S}$ be a family of statistics and let $d = |\mathcal{S}|$. Let $\mathcal{C}$ be a class of functions in $\mathcal{X} \to [0,1]^d$. Let $\epsilon \in (0,1)$, $\rho, \sigma \in (0,1)$. Let p be a $\mathcal{S}$ predictor that is that is $(\mathcal{C}, \epsilon/6d)$-multiaccurate and $\epsilon/12d$-calibrated with respect to $\mathcal{S}$. Then, p is an $(\mathcal{L}_{\mathcal{S},\mathrm{GLM}}, \mathcal{C}, \epsilon)$ omnipredictor. Further, given access to a (ρ, σ) weak learner for class $\mathcal{C}$ with $\sigma \leq \rho \leq \varepsilon/12d$ and sample complexity Z, such an omnipredictor can be computed in time and sample complexity*

$$O\left(d\left(\frac{d}{\epsilon}\right)^{d+5} + \frac{dZ}{\sigma^2}\right).$$

Proof From Theorem 40, we have that $\mathcal{S}$ forms a $(d, d+1, 0)$-uniform approximation for the class of losses $\mathcal{L}_{\mathcal{S},\mathrm{GLM}}$ with coefficient functions $r_i(t) = -t_i$ and $r_0(t) = g(t)$. Further, note that from Theorem 27 and Theorem 28, we have that a $\mathcal{S}$ predictor that is $\mathcal{R} \circ \mathcal{C} = \mathcal{C}$ multiaccurate and calibrated is an omnipredictor for the class of losses $\mathcal{L}_{\mathcal{S},\mathrm{GLM}}$. The algorithmic claim then follows from Theorem 32. ■

Note that this theorem generalizes the result from Gopalan et al. (2023a) which corresponds to the one-dimensional case where the set of statistics was $\mathcal{S} = \{1, y\}$. Gollakota et al. (2023) relate the problem of omniprediction for the one-dimensional GLM case to the problem of agnostically learning single index models. In independent and concurrent work, Noarov et al. (2023) obtain results for omniprediction in the multidimensional GLM case where the means are the sufficient statistic. The approach by Noarov et al. (2023) does not focus on omniprediction for general loss classes but can be used to obtain online omniprediction results for the multidimensional GLM case. For an extended discussion, see (Noarov et al., 2023, Section 6.3.2).

Appendix E. Calibrated multiaccuracy for statistic predictors

In this section, we will address the algorithmic question of designing omnipredictors for loss functions approximated by families of statistics. As we saw in Appendix C, in order to achieve omniprediction, we need to find a predictor that is both calibrated and multiaccurate. In Appendix E.1, we will design an algorithm that produces a calibrated predictor for a family of statistics. In Appendix E.2, we will design an algorithm that produces a predictor that in addition is multiaccurate with respect to a class of tests $\mathcal{B}$.

E.1. Calibrating d-dimensional statistics

As before, we will denote by $\mathcal{S}$ the family of statistics that we would like to produce calibrated predictors for. Denote by d the cardinality of the family of statistics $\mathcal{S}$. A predictor for $\mathcal{S}$ is a function $p : \mathcal{X} \to [-1, 1]^d$ where the ith coordinate corresponds to the prediction for the ith statistic.

Let $\delta > 0$ denote a discretization parameter. We will construct predictors that predict vectors of integer multiples of δ. We partition the range of the d-dimensional predictor $[-1, 1]^d$ into $m = \lceil 1/\delta \rceil^d$ distinct subsets, denoted by $\{\mathcal{V}_1, \ldots, \mathcal{V}_m\}$. For any d-dimensional vector $j = (j_1, \ldots, j_d)$, where the element j_i is an integer in the interval $[-\lceil 1/\delta \rceil, \lceil 1/\delta \rceil - 1]$, each subset $\mathcal{V}_j$ is the Cartesian product of intervals $[j_1\delta, (j_1 + 1)\delta] \times \cdots [j_d\delta, (j_d + 1)\delta]$. We will refer to the set of all such j by $\mathcal{J}_\delta$.

We can associate any $\mathcal{S}$-predictor p with two predictors. Denote by p^δ the predictor which rounds the predictions of p to integer multiples of δ, that is,

$$p^\delta(x) = j\delta, \text{ where } j \text{ is such that } p(x) \in \mathcal{V}_j$$

and a calibrated predictor $\bar{p}$

$$\bar{p}(x) = \mathbb{E}[s(\mathbf{y})|p(x) \in \mathcal{V}_j],$$

Note that the predictor p^δ predicts vectors that δ close (in ℓ_∞) to the predictions of p for every $x \in \mathcal{X}$. While $\bar{p}$ is the result of recalibrating p^δ, the entries of $\bar{p}(x)$ would not necessarily be multiples of δ.

Define $\mathsf{ECE}_{\mathcal{S}}(p)$ to be the expected calibration error of p. That is,

$$\mathsf{ECE}_{\mathcal{S}}(p) = \mathbb{E}_{\mathcal{D}^*}\left[\|\mathbb{E}[s(\mathbf{y}^*)|p(\mathbf{x})] - p(\mathbf{x})\|_\infty\right].$$

We define the following norms on the space of $\mathcal{S}$-predictors:

$$\ell_1(p,q) = \mathbb{E}\left[\|p(\mathbf{x}) - q(\mathbf{x})\|_1\right]$$
$$\ell_2(p,q) = \mathbb{E}\left[\|p(\mathbf{x}) - q(\mathbf{x})\|_2^2\right]^{1/2}$$
$$\ell_\infty(p,q) = \max_{x\in\mathcal{X}}\left[\|p(\mathbf{x}) - q(\mathbf{x})\|_\infty\right].$$

Then, $\ell_\infty(p,p^\delta) \le \delta$ and the expected calibration error of p^δ, $\mathsf{ECE}_{\mathcal{S}}(p^\delta) = \mathbb{E}[\|p^\delta(\mathbf{x}) - \bar{p}(\mathbf{x})\|_\infty]$. Therefore, we can estimate the $\mathsf{ECE}_{\mathcal{S}}(p^\delta)$ from an empirical estimate of $\bar{p}(\mathbf{x})$.

Lemma 42 *Let $\mu, \delta \in [0,1]$ and d be the dimension of the family of statistics $\mathcal{S}$. Given access to an $\mathcal{S}$-predictor p and random samples from $\mathcal{D}^*$,*

- *There exists an algorithm $\mathsf{estECE}_{\mathcal{S}}(p,\mu)$ which returns an estimate of $\mathsf{ECE}_{\mathcal{S}}(p^\delta)$ within additive error μ. The algorithm runs in time and sample complexity $\tilde{O}(d/(\delta^d\mu^3))$.*
- *There exists an algorithm $\mathsf{reCAL}_{\mathcal{S}}(p,\delta)$ which returns a predictor $\hat{p}$ which has $\mathsf{ECE}_{\mathcal{S}}(\hat{p}) \le \delta$ and $l_1(\bar{p},\hat{p}) \le \delta$. The algorithm has time and sample complexity $\tilde{O}(d/\delta^{d+3})$.*

The main idea for $\mathsf{estECE}_{\mathcal{S}}$ is to collect enough samples, $O(d\log^2(d/\delta)/\delta^d\mu^3))$, so that we can estimate the calibration error within each prediction bin $j \in \mathcal{J}_\delta$ with high accuracy. For bins that hold significant weight in the distribution $\mathcal{D}$, i.e $\Pr_{\mathcal{D}}[p^\delta(\mathbf{x}) = j\delta] \ge \mu\delta^d/4$, the collected samples are large enough such that the empirical statistics are within a $\mu/4$-margin of the true statistic with high probability. We ignore bins with smaller proportions since their total contribution to the overall calibration error is at most $\mu/4$.

Similarly, for $\mathsf{reCAL}_{\mathcal{S}}$, we collect enough samples, $O(d\log^2(d/\delta)/\delta^{d+3}))$, so that for prediction bins that hold significant weight in the distribution, the empirical statistics are within a $\delta/4$-margin of the true statistic with high probability. We construct the predictor $\hat{p}$ to simply output the empirical statistic for each prediction bin.

Proof [Proof of Lemma 42] We take $n = O(d\log^2(d/\delta)/(\delta^d\mu^3))$ random samples (x,y) from $\mathcal{D}^*$ to construct an empirical estimate $\hat{p}$ of $\bar{p}$. For each $j \in \mathcal{J}_\delta$, let T_j refer to the set of samples (x,y) such that $p(x) \in \mathcal{V}_j$ and n_j the number of such samples. Define the value

$$\bar{s}_j = \frac{1}{n_j}\sum_{(x,y)\in T_j} s(y)$$
$$\varepsilon_j = \|\bar{s}_j - j\delta\|_\infty$$
$$\mathsf{estECE}_{\mathcal{S}} = \sum_j^{[1/\delta]^d} \frac{n_j}{n}\varepsilon_j$$

The algorithm returns the value $\mathsf{estECE}_{\mathcal{S}}$ as estimate for $\mathsf{ECE}_{\mathcal{S}}(p^\delta)$.

Now we show that $\mathsf{estECE}_{\mathcal{S}}$ is μ-close to $\mathsf{ECE}_{\mathcal{S}}(p^\delta)$ with high probability. We ignore any values of j with small proportions, that is, $\Pr[p^\delta(\mathbf{x}) = j\delta] \le \mu\delta^d/4$, since all such values only contribute

Algorithm 1 Estimate Expected Calibration Error ($\mathsf{estECE}_{\mathcal{S}}$)

Input: Predictor $p : \mathcal{X} \to [-1, 1]^d$, Error parameter μ, Discretization parameter δ, Dimension d
Output: Estimate of $\mathsf{ECE}_{\mathcal{S}}(p^{\delta})$

$n \leftarrow O(d \log^2(d/\delta)/(\delta^d \mu^3))$
Collect n samples $\{(x, y)\}$ from $\mathcal{D}^*$
Initialize $\widehat{\mathsf{ECE}_{\mathcal{S}}} \leftarrow 0$
for each $j \in \mathcal{J}_{\delta}$ **do**
    Aggregate samples $T_j = \{(x, y) \mid p(x) \in \mathcal{V}_j\}$
    Compute $\bar{s}_j = \frac{1}{|T_j|} \sum_{(x,y) \in T_j} s(y)$
    Compute $\varepsilon_j = \|\bar{s}_j - j\delta\|_{\infty}$
    $\widehat{\mathsf{ECE}_{\mathcal{S}}} \leftarrow \widehat{\mathsf{ECE}_{\mathcal{S}}} + \frac{|T_j|}{n} \varepsilon_j$
end for
return $\widehat{\mathsf{ECE}_{\mathcal{S}}}$

Algorithm 2 Recalibrate Predictor ($\mathsf{reCAL}_{\mathcal{S}}$)

Input: Predictor $p : \mathcal{X} \to [-1, 1]^d$, Discretization Parameter δ
Output: Recalibrated predictor $\hat{p}$

$n \leftarrow O(d \log^2(d/\delta)/(\delta^d \mu^3))$
Collect n samples $\{(x, y)\}$ from $\mathcal{D}^*$
for each $j \in \mathcal{J}_{\delta}$ **do**
    Aggregate samples $T_j = \{(x, y) \mid p(x) \in \mathcal{V}_j\}$
    Compute $\bar{s}_j = \frac{1}{|T_j|} \sum_{(x,y) \in T_j} s(y)$
    for each x such that $p(x) \in \mathcal{V}_j$ **do**
        Set $\hat{p}(x) = \bar{s}_j$
    end for
end for
return $\hat{p}$

$\mu/4$ to $|\mathsf{estECE}_{\mathcal{S}} - \mathsf{ECE}_{\mathcal{S}}(p^{\delta})|$ with probability 0.1. Call the other values of j large. For every large j, we have by Chernoff bounds, we have

$$\Pr[n_j \leq C(d\log(d/\delta)/\mu^2\delta^d)] \leq \frac{\delta}{30}$$

Assuming this event holds, we have

$$\Pr\left[\left|\Pr[p(\mathbf{x}) \in \mathcal{V}_j] - \frac{n_j}{n}\right| \geq \frac{\mu}{4}\right] \leq \frac{\delta}{30}$$
$$\Pr\left[\|\bar{s}_j - \mathbb{E}[s(\mathbf{y}) \mid p(\mathbf{x}) \in \mathcal{V}_j]\|_\infty \geq \frac{\mu}{4}\right] \leq \frac{\delta}{30}.$$

We take a union bound over all $[1/2\delta]^d$ possible large values. Except with error probability 0.2, none of the bad events considered above occur, and we have $|\mathsf{estECE}_{\mathcal{S}} - \mathsf{ECE}_{\mathcal{S}}(p^{\delta})| \leq \mu$. We can reduce the failure probability by repeating the estimator and taking the median. For simplicity, we ignore the failure probability.

To define the predictor $\hat{p}$, we repeat the analysis above with $\mu = \delta$. We define $\hat{p}(x) = \bar{s}_j$ if $p(x) \in \mathcal{V}_j$. We show that it is close to $\bar{p}$ in ℓ_1. The contribution of small values of j to $\mathbb{E}[|\bar{p}(\mathbf{x}) - \hat{p}(\mathbf{x}))|]$ is no more than $\mu/4$. For large buckets, we have

$$|\bar{s}_j - \bar{p}(x)| \leq |\bar{s}_j - \mathbb{E}[s(\mathbf{y})|p(\mathbf{x}) \in \mathcal{V}_j]| \leq \delta/2 + \mu/4.$$

Thus overall, the distance is bounded by $(\delta/2 + \mu/4) \leq \delta$ by our choice of μ.

Lastly, we bound the calibration error, using the fact that $\bar{p}$ is perfectly calibrated, and $\hat{p}$ is close to it $\bar{p}$. Note that both $\bar{p}$ and $\hat{p}$ are constant on all $x \in p^{-1}(\mathcal{V}_j)$. Hence

$$\begin{aligned}
\mathsf{ECE}_{\mathcal{S}}(\hat{p}) &= \mathbb{E}_{\mathcal{V}_\mathbf{j}}\left|\mathbb{E}_{p(\mathbf{x})\in\mathcal{V}_\mathbf{j}}[s(\mathbf{y}) - \hat{p}(\mathbf{x})]\right| \\
&\leq \mathbb{E}_{\mathcal{V}_\mathbf{j}}\left|\mathbb{E}_{p(\mathbf{x})\in\mathcal{V}_\mathbf{j}}[s(\mathbf{y}) - \bar{p}(\mathbf{x})]\right| + \mathbb{E}_{\mathcal{V}_\mathbf{j}}\left|\mathbb{E}_{p(\mathbf{x})\in\mathcal{V}_\mathbf{j}}[\bar{p}(\mathbf{x}) - \hat{p}(\mathbf{x})]\right| \\
&= \mathbb{E}[|\bar{p}(\mathbf{x}) - \hat{p}(\mathbf{x})|] \\
&\leq \delta
\end{aligned}$$

as required.

■

E.2. Calibrated Multiaccuracy for d-dimensional statistics

In this section, we will design algorithms that produce multiaccurate predictors for a class of tests $\mathcal{B}$. The algorithm will assume access to a weak learning oracle.

Definition 43 (Weak agnostic learning) *Let $\mathcal{B}$ be a class of tests. For parameters $\rho > \sigma > 0$, a (ρ, σ)-weak learner $\mathsf{WL}_{\mathcal{B}}$ is an algorithm $\mathsf{WL}_{\mathcal{B}}$ specified with the following input-output behavior. The input is a function $f : \mathcal{X} \to [-1, 1]$ which $\mathsf{WL}_{\mathcal{B}}$ is given access to through samples $(x, z) \sim \mathcal{D}_f$ where $\mathcal{D}_f$ corresponds to a distribution such that $\mathbf{x} \sim \mathcal{D}_{\mathcal{X}}$ and $\mathbb{E}[\mathbf{z}|\mathbf{x} = x] = f(x)$. $\mathsf{WL}_{\mathcal{B}}$ outputs either $b \in \mathcal{B}$ or $\perp$ such that the following conditions hold:*

- *If the output is $b \in \mathcal{B}$, then $\mathbb{E}[b(\mathbf{x})f(\mathbf{x})] \geq \sigma$.*

- *If there exists any $b' \in \mathcal{B}$ such that $\mathbb{E}[b'(\mathbf{x})f(\mathbf{x})] \geq \rho$, then the output cannot be $\perp$.*

The number of samples drawn from $\mathcal{D}_f$ during the execution of the algorithm is referred to as the sample complexity of $\mathsf{WL}_{\mathcal{B}}$ and the running time of the algorithm is defined in the natural way.

In our algorithms for calibrated multiaccuracy, $f(x)$ will take the form of $\frac{1}{2}(\mathbb{E}[s_i(\mathbf{y})|\mathbf{x}]-p_{t,i}(x))$ for $i \in [d]$ and $(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}$ where $\mathcal{D}$ is a corresponding to the original problem. Here $p_{t,i}(x)$ refer to the predictions of the i-th statistic s_i of a predictor $p_t(x)$. Note that $\frac{1}{2}(\mathbb{E}[s_i(\mathbf{y})|\mathbf{x}] - p_{t,i}(x)) \in [-1, 1]$. In order to simulate sample access to $\mathcal{D}_f$, we draw a sample $(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}$. Then we generate $\mathbf{z} \in \{\pm 1\}$ so that $\mathbb{E}[\mathbf{z}] = \frac{1}{2}(s_i(\mathbf{y}) - p_{t,i}(\mathbf{x}))$. Since $\frac{1}{2}(s_i(\mathbf{y}) - p_{t,i}(\mathbf{x})) \in [-1, 1]$, this uniquely specifies the distribution of $\mathbf{z}$. Moreover

$$\mathbb{E}[\mathbf{z}|\mathbf{x}] = \frac{1}{2}(\mathbb{E}[s_i(\mathbf{y})|\mathbf{x}] - p_{t,i}(\mathbf{x})) = f(\mathbf{x}).$$

Though we don't explicit allow for this in the above definition, some weak learners take as input real-valued labels; in this case, we can use $\mathbf{z} = \frac{1}{2}(s_i(\mathbf{y}) - p_{t,i}(\mathbf{x}))$ to label $\mathbf{x} \sim \mathcal{D}_{\mathcal{X}}$. Since, we are alternate between the two models of access through sampling, we will not elaborate on this further.

Multiaccuracy. A main ingredient in our algorithm is the algorithm for multiaccuracy provided in Hébert-Johnson et al. (2018). Although it is designed to achieve multiaccuracy for a single mean predictor in the boolean setting, it works for any one-dimensional statistic predictor $q : \mathcal{X} \to [-1, 1]$. The algorithm in Hébert-Johnson et al. (2018) assumes access to a (ρ, σ)-weak learner for $\mathcal{B}$ and behaves iteratively as follows: Starting with an arbitrary predictor q_0, it iteratively updates the predictor using a tests $b \in \mathcal{B}$ that correlates with the predictor. This step is repeated until no such hypothesis exists.

Algorithm 3 Multiaccuracy for one-dimensional statistic predictors (MA)

Input: Predictor $q_0 : \mathcal{X} \to [-1, 1]$
Error parameter $\alpha \in [0, 1]$.
Oracle access to a (ρ, σ) Weak learner WL for $\mathcal{B}$ under $\mathcal{D}_{\mathcal{X}}$ where $\alpha \geq \rho$.
Output: Predictor q_T.

```
t ← 0
ma ← false
while ¬ma do
  b_{t+1} ← WL(½(q* − q_t)).
  if b_{t+1} = ⊥ then
    ma ← true
  else
    h_{t+1} ← q_t + σ b_{t+1}.
    q_{t+1} ← Π(h_{t+1}) (where Π projects h_{t+1} onto [−1, 1]).
    t ← t + 1.
  end if
end while
return q_t.
```

Lemma 44 (Hébert-Johnson et al. (2018)) *Let $\mathcal{B}$ be a class of tests and $\alpha \in [0,1]$ be an accuracy parameter. There exists an algorithm $\mathsf{MA}(q_0, \mathcal{B}, \alpha)$ that takes a one-dimensional predictor $q_0 : \mathcal{X} \to [-1,1]$ and returns a predictor q_T for $T \geq 0$ where q_T is $(\mathcal{B}, \alpha)$-multiaccurate and*

$$l_2(q^*, q_0)^2 - l_2(q^*, q_T)^2 \geq T\sigma^2,$$

where $q^*(x) = \mathbb{E}[s(\mathbf{y})|\mathbf{x}]$ is the true predictor of the one-dimensional statistic.

Calibrated Multiaccuracy We now present our algorithm for finding a predictor that is both calibrated and $(\mathcal{B}, \alpha)$-multiaccurate with respect to a family of statistics $\mathcal{S}$. It follows the same outline as the algorithm for calibrated multiaccuracy in Gopalan et al. (2023a). We will set the discretization parameter to δ to be small compared to α (concretely we will choose $\delta = \alpha^2/32$). Informally, the algorithm may be viewed as starting with an arbitrary predictor $p_0 : \mathcal{X} \to [-1,1]^d$ and iteratively improving it by alternating the following steps.

1. Multiaccuracy:
   (a) For each dimension $i \in d$, run $\mathsf{MA}(p_{t,i}, \mathcal{B}, \alpha)$ to obtain $p_{t+1,i}$
   (b) Set $p_{t+1} = [p_{t+1,1}, p_{t+1,2}, \ldots, p_{t+1,d}]$ and compute the discretization p^δ_{t+1}
2. Calibration:
   (a) Estimate the calibration error of p^δ_{t+1} using $\mathsf{estECE}_{\mathcal{S}}$.
   (b) If the calibration error is low, return the predictor p^δ_{t+1}.
   (c) If the calibration error is large, recalibrate it to $\hat{p}_{t+1}$, using $\mathsf{reCAL}_{\mathcal{S}}$ and return to the multiaccuracy step.

We formally present this as Algorithm learnOmni below.

Note that if we terminate, we output a predictor that achieves both multiaccuracy and calibration, as required. The main part of the analysis is showing that the algorithm terminates in a small number of iterations. The key observation is that both steps reduce the potential function $\ell_2(p^*, p_t)^2$, which (for suitable choices of parameters) allows us to bound the overall number of iterations. We capture the overall complexity of the algorithm in the following theorem.

Theorem 32 *Let $\mathcal{S}$ be family of statistics and $\mathcal{L}$ be a family of loss functions such that that $\mathcal{S}$ gives $(d, \lambda, \varepsilon)$-uniform approximations to $\mathcal{L}$ with coefficient family $\mathcal{R}$, there exists an algorithm that returns an $(\mathcal{L}, \mathcal{C}, 4\varepsilon)$-omnipredictor[10], satisfying the following properties.*

- *The algorithm makes $O(d/\sigma^2)$ calls to a (ρ, σ)-weak learner for $\mathcal{R} \circ \mathcal{C}$ where $\sigma \leq \rho \leq \varepsilon/12\lambda$.*
- *The algorithm has time and sample complexity $\tilde{O}\left(d\left(\frac{\lambda}{\varepsilon}\right)^{d+5} + \frac{d}{\sigma^2}Z\right)$ where Z is the runtime/sample complexity of the weak learner.*

Notably, the number of calls to the weak learner is polynomial in d. This is in contrast to achieving multicalibration, for which the best known algorithm requires an exponential (in d) number of calls to the weak learner.

Our proof of Theorem 32 will rely on some results from Gopalan et al. (2023a).

10. Note that once $\mathcal{S}$ is fixed, the 3ε factor in the omniprediction slack is unavoidable.

Algorithm 4 learnOmni

Input: Parameters d, λ, ε for which $\mathcal{S}$ gives $(d, \lambda, \varepsilon)$ uniform approximations to $\mathcal{L}$
$\mathcal{S}$-predictor $p_0 : \mathcal{X} \to [-1, 1]^d$
Coefficient family $\mathcal{R} : \{r : \mathcal{A} \to \mathbb{R}\}$
Hypothesis Class $\mathcal{C} = \{c : \mathcal{X} \to \mathcal{A}\}$
Oracle access to a (ρ, σ)-Weak learner WL for $\mathcal{R} \circ \mathcal{C}$ under $\mathcal{D}_{\mathcal{X}}$ where $\sigma \leq \rho \leq \varepsilon/12\lambda$.
Output: $\mathcal{S}$-predictor q_T.

$\alpha \leftarrow \varepsilon/6d$
$\beta \leftarrow \varepsilon/6\lambda$
$\delta \leftarrow \beta^2/32$
$q_0 \leftarrow p_0$
$ma \leftarrow$ false
$t \leftarrow 0$
while $\neg ma$ **do**
  $t \leftarrow t + 1$
  for each dimension $i \in d$ **do**
    $p_{t,i} \leftarrow \mathsf{MA}(q_{t,i}, \mathcal{R} \circ \mathcal{C}, \alpha - \delta)$
  end for
  $p_t \leftarrow [p_{t,1}, p_{t,2}, \ldots, p_{t,d}]$
  if $\mathsf{estECE}_{\mathcal{S}}(p_t, \beta/4) > 3\beta/4$ **then**
    $q_t \leftarrow \mathsf{reCAL}_{\mathcal{S}}(p_t, \delta)$
  else
    $q_t \leftarrow p_t^{\delta}$
    $ma \leftarrow$ true
  end if
end while
return q_t

Corollary 45 (Generalization of Corollary 7.5 in Gopalan et al. (2023a)) *For the predictors $p^\delta, \bar{p}$ defined above,*

$$l_2(p^*, p^\delta)^2 - l_2(p^*, \bar{p})^2 \geq \mathsf{ECE}_{\mathcal{S}}(p^\delta)^2. \tag{4}$$

where $p^*(x) = \mathbb{E}[s(\mathbf{y})|\mathbf{x} = x]$

Since this is a generalization of the result from Gopalan et al. (2023a), we present its proof in Appendix F for completeness.

Lemma 46 (Gopalan et al. (2023a)) *For any predictors p_1, p_2 such that $l_1(p_1, p_2) \leq \delta$,*

$$\left|l_2(p^*, p_1)^2 - l_2(p^*, p_2)^2\right| \leq 2\delta.$$

Further, if p_1 is $(\mathcal{C}, \alpha)$-multiaccurate, then p_2 is $(\mathcal{C}, \alpha + \delta)$-multiaccurate.

Proof [Proof of Theorem 32] First we show that the predictor q_t returned by the learnOmni algorithm is $(R \circ C, \alpha)$-multiaccurate. By construction, q_t only predicts multiples of δ and is δ-close (in ℓ_∞ norm) to p_t which is $(R \circ C, \alpha - \delta)$-multiaccurate. Thus, by Theorem 46, q_t is $(\mathcal{R} \circ \mathcal{C}, \alpha)$-multiaccurate.

Observe that the predictor q_t is β-calibrated. learnOmni terminates if $\mathsf{estECE}_{\mathcal{S}}(q_t, \beta/4) \leq 3\beta/4$. Thus, the calibration error of q_t is at most β.

Since q_t is both β-calibrated and $(R \circ C, \alpha)$-multiaccurate, by Theorem 28, it is an $(\mathcal{L}, \mathcal{C}, 3d\alpha + 3\lambda\beta + 3\varepsilon)$-omnipredictor. By our choice of $\alpha = \varepsilon/6d, \beta = \varepsilon/6\lambda$, q_t is an $(\mathcal{L}, \mathcal{C}, 4\varepsilon)$-omnipredictor.

Now we show that the number of calls to the (ρ, σ)-weak learner is bounded by $O(d/\sigma^2)$. When we set $p_{t,i} = \mathsf{MA}(q_{t,i}, \mathcal{C}_i, \alpha - \delta)$, this results in $N_{t,i}$ calls to the weak learner. Thus, by Theorem 44,

$$l_2(p^*, q_{t-1,i})^2 - l_2(p^*, p_{t,i})^2 \geq N_{t,i}\sigma^2.$$

Summing over $i \in [d]$, we have

$$l_2(p^*, q_{t-1})^2 - l_2(p^*, p_t)^2 \geq \sum_{i \in [d]} N_{t,i}\sigma^2. \tag{5}$$

since $\ell_2(p^*, p_t)^2 = \sum_{i \in [d]} \ell_2(p^*, p_{t,i})^2$. In total, we make $\sum_{i \in [d]} N_{t,i}$ calls to the weak learner to obtain p_t from q_{t-1}. We wish to bound the number of loops T. To do so, we use the fact that every time the calibration error of p_t^δ is large and we have to recalibrate, our potential function $\ell_2(p^*, p_t)$ increase by a good amount. Concretely, if $\mathsf{estECE}_{\mathcal{S}}(p_t, \beta/4) \geq 3\beta/4$, then by Theorem 42, $\mathsf{ECE}_{\mathcal{S}}(p_t^\delta) \geq 3\beta/4 - \beta/4 = \beta/2$. Since $q_t = p_t^\delta$, applying Theorem 45 gives

$$l_2(p^*, p_t)^2 - l_2(p^*, q_t)^2 \geq \mathsf{ECE}_{\mathcal{S}}(p_t^\delta)^2 - 4\delta \geq \frac{\beta^2}{8}. \tag{6}$$

Adding Equations (5) and (6), for $t \in \{1, \ldots, T-1\}$,

$$l_2(p^*, q_{t-1})^2 - l_2(p^*, q_t)^2 \geq \frac{\beta^2}{8}.$$

Summing this over all t,

$$l_2(p^*, q_0)^2 - l_2(p^*, q_{T-1})^2 \geq (T-1)\frac{\beta^2}{8}.$$

Since $q_0 = p_0$ and $l_2(p^*, q_{T-1})^2 \geq 0$, we have

$$T \leq 1 + \frac{8}{\beta^2} l_2(p^*, p_0)^2 = O(1/\beta^2).$$

To bound the number of calls to the weak learner, we sum Equation (5) over all $t \in [T]$, and Equation (6) over all $t \leq T-1$ to get

$$l_2(p^*, p_T)^2 - l_2(p^*, p_0)^2 \geq \sum_{t,i} N_{t,i}\sigma^2 + (T-1)\frac{\beta^2}{8}.$$

This implies that

$$\sum_{t,i} N_{t,i} \leq d/\sigma^2$$

Since the number of calls to the weak learner in loop t is bounded by $\sum_{i\in[d]} N_{t,i} + d$, we bound the number of calls by

$$\sum_{t,i}(N_{t,i}+1) \leq \frac{d}{\sigma^2} + T = O(d/\sigma^2)$$

since $T = O(1/\beta^2)$ and $\beta \geq \rho \geq \sigma$.

■

Appendix F. Proof of Theorem 21

Theorem 21 follows immediately by applying the following proposition. We provide a proof, which uses John's theorem (Ball, 1997, Theorem 3.1) for completeness.

Fact 47 *Let $\mathcal{F}$ be a family of functions with domain $\mathcal{X}$ such that for all $f \in \mathcal{F}$, we have $|f(x)| \leq 1$. Let $\{g_i'\}_{i\leq d}$ be a family of functions that ϵ-approximately spans $\mathcal{F}$. Then, there exists $\{g_i\}_{i\leq d}$ with $|g_i(x)| \leq 1$ that ϵ-approximately span $\mathcal{F}$ with coefficients α_i satisfying*

$$\sum_i |\alpha_i(f)| \leq (1+\epsilon)d$$

for all $f \in \mathcal{F}$.

Proof [Proof of Fact 47] Let $f \in \mathcal{F}$. Let $\alpha'(f)$ be the vector such that

$$\left| f(x) - \sum_i \alpha_i'(f) g_i'(x) \right| \leq \epsilon.$$

Note that $\alpha'(f) \in \mathbb{R}^d$ for each f. Let K be the convex hull of $\{\pm\alpha'(f)\}_{f\in\mathcal{F}}$. This is a symmetric convex body in $\mathbb{R}^d$. Thus, by John's theorem (Ball, 1997, Theorem 3.1), there exists a linear

transformation T such that $B_2 \subset T(K) \subset \sqrt{d}B_2$ where $B_2 = \{x \in \mathbb{R}^k : \|x\|_2 \leq 1\}$ is the unit ball.

Consider the coefficients $\alpha(f) = T\alpha'(f) \in TK$ and functions $g_i(x) = \sum_j (T^{-1})_{j,i} g'_j(x)$. The fact that $\{g_i\}_{i \leq d}$ ϵ-approximately span $\mathcal{F}$ with coefficients α follows by construction. Note that

$$\sum_i |\alpha_i| \leq \sqrt{d}\,\|\alpha\|_2 \leq d.$$

Further, note that

$$\begin{aligned}
|g_i(x)| &\leq \sqrt{\sum_i (g_i(x))^2} \\
&\leq \sup_{\gamma \in T(K)} |\langle \gamma_i, (g_1(x), \ldots, g_d(x))\rangle| \\
&\leq \sup_{\gamma \in K} \left|\left\langle T\gamma, T^{-1}(g'_1(x), \ldots, g'_d(x))\right\rangle\right| \\
&\leq \sup_{\gamma \in K} \left|\sum_i \gamma_i g'_i(x)\right| \\
&\leq \sup_{\alpha'(f)} \left|\sum_i \alpha'_i(f) g'_i(x)\right| \\
&\leq \sup_{f \in \mathcal{F}} |f(x) + \epsilon| \leq 1 + \epsilon
\end{aligned}$$

The inequality bounding the ℓ_2 norm by the supremum over $T(K)$ follows because $B_2 \subset T(K)$. We get the desired bound by scaling down g by $(1+\epsilon)$ and scaling up α_i by $(1+\epsilon)$. ■

Proof [Proof of Corollary 45]

We express the left-hand side of Equation (4) as the difference of expectations of the 2-norms involving $p^*, p^\delta, \bar{p}$

$$\mathbb{E}\left[\left\|p^*(\mathbf{x}) - p^\delta(\mathbf{x})\right\|_2^2\right] - \mathbb{E}\left[\|p^*(\mathbf{x}) - \bar{p}(\mathbf{x})\|_2^2\right] = \mathbb{E}\left[\langle \bar{p}(\mathbf{x}) - p^\delta(\mathbf{x}), 2p^*(\mathbf{x}) - p^\delta(\mathbf{x}) - \bar{p}(\mathbf{x})\rangle\right]$$

To see this, observe that

$$\begin{aligned}
&\mathbb{E}\left[\left\|p^*(\mathbf{x}) - p^\delta(\mathbf{x})\right\|_2^2\right] - \mathbb{E}\left[\|p^*(\mathbf{x}) - \bar{p}(\mathbf{x})\|_2^2\right] \\
&= \mathbb{E}\left[\langle p^*(\mathbf{x}) - p^\delta(\mathbf{x}), p^*(\mathbf{x}) - p^\delta(\mathbf{x})\rangle\right] - \mathbb{E}\left[\langle p^*(\mathbf{x}) - \bar{p}(\mathbf{x}), p^*(\mathbf{x}) - \bar{p}(\mathbf{x})\rangle\right] \\
&= \mathbb{E}\left[\langle p^*(\mathbf{x}), p^*(\mathbf{x})\rangle - 2\langle p^*(\mathbf{x}), p^\delta(\mathbf{x})\rangle + \langle p^\delta(\mathbf{x}), p^\delta(\mathbf{x})\rangle\right] \\
&\quad - \mathbb{E}\left[\langle p^*(\mathbf{x}), p^*(\mathbf{x})\rangle - 2\langle p^*(\mathbf{x}), \bar{p}(\mathbf{x})\rangle + \langle \bar{p}(\mathbf{x}), \bar{p}(\mathbf{x})\rangle\right] \\
&= \mathbb{E}\left[\langle \bar{p}(\mathbf{x}) - p^\delta(\mathbf{x}), 2p^*(\mathbf{x}) - p^\delta(\mathbf{x}) - \bar{p}(\mathbf{x})\rangle\right]
\end{aligned}$$

We consider the distribution on subspaces induced by choosing $\mathbf{x} \sim \mathcal{D}$ and $\mathcal{V}_\mathbf{j} \ni p(\mathbf{x})$. Since p^δ and $\bar{p}$ are constant for each subspace $\mathcal{V}_j$, we can write $p^\delta(\mathcal{V}_j)$ and $\bar{p}(\mathcal{V}_j)$ for their values in this

subspace without ambiguity. Hence by first taking expectations over $\mathcal{V}_\mathbf{j}$ and then $p(\mathbf{x}) \in \mathcal{V}_\mathbf{j}$

$$\begin{aligned}\mathbb{E}\left[\langle \bar{p}(\mathbf{x}) - p^\delta(\mathbf{x}), 2p^*(\mathbf{x}) - p^\delta(\mathbf{x}) - \bar{p}(x)\rangle\right] &= \mathbb{E}_{\mathcal{V}_\mathbf{j}}\left[\langle \bar{p}(\mathcal{V}_\mathbf{j}) - p^\delta(\mathcal{V}_\mathbf{j}), \mathbb{E}_{\mathbf{x}|p(\mathbf{x})\in\mathcal{V}_j}[2p^*(\mathbf{x}) - p^\delta(\mathcal{V}_\mathbf{j}) - \bar{p}(\mathcal{V}_\mathbf{j})]\rangle\right] \\ &= \mathbb{E}_{\mathcal{V}_\mathbf{j}}\left[\left\|\bar{p}(\mathcal{V}_\mathbf{j}) - p^\delta(\mathcal{V}_\mathbf{j})\right\|_2^2\right] \\ &= \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\left\|\bar{p}(\mathbf{x}) - p^\delta(\mathbf{x})\right\|_2^2\right].\end{aligned}$$

where the final line uses $\mathbb{E}[p^*(\mathbf{x})|\mathbf{x} \in \mathcal{V}_j] = \bar{p}(\mathcal{V}_j)$.
Since $p^\delta, \bar{p}$ are both constant one each subspace $\mathcal{V}_j$, we have

$$\begin{aligned}\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\left\|\bar{p}(\mathbf{x}) - p^\delta(\mathbf{x})\right\|_2^2\right] &= \mathbb{E}_{\mathcal{V}_\mathbf{j}}\left[\mathbb{E}\left[\left\|\bar{p}(\mathbf{x}) - p^\delta(\mathbf{x})\right\|_2^2 \mid p(\mathbf{x}) \in \mathcal{V}_j\right]\right] \\ &= \mathbb{E}_{\mathcal{V}_\mathbf{j}}\left[\mathbb{E}\left[\left\|s(\mathbf{y}) - p^\delta(\mathbf{x})\right\|_2^2 \mid p(\mathbf{x}) \in \mathcal{V}_j\right]\right] \\ &\geq \mathbb{E}_{\mathcal{V}_\mathbf{j}}\left[\mathbb{E}\left[\left\|s(\mathbf{y}) - p^\delta(\mathbf{x})\right\|_2 \mid p(\mathbf{x}) \in \mathcal{V}_j\right]\right]^2 \\ &\geq \mathsf{ECE}_\mathcal{S}(p^\delta)^2\end{aligned}$$

where the first inequality uses the convexity of x^2. ■

Appendix G. Proof of Theorem 36

Proof Let T_j be the degree j Chebyshev polynomial. Recall that

$$T_j(x) = j\sum_{k=0}^{j}(-2)^k\frac{(j+k-1)!}{(j-k)!(2k)!}(1-x)^k$$

and that for $|x| \leq 1$, we have

$$|T_j(x)| \leq 1.$$

Further, we have the representation (see Szego (1939); Cody (1970))

$$x^n = 2^{1-n}\left[\sum_{j\equiv n \pmod 2; j\neq 0}\binom{n}{\frac{n-j}{2}}\cdot T_j(x) + \mathbb{I}\left[n \equiv 0 \pmod 2\right]\cdot\frac{T_0(x)}{2}\binom{n}{\frac{n}{2}}\right]$$

We truncate this up to degree $d = O(\sqrt{n\log(1/\epsilon)})$. The residual is

$$\left|2^{1-n}\sum_{j\geq d}\binom{n}{\frac{n-j}{2}}\cdot T_j(x)\right| \leq 2^{1-n}\sum_{j\geq d}\binom{n}{\frac{n-j}{2}} \leq \epsilon$$

The last step follows from the Chernoff bound. ■