On Computationally Efficient Multi-Class Calibration

Parikshit Gopalan PARIK.G@GMAIL.COM
Apple

Lunjia Hu LUNJIA@STANFORD.EDU
Stanford University

Guy N. Rothblum ROTHBLUM@ALUM.MIT.EDU
Apple

Editors: Shipra Agrawal and Aaron Roth

Abstract

Consider a multi-class labelling problem, where the labels can take values in $[k]$, and a predictor predicts a distribution over the labels. In this work, we study the following foundational question: *Are there notions of multi-class calibration that give strong guarantees of meaningful predictions and can be achieved in time and sample complexities polynomial in k?* Prior notions of calibration exhibit a tradeoff between computational efficiency and expressivity: they either suffer from having sample complexity exponential in k, or needing to solve computationally intractable problems, or give rather weak guarantees.

Our main contribution is a notion of calibration that achieves all these desiderata: we formulate a robust notion of *projected smooth calibration* for multi-class predictions, and give new recalibration algorithms for efficiently calibrating predictors under this definition with complexity polynomial in k. Projected smooth calibration gives strong guarantees for all downstream decision makers who want to use the predictor for binary classification problems of the form: does the label belong to a subset $T \subseteq [k]$: *e.g. is this an image of an animal?* It ensures that the probabilities predicted by summing the probabilities assigned to labels in T are close to some perfectly calibrated binary predictor for that task. We also show that natural strengthenings of our definition are computationally hard to achieve: they run into information theoretic barriers or computational intractability.

Underlying both our upper and lower bounds is a tight connection that we prove between multi-class calibration and the well-studied problem of agnostic learning in the (standard) binary prediction setting. This allows us to use kernel methods to design efficient algorithms, and also to use known hardness results for agnostic learning based on the hardness of refuting random CSPs to show lower bounds.

Keywords: Calibration, multi-class prediction, agnostic learning

1. Introduction

The ubiquitous use of machine learning for making consequential decisions has resulted in a renewed interest in the question *what should probabilistic predictions mean*? This question has a long history going back at least as far as the literature on forecasting (Dawid, 1982, 1984). Calibration is a classical interpretability notion for binary predictions originating in this setting that is widely used in modern machine learning. In the binary classification setting, denoting the label $\mathbf{y} \in \{0, 1\}$ and the predicted probability of 1 by $\mathbf{v} \in [0, 1]$, (perfect) calibration requires $\mathbf{E}[\mathbf{y}|\mathbf{v}] = \mathbf{v}$.

There has been renewed research interest both in the calibration of modern DNNs (Guo et al., 2017) and in foundational questions about how best to define and measure calibration to ensure

© 2024 P. Gopalan, L. Hu & G.N. Rothblum.

robustness and efficiency (Błasiok et al., 2023; Kleinberg et al., 2023) building on earlier work of (Kakade and Foster, 2008). We study calibration notions in the context of multi-class classification, where the goal is to assign one of k possible labels to each input. A predictor assigns to each input a distribution over the labels, which allows it to convey uncertainty in its predictions. Values of k in the thousands are increasingly common, especially for vision tasks (Deng et al., 2009), so the efficiency in terms of the parameter k is increasingly relevant. In this setting, even the right definition of calibration is not immediate. There are a multitude of existing definitions in theory and practice, such as confidence (Guo et al., 2017), class-wise (Kull et al., 2019), distribution (Kull and Flach, 2015) and decision (Zhao et al., 2021) calibration. However, existing notions either provide only weak guarantees for meaningful predictions, are computationally hard to achieve, or are even information theoretically hard to achieve, requiring exponential sample complexity in k.

In this work, we study the following foundational question:

Are there notions of multi-class calibration that give strong guarantees of meaningful predictions and can also be achieved with time and sample complexities polynomial in k?

Our main contribution is answering this question in the affirmative: we formulate a robust notion of *projected smooth calibration* for multi-class predictions, and give new recalibration algorithms[1] for efficiently calibrating predictors under this definition (and variants of it). We also show that natural strengthenings of this definition are computationally or information-theoretically hard to achieve. An important ingredient in showing these new upper and lower bounds is a tight connection between multi-class calibration and the well-studied problem of agnostic learning in the (standard) binary prediction setting. We proceed to elaborate on the setting, prior work, and our contributions.

Multi-class calibration. In the k-class prediction setting, we have an underlying distribution over instance-outcome pairs, where we view the outcome $\mathbf{y}$ as the one-hot encoding of a label from $[k]$. A prediction vector $\mathbf{v} \in \Delta_k$ describes a distribution in the k-dimensional simplex, where a perfect prediction would describe the exact distribution of the outcome $\mathbf{y}$ for that instance. *Canonical calibration* (Kull and Flach, 2015), also called distribution calibration, is the most stringent notion, which requires that $\mathbf{E}[\mathbf{y}|\mathbf{v}] = \mathbf{v}$ (the expectation averages over all instances for which the prediction is $\mathbf{v}$). The naive procedure for checking whether canonical calibration holds even approximately requires (after suitable discretization) conditioning on $\exp(k)$ many possible predictions in Δ_k. Indeed, we show that even the easier problem of distinguishing a perfectly calibrated predictor from one that is far from calibrated requires $\exp(k)$ samples. At the other extreme, *class-wise calibration* (Kull et al., 2019) only requires that for every $i \in [k]$, $\mathbf{E}[\mathbf{y}_i|\mathbf{v}_i] = \mathbf{v}_i$. This notion can be achieved efficiently, but we argue below that it is not sufficiently expressive.

Assume that we have a class-wise calibrated predictor and we wish to use it for downstream binary classification tasks. For instance, we might want to classify images as being those of animals, where *animals* is a subset of labels. Assume for simplicity that c for *cat* and d for *dog* are the only animals in our k labels. Class-wise calibration ensures that the predicted probabilities $\mathbf{v}_c, \mathbf{v}_d \in [0, 1]$ are each calibrated on their own: conditioned on, say, the predicted probability of cat being 0.2, the outcome should be a cat w.p. roughly 0.2. Suppose, however, that we want to predict whether the image is a cat *or* a dog. The natural probability to predict is $\mathbf{v}_c + \mathbf{v}_d$, but this might

1. The exact notion of calibrating a predictor has to be defined carefully to avoid trivial solutions (for example, the constant predictor that always outputs the empirical mean is perfectly calibrated). Following much of the literature, our algorithms post-process a given predictor to make it calibrated while not increasing the squared loss.

be far from calibrated w.r.t the actual probability that the outcome is a cat or a dog (even though the predictor is class-wise calibrated). This reveals a weakness of class-wise calibration that is also shared by other guarantees that we know how to achieve efficiently (such as *confidence calibration* (Guo et al., 2017), see below): their calibration guarantees are rather fragile, and break down when used in downstream tasks.

Aiming to achieve rigorous downstream guarantees, Zhao *et al.* (Zhao et al., 2021) introduced *decision calibration*, which can be achieved in $\mathrm{poly}(k)$ sample complexity.[2] However we show that the algorithmic task they aim to solve is as hard as agnostically learning halfspaces, and hence is unlikely to be achievable in time $\mathrm{poly}(k)$ by results of (Daniely, 2016).

To summarize, the state of the art for multiclass calibration notions:

- There are **efficient** notions, such as classwise and confidence calibration, but they are not very expressive. In particular their calibration guarantees are rather fragile and do not imply good guarantees for downstream tasks.

- There are **expressive** notions, such as canonical calibration and the recently proposed notion of decision calibration, but they are inefficient. These notions run into information or complexity theoretic barriers, which prevent them from being achievable in running time and sample complexity $\mathrm{poly}(k)$.

This motivates our foundational question: is there an *expressive and efficient* notion of multiclass calibration? Such a notion should give robust calibration guarantees for downstream tasks, and should be achievable in $\mathrm{poly}(k)$ time and sample complexity. More broadly, is there a general framework for understanding the complexity of various calibration notions? Ideally, such a framework would let us identify broad classes of notions that are efficiently achievable and identify computational and information-theoretic barriers to other notions.

These questions are motivated not only by the use of calibration as an notion of interpretability for probabilistic predictions in machine learning, but also by the recent applications of calibration to fairness (Hébert-Johnson et al., 2018), loss minimization (Gopalan et al., 2022a, 2023) and indistinguishability (Dwork et al., 2021, 2022, 2023). In the multi-class setting with k labels, algorithms for all of these notions become exponential in k, which stems from the fact that they try to achieve canonical calibration or similarly expressive notions (see for example (Gopalan et al., 2022a; Dwork et al., 2022)). We see formulating more efficient notions of calibration as a step towards more efficient algorithms for these applications in the multiclass setting.

1.1. Our Contributions

We start by describing a unifying framework from (Gopalan et al., 2022b) for various notions of multiclass calibration, for which we need some notation. Let $\Delta_k \subseteq \mathbb{R}^k$ denote the probability simplex for k outcomes. Given a distribution $\mathcal{D}_0$ on $(\mathbf{x}, \mathbf{y})$ pairs where $\mathbf{y} \in \{0,1\}^k$ is the one-hot encoding of a label and a predictor p, let $\mathbf{v} = p(\mathbf{x}) \in \Delta_k$ be the prediction of p. Let $\mathcal{D}$ denote the induced distribution of $(\mathbf{v}, \mathbf{y})$.

2. The paper claims the notion is both time and space efficient, but their main result (Zhao et al., 2021, Theorem 2) only proves a bound on sample complexity. See the discussion in Sections 1.2 and A and D.

Weighted calibration. As observed by various works (Dwork et al., 2021; Gopalan et al., 2022b; Dwork et al., 2022; Błasiok et al., 2023), calibration is essentially a notion of indistinguishability of distributions. For multiclass learning, perfect canonical calibration requires that for every $\mathbf{v} \in \Delta_k$, $\mathbf{E}[\mathbf{y}|\mathbf{v}]$ (which completely describes the distribution of $\mathbf{y}$ conditioned on $\mathbf{v}$) equals $\mathbf{v}$. If we relax equality to expected closeness in ℓ_1 distance $\mathbf{E}_{\mathbf{v}}[|\mathbf{E}[\mathbf{y}|\mathbf{v}] - \mathbf{v}|]$, we arrive at the notion of the expected calibration error or ECE. This notion requires $\exp(k)$ samples to estimate (see Theorem 15); it is also not robust to small perturbations of the predictor p (Kakade and Foster, 2008; Błasiok et al., 2023). We aim for relaxed calibration notions that capture the same underlying principle that $\mathbf{E}[\mathbf{y}|\mathbf{v}]$ and $\mathbf{v}$ are "close" under $\mathcal{D}$, but which are efficient to estimate, and also do not suffer from the same kind of non-robustness.

Following (Gopalan et al., 2022b), we work with the definition of weighted calibration, which is general enough to capture all the aforementioned notions of calibration. For a hypothesis class $\mathcal{H} := \{h : \Delta_k \to [-1, 1]\}$, we consider the family of *weight functions* $\mathcal{H}^k$ mapping $\Delta_k \to [-1, 1]^k$, where for every $i \in [k]$, coordinate i of the output is a function $h_i \in \mathcal{H}$ of the input. Define the weighted calibration error as

$$\mathsf{CE}_{\mathcal{H}^k}(\mathcal{D}) := \max_{w \in \mathcal{H}^k} \left| \mathop{\mathbf{E}}_{(\mathbf{v},\mathbf{y})\sim\mathcal{D}} [\langle w(\mathbf{v}), \mathbf{y} - \mathbf{v}\rangle] \right| = \max_{w \in \mathcal{H}^k} \left| \mathop{\mathbf{E}}_{\mathbf{v}} [\langle w(\mathbf{v}), \mathbf{E}[\mathbf{y}|\mathbf{v}] - \mathbf{v}\rangle] \right|.$$

This can be seen as requiring closeness of the distributions of $\mathbf{v}$ and $\mathbf{E}[\mathbf{y}|\mathbf{v}]$ to the class of distinguishers $\mathcal{H}^k$ in the spirit of pseudorandomness. Taking $\mathcal{H}$ to be all functions on Δ_k bounded by 1 in absolute value recovers the notion of ECE. Relaxing the space of distinguishers weakens the definition. Are there distinguisher families where the calibration guarantee remains meaningful, while simultaneously allowing for efficient *auditing*: deciding whether a given predictor satisfies $\mathsf{CE}_{\mathcal{H}^k}(\mathcal{D}) \le \alpha$?

Projected smooth calibration. We now formulate projected smooth calibration, a weighted calibration notion that satisfies our desiderata. As discussed above, we want to ensure the following **subset calibration** guarantee: for every subset $T \subseteq [k]$ of labels, the probabilities assigned by our predictor to the event that the label belongs to T should be calibrated. Let $\mathbf{v} \in \Delta_k$ denote the prediction of our predictor. Letting $\mathbf{1}_T \subseteq \{0, 1\}^k$ denote the indicator vector of T, the indicator for the event that the outcome $\mathbf{y}$ is in T is $\mathbf{1}_T \cdot \mathbf{y}$, whereas the predicted probability is $\mathbf{1}_T \cdot \mathbf{v}$. Say we want to enforce the calibration condition that when the predicted probability of belonging to T exceeds $v \in [0, 1]$, the label indeed lies in T with roughly the predicted probability. We can view this as requiring a bound on

$$\left|\mathbf{E}[\mathbb{I}(\mathbf{1}_T \cdot \mathbf{v} \ge v)(\mathbf{1}_T \cdot \mathbf{y} - \mathbf{1}_T \cdot \mathbf{v})]\right| = \left|\mathbf{E}[\mathbb{I}(\mathbf{1}_T \cdot \mathbf{v} \ge v)\mathbf{1}_T \cdot (\mathbf{y} - \mathbf{v})]\right|$$

where $(\mathbb{I}(\mathbf{1}_T \cdot \mathbf{v} \ge v)\mathbf{1}_T) \in \{0, 1\}^k$ is a vector-valued function on Δ_k, which takes the value $\mathbb{I}(\mathbf{1}_T \cdot \mathbf{v} \ge v)$ for coordinates in T, and the value 0 for coordinates outside T. The good news is that this setup fits the template of weighted calibration, where the class $\mathcal{H}$ contains all functions the form $\mathbb{I}(\mathbf{1}_T \cdot \mathbf{v} \ge v)$ for $T \in \{0, 1\}^k, v \in \mathbb{R}$. The bad news is that we will show this problem is as hard as agnostically learning halfspaces. Daniely (Daniely, 2016) showed that, assuming the hardness of refuting random Xor CSPs, this problem cannot be solved in polynomial time.

In projected smooth calibration, we replace the hard thresholds $\mathbb{I}(\mathbf{1}_T \cdot \mathbf{v} \ge v)$ with the class $\mathcal{H}_{\text{pLip}} = \{\phi(\mathbf{a}\cdot\mathbf{v})\}$ where $\phi : [-1, 1] \to [-1, 1]$ is a Lipschitz continuous function and $\mathbf{a} \in [-1, 1]^k$.

In particular, this includes indicator vectors for subsets. Projected smooth calibration requires that the weighted calibration error for the weight family $\mathcal{H}^k_{\text{pLip}}$ is bounded.

Definition 1 (Projected smooth calibration, informal statement of Theorem 11) *For a joint distribution $\mathcal{D}$ on predictions $\mathbf{v}$ and true outcomes $\mathbf{y}$, the* projected smooth calibration error *is*

$$\text{psCE}(\mathcal{D}) := \sup_{w \in \mathcal{H}^k_{\text{pLip}}} \left| \mathop{\mathbf{E}}_{(\mathbf{v},\mathbf{y})\sim\mathcal{D}} [(\mathbf{y} - \mathbf{v})w(\mathbf{v})] \right|,$$

where $\mathcal{H}^k_{\text{pLip}}$ is the class of functions $w : \Delta_k \to [-1,1]^k$ such that for every coordinate i in w's output, denoted by $w^{(i)}$, there are a 1-Lipschitz function $\phi^{(i)} : [-1,1] \to [-1,1]$ and a vector $\mathbf{a}_i \in [-1,1]^k$ s.t. $w^{(i)}(\mathbf{v}) = \phi^{(i)}(\mathbf{a}_i \cdot \mathbf{v})$ for every $\mathbf{v} \in \Delta_k$.

A predictor satisfies projected smooth calibration if its error psCE is bounded. We show that this definition satisfies all our desiderata:

Property (1): expressive power. Projected smooth calibration guarantees that for every subset $T \subseteq [k]$, the predicted probabilities for the binary classification task (namely, the outcome is in T) satisfy smooth calibration, a well-studied calibration notion with several desirable properties (Kakade and Foster, 2008; Gopalan et al., 2022b; Błasiok et al., 2023). In particular, this implies that the predicted probabilities $\mathbf{1}_T \cdot \mathbf{v}$ that the outcome will be in T are close to being perfectly calibrated. The proof builds on the work of (Błasiok et al., 2023). In particular, for each such subset T, there exists a perfectly calibrated predictor p^*_T for the binary classification task of determining whether the outcome will land in T, whose predictions are close to $\mathbf{1}_T \cdot \mathbf{v}$ in earthmover distance. Thus, we get meaningful guarantees for a rich collection of downstream binary classification tasks (including subset membership and more).

Property (2): Computational efficiency. We show an efficient algorithm for auditing whether the projected smooth calibration error of a predictor is bounded. Here we state our result informally, the formal statement is in Theorem 38.

Theorem 2 (Efficient auditing, informal statement of Theorem 38) *There is an algorithm for deciding whether the projected smooth calibration error is at most α, with sample complexity and running time $O(k^{O(1/\alpha)})$.*

The work of (Shalev-Shwartz et al., 2011) showed that agnostic learning halfspaces becomes tractable if we replace the hard thresholds used in halfspaces with Lipschitz transfer functions. Building on their techniques, and using Jackson's Theorem on low-degree uniform approximations for Lipschitz functions, we show that auditing for projected smooth calibration is polynomial time solvable. Moreover, our auditing algorithm is quite simple and does not need to solve a convex program, generalizing results in (Kumar et al., 2018; Błasiok et al., 2023). Our algorithm in fact solves the associated search problem (see Definition 21): if p is not calibrated, it finds a witness to the lack of calibration, which can be used to post-process p and reduce its calibration error without increasing the squared loss (see Theorem 22). Defining a recalibration algorithm correctly is subtle, see Definition 21, Theorem 22 and the discussion around them.

Our algorithm has running time polynomial in k for every fixed constant α, in contrast with previous results for expressive notions of calibration. If we only care about auditing using $\phi(w \cdot \mathbf{v})$

for vectors w where $\|w\|_2^2 \leq m$ then the sample complexity can be bounded by $m^{O(1/\alpha)}$. This gives a running time fixed polynomial in k (but exponential in α) for subset calibration where we only care about bounded size subsets. One can also get better run times by restricting the family of Lipschitz functions ϕ. By restricting to auditors of the form $\tanh(w \cdot \mathbf{v})$ we get the weaker notion of **sigmoid calibration**, for which the auditor runs in time $k^{O(\log(1/\alpha))}$. This can be seen as a smooth relaxation of the intractable notion of halfspace auditing. However, the improved efficiency comes at the price of some expressivity, we do not get closeness to perfect calibration for downstream subset classification tasks.

It is interesting to investigate whether the exponential dependence on $1/\alpha$ in Theorem 2 can be avoided. As a result in this direction, we show that the running time cannot be improved to $\mathrm{poly}(k, 1/\alpha)$:

Theorem 3 (Informal statement of Theorem 47) *Under standard complexity-theoretic assumptions, there is no algorithm that can decide whether the projected smooth calibration error is at most α with sample complexity and running time $k^{O(\log^{0.99}(1/\alpha))}$.*

We prove the theorem by showing a reduction from the task of refuting random XOR formulas. Getting the right exponent for k as a function of $1/\alpha$ is an interesting question for future work.

Property (3): Robustness. The works of (Kakade and Foster, 2008; Foster and Hart, 2018; Błasiok et al., 2023) advocate the use of Lipschitz functions in defining calibration since it results in robust measures that do not change drastically under small perturbations of the predictor. Since projected smooth calibration is defined using Lipschitz functions, it is a robust calibration measure.

Lower bounds for stronger notions. The discussion above suggests possible strengthenings of the notion of projected smooth calibration. The weight function family $\phi(w \cdot \mathbf{v})$ is a subset of the family fLip of all Lipschitz functions $\psi : \Delta_k \to [-1, 1]$. We could imagine using fLip^k as our weight function family to get a stronger notion of calibration, which we call *full smooth calibration*. We show in Theorem 17 that this notion is information-theoretically intractable and requires $\exp(k)$ samples.

Theorem 4 (Informal statement of Theorem 17) *Any algorithm to decide whether the full smooth calibration error is 0 or exceeds a positive absolute constant requires $\exp(k)$ samples.*

In our closeness to calibration guarantee, for every $T \subset [k]$ there exists a binary predictor p^*_T whose predictions are close to $\mathbf{1}_T \cdot \mathbf{v}$ and which is perfectly calibrated. But the different p^*_T for various Ts might not be consistent, meaning they need not arise as $p^*_T = \mathbf{1}_T \cdot p^*$ where p^* is a perfectly calibrated predictor (independent of the choice of T). Could we instead measure calibration error by comparing our predictions to those made by a single calibrated predictor p^*? Put differently, projected smooth calibration guarantees that our predictions on each subset T are *locally* close to the predictions of a calibrated binary predictor p^*_T. We are now asking whether one can measure *global* closeness to a single perfectly calibrated predictor p^*.

There is prior work that suggests measuring closeness to calibration in terms of distance of its predictions from the nearest perfectly calibrated predictor $\mathbf{v}^*$. This notion, called distance to calibration, was studied in the binary setting by (Błasiok et al., 2023), where it plays a central role in their theory of *consistent calibration measures*. Such measures are ones that approximate the distance to calibration within polynomial factors. They identified several efficiently computable

consistent calibration measures in the binary setting, including smooth calibration and Laplace kernel calibration.

We show a strong negative result for measuring or even weakly approximating the distance to calibration in the multiclass prediction setting:

Theorem 5 (Informal statement of Theorem 15) *Any algorithm to decide whether the distance to calibration is* 0 *or exceeds a positive absolute constant requires* $\exp(k)$ *samples.*

In contrast to the work of (Błasiok et al., 2023), Theorem 5 shows that in the multiclass setting, any consistent calibration measure requires $\exp(k)$ samples.

These lower bounds stem from an indistinguishability argument that we sketch below. We take $V \subset \Delta_k$ of size $\exp(k)$ of predictions that are $\Omega(1)$ far from each other. We construct two distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ on predictions and labels. In either distribution, the marginal distribution on predictions $\mathbf{v}$ is uniform on V. In $\mathcal{D}_1$, the distribution on labels $\mathbf{y}_1$ conditioned on $\mathbf{v}$ is perfectly calibrated, so that $\mathbf{E}[\mathbf{y}_1|\mathbf{v}] = \mathbf{v}$. In $\mathcal{D}_2$, for each $\mathbf{v} \in V$, the label $\mathbf{y}_2$ is fixed to be some single value. We imagine this label $\mathbf{y}_2$ being picked at random with $\mathbf{E}[\mathbf{y}_2|\mathbf{v}] = \mathbf{v}$ the very first time we predict $\mathbf{v}$. Every subsequent time we see the prediction $\mathbf{v}$, we will see this same label $\mathbf{y}_2|\mathbf{v}$. A sampling algorithm cannot tell the difference between these distributions until it sees multiple samples with the same value of $\mathbf{v} \in V$, since until that point, samples from the two distributions are identically distributed. By the birthday paradox, this requires $\Omega(\sqrt{|V|}) = \exp(k)$ samples.

Equivalence between auditing and agnostic learning. Underlying our algorithms and hardness results is a tight characterization of efficient auditing in terms of agnostic learning. We elaborate on these two computational tasks. The auditing task for the class of weight functions $\mathcal{H}^k$ gets as input a predictor p, and needs to decide whether it has large calibration error for $\mathcal{H}^k$. If so, then the auditor should also return a weight function w' that has large calibration error (in the spirit of weak agnostic learning (Ben-David et al., 2001; Kalai et al., 2008b), we allow for a gap between the largest calibration error in $\mathcal{H}^k$ and the error of the weight function found by the auditor). As noted in the discussion following Theorem 2, solving the auditing task also allows us to efficiently recalibrate a given predictor to achieve low weighted calibration error for the class $\mathcal{H}^k$. Weak agnostic learning for a class $\mathcal{H}$ is a standard learning problem in the binary (not multi-class) classification setting, where we have a distribution on $\Delta_k \times \{\pm 1\}$ and the goal is to find a witness given the existence of $h \in \mathcal{H}$ with correlation $\mathbf{E}_{(\mathbf{v},z)\sim\mathcal{D}}[h(\mathbf{v})z]$ at least γ. We show that *auditing for $\mathcal{H}^k$ is efficient iff the class $\mathcal{H}$ is efficiently weakly agnostically learnable.*

Theorem 6 (Informal statement of Theorems 24 and 25) *Auditing for $\mathcal{H}^k$ and agnostic learning for $\mathcal{H}$ reduce to each other efficiently.*[3] *The calibration error parameter in auditing corresponds to the correlation parameter in learning up to a constant factor.*

Connections between auditing for calibration and agnostic learning have appeared in (Hébert-Johnson et al., 2018) and subsequent works. The focus was on binary or scalar prediction tasks, where the challenge is guaranteeing calibration for many different subsets of the feature vectors. The challenge in our work is different: we aim to guarantee calibration w.r.t. the k-dimensional

3. We use an auditor for a slightly different class $\tilde{\mathcal{H}}^k$ to solve the learning task for $\mathcal{H}$, where $\tilde{\mathcal{H}}$ is obtained from $\mathcal{H}$ by taking a simple affine transformation of the input. In particular, the two classes are the same when $\mathcal{H}$ is the class of halfspaces.

multi-class outcome vector $\mathbf{y}$, and to relate this task to agnostic learning with binary labels. As we show in Appendix C.1, applying a learning algorithm to an auditing task in a coordinate-wise manner would result in losing a factor of k in the calibration error. This loss of k would result in auditing algorithms that do not run in time $\mathrm{poly}(k)$ even for constant α. We show that this loss can always be avoided by applying the learning algorithm on carefully constructed conditional distributions, giving a tight connection up to constant factors in the calibration error.

The equivalence between auditing and learning allows us to apply a rich set of techniques from the literature for agnostic learning to show both hardness results and efficient algorithms for auditing tasks. In particular, our hardness result for auditing decision calibration (Theorem 27) is based on the hardness of agnostically learning halfspaces shown in previous work (Daniely, 2016). In general, our auditing algorithms can be instantiated with weight functions that have bounded norm in any reproducing kernel Hilbert space over Δ_k, as long as the corresponding kernel can be evaluated efficiently. We apply polynomial approximation theorems and the multinomial kernel used in learning algorithms (Shalev-Shwartz et al., 2011; Goel et al., 2017, 2020) to give efficient auditors for projected smooth calibration and sigmoid calibration.

1.2. Further Discussion of Related Work

As discussed above, many works have discussed notions of calibration for multi-class prediction. These either offer limited expressiveness, or require super-polynomial runtime or sample complexities. We further elaborate on two recent works (Zhao et al., 2021; Dwork et al., 2022) that achieve polynomial sample complexity, but suffer from computational intractability. We also discuss the work of (Gopalan et al., 2022b; Kleinberg et al., 2023; Noarov et al., 2023).

Perhaps the most closely related work to ours is Zhao *et al.*'s (Zhao et al., 2021) work on decision calibration. They imagine a down-stream decision maker using the predictions to choose between a finite set of actions, subject to a loss function that depends only on the action and on the outcome. The predicted distribution should be indistinguishable from the true distribution in terms of the loss experienced by the decision maker (and this should hold for any such decision maker and any loss function). We view this as an expressive calibration notion: in particular, even if we only allow for two possible actions, decision calibration (see Definition 9) guarantees a sharp flavor of subset calibration: for any subset $T \subseteq [k]$ and any threshold $b \in [0, 1]$, conditioning on instances where the predictor assigns total probability at least b to the set T, the probability that the outcome lands in T is at least b (up to a small error).[4] They showed that this strong guarantee can be obtained using only $\mathrm{poly}(k)$ samples. We show, however, that the runtime complexity of obtaining decision calibration cannot be $\mathrm{poly}(k)$ (assuming the hardness of refuting random CSPs). Intuitively, the hardness is due to the "sharpness" of the guarantee: conditioning on the event that the probability of T is *exactly* above the threshold b. This has the flavor of a halfspace learning guarantee, and this underlies our intractability result. In contrast, our notion of projected smooth calibration (and our results on sigmoids) enforces a "softer" Lipschitz condition, which makes the problem computationally tractable and allows us to construct efficient algorithms. On a more technical level, Zhao *et al.* (Zhao et al., 2021) require solving an optimization problem over the class of halfspaces. Noting that the objective is not differentiable, they present a heuristic gradient-based algorithm after relaxing the hard halfspace threshold using a differentiable sigmoid function. They allow the Lipschitz constant of the sigmoid function to grow arbitrarily large in order to recover the halfspaces in the

4. The guarantee is even stronger: the conditional expectation of the predictions and the outcomes should be close.

limit. However, they do not provide a provable guarantee on the correctness or efficiency of their algorithm. We show that such guarantees are unlikely to be established due to inherent intractability of the problem (see above and in Appendix D).

In their work on outcome indistinguishability "beyond Bernoulli", Dwork *et al.* (Dwork et al., 2022) also study meaningful predictions over non-Boolean outcome spaces. Their notion of Generative OI guarantees indistinguishability for a rich class of distinguishers that can examine the prediction and also features of the particular instance. This is quite expressive, and in particular, by formulating an appropriate class of distinguishers, their framework can capture all notions considered in this work. Their most general algorithm, for an outcome space of size k and a (finite) class of distinguishers $\mathcal{A}$, requires sample complexity that is logarithmic in k and in $|\mathcal{A}|$. The runtime is at least linear in the number of distinguishers $|\mathcal{A}|$. Guaranteeing subset calibration would require (at least) $\exp(k)$ distinguishers, so while their algorithm would be sample-efficient, its runtime is exponential in k.

The work of (Gopalan et al., 2022b) formulated the general notion of weighted calibration that we use. Their focus is on a particular instantiation of this notion they call low degree calibration, where the weight family is $P(d,1)^k$, where $P(d,1)$ contains all degree d polynomials in $\mathbf{v}$ with absolute values of coefficients summing to 1. They do not consider the downstream calibration guarantees for binary classification tasks, rather their focus is on multicalibration and multigroup fairness. They present an auditing algorithm that runs in time $O(k^d)$. We show that by using kernel methods, one can obtain an auditor with running time $\mathrm{poly}(k,d)$ (Lemma 35).

Recently, (Kleinberg et al., 2023) and (Noarov et al., 2023) studied relaxations of canonical calibration and gave algorithms for achieving them in the online setting. Similar to our work, they were motivated by giving meaningful guarantees for downstream tasks while avoiding the inefficiency inherent in canonical calibration. However, their goal is to make calibrated predictions, which is challenging in the online setting, but becomes trivial in our offline setting (the constant predictor that always outputs the expectation of $\mathcal{D}$'s outcomes is calibrated). Therefore, we focus instead on the auditing task of post-processing a given predictor. This auditing task is also considered in (Noarov et al., 2023), which gives online algorithms with running time growing polynomially with the size of the family of weight functions. We study the offline setting, where achieving running time that is linear in the size of the family of weight functions follows from (Gopalan et al., 2022b), and our focus is on achieving polynomial running time even when the family of weight functions is exponential or infinite.

As in the standard setup of calibration, each prediction is a probability distribution over the possible labels. This distribution conveys the uncertainty of the predictor about the true label, and calibration can be viewed as a guarantee of accurate uncertainty quantification. Another common method for uncertainty quantification is conformal prediction (see e.g. (Shafer and Vovk, 2008; Angelopoulos and Bates, 2021)), where the predictor outputs *prediction sets* (sets of labels) aiming to provide a *coverage guarantee*: the true label belongs to the prediction set with a certain, prespecified probability. A recent line of work applies techniques from multicalibration to get robust conformal prediction algorithms that give coverage guarantees that hold not just on average, but *conditionally* on every important subpopulation and beyond (Gupta et al., 2022; Bastani et al., 2022; Jung et al., 2023).

1.3. Organization

The rest of the paper is organized as follows. We start by defining old and new notions of multi-class calibration and discussing the connections among them in Appendix A. We prove an exponential sample complexity lower bound for canonical calibration in Appendix B. We define the auditing task and show a tight connection to agnostic learning in Appendix C. We apply the connection to show hardness of auditing for decision calibration and halfspaces in Appendix D. We describe a general kernel method for auditing in Appendix E and apply it to give efficient auditors for projected smooth calibration and sigmoid calibration in Appendix F. We show barriers to further improving the efficiency of our algorithms by proving additional computational lower bounds in Appendix G.

Acknowledgments

Part of this work done while LH was interning at Apple. LH is also supported by Moses Charikar's and Omer Reingold's Simons Investigators awards, Omer Reingold's NSF Award IIS-1908774, and the Simons Foundation Collaboration on the Theory of Algorithmic Fairness.

References

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Osbert Bastani, Varun Gupta, Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Practical adversarial multivalid conformal prediction. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 29362–29373. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/bcdaaa1aec3ae2aa39542acefdec4e4b-Paper-Conference.pdf.

Shai Ben-David, Philip M. Long, and Yishay Mansour. Agnostic boosting. In *14th Annual Conference on Computational Learning Theory, COLT*, 2001.

Jarosław Błasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. A unifying theory of distance from calibration. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, page 1727–1740, 2023.

E. Cheney. *Introduction to approximation theory*. McGraw-Hill, New York, 1966.

Amit Daniely. Complexity theoretic limitations on learning halfspaces. In *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*, page 105–117, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450341325.

Amit Daniely, Nati Linial, and Shai Shalev-Shwartz. From average case complexity to improper learning complexity. In *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing*, STOC '14, page 441–448, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450327107. doi: 10.1145/2591796.2591820. URL https://doi.org/10.1145/2591796.2591820.

A. P. Dawid. Objective probability forecasts. *University College London, Dept. of Statistical Science. Research Report 14*, 1982.

A Philip Dawid. Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, 147 (2):278–290, 1984.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Cynthia Dwork, Michael P. Kim, Omer Reingold, Guy N. Rothblum, and Gal Yona. Outcome indistinguishability. In *ACM Symposium on Theory of Computing (STOC'21)*, 2021. URL https://arxiv.org/abs/2011.13426.

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 *The 33rd International Conference on Algorithmic Learning Theory*, 2022.

Cynthia Dwork, Daniel Lee, Huijia Lin, and Pranay Tankala. From pseudorandomness to multi-group fairness and back. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 3566–3614. PMLR, 12–15 Jul 2023. URL https://proceedings.mlr.press/v195/dwork23a.html.

Uriel Feige. Relations between average case complexity and approximation complexity. In John H. Reif, editor, *Proceedings on 34th Annual ACM Symposium on Theory of Computing, May 19-21, 2002, Montréal, Québec, Canada*, pages 534–543. ACM, 2002. doi: 10.1145/509907.509985. URL https://doi.org/10.1145/509907.509985.

Vitaly Feldman. Distribution-specific agnostic boosting. In Andrew Chi-Chih Yao, editor, *Innovations in Computer Science - ICS 2010, Tsinghua University, Beijing, China, January 5-7, 2010. Proceedings*, pages 241–250. Tsinghua University Press, 2010. URL http://conference.iiis.tsinghua.edu.cn/ICS2010/content/papers/20.html.

Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. On agnostic learning of parities, monomials, and halfspaces. *SIAM J. Comput.*, 39(2):606–645, 2009.

Dean P. Foster and Sergiu Hart. Smooth calibration, leaky forecasts, finite recall, and nash dynamics. *Games Econ. Behav.*, 109:271–293, 2018. URL https://doi.org/10.1016/j.geb.2017.12.022.

Surbhi Goel, Varun Kanade, Adam R. Klivans, and Justin Thaler. Reliably learning the relu in polynomial time. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, volume 65 of *Proceedings of Machine Learning Research*, pages 1004–1042. PMLR, 2017.

Surbhi Goel, Adam R. Klivans, and Frederic Koehler. From boltzmann machines to neural networks and back again. In *Annual Conference on Neural Information Processing Systems 2020*, 2020.

Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In *Innovations in Theoretical Computer Science (ITCS'2022)*, 2022a. URL https://arxiv.org/abs/2109.05389.

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, 2022b.

Parikshit Gopalan, Lunjia Hu, Michael P. Kim, Omer Reingold, and Udi Wieder. Loss Minimization Through the Lens Of Outcome Indistinguishability. In Yael Tauman Kalai, editor, *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*, volume 251 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 60:1–60:20, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-263-1. doi: 10.4230/LIPIcs.ITCS.2023.60. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ITCS.2023.60.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

Chirag Gupta and Aaditya Ramdas. Top-label calibration and multiclass-to-binary reductions. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=WqoBaaPHS-.

Varun Gupta, Christopher Jung, Georgy Noarov, Mallesh M. Pai, and Aaron Roth. Online Multivalid Learning: Means, Moments, and Prediction Intervals. In Mark Braverman, editor, *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*, volume 215 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 82:1–82:24, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-217-4. doi: 10.4230/LIPIcs.ITCS.2022.82. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ITCS.2022.82.

Venkatesan Guruswami and Prasad Raghavendra. Hardness of learning halfspaces with noise. In *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2006), 21-24 October 2006, Berkeley, California, USA, Proceedings*, pages 543–552. IEEE Computer Society, 2006.

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

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Dk7QQp8jHEo.

Sham Kakade and Dean Foster. Deterministic calibration and nash equilibrium. *Journal of Computer and System Sciences*, 74(1):115–130, 2008.

Adam Kalai and Varun Kanade. Potential-based agnostic boosting. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.,

2009. URL https://proceedings.neurips.cc/paper/2009/file/13f9896df61279c928f19721878fac41-Paper.pdf.

Adam Tauman Kalai, Adam R. Klivans, Yishay Mansour, and Rocco A. Servedio. Agnostically learning halfspaces. *SIAM J. Comput.*, 37(6):1777–1805, 2008a. doi: 10.1137/060649057. URL https://doi.org/10.1137/060649057.

Adam Tauman Kalai, Yishay Mansour, and Elad Verbin. On agnostic boosting and parity learning. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, pages 629–638. ACM, 2008b. doi: 10.1145/1374376.1374466. URL https://doi.org/10.1145/1374376.1374466.

Bobby Kleinberg, Renato Paes Leme, Jon Schneider, and Yifeng Teng. U-calibration: Forecasting for an unknown agent. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 5143–5145. PMLR, 12–15 Jul 2023.

Meelis Kull and Peter Flach. Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 68–85. Springer, 2015.

Meelis Kull, Miquel Perello-Nieto, Markus Kängsepp, Hao Song, Peter Flach, et al. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration. *arXiv preprint arXiv:1910.12656*, 2019.

Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2805–2814. PMLR, 2018.

Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. On the computational efficiency of training neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 855–863, 2014.

Rajeev Motwani and Prabhakar Raghavan. *Randomized algorithms*. Cambridge university press, 1995.

Georgy Noarov, Ramya Ramalingam, Aaron Roth, and Stephan Xie. High-dimensional prediction for sequential decision making. *arXiv preprint arXiv:2310.17651*, 2023.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12):371–421, 2008. URL http://jmlr.org/papers/v9/shafer08a.html.

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

Shai Shalev-Shwartz, Ohad Shamir, and Karthik Sridharan. Learning kernel-based halfspaces with the 0-1 loss. *SIAM J. Comput.*, 40(6):1623–1646, 2011.

Alexander A. Sherstov. Making polynomials robust to noise. *Theory of Computing*, 9(18):593–615, 2013. doi: 10.4086/toc.2013.v009a018.

Martin Wainwright. Lecture 6 in EECS 281B / STAT 241B: Advanced topics in statistical learning. https://people.eecs.berkeley.edu/~wainwrig/stat241b/lec6.pdf, 2009.

Shengjia Zhao, Michael P. Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. Calibrating predictions to decisions: A novel approach to multi-class calibration. In *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=iFF-zKCgzS.

Appendix A. Multi-Class Calibration

In this section, we discuss prior notions of multi-class calibration as well as their relationships, strengths, and drawbacks. We show that prior notions lack either expressivity or efficiency, and we introduce new notions to achieve a better balance between the two desiderata.

For a classification task with k categories, we use $\mathcal{E}_k = \{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ to denote the set of one-hot encodings of the categories. Here each $\mathbf{e}_i$ is the unit vector in $\mathbb{R}^k$ with the i-th coordinate being 1.

Throughout the paper, we use boldface letters to represent vectors in $\mathbb{R}^k$. For a vector $\mathbf{v} \in \mathbb{R}^k$, we use $\mathbf{v}^{(j)} \in \mathbb{R}$ to denote its j-th coordinate for every $j = 1, \ldots, k$. We use Δ_k to denote the set of all vectors $\mathbf{v} \in \mathbb{R}^k$ such that $\mathbf{v}^{(j)} \geq 0$ for every $j = 1, \ldots, k$ and $\mathbf{v}^{(1)} + \cdots + \mathbf{v}^{(k)} = 1$.

For a set $\mathcal{X}$ of individuals, a predictor is a function $p : \mathcal{X} \to \Delta_k$ that assigns every individual $x \in \mathcal{X}$ a prediction vector $\mathbf{v} = p(x) \in \Delta_k$, where each coordinate $\mathbf{v}^{(j)}$ is the predicted probability that the label of x falls in the j-th category.

Canonical Calibration. For a ground-truth distribution $\mathcal{D}_0$ of labeled examples $(x, \mathbf{y}) \in \mathcal{X} \times \mathcal{E}_k$, we say a predictor $p : \mathcal{X} \to \Delta_k$ satisfies (perfect) *canonical calibration* if

$$\mathop{\mathbf{E}}_{(x,\mathbf{y})\sim\mathcal{D}_0}[\mathbf{y}|p(x) = \mathbf{v}] = \mathbf{v} \quad \text{for every } \mathbf{v} \in \Delta_k.$$

A simple but important observation is that the above definition only depends on the distribution of $(p(x), \mathbf{y}) \in \Delta_k \times \mathcal{E}_k$. As a consequence, we obtain the following simplified but equivalent definition:

Definition 7 *We say a distribution $\mathcal{D}$ of $(\mathbf{v}, \mathbf{y}) \in \Delta_k \times \mathcal{E}_k$ satisfies (perfect)* canonical calibration *if*

$$\mathop{\mathbf{E}}_{(\mathbf{v},\mathbf{y})\sim\mathcal{D}}[\mathbf{y}|\mathbf{v}] = \mathbf{v}.$$

In the definition above, we work with a distribution $\mathcal{D}$ of $(\mathbf{v}, \mathbf{y}) \in \Delta_k$ without explicitly stating that it is the distribution of $(p(x), \mathbf{y})$ where $(x, \mathbf{y})$ comes from the ground-truth distribution $\mathcal{D}_0$, and p is a predictor. We will use this convention throughout the paper.

It is folklore that the sample complexity of determining whether a distribution $\mathcal{D}$ satisfies perfect canonical calibration grows exponentially in k. In Appendix B, we prove a stronger result (Theorem 15), showing that distinguishing whether a distribution $\mathcal{D}$ satisfies perfect canonical calibration or it is $\Omega(1)$-far from canonical calibration (in ℓ_1 distance) requires sample complexity exponential in k.

Weighted Calibration. Due to the sample inefficiency of canonical calibration, many previous works considered relaxations of canonical calibration such as confidence calibration and top-label calibration. These notions can be framed as special cases of a general notion called *weighted calibration* studied in Gopalan et al. (2022b):

Definition 8 (Weighted calibration (Gopalan et al., 2022b)) *Let $\mathcal{W} : \Delta_k \to [-1,1]^k$ be a family of weight functions. We define the $\mathcal{W}$-calibration error of a distribution $\mathcal{D}$ of $(\mathbf{v}, \mathbf{y}) \in \Delta_k \times \mathcal{E}_k$ as*

$$\mathsf{CE}_{\mathcal{W}}(\mathcal{D}) = \sup_{w \in \mathcal{W}} \left| \mathop{\mathbf{E}}_{(\mathbf{v},\mathbf{y})\sim\mathcal{D}} [\langle \mathbf{y} - \mathbf{v}, w(\mathbf{v}) \rangle] \right|.$$

We say that $\mathcal{D}$ is $(\mathcal{W}, \alpha)$-calibrated if $\mathsf{CE}_{\mathcal{W}}(\mathcal{D}) \leq \alpha$. [5]

If a distribution $\mathcal{D}$ satisfies perfect canonical calibration, then it is $(\mathcal{W}, 0)$-calibrated for any class $\mathcal{W}$. When the class $\mathcal{W}$ consists of all functions $w : \Delta_k \to [-1,1]^k$, $(\mathcal{W}, 0)$-calibration becomes equivalent to perfect canonical calibration.

Class-wise, Confidence, and Top-label Calibration. The notion of weighted calibration is very general. By choosing the class $\mathcal{W}$ appropriately, it recovers many concrete notions of calibration.

Class-wise calibration (Kull et al., 2019) is the following requirement:

$$\mathop{\mathbf{E}}_{(\mathbf{v},\mathbf{y})\sim\mathcal{D}} [\mathbf{y}^{(\ell)} | \mathbf{v}^{(\ell)}] = \mathbf{v}^{(\ell)} \quad \text{for every } \ell = 1, \ldots, k.$$

This is equivalent to $(\mathcal{W}, 0)$-calibration where $\mathcal{W}$ consists of all functions w mapping $\mathbf{v} \in \Delta_k$ to $w(\mathbf{v}) = \phi(\mathbf{v}^{(\ell)})\mathbf{e}_\ell$ for every $\phi : [0,1] \to [-1,1]$ and $\ell = 1, \ldots, k$.

Confidence calibration (Guo et al., 2017) is also a special case of weighted calibration. For any $\mathbf{v} \in \Delta_k$, let $\ell_{\mathbf{v}}$ denote the coordinate $\ell \in \{1, \ldots, k\}$ that maximizes $\mathbf{v}^{(\ell)}$. Confidence calibration is the following requirement:

$$\mathop{\mathbf{E}}_{(\mathbf{v},\mathbf{y})\sim\mathcal{D}} [\mathbf{y}^{(\ell_{\mathbf{v}})} | \mathbf{v}^{(\ell_{\mathbf{v}})}] = \mathbf{v}^{(\ell_{\mathbf{v}})}.$$

This is equivalent to $(\mathcal{W}, 0)$-calibration where $\mathcal{W}$ consists of all functions w mapping $\mathbf{v} \in \Delta_k$ to $w(\mathbf{v}) = \phi(\mathbf{v}^{(\ell_{\mathbf{v}})})\mathbf{e}_{\ell_{\mathbf{v}}}$ for every $\phi : [0,1] \to [-1,1]$.

Top-label calibration (Gupta and Ramdas, 2022) is defined to be the following requirement:

$$\mathop{\mathbf{E}}_{(\mathbf{v},\mathbf{y})\sim\mathcal{D}} [\mathbf{y}^{(\ell_{\mathbf{v}})} | \mathbf{v}^{(\ell_{\mathbf{v}})}, \ell_{\mathbf{v}}] = \mathbf{v}^{(\ell_{\mathbf{v}})}.$$

This is equivalent to $(\mathcal{W}, 0)$-calibration where $\mathcal{W}$ consists of all functions w mapping $\mathbf{v} \in \Delta_k$ to $w(\mathbf{v}) = \phi(\mathbf{v}^{(\ell_{\mathbf{v}})}, \ell_{\mathbf{v}})\mathbf{e}_{\ell_{\mathbf{v}}}$ for every $\phi : [0,1] \times \{1, \ldots, k\} \to [-1,1]$.

Decision Calibration. A drawback of class-wise, confidence, and top-label calibration is that they do not imply good calibration performance if the predictions are used for downstream tasks. To improve the expressivity while avoiding the exponential sample complexity of canonical calibration, (Zhao et al., 2021) introduced the notion of *decision calibration*, where they studied downstream loss-minimization tasks of deciding which action to choose among a fixed set of actions based on the predictions. We focus on the special case of two actions, where the definition of decision calibration is as follows:

5. The original definition in (Gopalan et al., 2022b) was presented in the more general context of multicalibration. Their definition allows for weight families whose range is $[0,1]^k$ rather than $[-1,1]^k$, but this is a technical issue.

Definition 9 (Decision Calibration (Zhao et al., 2021)) *For a distribution $\mathcal{D}$ of $(\mathbf{v}, \mathbf{y}) \in \Delta_k \times \mathcal{E}_k$, the decision calibration error of $\mathcal{D}$ is defined to be* [6]

$$\mathsf{decCE}(\mathcal{D}) := \sup_{\mathbf{a}\in\mathbb{R}^k, b\in\mathbb{R}} \|(\mathbf{y}-\mathbf{v})\mathbb{I}(\langle \mathbf{a}, \mathbf{v}\rangle > b)\|_2 + \|(\mathbf{y}-\mathbf{v})\mathbb{I}(\langle \mathbf{a}, \mathbf{v}\rangle \leq b)\|_2.$$

Equivalently, this is the weighted calibration error $\mathsf{CE}_{\mathcal{W}}(\mathcal{D})$, where $\mathcal{W}$ consists of functions mapping $\mathbf{v}$ to $\mathbb{I}(\langle \mathbf{a}, \mathbf{v}\rangle > b)\mathbf{g} + \mathbb{I}(\langle \mathbf{a}, \mathbf{v}\rangle > b)\mathbf{g}' \in \mathbb{R}^k$ for every $\mathbf{a} \in \mathbb{R}^k, b \in \mathbb{R}$, and $\mathbf{g}, \mathbf{g}' \in \mathbb{R}^k$ satisfying $\|\mathbf{g}\|_2 \leq 1$, $\|\mathbf{g}'\|_2 \leq 1$.

The work of (Zhao et al., 2021) showed that decision calibration ensures a desirable indistinguishability property for downstream loss minimization tasks, demonstrating the expressivity of the notion. However, we show that decision calibration is a computationally inefficient notion. Here, efficiency is evaluated on the *auditing* task (defined formally in Appendix C), where the goal is to recalibrate a given mis-calibrated predictor while reducing its squared loss. (Zhao et al., 2021) showed that auditing for decision calibration has *sample* complexity polynomial in k, improving over the exponential sample complexity of canonical calibration, but they fell short of proving a computational efficiency guarantee. Instead, they provided a heuristic algorithm for the auditing task without correctness or running time analyses. Our computational hardness results in Appendix D show that under standard complexity-theoretic assumptions, there is no $\mathrm{poly}(k)$-time algorithm for auditing decision calibration.

Smooth Calibration. We now introduce new notions of multi-class calibration inspired by a recent theory of Błasiok et al. (2023) on calibration measures in the binary setting.

Consider the downstream binary prediction task of predicting whether the true label belongs to a set $T \subset [k]$ of labels. Let $\mathbf{a} := \mathbf{1}_T \in \{0,1\}^k$ denote the indicator of the subset T. Given the distribution $\mathcal{D}$ of $(\mathbf{v}, \mathbf{y})$, the predicted probability of this event is $\langle \mathbf{1}_T, \mathbf{v}\rangle$ while the true label is given by $\langle \mathbf{1}_T, \mathbf{y}\rangle$. A natural approach to defining calibration notions for the multi-class setting is to seek good calibration guarantees for every such binary prediction tasks.

A well studied notion of calibration for binary classification is the notion of smooth calibration (Kakade and Foster, 2008; Foster and Hart, 2018). A key advantage of this notion is that it is robust to perturbations of the predictor, unlike notions such as ECE. More recently, it plays a central role in the the work of (Błasiok et al., 2023) and their theory of consistent calibration measures for binary classification. At a high level, these are calibration measures that are polynomially related to the (earthmover) distance to the closest perfectly calibrated predictor. Applying the notion of smooth calibration error to the downstream binary prediction tasks for subsets $T \subseteq [k]$, we get the following definition:

6. The original definition of decision calibration in (Zhao et al., 2021) takes a slightly different form:

$$\mathsf{decCE}(\mathcal{D}) := \sup_{\mathbf{r},\mathbf{r}'\in\mathbb{R}^k, b\in\mathbb{R}} \|(\mathbf{y}-\mathbf{v})\mathbb{I}(\langle \mathbf{r}, \mathbf{v}\rangle > \langle \mathbf{r}', \mathbf{v}\rangle)\|_2 + \|(\mathbf{y}-\mathbf{v})\mathbb{I}(\langle \mathbf{r}, \mathbf{v}\rangle \leq \langle \mathbf{r}', \mathbf{v}\rangle)\|_2.$$

Here $\langle \mathbf{r}, \mathbf{v}\rangle$ (resp. $\langle \mathbf{r}', \mathbf{v}\rangle$) is the expected *loss* of taking action 1 (resp. action 2) over the randomness in an outcome $\widehat{\mathbf{y}} \in \mathcal{E}_k$ distributed with mean $\mathbf{v}$. In fact Theorem 9 is equivalent to this definition. For any $\mathbf{v} \in \Delta_k$, the sum of the coordinates of $\mathbf{v}$ is 1, i.e., $\langle \mathbf{1}, \mathbf{v}\rangle = 1$, where $\mathbf{1}$ is the all-ones vector. Therefore, in Theorem 9, we have $\mathbb{I}(\langle \mathbf{a}, \mathbf{v}\rangle > b) = \mathbb{I}(\langle \mathbf{a} - b\mathbf{1}, \mathbf{v}\rangle > 0)$, and thus restricting $b = 0$ does not change Theorem 9. Under this restriction, the equivalence between the two definitions follows by taking $\mathbf{a} = \mathbf{r} - \mathbf{r}'$.

Definition 10 (Subset Smooth Calibration) *Let* Lip *be the class of* 1*-Lipschitz functions* $\phi : \mathbb{R} \to [-1, 1]$*. For a distribution* $\mathcal{D}$ *of* $(\mathbf{v}, \mathbf{y}) \in \Delta_k \times \mathcal{E}_k$*, we define the smooth calibration error of* $\mathcal{D}$ *on the subset* T *to be*

$$\begin{aligned}\mathrm{smCE}_T(\mathcal{D}) &= \sup_{\phi \in \mathrm{Lip}} \mathop{\mathbf{E}}_{\mathcal{D}}[(\langle \mathbf{1}_T, \mathbf{y}\rangle - \langle \mathbf{1}_T, \mathbf{v}\rangle)\phi(\langle \mathbf{1}_T, \mathbf{v}\rangle)] \\ &= \sup_{\phi \in \mathrm{Lip}} \mathop{\mathbf{E}}_{\mathcal{D}}[\langle \mathbf{1}_T, \mathbf{y} - \mathbf{v}\rangle\phi(\langle \mathbf{1}_T, \mathbf{v}\rangle)].\end{aligned}$$

We define the subset smooth calibration error of $\mathcal{D}$ *as*

$$\mathrm{ssCE}(\mathcal{D}) = \sup_{T \subseteq [k]} \mathrm{smCE}_T(\mathcal{D}).$$

More generally, for $m \geq 0$*, we define the* m*-subset smooth calibration error of* $\mathcal{D}$ *as*

$$\mathrm{ssCE}_m(\mathcal{D}) = \sup_{T \subseteq [k], |T| \leq m} \mathrm{smCE}_T(\mathcal{D}).$$

Note that we can define subset smooth calibration as a special case of weighted calibration. We define $\mathcal{W}_{m\text{-}\mathsf{ss}}$ to be the set of all functions $w : \Delta_k \to [-1, 1]^k$ such that there exist $T \subseteq [k]$ and $\phi \in \mathrm{Lip}$ satisfying $|T| \leq m$ and $w(\mathbf{v}) = \mathbf{1}_T\phi(\langle \mathbf{1}_T, \mathbf{v}\rangle)$ for every $\mathbf{v} \in \Delta_k$. Then

$$\mathrm{ssCE}_m(\mathcal{D}) = \mathsf{CE}_{\mathcal{W}_{m\text{-}\mathsf{ss}}}(\mathcal{D}).$$

In the binary setting, a result of (Błasiok et al., 2023) shows that the smooth calibration error is polynomially related to the (earthmover) distance to the nearest perfectly calibrated predictor. Therefore, a small subset smooth calibration error in our multi-class setting implies that for every subset $T \subseteq [k]$, the prediction $\langle \mathbf{1}_T, \mathbf{v}\rangle$ is close to perfect calibration for the corresponding downstream binary prediction task.

Having demonstrated the expressivity of subset smooth calibration, we move on to establish its efficiency. The main algorithmic result of our paper is that auditing for subset smooth calibration can be achieved in time polynomial in k (for any fixed error parameter α, see Appendix C for formal definition of auditing). That is, subset smooth calibration simultaneously achieves strong expressivity and computational efficiency. In fact, the efficiency of our auditing algorithm extends to a more expressive notion which we call *projected smooth calibration*, where we generalize indicators of sets that are vectors in $\{0, 1\}^k$ to allow vectors in $[-1, 1]^k$.

Definition 11 (Projected Smooth Calibration) *For* $m \geq 0$*, let* $\mathcal{H}_{m\text{-pLip}}$ *denote the set of all functions* $h : \Delta_k \to [-1, 1]$ *such that there exist* $\phi \in \mathrm{Lip}$ *and* $\mathbf{a} \in [-1, 1]^k$ *with* $\|\mathbf{a}\|_2^2 \leq m$ *satisfying*

$$h(\mathbf{v}) = \phi(\langle \mathbf{a}, \mathbf{v}\rangle) \quad \textit{for every } \mathbf{v} \in \Delta_k.$$

Define the m*-projected smooth calibration error as*

$$\mathrm{psCE}_m(\mathcal{D}) = \mathsf{CE}_{\mathcal{H}^k_{m\text{-pLip}}}(\mathcal{D}).$$

In measuring psCE, we audit each coordinate $i \in [k]$ using a distinct function $h^{(i)} \in \mathcal{H}_{m\text{-pLip}}$. We also consider a further strengthening of projected smooth calibration by allowing arbitrary ℓ_1-Lipshcitz functions in each coordinate:

Definition 12 (Full Smooth Calibration) *Let $\mathcal{H}_{\text{fLip}}$ denote the set of all functions $h : \Delta_k \to [-1, 1]$ such that*

$$|h(\mathbf{v}) - h(\mathbf{v}')| \leq \left\|\mathbf{v} - \mathbf{v}'\right\|_1 \quad \text{for every } \mathbf{v}, \mathbf{v}' \in \Delta_k.$$

Define the full smooth calibration error of a distribution $\mathcal{D}$ of $(\mathbf{v}, \mathbf{y}) \in \Delta_k \times \mathcal{E}_k$ as

$$\text{fsCE}(\mathcal{D}) = \mathsf{CE}_{\mathcal{H}^k_{\text{fLip}}}(\mathcal{D}).$$

Lemma 13 *For any $m \geq 0$, for any distribution $\mathcal{D}$ of $(\mathbf{v}, \mathbf{y}) \in \Delta_k \times \mathcal{E}_k$,*

$$\text{ssCE}_m(\mathcal{D}) \leq \text{psCE}_m(\mathcal{D}) \leq \text{fsCE}(\mathcal{D}).$$

Proof To prove the first inequality, let $T \subset [k]$ be the set of size bounded by m that maximizes $\text{smCE}_T(\mathcal{D})$, and $\phi \in \text{Lip}$ the Lipschitz function that witnesses it, so that

$$\text{ssCE}_m(\mathcal{D}) = \mathop{\mathbf{E}}_{\mathcal{D}}[\langle \mathbf{1}_T, \mathbf{y} - \mathbf{v}\rangle \phi(\langle \mathbf{1}_T, \mathbf{v}\rangle)] = \mathop{\mathbf{E}}_{\mathcal{D}}[\langle \phi(\langle \mathbf{1}_T, \mathbf{v}\rangle)\mathbf{1}_T, \mathbf{y} - \mathbf{v}\rangle].$$

We define the auditor function $w \in \mathcal{H}^k_{m\text{-pLip}}$ where

$$w^{(i)}(\mathbf{v}) = \mathbf{1}_T \phi(\langle \mathbf{1}_T, \mathbf{v}\rangle) = \begin{cases} \phi(\langle \mathbf{1}_T, \mathbf{v}\rangle) \text{ for } i \in T \\ 0 \text{ otherwise} \end{cases}$$

Hence

$$\begin{aligned} \text{psCE}_m(\mathcal{D}) &= \max_{w' \in \mathcal{H}^k_{m\text{-pLip}}} \mathop{\mathbf{E}}_{\mathcal{D}}[\langle \mathbf{y} - \mathbf{v}, w'(\mathbf{v})\rangle] \\ &\geq \mathop{\mathbf{E}}_{\mathcal{D}}[\langle \mathbf{y} - \mathbf{v}, w(\mathbf{v})\rangle] \\ &= \langle \mathbf{y} - \mathbf{v}, \mathbf{1}_T\rangle \phi(\langle \mathbf{1}_T, v\rangle) \\ &= \text{ssCE}_m(\mathcal{D}). \end{aligned}$$

The second inequality is implied by the inclusion $\mathcal{H}_{m\text{-pLip}} \subseteq \mathcal{H}_{\text{fLip}}$. To prove this inclusion, note that for any function $h \in \mathcal{H}_{m\text{-pLip}}$, there exists $\phi \in \text{Lip}, \mathbf{a} \in [-1, 1]^k$ such that $h(\mathbf{v}) = \phi(\langle \mathbf{a}, \mathbf{v}\rangle)$ for every $\mathbf{v} \in \Delta_k$. We have

$$\begin{aligned} |h(\mathbf{v}) - h(\mathbf{v}')| &= |\phi(\langle \mathbf{a}, \mathbf{v}\rangle) - \phi(\langle \mathbf{a}, \mathbf{v}'\rangle)| \\ &\leq |\langle \mathbf{a}, \mathbf{v}\rangle - \langle \mathbf{a}, \mathbf{v}'\rangle| \\ &= |\langle \mathbf{a}, \mathbf{v} - \mathbf{v}'\rangle| \\ &\leq \|\mathbf{a}\|_\infty \left\|\mathbf{v} - \mathbf{v}'\right\|_1 \\ &\leq \left\|\mathbf{v} - \mathbf{v}'\right\|_1 \end{aligned}$$

where the first inequality uses the Lipschitz property of ϕ. This shows $h \in \mathcal{H}_{\text{fLip}}$, which completes the proof. ■

In Appendix F we show that both subset smooth calibration and projected smooth calibration allow efficient auditing, whereas in Theorem 17 we show that full smooth calibration requires sample complexity exponential in k.

Appendix B. Sample Complexity of Canonical Calibration

The main goal of this section is to prove that distinguishing whether a distribution $\mathcal{D}$ of $(\mathbf{v}, \mathbf{y})$ satisfies perfect canonical calibration or $\mathcal{D}$ is far from canonical calibration requires sample complexity exponential in k.

We use the following definition of distance to canonical calibration, generalizing the *lower distance to calibration* in (Błasiok et al., 2023) from the binary setting to the multi-class setting.

Definition 14 (Distance to Canonical Calibration) *Consider a distribution $\mathcal{D}$ of $(\mathbf{v}, \mathbf{y}) \in \Delta_k \times \mathcal{E}_k$. We define $\mathsf{ext}(\mathcal{D})$ to be the set of distributions Π of $(\mathbf{u}, \mathbf{v}, \mathbf{y})$ where the marginal distribution of $(\mathbf{v}, \mathbf{y})$ is $\mathcal{D}$, and the marginal distribution of $(\mathbf{u}, \mathbf{y})$ satisfies perfect canonical calibration. We define the distance to calibration, denoted by $\mathsf{dCE}(\mathcal{D})$, as follows:*

$$\mathsf{dCE}(\mathcal{D}) := \inf_{\Pi \in \mathsf{ext}(\mathcal{D})} \mathop{\mathbf{E}}_{\Pi} \|\mathbf{u} - \mathbf{v}\|_1.$$

Here is our sample complexity lower bound:

Theorem 15 *Let A be an algorithm that takes examples $(\mathbf{v}_1, \mathbf{y}_1), \ldots, (\mathbf{v}_n, \mathbf{y}_n) \in \Delta_k \times \mathcal{E}_k$ drawn i.i.d. from a distribution $\mathcal{D}$ as input, and outputs "accept"or "reject". Assume that for any distribution $\mathcal{D}$ satisfying perfect canonical calibration, algorithm A outputs "accept" with probability at least $2/3$. Also, for some $\alpha > 0$, assume that for any distribution $\mathcal{D}$ satisfying $\mathsf{dCE}(\mathcal{D}) \geq \alpha$, algorithm A outputs "reject" with probability at least $2/3$. Then for some absolute constants $k_0 > 0$ and $c > 0$, assuming $k \geq k_0$, we have $n \geq (c/\alpha)^{(k-1)/2}$.*

To prove Theorem 15, we use the following lemma to connect the distance to canonical calibration $\mathsf{dCE}(\mathcal{D})$ with the full smooth calibration error $\mathrm{fsCE}(\mathcal{D})$.

Lemma 16 *For any distribution $\mathcal{D}$ over $\Delta_k \times \mathcal{E}_k$, $\mathrm{fsCE}(\mathcal{D}) \leq 4\mathsf{dCE}(\mathcal{D})$.*

Proof Consider any function $w \in \mathcal{H}^k_{\mathrm{fLip}}$ and any distribution $\Pi \in \mathsf{ext}(\mathcal{D})$. By the definition of $\mathcal{H}^k_{\mathrm{fLip}}$, for any $\mathbf{u}, \mathbf{v} \in \Delta_k$, we have

$$\|w(\mathbf{u}) - w(\mathbf{v})\|_\infty \leq \|\mathbf{u} - \mathbf{v}\|_1. \tag{1}$$

By the definition of $\mathsf{ext}(\mathcal{D})$, for $(\mathbf{u}, \mathbf{v}, \mathbf{y}) \sim \Pi$, the distribution of $(\mathbf{u}, \mathbf{y})$ satisfies perfect canonical calibration, and thus

$$\mathop{\mathbf{E}}_{\Pi}[\langle \mathbf{y} - \mathbf{u}, w(\mathbf{u}) \rangle] = 0. \tag{2}$$

Therefore,

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathcal{D}}[\langle \mathbf{v} - \mathbf{y}, w(\mathbf{v}) \rangle] &\leq |\mathop{\mathbf{E}}_{\Pi}[\langle \mathbf{v} - \mathbf{y}, w(\mathbf{v}) - w(\mathbf{u}) \rangle]| + \mathop{\mathbf{E}}_{\Pi}[\langle \mathbf{v} - \mathbf{y}, w(\mathbf{u}) \rangle] \\
&\leq 2 \mathop{\mathbf{E}}_{\Pi} \|\mathbf{u} - \mathbf{v}\|_1 + \mathop{\mathbf{E}}_{\Pi}[\langle \mathbf{v} - \mathbf{y}, w(\mathbf{u}) \rangle] && \text{(by (1))} \\
&= 2 \mathop{\mathbf{E}}_{\Pi} \|\mathbf{u} - \mathbf{v}\|_1 + \mathop{\mathbf{E}}_{\Pi}[\langle \mathbf{v} - \mathbf{u}, w(\mathbf{u}) \rangle] && \text{(by (2))} \\
&\leq 4 \mathop{\mathbf{E}}_{\Pi} \|\mathbf{u} - \mathbf{v}\|_1,
\end{aligned}$$

where the last inequality holds because $\|w(\mathbf{u})\|_\infty \leq 1$. The proof is completed by taking supremum over $w \in \mathcal{H}^k_{\mathrm{fLip}}$ and infimum over $\Pi \in \mathsf{ext}(\mathcal{D})$. ■

By Theorem 16, Theorem 15 is a direct corollary of the following theorem which gives a sample complexity lower bound for distinguishing perfect canonical calibration from having a large full smooth calibration error:

Theorem 17 *Let A be an algorithm that takes examples $(\mathbf{v}_1, \mathbf{y}_1), \ldots, (\mathbf{v}_n, \mathbf{y}_n) \in \Delta_k \times \mathcal{E}_k$ drawn i.i.d. from a distribution $\mathcal{D}$ as input, and outputs "accept"or "reject". Assume that for any distribution $\mathcal{D}$ satisfying perfect canonical calibration, algorithm A outputs "accept" with probability at least $2/3$. Also, for some $\alpha > 0$, assume that for any distribution $\mathcal{D}$ satisfying $\mathrm{fsCE}(\mathcal{D}) \geq \alpha$, algorithm A outputs "reject" with probability at least $2/3$. Then for some absolute constants $k_0 > 0$ and $c > 0$, for all $k \geq k_0$ we have $n \geq (c/\alpha)^{(k-1)/2}$.*

Our proof of Theorem 17 starts with the following lemma which can be proved by a standard greedy algorithm:

Lemma 18 *There exist absolute constants $c > 0$ and $k_0 > 0$ with the following property. For any positive integer $k > k_0$ and any $\varepsilon > 0$, there exists a set $V \subseteq \Delta_k$ with the following properties:*

1. $|V| \geq (c/\varepsilon)^{k-1}$;
2. $\|\mathbf{v}_1 - \mathbf{v}_2\|_1 \geq \varepsilon$ *for any distinct* $\mathbf{v}_1, \mathbf{v}_2 \in V$;
3. $\|\mathbf{v} - \mathbf{e}_i\|_1 \geq 1/3$ *for any* $\mathbf{v} \in V$ *and* $i \in \{1, \ldots, k\}$.

Proof The lemma can be proved by a simple greedy algorithm. Let us start with $V = \emptyset$ and repeat the following step: if there exists $\mathbf{v}' \in \Delta_k$ such that $\|\mathbf{v}' - \mathbf{v}\|_1 \geq \varepsilon$ for every $u \in U$ and $\|\mathbf{v}' - \mathbf{e}_i\|_1 \geq 1/3$ for every $i = 1, \ldots, k$, we add $\mathbf{v}'$ to V. We repeat the step until no such $\mathbf{v}'$ exists to obtain the final V. Clearly, V satisfies properties 2 and 3 required by the lemma. It remains to prove that V also satisfies property 1.

Consider the final V in the process of the algorithm. For any $\mathbf{v} \in V$, consider a set $S_{\mathbf{v}}$ consisting of all points $\mathbf{s} \in \mathbb{R}^{k-1}$ such that $\|\mathbf{s} - \mathbf{v}|_{1,\ldots,k-1}\|_1 \leq \varepsilon$. Similarly, for every $i = 1, \ldots, k$, consider a set S_i consisting of all points $\mathbf{s} \in \mathbb{R}^{k-1}$ such that $\|\mathbf{s} - \mathbf{e}_i|_{1,\ldots,k-1}\|_1 \leq 1/3$. Also, consider the set S consisting of all points $\mathbf{s} \in \mathbb{R}^{k-1}_{\geq 0}$ such that $\|\mathbf{s}\|_1 \leq 1$. If $S \setminus ((\bigcup_{\mathbf{v} \in V} S_{\mathbf{v}}) \cup (\bigcup_{i=1}^{k} S_i))$ is non-empty, then we can take any $\mathbf{s}$ in that set and construct a vector $\mathbf{v}' = (\mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(k-1)}, 1 - \mathbf{s}^{(1)} - \cdots - \mathbf{s}^{(k-1)}) \in \Delta_k$. Since $\mathbf{s} \notin S_v$, it is easy to see that $\|\mathbf{v}' - \mathbf{v}\|_1 > \varepsilon$ for every $\mathbf{v} \in V$. Similarly, $\|\mathbf{v}' - \mathbf{e}_i\| > 1/3$ for every $i = 1, \ldots, k$. Therefore, the iterative steps of the algorithm can be continued. For the final V, it must hold that $S \setminus ((\bigcup_{\mathbf{v} \in V} S_{\mathbf{v}}) \cup (\bigcup_{i=1}^{k} S_i))$ is empty. The volume of each $S_{\mathbf{v}}$ is $(2\varepsilon)^{k-1}$ times the volume of S, and the volume of each S_i is $(2/3)^{k-1}$ times the volume of S. Therefore,

$$(2\varepsilon)^{k-1}|V| + (2/3)^{k-1}k \geq 1.$$

When k is sufficiently large, we have $(2/3)^{k-1}k \leq 1/2$, in which case $|V| \geq (1/2)(1/(2\varepsilon))^{k-1} \geq (c/\varepsilon)^{k-1}$, where the last inequality holds whenever k is sufficiently large and $c > 0$ is sufficiently small. ■

In the lemma below, we use the set V from Theorem 18 to construct candidate distributions with large full smooth calibration error. Later in Theorem 20 we combine these distributions to achieve indistinguishability from a distribution with no calibration error, unless given at least $\exp(k)$ examples.

Lemma 19 *For a sufficiently large positive integer k and $\varepsilon \in (0, 1/2)$, let $V \subseteq \Delta_k$ be the set guaranteed by Theorem 18. For a function $w : V \to \mathcal{E}_k$, define distribution $\mathcal{D}_w$ of $(\mathbf{v}, \mathbf{y}) \in V \times \mathcal{E}_k$ such that $\mathbf{v}$ is distributed uniformly over V and $\mathbf{y} = w(\mathbf{v})$. Then $\mathrm{fsCE}(\mathcal{D}_w) \geq \varepsilon/12$.*

Proof For any $\mathbf{v} \in V$, by property 3 in Theorem 18 and the fact that $w(\mathbf{v}) \in \{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$, we have $\|\mathbf{v} - w(\mathbf{v})\|_1 \geq 1/3$. Since $\mathbf{v} \in \Delta_k$ and $w(\mathbf{v}) \in \mathcal{E}_k$, we can separately consider the unique non-zero coordinate of $w(\mathbf{v})$ and the other zero coordinates to get

$$1/3 \leq \|\mathbf{v} - w(\mathbf{v})\|_1 = (1 - \langle \mathbf{v}, w(\mathbf{v})\rangle) + \langle \mathbf{v}, \mathbf{1} - w(\mathbf{v})\rangle = 2(1 - \langle \mathbf{v}, w(\mathbf{v})\rangle),$$

where $\mathbf{1} \in \mathbb{R}^k$ is the all-ones vector. Therefore, $\langle w(\mathbf{v}) - \mathbf{v}, w(\mathbf{v})\rangle = (1 - \langle \mathbf{v}, w(\mathbf{v})\rangle) \geq 1/6$, and thus

$$\mathop{\mathbf{E}}_{(\mathbf{v},\mathbf{y})\sim\mathcal{D}_w}[\langle \mathbf{y} - \mathbf{v}, w(\mathbf{v})\rangle] \geq 1/6.$$

To complete the proof, it remains to show that w is $(2/\varepsilon)$-Lipschitz over V (we can then extend w to a $(2/\varepsilon)$-Lipschitz function over Δ_k by standard construction). For any distinct $\mathbf{v}, \mathbf{v}' \in V$, we have

$$\|w(\mathbf{v}) - w(\mathbf{v}')\|_\infty \leq \|w(\mathbf{v}) - w(\mathbf{v}')\|_1 \leq 2 \leq (2/\varepsilon)\|\mathbf{v} - \mathbf{v}'\|_1,$$

where the last inequality uses property 2 in Theorem 18. ■

Lemma 20 *Let A be any algorithm that takes $(\mathbf{v}_1, \mathbf{y}_1), \ldots, (\mathbf{v}_n, \mathbf{y}_n) \in V \times \mathcal{E}_k$ as input, and outputs "accept" or "reject". Let p_1 be the acceptance probability when we first draw v_i independently and uniformly from V, and then draw each $\mathbf{y}_i$ independently with $\mathbf{E}[\mathbf{y}_i] = \mathbf{v}_i$. Let p_2 be the acceptance probability where we first draw $w : V \to \mathcal{E}_k$ such that for every $\mathbf{v} \in V$, $w(\mathbf{v})$ is distributed independently with mean $\mathbf{v}$, and then draw each $(\mathbf{v}_i, \mathbf{y}_i)$ independently from $\mathcal{D}_w$. Then,*

$$|p_1 - p_2| \leq O(n^2/|V|).$$

Proof Assume without loss of generality that $n < |V|$. Let p_3 denote the acceptance probability when we first draw $\mathbf{v}_1, \ldots, \mathbf{v}_n$ uniformly from V *without replacement*, and then draw each $\mathbf{y}_i \in \mathcal{E}_k$ independently with mean $\mathbf{v}_i$. We relate p_1 and p_2 to p_3 as follows.

Suppose we first draw each $\mathbf{v}_i$ independently and uniformly from V, and then draw each $\mathbf{y}_i$ independently with $\mathbf{E}[\mathbf{y}_i] = \mathbf{v}_i$. The probability that $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are distinct is

$$p_4 := (1 - 1/|V|) \cdots (1 - (n-1)/|V|) \geq 1 - 1/|V| - \cdots - (n-1)/|V| \geq 1 - O(n^2/|V|).$$

Conditioned on that event, the acceptance probability is exactly p_3. Conditioned on the complement of that event, the acceptance probability is bounded in $[0, 1]$. Therefore,

$$p_3 p_4 \leq p_1 \leq p_3 p_4 + (1 - p_4).$$

Similarly, we can show that

$$p_3 p_4 \leq p_2 \leq p_3 p_4 + (1 - p_4).$$

Combining these inequalities, we get $|p_1 - p_2| \leq 1 - p_4 \leq O(n^2/|V|)$. ■

Proof [Proof of Theorem 17] Consider the set V from Theorem 18 where we choose ε to be 12α. If $\mathcal{D}$ is the distribution of $(\mathbf{v}, \mathbf{y}) \in V \times \mathcal{E}_k$ where $\mathbf{v}$ is chosen uniformly at random from V and

$\mathbf{E}[\mathbf{y}|\mathbf{v}] = \mathbf{v}$, then algorithm A outputs "accept" with probability at least $2/3$. If $\mathcal{D}$ is $\mathcal{D}_w$ for some $w : V \to \mathcal{E}_k$, then by Theorem 19, algorithm A outputs "accept" with probability at most $1/3$.

By Theorem 20, we have $n \geq \Omega(\sqrt{|V|})$. By Property 1 in Theorem 18, we get $n \geq \Omega(\sqrt{|V|}) \geq (c/\alpha)^{(k-1)/2}$ for a sufficiently small absolute constant $c > 0$ assuming k is sufficiently large. ■

Appendix C. Auditing for Weighted Calibration and Agnostic Learning

In this section, we study the sample and computational complexity of weighted calibration (Theorem 8), where the complexity is measured in an *auditing* task we define below. Specifically, for any weight family $\mathcal{W}$, we show an equivalence between the auditing task and the well-studied agnostic learning task in the learning theory literature. This equivalence allows us to establish both computational lower bounds and efficient algorithms for specific weight families $\mathcal{W}$ in Appendices D to G.

Auditing for weighted calibration. The notion of weighted calibration gives rise to a natural decision problem, which we call the decision version of auditing calibration: given a predictor p, can we decide whether or not it is $(\mathcal{W}, \alpha)$ calibrated? In the event that p is not calibrated, we would ideally like to post-process its predictions to get a new predictor $\kappa(p)$ for $\kappa : \Delta_k \to \Delta_k$, so that $\kappa(p)$ is $(\mathcal{W}, \alpha)$-calibrated. This post-processing goal needs to be formulated carefully, since one can always get perfect calibration using a trivial predictor that constantly predicts $\mathbf{E}[\mathbf{y}]$. A natural formulation that avoids such trivial solutions is to require that the post-processing does not harm some measure of accuracy such as the expected squared loss of p.

One can achieve both these goals by solving a search problem which we call *auditing with a witness* defined below.

Definition 21 (Auditing with a witness) *An (α, β) auditor for $\mathcal{W}$ is an algorithm that when given access to a distribution $\mathcal{D}$ where $\mathsf{CE}_{\mathcal{W}}(\mathcal{D}) > \alpha$ returns a function $w' : \Delta_k \to [-1, 1]^k$ (which need not belong to $\mathcal{W}$) such that*

$$\mathop{\mathbf{E}}_{(\mathbf{v},\mathbf{y})\sim\mathcal{D}}[\langle \mathbf{y} - \mathbf{v}, w'(\mathbf{v})\rangle] \geq \beta. \tag{3}$$

Concretely, the auditor takes i.i.d. examples $(\mathbf{v}_1, \mathbf{y}_1), \ldots, (\mathbf{v}_n, \mathbf{y}_n)$ drawn from $\mathcal{D}$, and the output function w' should satisfy the inequality above with probability at least $1 - \delta$ over randomness in the examples and the auditor itself, where $\delta \in (0, 1/3)$ is the failure probability parameter.

As demonstrated in previous work (for instance (Hébert-Johnson et al., 2018; Gopalan et al., 2022b)), a solution to this search problem allows us to solve both the decision problem of auditing for calibration, and in the case when p is not $(\mathcal{W}, \alpha)$-calibrated, we can use the witness to post-process p and produce a predictor $\kappa(p)$ with lower squared loss, that is $(\mathcal{W}, \alpha)$-calibrated.

Lemma 22 *Given a predictor $p : \mathcal{X} \to \Delta_k$ and access to an (α, β)-auditor for $\mathcal{W}$, there is an algorithm that computes a post-processing function $\kappa : \Delta_k \to \Delta_k$ so that $\kappa(p)$ is $(\mathcal{W}, \alpha)$-calibrated and its squared loss is not larger than that of p. The algorithm uses at most $O(k/\beta^2)$ calls to the (α, β)-auditor.*

Proof We start off with $p_0 = p$. If p is not $(\mathcal{W}, \alpha)$-calibrated, then the auditor produces w' satisfying Equation (3). Following the proof of (Gopalan et al., 2022b, Lemma 33), we can now update p to $\kappa(p)$ using w' so that we get a decrease in the expected squared loss:

$$\mathbf{E}[\|\kappa(\mathbf{v}) - \mathbf{y}\|_2^2] \leq \mathbf{E}[\|\mathbf{v} - \mathbf{y}\|_2^2] - \Omega(\beta^2/k).$$

Note that the squared loss is bounded in the interval $[0, 4]$ because $\|\mathbf{v} - \mathbf{y}\|_2 \leq \|\mathbf{v}\|_2 + \|\mathbf{y}\|_2 \leq \|\mathbf{v}\|_1 + \|\mathbf{y}\|_1 \leq 2$. Thus by repeatedly using the auditor and applying the update at most $O(k/\beta^2)$ times, we can eventually achieve $(\mathcal{W}, \alpha)$ calibration with decreased expected squared loss. ■

We make some observations about the role that the different parameters α, β and $\mathcal{W}$ play in the complexity of auditing with a witness.

- Auditing becomes easier for smaller β. The β parameter affects the running time, but not the final calibration guarantee. Thus an $(\alpha, \beta/10)$ auditor will result in the same guarantee as an (α, β) auditor, but at the cost of more iterations. Since we are interested in the question of whether auditing can be done in time $\mathrm{poly}(k)$ versus $\exp(k)$, we do not optimize too much for β, and are fine with losing polynomial factors in it.

- In contrast, auditing gets harder for smaller α, since the auditor is required to detect smaller violations of calibration. The final guarantee is also much more sensitive to α: a $(2\alpha, \beta)$ auditor can only be used to guarantee $(\mathcal{W}, 2\alpha)$ calibration, but not $(\mathcal{W}, \alpha)$ calibration.

- The complexity of auditing increases as the the weight function family becomes larger. If $\mathcal{W}_1 \subseteq \mathcal{W}_2$, then an (α, β)-auditor for $\mathcal{W}_2$ is also an (α, β)-auditor for $\mathcal{W}_1$, since $\mathsf{CE}_{\mathcal{W}_2}(\mathcal{D}) \geq \mathsf{CE}_{\mathcal{W}_1}(\mathcal{D})$ so the auditor is guaranteed to produce a witness whenever p is not $(\mathcal{W}_1, \alpha)$-calibrated. It might happen that $\mathsf{CE}_{\mathcal{W}_1}(\mathcal{D}) \leq \alpha$ whereas $\mathsf{CE}_{\mathcal{W}_2}(\mathcal{D}) > \alpha$. In such a scenario, an auditor for $\mathcal{W}_2$ will still find a witness to miscalibration. This is not required by our definition of auditor for $\mathcal{W}_1$, but it is allowed.

Agnostic learning. We understand the complexity of auditing for multiclass calibration by connecting it to the well-studied problem of agnostic learning in the standard binary classification setting. For a distribution $\mathcal{U}$ of $(\mathbf{v}, z) \in \Delta_k \times [-1, 1]$ and a class $\mathcal{H}$ of functions $\Delta_k \to [-1, 1]$, we define

$$\mathrm{Opt}(\mathcal{H}, \mathcal{U}) := \sup_{h \in \mathcal{H}} |\mathbf{E}[h(\mathbf{v})z]|.$$

Definition 23 (Weak agnostic learner) *(Ben-David et al., 2001; Kalai et al., 2008b) Let $\alpha \geq \beta \in [0, 1]$. An (α, β) agnostic learner for $\mathcal{H}$ is an algorithm that when given sample access to a distribution $\mathcal{U}$ over $\Delta_k \times [-1, 1]$ such that $\mathrm{Opt}(\mathcal{H}, \mathcal{U}) \geq \alpha$ returns $h' : \Delta_k \to [-1, 1]$ such that*

$$\mathbf{E}_{(\mathbf{v}, z) \sim \mathcal{U}}[h'(\mathbf{v})z] \geq \beta.$$

More concretely, the learner takes i.i.d. examples $(\mathbf{v}_1, z_1), \ldots, (\mathbf{v}_n, z_n)$ drawn from $\mathcal{U}$, and the output function h' should satisfy the inequality above with probability at least $1 - \delta$ over randomness in the examples and the learner itself, where $\delta \in (0, 1/3)$ is the failure probability parameter.

Similarly to auditing, the strength of an agnostic learner is more sensitive to the α parameter than the β parameter. Known results on agnostic boosting (Kalai et al., 2008b; Feldman, 2010; Kalai and Kanade, 2009) show that the existence of an (α, β)-weak agnostic learner implies the existence of an strong agnostic learner with polynomially increased time and sample complexity depending on $1/\beta$ (see the citations for a precise statement).

In the rest of the section we present our main result connecting the agnostic learning task for a class $\mathcal{H}$ and the auditing task for $\mathcal{H}^k$.

C.1. Auditing from Agnostic Learning

Theorem 24 *Given an $(\alpha/3, \beta)$ weak agnostic learner for $\mathcal{H}$ with sample complexity n_0, running time T_0 and failure probability parameter $\delta/2$, we can construct an $(\alpha, \alpha\beta/6k)$ auditor for $\mathcal{H}^k$ with sample complexity $n = O(kn_0/\alpha + k^2\alpha^{-2}\beta^{-2}\log(k/\delta))$, time complexity $O(kT_0 + kn)$, and failure probability parameter δ.*

A natural idea for proving the theorem above is to apply the agnostic learner on each coordinate of the residual $\mathbf{z} := \mathbf{y} - \mathbf{v}$ in the auditing task. Specifically, in the auditing task, we assume

$$\mathbf{E}[\langle \mathbf{z}, w(\mathbf{v})\rangle] = \mathbf{E}[\langle \mathbf{y} - \mathbf{v}, w(\mathbf{v})\rangle] > \alpha$$

for some $w \in \mathcal{H}^k$. Expressing $w(\mathbf{v})$ as $(w^{(1)}(\mathbf{v}), \ldots, w^{(k)}(\mathbf{v}))$ where each $w^{(j)} \in \mathcal{H}$, we have

$$\sum_{j=1}^{k} \mathbf{E}[\mathbf{z}^{(j)} w^{(j)}(\mathbf{v})] > \alpha,$$

which implies that there exists $j \in \{1, \ldots, k\}$ such that

$$\mathbf{E}[\mathbf{z}^{(j)} w^{(j)}(\mathbf{v})] > \alpha/k. \tag{4}$$

If we only use (4), we would need an $(\alpha/k, \beta)$ agnostic learner to prove Theorem 24, but we only have an $(\alpha/3, \beta)$ agnostic learner.

To avoid the loss of a factor of k, we define $\mathbf{z}$ in a better way that leverages the fact that $\mathbf{y}, \mathbf{v} \in \Delta_k$. Specifically, we note that the vector $\frac{1}{2}(\mathbf{y} - \mathbf{v})$ has ℓ_1 norm at most 1, and thus it is the mean of a distribution over $\mathcal{E}_k \cup (-\mathcal{E}_k)$. Given $\mathbf{y}$ and $\mathbf{v}$, we draw $\mathbf{z}$ randomly from that distribution. We have

$$\mathbf{E}[\langle \mathbf{z}, w(\mathbf{v})\rangle] = \mathbf{E}\left[\langle (\mathbf{y} - \mathbf{v})/2, w(\mathbf{v})\rangle\right] > \alpha/2.$$

Given $\mathbf{z} \in \mathcal{E}_k \cup (-\mathcal{E}_k)$, we use $\ell_{\mathbf{z}} \in [k]$ to denote the unique index such that $\mathbf{z}^{(\ell_{\mathbf{z}})} \neq 0$. We have $\langle \mathbf{z}, w(\mathbf{v})\rangle = \mathbf{z}^{(\ell_{\mathbf{z}})} w^{(\ell_{\mathbf{z}})}(\mathbf{v})$ and thus

$$\mathbf{E}[\mathbf{z}^{(\ell_{\mathbf{z}})} w^{(\ell_{\mathbf{z}})}(\mathbf{v})] > \alpha/2.$$

Therefore, there exists $j \in \{1, \ldots, k\}$ such that

$$\mathbf{E}[\mathbf{z}^{(j)} w^{(j)}(\mathbf{v}) | \ell_{\mathbf{z}} = j] > \alpha/2 > \alpha/3.$$

This allows us to use an $(\alpha/3, \beta)$ agnostic learner.

Proof [Proof of Theorem 24] In the auditing task, we assume that the input data points $(\mathbf{v}_i, \mathbf{y}_i)$ are drawn i.i.d. from a distribution $\mathcal{D}$ satisfying

$$\mathop{\mathbf{E}}_{(\mathbf{v},\mathbf{y})\sim\mathcal{D}}[\langle \mathbf{y}-\mathbf{v}, w(\mathbf{v})\rangle] > \alpha \quad \text{for some } w \in \mathcal{H}^k. \tag{5}$$

Given $(\mathbf{v}, \mathbf{y})$ drawn from $\mathcal{D}$, we draw $\mathbf{z}$ randomly from $\mathcal{E}_k \cup (-\mathcal{E}_k)$ such that $\mathbf{E}[\mathbf{z}|\mathbf{v}, \mathbf{y}] = (\mathbf{y}-\mathbf{v})/2$. This is possible because $\|\mathbf{y}-\mathbf{v}\|_1 \le 2$. A concrete way to draw $\mathbf{z}$ is the following. With probability $1/2$, we set $\mathbf{z}$ to be $\mathbf{y} \in \mathcal{E}_k$, and with the remaining probability $1/2$, we draw $\mathbf{z}$ randomly from $-\mathcal{E}_k$ with expectation $-\mathbf{v}$.

Given $\mathbf{z} \in \mathcal{E}_k \cup (-\mathcal{E}_k)$, we define a random variable $\ell_{\mathbf{z}} \in \{1, \dots, k\}$ such that $\ell_{\mathbf{z}}$ is the unique index satisfying $\mathbf{z}^{(\ell_{\mathbf{z}})} \neq 0$. For any $w \in \mathcal{H}^k$, there exists $w^{(1)}, \dots, w^{(k)} \in \mathcal{H}$ such that $w(\mathbf{v}) = (w^{(1)}(\mathbf{v}), \dots, w^{(k)}(\mathbf{v}))$ for every $\mathbf{v} \in \Delta_k$. We have $\langle \mathbf{z}, w(\mathbf{v})\rangle = \mathbf{z}^{(\ell_{\mathbf{z}})} w^{(\ell_{\mathbf{z}})}(\mathbf{v})$ and thus (5) implies

$$\mathbf{E}[\mathbf{z}^{(\ell_{\mathbf{z}})} w^{(\ell_{\mathbf{z}})}(\mathbf{v})\rangle] = \mathbf{E}[\langle \mathbf{z}, w(\mathbf{v})\rangle] = \mathbf{E}\left[\langle (\mathbf{y}-\mathbf{v})/2, w(\mathbf{v})\rangle\right] > \alpha/2.$$

Let $\mathcal{U}_j$ denote the conditional distribution of $(\mathbf{v}, \mathbf{z}^{(j)}) \in \Delta_k \times \mathcal{E}_k$ given $\ell_{\mathbf{z}} = j$. We have

$$\sum_{j=1}^{k} \Pr[\ell_{\mathbf{z}} = j] \mathop{\mathbf{E}}_{(\mathbf{v},z)\sim\mathcal{U}_j}[z w^{(j)}(\mathbf{v})] > \alpha/2. \tag{6}$$

Now we show that there exists $j \in \{1, \dots, k\}$ such that $\Pr[\ell_{\mathbf{z}} = j] \ge \alpha/6k$ and $\mathbf{E}_{(\mathbf{v},z)\sim\mathcal{U}_j}[z w^{(j)}(\mathbf{v})] > \alpha/3$. If this is not the case, then

$$\begin{aligned}
&\sum_{j=1}^{k} \Pr[\ell_{\mathbf{z}} = j] \mathop{\mathbf{E}}_{(\mathbf{v},z)\sim\mathcal{U}_j}[z w^{(j)}(\mathbf{v})] \\
&= \sum_{j:\Pr[\ell_{\mathbf{z}}=j]<\alpha/6k} \Pr[\ell_{\mathbf{z}} = j] \mathop{\mathbf{E}}_{(\mathbf{v},z)\sim\mathcal{U}_j}[z w^{(j)}(\mathbf{v})] + \sum_{j:\Pr[\ell_{\mathbf{z}}=j]\ge\alpha/6k} \Pr[\ell_{\mathbf{z}} = j] \mathop{\mathbf{E}}_{(\mathbf{v},z)\sim\mathcal{U}_j}[z w^{(j)}(\mathbf{v})] \\
&\le \alpha/6 + \alpha/3 \\
&= \alpha/2,
\end{aligned}$$

giving a contradiction with (6).

We have shown that there exists $j^* \in \{1, \dots, k\}$ and $h \in \mathcal{H}$ such that $\Pr[\ell_{\mathbf{z}} = j^*] \ge \alpha/6k$ and $\mathbf{E}_{(z,\mathbf{v})\sim\mathcal{U}_{j^*}}[z h(\mathbf{v})] > \alpha/3$. To solve the auditing task given examples $(\mathbf{v}_1, \mathbf{y}_1), \dots, (\mathbf{v}_n, \mathbf{y}_n)$, we first draw $\mathbf{z}_1, \dots, \mathbf{z}_n \in \mathcal{E}_k \cup (-\mathcal{E}_k)$ independently such that $\mathbf{E}[\mathbf{z}_i|\mathbf{v}_i, \mathbf{y}_i] = \mathbf{y}_i - \mathbf{v}_i$. Now $(\mathbf{v}_1, \mathbf{y}_1, \mathbf{z}_1), \dots, (\mathbf{v}_n, \mathbf{y}_n, \mathbf{z}_n)$ are distributed independently from the joint distribution of $(\mathbf{v}, \mathbf{y}, \mathbf{z})$. For every j, we define $I_j := \{i \in \{1, \dots, n\} : \ell_{\mathbf{z}_i} = j\}$. If $|I_j| \ge n_0$, we run the agnostic learner on the data points $((\mathbf{v}_i, \mathbf{z}_i^{(j)}))_{i\in I_j}$ to obtain a function $h^{(j)} : \Delta_k \to [-1, 1]$. We define $w_j : \Delta_k \to [-1, 1]^k$ such that $(w_j(\mathbf{v}))^{(j')} = 0$ if $j' \neq j$ and $(w_j(\mathbf{v}))^{(j')} = h^{(j)}(\mathbf{v})$ if $j' = j$.

When $n = O(kn_0/\alpha + k^2\alpha^{-2}\beta^{-2}\log(1/\delta))$ is sufficiently large, with probability at least $1 - \delta/4$, we have $|I_{j^*}| \ge n_0$. Conditioned on I_{j^*}, the data points $((\mathbf{v}_i, \mathbf{z}_i^{(j)}))_{i\in I_{j^*}}$ are distributed independently from $\mathcal{D}_{j^*}$, and thus by the guarantee of the agnostic learner, with probability at least $1 - \delta/2$,

$$\mathop{\mathbf{E}}_{(\mathbf{v},z)\sim\mathcal{U}_{j^*}}[z h^{(j^*)}(\mathbf{v})] \ge \beta,$$

which implies

$$\mathop{\mathbf{E}}_{(\mathbf{v},\mathbf{y})\sim\mathcal{D}}[\langle \mathbf{y}-\mathbf{v}, w_{j^*}(\mathbf{v})\rangle] = 2\Pr[\ell_{\mathbf{z}} = j^*] \mathop{\mathbf{E}}_{(\mathbf{v},z)\sim\mathcal{U}_{j^*}}[zh^{(j^*)}(\mathbf{v})] \geq \alpha\beta/3k.$$

We have thus shown that with probability at least $1 - 3\delta/4$, there exists j such that

$$\mathop{\mathbf{E}}_{(\mathbf{v},\mathbf{y})\sim\mathcal{D}}[(\mathbf{y}-\mathbf{v})w_j(\mathbf{v})] \geq \alpha\beta/3k.$$

By estimating the values of $\mathbf{E}_{(\mathbf{v},\mathbf{y})\sim\mathcal{D}}[\langle \mathbf{y}-\mathbf{v}, w_j(\mathbf{v})\rangle]$ using $O(\alpha^{-2}\beta^{-2}k^2\log(k/\delta))$ fresh examples, we can make sure that with probability at least $1-\delta$, we output a $\tilde{w}$ among the w_j's such that

$$\mathop{\mathbf{E}}_{(\mathbf{v},\mathbf{y})\sim\mathcal{D}}[\langle \mathbf{y}-\mathbf{v}, \tilde{w}(\mathbf{v})\rangle] \geq \alpha\beta/6k.$$

■

C.2. Agnostic Learning from Auditing

Now we prove the reverse direction of the reduction by constructing an agnostic learner for a class $\mathcal{H}$ using an auditor (Theorem 25). For the most general statement, instead of considering the auditing task for $\mathcal{H}^k$ as in Theorem 24, we need to consider a slightly different class $\tilde{\mathcal{H}}^k$. But as long as $\mathcal{H}$ is closed under coordinate-wise affine transformations of the inputs, we can choose $\tilde{\mathcal{H}}$ to be the same as $\mathcal{H}$. In particular, when $\mathcal{H}$ is the class of halfspaces, by our reduction, classic hardness results on agnostically learning halfspaces implies hardness of auditing for halfspaces (Theorem 28).

For a vector $\mathbf{v} \in \Delta_k$ with $k \geq 2$, define $\mathsf{lift}(\mathbf{v}) \in \Delta_k$ by

$$\mathsf{lift}(\mathbf{v}) := \frac{1}{3}\mathbf{v} + \frac{1}{3}\mathbf{e}_1 + \frac{1}{3}\mathbf{e}_2. \tag{7}$$

Theorem 25 *For $k \geq 2$, let $\mathcal{H}$ be a family of functions $h : \Delta_k \to [-1,1]$ closed under negation. Let $\tilde{\mathcal{H}}$ be a family of functions $\tilde{h} : \Delta_k \to [-1,1]$ such that for every $h \in \mathcal{H}$, there exists $\tilde{h} \in \tilde{\mathcal{H}}$ satisfying $\tilde{h}(\mathsf{lift}(\mathbf{v})) = h(\mathbf{v})$ for every $\mathbf{v} \in \Delta_k$. Given any $(2\alpha/3, 2\beta/3)$ auditor for $\tilde{\mathcal{H}}^k$, we can construct an (α,β) weak agnostic learner for $\mathcal{H}$ with the same sample complexity, time complexity, and failure probability parameter.*

We will in fact derive this result from a more general statement where the class of auditors is not necessarily a product set.

Theorem 26 *Let $\mathcal{H}$ be a family of functions $h : \Delta_k \to [-1,1]$ and let $\mathcal{W}$ be a family of functions $w : \Delta_k \to [-1,1]^k$. Let λ be a positive real number. Assume that for every $h \in \mathcal{H}$ there exists $w \in \mathcal{W}$ such that*

$$w(\mathsf{lift}(\mathbf{x}))_1 - w(\mathsf{lift}(\mathbf{x}))_2 = \lambda h(\mathbf{x}) \quad \textit{for every } \mathbf{x} \in \Delta_k. \tag{8}$$

Given any $(\lambda\alpha/3, 2\beta/3)$ auditor for $\tilde{\mathcal{W}}$, we can construct an (α,β) weak agnostic learner for $\mathcal{H}$ with the same sample complexity, time complexity, and failure probability parameter.

Proof We construct a weak agnostic learner for $\mathcal{H}$ using an auditor for $\mathcal{W}$. Let $(\mathbf{x}_1, z_1), \ldots, (\mathbf{x}_n, z_n)$ be the input data points in the weak agnostic learning task drawn i.i.d. from a distribution $\mathcal{U}$. For every input data point $(\mathbf{x}_i, z_i) \in \Delta_k \times [-1, 1]$, the learner generates a corresponding data point $(\mathbf{v}_i, \mathbf{y}_i) \in \Delta_k \times \{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ for the auditing task by setting

$$\mathbf{v}_i = \mathsf{lift}(\mathbf{x}_i),$$
$$\mathbf{y}_i \sim \mathbf{v}_i^* := \mathbf{v}_i + \frac{1}{3} z_i (\mathbf{e}_1 - \mathbf{e}_2).$$

Note that $\mathbf{v}_i^* \in \Delta_k$, and thus it can be interpreted as a distribution over $\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$. Let $\mathcal{D}$ denote the distribution of $(\mathbf{v}_i, \mathbf{y}_i)$. The intuition is that since $\mathbf{v}^*$ favors either $\mathbf{e}_1$ or $\mathbf{e}_2$ over $\mathbf{v}$ depending on $\mathbf{z}$, telling the difference between $\mathbf{v}_i$ and $\mathbf{v}_i^*$ for an auditor requires learning $\mathbf{z}$.

Formally, by our assumption, for any $h \in \mathcal{H}$, there exists $w \in \mathcal{W}$ satisfying (8), and thus

$$\mathop{\mathbf{E}}_{\mathcal{D}}[\langle \mathbf{y} - \mathbf{v}, w(\mathbf{v}) \rangle] = \frac{1}{3} \mathop{\mathbf{E}}_{\mathcal{U}}[z \langle \mathbf{e}_1 - \mathbf{e}_2, w(\mathsf{lift}(\mathbf{x})) \rangle] = \frac{\lambda}{3} \mathop{\mathbf{E}}_{\mathcal{U}}[z h(\mathbf{x})].$$

Therefore, if $\mathbf{E}_{\mathcal{U}}[z h(\mathbf{x})] \geq \alpha$ for some $h \in \mathcal{H}$, then $\mathbf{E}_{\mathcal{D}}[\langle \mathbf{y} - \mathbf{v}, w(\mathbf{v}) \rangle] \geq \lambda\alpha/3$ for some $w \in \mathcal{W}$, and with high probability, the auditing algorithm will produce some $w' : \Delta_k \to [-1, 1]^k$ such that $\mathbf{E}_{\mathcal{D}}[\langle \mathbf{y} - \mathbf{v}, w'(\mathbf{v}) \rangle] \geq 2\beta/3$. Defining $h' : \Delta_k \to [-1, 1]$ such that

$$h'(\mathbf{x}) = \frac{1}{2}(w(\mathsf{lift}(\mathbf{x}))|_1 - w(\mathsf{lift}(\mathbf{x}))|_2) \quad \text{for every } \mathbf{x} \in \Delta_k,$$

we have

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathcal{D}}[\langle \mathbf{y} - \mathbf{v}, w'(\mathbf{v}) \rangle] &= \frac{1}{3} \mathop{\mathbf{E}}_{\mathcal{U}}[z \langle \mathbf{e}_1 - \mathbf{e}_2, w'(\mathsf{lift}(\mathbf{x})) \rangle] \\
&= \frac{1}{3} \mathop{\mathbf{E}}_{\mathcal{U}}[z \langle w'(\mathsf{lift}(\mathbf{x}))|_1 - w'(\mathsf{lift}(\mathbf{x}))|_2 \rangle] \\
&= \frac{2}{3} \mathop{\mathbf{E}}_{\mathcal{U}}[z h'(\mathbf{x})].
\end{aligned}$$

Therefore, $\mathbf{E}_{\mathcal{D}}[\langle \mathbf{y} - \mathbf{v}, w'(\mathbf{v}) \rangle] \geq 2\beta/3$ implies $\mathbf{E}_{\mathcal{U}}[z h'(\mathbf{x})] \geq \beta$. We have thus constructed an (α, β)-weakly agnostic learning algorithm which returns h' as output. ■

We now complete the proof of Theorem 25.

Proof [Proof of Theorem 25] By our assumption about $\tilde{\mathcal{H}}$, for every $h \in \mathcal{H}$ there exist $\tilde{h}_1, \tilde{h}_2$ such that

$$\tilde{h}_1(\mathsf{lift}(\mathbf{v})) = h(\mathbf{v}), \ \tilde{h}_2(\mathsf{lift}(\mathbf{v})) = -h(\mathbf{v}).$$

We consider any $\tilde{h} \in (\tilde{\mathcal{H}})^k$ whose first two co-ordinates are $\tilde{h}_1$ and $\tilde{h}_2$, so that their difference is $2h(\mathbf{v})$. We now apply Theorem 26 with $\mathcal{W} = (\tilde{\mathcal{H}})^k$ and $\lambda = 2$. ■

Appendix D. Hardness of Auditing for Decision Calibration

Our tight connection between auditing and learning established in the previous section allows us to transfer hardness results from learning to auditing. We apply this machinery to show hardness of auditing for specific function classes. Under standard complexity-theoretic assumptions, we show that auditing for decision calibration (Theorem 9) cannot be solved in time $\mathrm{poly}(k)$.

Theorem 27 (Hardness of Decision Calibration) *For $k \in \mathbb{Z}_{>0}$, let W_k be the class W used in the definition of decision calibration (Theorem 9). Under standard hardness assumption on refuting random t-XOR (Theorem 29 below), for any $C > 2$ and any sufficiently large k, there is no $(1/3 - 1/C, 1/k^C)$-auditing algorithm for W_k that runs in time $O(k^C)$ and achieves success probability at least $3/4$.*

We also prove a related result showing hardness of auditing for the product class of halfspaces. Let $\mathcal{H}_{\mathsf{hs}}$ be the class of half-spaces over Δ_k. That is, $\mathcal{H}_{\mathsf{hs}}$ consists of all functions $h : \Delta_k \to [-1, 1]$ that can be written as $h(\mathbf{v}) = \mathrm{sign}(\mathbf{a} \cdot \mathbf{v} + b)$ for some $\mathbf{a} \in \mathbb{R}^k$ and $b \in \mathbb{R}$. We prove the following theorem showing that $\mathcal{H}_{\mathsf{hs}}^k$ does not allow $\mathrm{poly}(k)$-time auditing:

Theorem 28 (Hardness of Halfspace Calibration) *Under standard hardness assumption on refuting random t-XOR (Theorem 29 below), for any $C > 2$, there is no algorithm that, for every sufficiently large $k \in \mathbb{Z}_{>0}$, solves $(2/3 - 1/C, 1/k^C)$-auditing for $\mathcal{H}_{\mathsf{hs}}^k$ in time $O(k^C)$ and achieves success probability at least $3/4$.*

We combine reductions from Appendix C.2 with existing hardness results of agnostically learning halfspaces to prove the two theorems above. There are many results showing hardness of agnostic learning for halfspaces under various assumptions, for instance see (Feldman et al., 2009; Guruswami and Raghavendra, 2006). The strongest results for improper learning are due to Daniely based on the hardness of refuting random t-XOR-Sat (Daniely, 2016).

Assumption 29 (Random t-XOR Assumption (Daniely, 2016)) *There exist constants $\eta \in (0, 1/2)$ and $c > 0$ such that for any $t \in \mathbb{Z}_{>0}$, there is no $\mathrm{poly}(m)$-time algorithm A that satisfies the following properties for any sufficiently large $n \in \mathbb{Z}_{>0}$ and $m = \lfloor n^{c\sqrt{t}\log t} \rfloor$:*

- *given any size-m collection of t-XOR clauses on n variables where at least $1 - \eta$ fraction of the clauses are satisfiable, algorithm A outputs "accept" with probability at least $3/4$;*
- *with probability at least $q(n) = 1 - o(1)$ over a uniformly randomly chosen size-m collection of t-XOR clauses on n variables, given the collection as input, algorithm A outputs "reject" with probability at least $3/4$.*

Theorem 30 ((Daniely, 2016)) *Under Theorem 29, for any $C > 2$, there is no algorithm that, for every sufficiently large $k \in \mathbb{Z}_{>0}$, solves $(1 - 1/C, 1/k^C)$-agnostic learning for $\mathcal{H}_{\mathsf{hs}}$ with success probability at least $3/4$ and runs in time $O(k^C)$.*

The original result by Daniely (2016) was stated for the Boolean cube instead of Δ_k, but the result extends to Δ_k by taking an affine injection from the Boolean cube $\{-1, 1\}^{k-1}$ to Δ_k.

We prove Theorem 27 and Theorem 28 by combining Theorem 30 with Theorem 25 and Theorem 26 from Appendix C.2. The following simple claim is convenient for our proof and it follows immediately from the definition of $\mathsf{lift}(\cdot)$ in (7).

Claim 31 *For $\mathbf{a} \in \mathbb{R}^k, b \in \mathbb{R}$, define $\mathbf{a}' = 3\mathbf{a}$, $b' = b - \mathbf{a}^{(1)} - \mathbf{a}^{(2)}$. Then for every $\mathbf{v} \in \Delta_k$,*

$$\langle \mathbf{a}', \mathsf{lift}(\mathbf{v}) \rangle + b' = \langle \mathbf{a}, \mathbf{v} \rangle + b.$$

Proof [Proof of Theorem 27] Any function $h \in \mathcal{H}_{\mathsf{hs}}$ can be expressed as $h(\mathbf{v}) = \mathrm{sign}(\langle \mathbf{a}, \mathbf{v} - b \rangle)$ for $\mathbf{a} \in \mathbb{R}^k$ and $b \in \mathbb{R}$. Define $\mathbf{a}' = 3\mathbf{a}$, $b' = b - \mathbf{a}^{(1)} - \mathbf{a}^{(2)}$, $\mathbf{g} = \mathbf{e}_1$ and $\mathbf{g} = -\mathbf{e}_1$. The function w mapping $\mathbf{v}'$ to $\mathbb{I}(\langle \mathbf{a}', \mathbf{v}' \rangle > b')\mathbf{g} + \mathbb{I}(\langle \mathbf{a}', \mathbf{v}' \rangle \le b')\mathbf{g}'$ belongs to W_k, and

$$w(\mathsf{lift}(\mathbf{v}))^{(1)} - w(\mathsf{lift}(\mathbf{v}))^{(2)} = \mathbb{I}(\langle \mathbf{a}', \mathsf{lift}(\mathbf{v}) \rangle > b') - \mathbb{I}(\langle \mathbf{a}', \mathsf{lift}(\mathbf{v}) \rangle \le b') = \mathrm{sign}(\langle \mathbf{a}', \mathsf{lift}(\mathbf{v}) - b' \rangle) = h(\mathbf{v}).$$

Therefore, by Theorem 26, any $(1/3 - 1/C, 1/k^C)$-auditing algorithm for W_k implies a $(1 - 3/C, 3/2k^C)$-agnostic learning algorithm for $\mathcal{H}_{\mathsf{hs}}$ with the same sample complexity, running time, and failure probability. The proof is completed by Theorem 30. ■

Proof [Proof of Theorem 28] By Theorem 31, $\mathcal{H}_{\mathsf{hs}}$ is closed under the lift operation, namely for every $h \in \mathcal{H}_{\mathsf{hs}}$ we can construct $h' \in \mathcal{H}_{\mathsf{hs}}$ which satisfies $h'(\mathsf{lift}(\mathbf{v})) = h(\mathbf{v})$ for every $\mathbf{v} \in \Delta_k$. Assume for the sake of contradiction that an auditing algorithm for $\mathcal{H}^k_{\mathsf{hs}}$ as described in the theorem exists. By Theorem 25, such an algorithm implies a $(1 - 3C/2, 3/2k^C)$-weak agnostic learner for $\mathcal{H}_{\mathsf{hs}}$ that runs in time $O(k^C)$ for any sufficiently large k, contradicting Theorem 30. ■

Appendix E. Kernel Algorithms for Auditing Calibration

In this section, we give efficient auditing algorithms for weighted calibration where the weight family $\mathcal{W}$ consists of functions from a reproducing kernel Hilbert space (RKHS). In Appendix E.1, we discuss a special case using the multinomial kernel, which is important for our efficient auditors for projected smooth calibration in Appendix F.

It is well known that learning for functions with bounded norm in an RKHS with convex losses is feasible by solving a convex program. Here we observe that the simple structure of the correlation objective $\mathbf{E}[zw(\mathbf{v})]$ in agnostic learning makes it possible to optimize, even without solving a convex program, just using $O(n^2)$ kernel evaluations, via Algorithm 1. The algorithm and its analysis are not novel and are similar in nature to the kernel ridge regression algorithm (see e.g. (Wainwright, 2009)). Based on our connection between auditing and learning shown in Appendix C, we give a similar kernel evaluation based algorithm for multi-class auditing, which we present in Algorithm 2.

Let $\mathcal{D}$ be a distribution over $\Delta_k \times [-1, 1]$. Consider a positive definite kernel $\ker : \Delta_k \times \Delta_k \to \mathbb{R}$ and the corresponding RKHS Γ consisting of functions $w : \Delta_k \to \mathbb{R}$. We assume that the kernel can be evaluated efficiently. Let $\varphi_{\mathbf{v}} \in \Gamma$ denote the function $\ker(\mathbf{v}, \cdot)$. By the reproducing property,

$$w(\mathbf{v}) = \langle w, \varphi_{\mathbf{v}} \rangle_\Gamma \quad \text{for every } w \in \Gamma \text{ and } \mathbf{v} \in \Delta_k.$$

Define $B_\Gamma(r)$ to be the set of $w \in \Gamma$ satisfying $\|w\|_\Gamma^2 := \langle w, w \rangle_\Gamma \le r^2$. For $s > 0$, assume that $\ker(\mathbf{v}, \mathbf{v}) \le s^2$ for every $\mathbf{v} \in \Delta_k$. That is, $\|\varphi_{\mathbf{v}}\|_\Gamma \le s$. Under this assumption, for any $w \in B_\Gamma(1/s)$ and $\mathbf{v} \in \Delta_k$, we have

$$|w(\mathbf{v})| = |\langle w, \varphi_{\mathbf{v}} \rangle_\Gamma| \le \|w\|_\Gamma \|\varphi_{\mathbf{v}}\|_\Gamma \le (1/s) \cdot s \le 1.$$

The following theorems are proved in Appendix H.

Theorem 32 *For $n = O(r^2 s^2 \alpha^{-2} \log(1/\delta))$, Algorithm 1 is an $(\alpha, \alpha/3rs)$ agnostic learner for the class $B_\Gamma(r)$ with failure probability at most δ. Moreover, it always returns a function from $B_\Gamma(1/s)$.*

Theorem 33 *For $n = O(kr^2 s^2 \alpha^{-2} \log(1/\delta))$, Algorithm 2 is an $(\alpha, \alpha/3rs)$ auditor for the class $B_\Gamma(r)^k$ with failure probability at most δ. Moreover, it always returns a function from $B_\Gamma(1/s)^k$.*

Algorithm 1: Kernel Algorithm for Weak Agnostic Learning
Input : Data points $(\mathbf{v}_1, z_1), \ldots, (\mathbf{v}_n, z_n) \in \Delta_k \times [-1, 1]$.
Output: Function $w_2 : \Delta_k \to [-1, 1]$.
begin
$\lambda \leftarrow \left(\sum_{i=1}^n \sum_{j=1}^n z_i z_j \ker(\mathbf{v}_i, \mathbf{v}_j)\right)^{1/2}$ $w_2(\mathbf{v}) \leftarrow \frac{1}{\lambda s} \sum_{i=1}^n z_i \ker(\mathbf{v}_i, \mathbf{v})$ for every $\mathbf{v} \in \Delta_k$ (in the degenerate case where $\lambda = 0$, set $w_2(\mathbf{v}) \leftarrow 0$) **return** w_2
end

Algorithm 2: Kernel Algorithm for Auditing
Input : Data points $(\mathbf{v}_1, \mathbf{y}_1), \ldots, (\mathbf{v}_n, \mathbf{y}_n) \in \Delta_k \times \mathcal{E}_k$.
Output: Function $w_2 : \Delta_k \to [-1, 1]^k$.
begin
$\mathbf{z}_i \leftarrow \mathbf{y}_i - \mathbf{v}_i$ for every $i = 1, \ldots, n$ For $i = 1, \ldots, n$ and $\ell = 1, \ldots, k$, let $\mathbf{z}_i^{(\ell)}$ denote the ℓ-th coordinate of z_i $\lambda^{(\ell)} \leftarrow \left(\sum_{i=1}^n \sum_{j=1}^n \mathbf{z}_i^{(\ell)} \mathbf{z}_j^{(\ell)} \ker(\mathbf{v}_i, \mathbf{v}_j)\right)^{1/2}$ for every $\ell = 1, \ldots, k$ $w_2^{(\ell)}(\mathbf{v}) \leftarrow \frac{1}{\lambda^{(\ell)} s} \sum_{i=1}^n \mathbf{z}_i^{(\ell)} \ker(\mathbf{v}_i, \mathbf{v})$ for every $\ell = 1, \ldots, k$ and $\mathbf{v} \in \Delta_k$ (in the degenerate case where $\lambda^{(\ell)} = 0$, set $w_2^{(\ell)}(\mathbf{v}) \leftarrow 0$) **return** w_2 such that $w_2(\mathbf{v}) = \left(w_2^{(1)}(\mathbf{v}), \ldots, w_2^{(k)}(\mathbf{v})\right)$ for every $\mathbf{v} \in \Delta_k$
end

E.1. Auditing for the Multinomial Kernel

A kernel that will be of particular importance for us is the multinomial kernel. We follow the elegant formulation from (Goel et al., 2017).

Definition 34 *(Goel et al., 2017) For any vector $\mathbf{v} = (v_1, \ldots, v_k) \in \Delta_k$ and tuple $t = (t_1, \ldots, t_d) \in [k]^d$, define $\mathbf{v}^t$ to be the product $v_{t_1} \cdots v_{t_d}$. Define $\psi : \Delta_k \to \mathbb{R}^{1+k+\cdots+k^d}$ such that $\psi(\mathbf{v})$ is a vector whose coordinate indexed by $t \in T_d := [k]^0 \cup [k]^1 \cup \cdots \cup [k]^d$ is $\mathbf{v}^t$. The degree d multinomial kernel is given by*

$$\ker_d(\mathbf{v}, \mathbf{v}') = \sum_{i=0}^{d} (\mathbf{v} \cdot \mathbf{v}')^i = \psi(\mathbf{v}) \cdot \psi(\mathbf{v}').$$

We denote its RKHS as $\Gamma(d)$.

Instantiating Theorem 33 for the degree d multinomial kernel gives the following result.

Lemma 35 *For all, $r \geq 0$ and $d \geq 0$, Algorithm 2 with $n = O(kr^2 d \log(1/\delta)/\alpha^2)$ samples is an $(\alpha, \alpha/(3r\sqrt{d}))$-auditor for the class $(B_{\Gamma(d)}(r))^k$ with failure probability at most δ. Moreover, it always returns a function from $(B_{\Gamma(d)}(1/\sqrt{d}))^k$ in time $\mathrm{poly}(n, k, d)$.*

Proof Observe that for $\mathbf{v} \in \Delta_k$,

$$\|\varphi_{\mathbf{v}}\|_{\Gamma(d)}^2 = \ker_d(\mathbf{v}, \mathbf{v}) = \sum_{i=0}^{d} (\mathbf{v} \cdot \mathbf{v})^i \leq d$$

since $\|\mathbf{v}\|_2^2 \leq 1$ for $\mathbf{v} \in \Delta_k$. We can thus apply Theorem 33 with $s = \sqrt{d}$ to get the claimed bound. ■

This gives a faster auditor for the notion of low-degree calibration defined by (Gopalan et al., 2022b).

Definition 36 *(Gopalan et al., 2022b) Let $P(d,1)$ denote the set of multivariate degree d polynomials*

$$p(v_1, \ldots, v_k) = \sum_{e:\deg(e)\leq d} w_e \prod_i v_i^{e_i}$$
$$\text{where } \forall \mathbf{v} \in \Delta_k,\ |p(\mathbf{v})| \leq 1,$$
$$\sum_{e:\deg(e)\leq d} |w_e| \leq 1.$$

We say that a predictor p is α degree-d calibrated if $\mathsf{CE}_{P(d,1)^k}(\mathcal{D}) \leq \alpha$.

(Gopalan et al., 2022b) give an $(\alpha, \alpha/k^d)$-auditor for $P(d,1)^k$ which runs in time $O(k^d)$ by enumerating over all k^d monomials. Algorithm 2 implies a better auditor which is polynomial in both k and d.

Corollary 37 *There is an $(\alpha, \alpha/3\sqrt{d})$-auditor for $P(d,1)^k$ that with success probability at least $1-\delta$, sample complexity $n = O(kd\log(1/\delta)/\alpha^2)$ and time complexity $\mathrm{poly}(n,k,d)$.*

Proof Any polynomial $p \in P(d,1)$ can be written as $p(\mathbf{v}) = \sum_{t\in T_d} w_t \mathbf{v}^t$ for every $\mathbf{v} \in \mathbb{R}^k$, where $w_t \in \mathbb{R}$ for every $t \in T_d$ and $\sum_{t\in T_d} |w_t| \leq 1$. We define a vector $\psi^p \in \mathbb{R}^{1+k+\cdots+k^d}$ whose coordinate indexed by $t \in T_d$ is w_t. It follows that

$$p(\mathbf{v}) = \sum_{t\in T_d} w_t \mathbf{v}^t = \psi^p \cdot \psi(\mathbf{v}),$$
$$\|p\|_{\Gamma(d)}^2 \leq \|\psi^p\|_2^2 = \sum_t w_t^2 \leq \left(\sum_t |w_t|\right)^2 \leq 1.$$

Hence $P(d,1) \subseteq \mathcal{B}_{\Gamma(d)}(1)$. Hence the claimed bound follows from Lemma 35 with $r = 1$. ■

Appendix F. Efficient Auditing for Projected Smooth Calibration

In this section, we prove the following theorem showing an efficient kernel-based auditing algorithm for projected smooth calibration (Theorem 11).

Theorem 38 *There exists $c > 0$ so that for any $\alpha, \delta \in (0, 1/2)$ and $m \in [2,k]$, there is an $(\alpha, 1/m^{O(1/\alpha)})$ auditor for m-projected smooth calibration (and hence also for m-subset smooth calibration), with success probability at least $1-\delta$, sample complexity $n = O(km^{O(1/\alpha)}\log(1/\delta))$, and running time $\mathrm{poly}(n,k,1/\alpha)$.*

Even when we consider subset calibration over arbitrary subsets, which corresponds to taking $m = k$, the running time of the auditor is $k^{O(1/\alpha)}$, which is polynomial in k for every fixed α. In the next section (Appendix G), we show that this running time cannot be improved to $\mathrm{poly}(k, 1/\alpha)$ under standard complexity-theoretic assumptions. At the end of this section, we show that the dependence on α can be improved if we consider sigmoid functions instead of all 1-Lipschitz functions.

We prove Theorem 38 using Algorithm 2, together with polynomial approximations. Low degree polynomial approximations have been used successfully for agnostic learning, starting with the work of (Kalai et al., 2008a). The important work of (Shalev-Shwartz et al., 2011) showed that one can improve the efficiency of such learning algorithms by kernelizing them.

Using results from (Goel et al., 2017) and (Sherstov, 2013), we will show the following bound on multivariate polynomials obtained by composing bounded univariate polynomials with innner products.

Lemma 39 *Let p be a univariate polynomial of degree d so that $|p(u)| \leq 1$ for $u \in [-1, 1]$. Let $p_{\mathbf{a}}(\mathbf{v}) = p(\mathbf{a} \cdot \mathbf{v})$ where $\mathbf{a} \in [-1, 1]^d$ and $\mathbf{v} \in \Delta_k$. Then $p_{\mathbf{a}} \in \Gamma(d)$ and*

$$\|p_{\mathbf{a}}\|^2_{\Gamma(d)} \leq \max(4, 4\|\mathbf{a}\|_2)^{2d}.$$

We prove Theorem 39 using the following two lemmas from the literature:

Lemma 40 *(Goel et al., 2017, Lemma 2.7) Let $p = \sum_{i=0}^{d} \eta_i u^i$ be a univariate polynomial of degree d and $p_{\mathbf{a}}(\mathbf{v}) = p(\mathbf{a} \cdot \mathbf{v})$ for $\mathbf{a} \in [-1, 1]^k$ and $\mathbf{v} \in \Delta_k$. Then*

$$\|p_{\mathbf{a}}\|^2_{\Gamma^d} \leq \sum_{i=0}^{d} \eta_i^2 \|\mathbf{a}\|_2^{2i} \leq \max(1, \|\mathbf{a}\|_2)^{2d} \sum_{i=0}^{d} \eta_i^2.$$

Lemma 41 *(Sherstov, 2013, Lemma 4.1) For a degree d polynomial $p(u) = \sum_{i=0}^{d} \eta_i u^i$ satisfying $|p(u)| \leq 1$ for $u \in [-1, 1]$, it holds that $\sum_{i=0}^{d} |\eta_i| \leq 4^d$.*

Proof [Proof of Lemma 39] By Lemma 41, we can bound

$$\sum_{i=0}^{d} |\eta_i|^2 \leq \left(\sum_{i=0}^{d} |\eta_i| \right)^2 \leq 4^{2d}.$$

We plug this bound into Lemma 40 to get

$$\|p_{\mathbf{a}}\|^2_{\Gamma^d} \leq \max(4, 4\|\mathbf{a}\|_2)^{2d}.$$

■

Let Lip denote the set of all bounded 1-Lipschitz functions $\phi : [0, 1] \to [-1, 1]$. We use an approximation result for arbitrary Lipschitz functions using Jackson's theorem (Cheney, 1966), together with a rescaling argument to ensure boundedness. A similar argument for the ReLU function appears in (Goel et al., 2017, Lemma 2.12).

Lemma 42 *There exists a constant $c' > 0$ such that for any $\phi \in \mathrm{Lip}$ and any $\varepsilon > 0$, there exists a univariate polynomial $p(t)$ with $\deg(p) \leq c'/\varepsilon$ such that for $t \in [-1, 1]$,*

- $|\phi(t) - p(t)| \le \varepsilon$.
- $p(t) \in [-1, 1]$.

Proof By Jackson's theorem (Cheney, 1966), there exist a polynomial $p(t)$ so that $|\phi(t) - p(t)| \le \varepsilon/2$ for $t \in [-1, 1]$ where $\deg(p) \le O(1/\epsilon)$. Since $|\phi(t)| \le 1$, $|p(t)| \le 1 + \varepsilon/2$. Now let $p_\phi(t) = p(t)/(1 + \varepsilon/2)$ so that $|p_\phi(t)| \le 1$. We then bound

$$\begin{aligned}|p_\phi(t) - \phi(t)| &= \frac{1}{1+\varepsilon/2}|(p(t) - (1+\varepsilon/2)\phi(t))| \\ &\le \frac{1}{1+\varepsilon/2}(|p(t) - \phi(t)| + \varepsilon/2|\phi(t)|) \\ &\le \frac{\varepsilon}{1+\varepsilon/2} \le \varepsilon.\end{aligned}$$

■

Combining Lemmas 42 and 39, we have the following corollary.

Corollary 43 *For any $\phi \in$ Lip, and $\varepsilon > 0$, let p be as in Lemma 42. For $\mathbf{a} \in [-1, 1]^k$, let $p_{\mathbf{a}}(\mathbf{v}) = p(\mathbf{a} \cdot \mathbf{v})$. Then $p_{\mathbf{a}} \in \Gamma(d)$ for $d = O(1/\varepsilon)$ and*

- $\left|p_{\mathbf{a}}(\mathbf{v}) - \phi(\mathbf{a} \cdot \mathbf{v})\right| \le \varepsilon$, *for every $v \in \Delta_k$.*
- $\|p_{\mathbf{a}}\|_{\Gamma(d)} \le c_1 \max(1, \|\mathbf{a}\|_2)^{c_2/\varepsilon}$.

We now complete the proof of Theorem 38.

Proof [Proof of Theorem 38] We claim that if $\mathrm{psCE}_m(\mathcal{D}) \ge \alpha$, then $\mathsf{CE}_{(\mathcal{B}_{\Gamma(d)}(r))^k}(\mathcal{D}) \ge \alpha/2$ for some $r = m^{O(1/\alpha)}$. To see this, take $\psi \in \mathrm{pLip}^k$ so that

$$\mathop{\mathbf{E}}_{\mathcal{D}}[\langle \mathbf{y}^* - \mathbf{v}, \psi(\mathbf{v})\rangle \ge \alpha.$$

For every $i = 1, \ldots, k$, there exists $\mathbf{a}_i \in [-1, 1]^k$ and $\phi \in$ Lip such that $\psi^{(i)}(\mathbf{v}) = \phi^{(i)}(\langle \mathbf{a}_i, \mathbf{v}\rangle)$ for $i \in [k]$ where $\|\mathbf{a}_i\|_2 \le \sqrt{m}$. By Corollary 43, there exists $p^{(i)} \in \Gamma(d)$ where $d = O(1/\alpha)$ such that

$$\begin{aligned}\left\|\psi_i(\mathbf{v}) - p^{(i)}(\mathbf{v})\right\|_\infty &\le \alpha/4, \\ \left\|p^{(i)}\right\|_{\Gamma(d)} &\le (c_1 \max(1, \|\mathbf{a}_i\|_2))^{c_2/\alpha} \le (c_1\sqrt{m})^{c_2/\alpha}.\end{aligned}$$

Hence $p^{(i)} \in \mathcal{B}_{\Gamma(d)}(r)$ for each i for $r = m^{O(1/\alpha)}$.

Define $p(\mathbf{v}) = (p^{(1)}(\mathbf{v}), \ldots, p^{(k)}(\mathbf{v})) \in \mathbb{R}^k$. By the triangle inequality

$$\begin{aligned}\left|\mathop{\mathbf{E}}_{\mathcal{D}}[\langle \mathbf{y}^* - \mathbf{v}, p(\mathbf{v})\rangle - \mathop{\mathbf{E}}_{\mathcal{D}}[\langle \mathbf{y}^* - \mathbf{v}, \psi(\mathbf{v})\rangle]\right| &= \left|\mathop{\mathbf{E}}_{\mathcal{D}}[\langle \mathbf{y}^* - \mathbf{v}, p(\mathbf{v}) - \psi(\mathbf{v})\rangle]\right| \\ &\le \mathop{\mathbf{E}}_{\mathcal{D}}\left[|\langle \mathbf{y}^* - \mathbf{v}, p(\mathbf{v}) - \psi(\mathbf{v})\rangle|\right] \\ &\le \mathop{\mathbf{E}}_{\mathcal{D}}\left[\|\mathbf{y}^* - \mathbf{v}\|_1 \|p(\mathbf{v}) - \psi(\mathbf{v})\|_\infty\right] \\ &\le 2 \cdot \frac{\alpha}{4} \le \frac{\alpha}{2}\end{aligned}$$

where we use $\|\mathbf{y}^* - \mathbf{v}\|_1 \leq 2$. As a result we have

$$\mathop{\mathbf{E}}_{\mathcal{D}}[\langle \mathbf{y}^* - \mathbf{v}, p(\mathbf{v})\rangle] \geq \mathop{\mathbf{E}}_{\mathcal{D}}[\langle \mathbf{y}^* - \mathbf{v}, \psi(\mathbf{v})\rangle] - \alpha/2 \geq \alpha - \alpha/2 = \alpha/2.$$

We now apply Lemma 35 with the weight functions $(\mathcal{B}_{\Gamma(d)}(r))^k$ where $d = O(1/\alpha), r = m^{O(1/\alpha)}$ to get an $(\alpha/2, \Omega(\alpha^{3/2}/m^{c/\alpha}))$-auditing algorithm. ■

Auditing for Sigmoids. We show additionally that the exponential dependence on $1/\alpha$ in Theorem 38 can be improved if we audit only for sigmoid functions. Formally we use the $\tanh$ function rather then the sigmoid, since we want the range to be $[-1, 1]$ in order to approximate the sign function. Nevertheless, we refer to the family as the family of sigmoid functions.

Definition 44 *For $L \geq 1$, define $\Sigma_L = \{g : \mathbb{R} \to [-1, 1]\}$ to be the family of functions of the form*

$$g(\mathbf{v}) = \tanh(L\langle \mathbf{a}, \mathbf{v}\rangle + b) \text{ for } \mathbf{a} \in [-1, 1]^k,\ b \in \mathbb{R}.$$

In Theorem 46 below we show an efficient auditor for Σ_L^k whose running time is polynomial in $1/\alpha$ for every fixed k and L.

Observe that Σ_L increase monotonically with L, since for $L' < L$, $L'\langle \mathbf{a}, \mathbf{v}\rangle = L\langle \mathbf{a}', v\rangle$ where $\mathbf{a}' = L'\mathbf{a}/L \in [-1, 1]^k$. The problem of agnostically learning Σ_L over Δ_k is given a distribution $\mathcal{U}$ on $\Delta_k \times \{\pm 1\}$, find $g \in \Sigma_L$ that maximizes $\mathbf{E}_{\mathcal{U}}[g(\mathbf{v})z]$. The problem of agnostically learning sigmoids over the unit sphere (rather than Δ_k) was considered in the influential work of (Shalev-Shwartz et al., 2011). They work with the objective function $\min_{g\in\Sigma_L} \mathbf{E}\,|z - g(\mathbf{v})|$, but this is seen to be equivalent to $\max_{g\in\Sigma_L}[\mathbf{E}[g(\mathbf{v})z]$ when $z \in \{-1, 1\}$. A more substantial difference is that they work in the ℓ_2 bounded setting where $\|\mathbf{v}\|_2 \leq 1$, $\|\mathbf{a}\|_2 \leq 1$, whereas we work with ℓ_1/ℓ_∞-bounded setting where $\|\mathbf{v}\|_1 \leq 1$ and $\|\mathbf{a}\|_\infty \leq 1$. Thus we cannot directly use their results, although our techniques are influenced by them.

Our algorithm will use the following results about univariate approximations to the $\tanh$ function was proved in the work of (Shalev-Shwartz et al., 2011), with subsequent proofs given by (Livni et al., 2014; Goel et al., 2020). We use the following version from (Goel et al., 2020).

Lemma 45 *(Goel et al., 2020) For $\varepsilon \in (0, 1/2), L \geq 1$ and $b \in \mathbb{R}$, there exists a univariate polynomial $p(t)$ with $\deg(p) \leq O(L\log(L/\varepsilon))$ so that for $t \in [-1, 1]$*

- $|\tanh(Lt + b) - p(t)| \leq \varepsilon$.
- $p(t) \in [-1, 1]$.

Following the same proof outline as Theorem 38 gives the following result.

Theorem 46 *For any $\alpha \in (0, 1/2), L > 1$, there is an efficient (α, β)-auditor for $(\Sigma_L)^k$ calibration for*

$$\beta = \frac{\alpha}{k^{O(L\log(L/\alpha))}}$$

which has time and sample complexity $k^{O(L\log(L/\alpha))}$ and success probability at least $1 - 2^{-k}$.

The same techniques also yield an algorithm for agnostically learning Σ_L under any distribution $\mathcal{U}$ on $\Delta_k \times [-1, 1]$ with similar parameters.

Appendix G. Computational Lower Bound for Projected Smooth Calibration

The sample and time complexity of our auditing algorithm for projected smooth calibration in Appendix F is $k^{O(1/\alpha)}$ (when setting $m = k$). In this section, we show that an improvement to $\mathrm{poly}(k, 1/\alpha)$ (or just to $k^{O(\log^{0.99}(1/\alpha))}$) would violate standard complexity-theoretic assumptions:

Theorem 47 *Under a standard hardness assumption on refuting t-XorSat (Theorem 51), for any $C > 0$, $\varepsilon > 0$, there is no algorithm solving $(\alpha, 1/k^C)$ auditing for k-projected smooth calibration for every sufficiently large k and every $\alpha \in (0, 1/3)$ with success probability at least $3/4$ and running time $k^{O((\log(1/\alpha))^{1-\varepsilon})}$.*

We use the following connection between auditing projected smooth calibration and auditing for sigmoids Σ_L^k.

Lemma 48 *For $\alpha, \beta \in (0, 1)$ and $L > 1$, any $(\alpha/L, \beta)$-auditing algorithm for Lip^k is an (α, β)-auditing algorithm for Σ_L^k.*

Proof For a class $\mathcal{W}$ of functions $h : \Delta_k \to [-1, 1]^k$, recall the following notion in our definition of auditing:

$$\mathsf{CE}_{\mathcal{W}}(\mathcal{D}) = \sup_{w \in \mathcal{W}} \left| \mathop{\mathbf{E}}_{(\mathbf{v}, \mathbf{y}) \sim \mathcal{D}} [\langle \mathbf{y} - \mathbf{v}, w(\mathbf{v}) \rangle] \right|.$$

It suffices to show that for any distribution $\mathcal{D}$ over $\Delta_k \times \mathcal{E}_k$,

$$\mathsf{CE}_{\mathrm{Lip}^k}(\mathcal{U}) \geq \frac{1}{L} \mathsf{CE}_{\Sigma_L^k}(\mathcal{U}). \tag{9}$$

Consider any function $g \in \Sigma_L$. By definition, there exist $\mathbf{a} \in [-1, 1]^k$ and $b \in \mathbb{R}$ such that $g(\mathbf{v}) = \tanh(L\langle \mathbf{a}, \mathbf{v} \rangle + b)$ for every $\mathbf{v} \in \Delta_k$. It is easy to verify that $\tanh$ is 1-Lipschitz, and thus for any $\mathbf{v}_1, \mathbf{v}_2 \in \Delta_k$,

$$|g(\mathbf{v}_1) - g(\mathbf{v}_2)| \leq L|\langle \mathbf{a}, \mathbf{v}_1 \rangle - \langle \mathbf{a}, \mathbf{v}_2 \rangle| \leq L\|\mathbf{v}_1 - \mathbf{v}_2\|_1.$$

Therefore, the function g/L belongs to Lip, confirming (9). ■

Theorem 49 *Under a standard hardness assumption on refuting t-XorSat (Theorem 51), for some fixed $\alpha > 0$, for any $C > 0$, $\varepsilon > 0$, $c \in (0, 1)$, there is no algorithm that solves $(\alpha, 1/k^C)$ auditor for Σ_L^k for every sufficiently large $k \in \mathbb{Z}_{>0}$ and $L := \exp(\log^c k)$ with success probability at least $3/4$ and running time $k^{(\log L)^{1-\varepsilon}}$.*

Proof [Proof of Theorem 47] Let α_0 denote the fixed constant α in Theorem 49. Theorem 47 follows immediately by combining Theorem 49 and Theorem 48, where we choose α in Theorem 47 to be α_0/L. ■

G.1. Projected Smooth Calibration and Sigmoids

Now we prove Theorem 49. Our reduction from auditing to agnostic learning lets us focus on the complexity of agnostic learning Σ_L to understand auditing with weight functions $\Sigma^k_{L/3}$. This is formally stated below.

Lemma 50 *Given any $(2\alpha/3, 2\beta/3)$ auditor for $\Sigma^k_{L/3}$, we can construct an (α, β) weak agnostic learner for Σ_L over Δ_k with the same sample complexity, time complexity, and failure probability parameter.*

Proof For $g \in \Sigma_{L/3}$, we claim there exists $g' \in \Sigma_L$ such that $g'(\mathsf{lift}(\mathbf{v})) = g(\mathbf{v})$. Indeed since $\mathsf{lift}(\mathbf{v}) = \mathbf{v}/3 + \mathbf{e}_1/3 + \mathbf{e}_2/3$,

$$\begin{aligned}\tanh(L\langle w, \mathbf{v}\rangle/3 + b) &= \tanh(L\langle w, 3\mathsf{lift}(\mathbf{v})\rangle/3 + b - Lw^{(1)}/3 - Lw^{(2)}/3) \\ &= \tanh(L\langle w, \mathsf{lift}(\mathbf{v})\rangle + b') \in \Sigma_L\end{aligned}$$

We now apply Theorem 25 to get the stated claim. ■

Our lower bound for agnostically learning sigmoids is obtained by tailoring Daniely's (Daniely, 2016) reduction from refuting random XorSat to the ℓ_∞/ℓ_1 bounded setting.

An instance of t-XorSat consists of m clauses on n variables $\{z_1, \ldots, z_n\}$ each taking values in $\{\pm 1\}$. Each clause consists of exactly t literals which might be variables or their negations, we assume that x_i and $-x_i$ do not occur in the same clause. Thus each clause c can be arithmetized as a vector in $\{0,1\}^{2n}$ of weight exactly t, interpreted as a subset of literals. We will let $C(t) \subset \{0,1\}^{2n}$ denote the set of valid clauses. Similarly, assignments to z can be (redundantly) arithmetized as vectors in $Z \subseteq \{\pm 1\}^{2n}$ where $|Z| = 2^n$. Given a clause $c_i \in C$ and $z \in Z$, $c_i \cdot z \in \{-t, \ldots, t\}$ equals the sum of literals in the clause c_i. An instance of t-XorSat is given by $I = \{(c_i, b_i)\}_{i=1}^m$ where $c_i \in C(t)$ and $b_i \in \{\pm 1\}$. For a clause $c \in C_t$, let $\mathrm{Xor}_c(z) = \prod_{i \in c} z_i$. For an instance I and $z \in Z$, we define

$$\mathrm{val}(z, I) = \frac{|i \in [m] : \mathrm{Xor}_{c_i}(z) = b_i|}{m}$$

and $\mathrm{val}(I) = \max_{z \in Z} \mathrm{val}(z, I)$ to be the maximum fraction of satisfiable clauses.

A random instance of t-XorSat is one where $c_i \leftarrow C$ and $b_i \leftarrow \{\pm 1\}$ are drawn uniformly and independently at random. We let $\mathcal{R}$ denote the distribution on instances that this defines. An algorithm $\mathcal{A}$ which maps t-XorSat instances to $\{0,1\}$ successfully refutes random t-XorSat if

$$\begin{aligned}\Pr[\mathcal{A}(I) = 1] &\geq \frac{3}{4} \text{ if } \mathrm{val}(I) \geq 1 - \eta \\ \Pr[\mathcal{A}(I) = 0]] &\geq \frac{3}{4} \text{ with probability } 1 - o_n(1) \text{ over } I \sim \mathcal{R}.\end{aligned}$$

We are interested in the asymptotics in both t and n. The best known algorithms for refutation require $m = \Omega(n^{t/2})$ and it is conjectured that there are no algorithms with running time $n^{o(t)}$.

Assumption 51 (Random t-XOR Assumption (Daniely, 2016)) *There exist constants $\eta \in (0, 1/2)$ and $\gamma > 0$ such that for any $s > 0$, there is no $\mathrm{poly}(m)$-time algorithm that refutes random t-XorSat with m clauses for any sufficiently large $n \in \mathbb{Z}_{>0}$, $m = \lfloor n^{\gamma t} \rfloor$, and $t = \lfloor \log^s(n) \rfloor$.*

Theorem 52 *Under Theorem 51, for some fixed $\alpha > 0$, any $C > 0$, $c \in (0,1)$, and $\varepsilon > 0$, there is no algorithm that solves (α, k^{-C})-weak agnostic learning for Σ_L over Δ_k for every sufficiently large $k \in \mathbb{Z}_{>0}$ and $L := \exp(\log^c k)$ with success probability at least $3/4$ and running time $k^{O(\log^{1-\varepsilon} L)}$.*

Proof [Proof of Theorem 49] Theorem 49 follows immediately by combining Theorem 52 and Theorem 50. ■

In preparation for proving Theorem 52, we prove a few preliminary results. Given a set of clauses $c = \{c_i\}_{i\in[m]}$, define the function:

$$q(c) = \max_{z\in Z} \frac{1}{m} \sum_{i\in[m]} (c_i \cdot z)^2$$

The following lemma is implicit in (Daniely, 2016)

Lemma 53 *There exists a constant a_1 such that*

$$\Pr_{c\leftarrow C^m}[q(c) \le a_1 t \log(t)] \ge 1 - o_m(1)$$

where the $o_m(1)$ is exponentially small in m.

Proof Fix an assignment $z \in Z$. We view choosing $c_i \leftarrow C$ and first choosing a subset $T_i \subseteq [n]$ of variables, and then choosing their polarities $p_i \in \{\pm 1\}^t$.

For every z and T, by a Chernoff bound (over the choice of p), there exists a constant a_2 so that

$$Pr_{p_i\leftarrow\{\pm1\}^t}[|c_i \cdot z| \ge a_2\sqrt{t\log(1/\delta)}] \le \delta.$$

By a Chernoff bound over the choice of T_i (Motwani and Raghavan, 1995, Theorem 4.1), we have that with probability $\exp(-a_3\delta m)$, the condition

$$|c_i \cdot z| \le a_2\sqrt{t\log(1/\delta)}$$

holds for $m(1-2\delta)$ clauses. For such a z, we can bound

$$\begin{aligned}\frac{1}{m}\sum_{i\in[m]}((c_i\cdot z)^2 - t) &\le \frac{1}{m}\sum_{i=1}^{m}|c_i\cdot z|^2 \\ &\le (1-2\delta)a_2 t\log(1/\delta) + 2\delta t^2 \le a_1 t\log(t)\end{aligned}$$

where we choose $\delta = \log(t)/t$.

By a union bound over all 2^n choices of z, this holds for every z, and hence for $q(c)$ with probability $2^n\exp(-a_2\delta m)$, which is exponentially small once $m \ge a_3 nt$. ■

The next lemma is also proved in (Daniely, 2016). We use a different technique based on semi-definite programming and Grothendieck's inequality, which is more along the lines of the reduction in Fiege's work (Feige, 2002).

Lemma 54 *There is an algorithm that accepts all $c \in C^m$ such that $q(c) \ge a_1 t\log(t)$ and rejects instances such that $q(c) \le 2a_1 t\log(t)$.*

Proof We can write

$$q(c) - t = \max_{z \in Z} \frac{1}{m} \sum_{i=1}^{m} ((c_i \cdot z)^2 - t)$$
$$= \max_{z \in Z} \frac{1}{m} \sum_{i \in [m]} \sum_{j \neq j' \in c_i} z_j z_{j'}$$

We consider the semi-definite relaxation over v_i which are unit vectors in a high-dimensional space (with the constraint that the vectors assigned to a literal and its negation sum to 0).

$$\tilde{q}(c) = \max_{\|v_j\|_2^2 = 1} \frac{1}{m} \sum_{i \in [m]} \sum_{j \neq j' \in c_i} v_j v_{j'}$$

We solve the semi-definite program efficiently (up to small additive error which we will ignore) and accept instances where

$$\tilde{q}(c) > K_G(a_1 t \log(t) - t).$$

Grothendieck's inequality implies that the integrality gap of this relaxation is a constant; there exist $K_G \in [1.5, 2]$ such that

$$q(c) - t \leq \tilde{q}(c) \leq K_G(q(c) - t). \tag{10}$$

For such instances, Equation (10) implies that $q(c) > a_1 t \log(t)$ since

$$q(c) - t \geq \frac{\tilde{q}(c)}{K_G} > a_1 t \log(t) - t.$$

For instances that we reject, it holds that

$$q(c) - t \leq \tilde{q}(c) < K_G(a_1 t \log(t) - t)$$

Since $K_G \in [1.5, 2]$, we have $q(c) \leq 2a_1 t \log(t)$. ■

We refer to instances rejected by the algorithm as *pseudorandom*. By Markov's inequality applied to the definition of $q(c)$, we have the following claim:

Lemma 55 *Given pseudorandom $c \in C^m$ and $z \in Z$, for every $\delta > 0$, there are at most δm clauses such that $|c_i \cdot z| \geq \sqrt{a_1 t \log(t)/\delta}$.*

We also have the following lemma which we state without proof

Lemma 56 *Every functions $g : \{-d, \ldots, d\} \to \{\pm 1\}$ can be written as a polynomial in x of degree $2d$ with coefficients bounded by $\exp(d \log(d))$.*

We now complete the proof of Theorem 52.

Proof Let $\eta \in (0, 1/2)$ be the constant guaranteed to exist by Theorem 51. We define $\alpha = 1/4 - \eta/2 \in (0, 1/4)$. We fix an arbitrary constant $\varepsilon \in (0, 1/3)$. Throughout the proof, we will treat η, α, and ε as fixed constants (that can hide in big-O notations). Consider an algorithm A for (α, k^{-C})-weak agnostic learning for Σ_L over Δ_k with running time $k^{(\log(L))^{1-\varepsilon}}$ and success

probability at least $3/4$ for any sufficiently large k and $L := \exp(\log^c k)$ for some $c \in (0,1)$. It suffices to use A to efficiently refute random t-XorSat with parameter η for any sufficiently large n and $t = \lfloor \log^s(n) \rfloor$ in time $n^{o(t)}$, where $s = 2c/(1-c+\varepsilon) > 0$.

We view the clauses in a t-XorSat problem as a distribution over $C \subseteq \{0,1\}^{2n}$, with the c_is being points and b_i their labels. Let $d = \Theta(\sqrt{(t \log t)/\alpha}) = \Theta(\sqrt{t \log t})$ be such that at most αm clauses fail to satisfy $|c_i \cdot z| \leq d/2$. We consider the low degree feature expansion of C denoted $C^{\otimes d}$ which contains a monomial $\prod_{j \in T} c_i^{(j)}$ for every $T \subseteq [2n]$ of size at most d, so that $C^{\otimes d} \subseteq \{0,1\}^k$ for

$$k = \binom{2n}{\leq d} := \sum_{j=0}^{d} \binom{2n}{j}. \tag{11}$$

Since every $c \in C$ has weight exactly t, every $c^{\otimes d} \in C^{\otimes d}$ has weight $B_1 = \binom{t}{\leq d} = \exp(O(d \log(d)))$. This lets us write $c_i^{\otimes d} = B_1 \mathbf{v}_i$ where $\mathbf{v}_i \in \Delta_k$. Thus an instance I which gives a distribution on (c_i, b_i) where $c_i \in C$ $b_i \in \{\pm 1\}$ also gives a distribution over $(\mathbf{v}_i, b_i) \in \Delta_k \times \{\pm 1\}$.

There exists a degree d polynomial

$$p(t) = \sum_{j=0}^{d} \alpha_j t^j$$

such that $|\alpha_j| = \exp(O(d \log(d)))$ and $p(c_i \cdot z) = \mathrm{Xor}_{c_i}(z)$ for $|c_i \cdot z| \leq d/2$. If $|c_i \cdot z| \geq d/2$ then $p(c_i \cdot z) \in \mathbb{R}$. Since $c_i \in \{0,1\}^n$, we can multilinearize the terms of the form $(c_i \cdot z)^j$ as

$$(c_i \cdot z)^j = \sum_{T \subseteq [n], |T| \leq j} w_T^j \prod_{i \in T} c_i$$

for coefficients $w_T^j = \exp(O(j \log j))$. So we can write

$$\begin{aligned}
p(c_i \cdot z) &= \sum_{j=0}^{2d} \alpha_j (c_i \cdot z)^j \\
&= \sum_{j=0}^{2d} \alpha_j \sum_{T \subseteq [n], |T| \leq j} w_T^j \prod_{i \in T} c_i \\
&= \sum_{T \subseteq [n], |T| \leq 2d} \left(\sum_{j \geq |T|} \alpha_j w_T^j \right) \prod_{i \in T} c_i \\
&= \sum_{T \subseteq [n], |T| \leq 2d} w'_T \prod_{i \in T} c_i \\
&= w' \cdot c^{\otimes d}
\end{aligned}$$

for coefficients w'_T bounded in absolute value by $|w'_T| = \exp(O(d \log d))$.

We renormalize w' to be bounded in $[-1,1]^k$. We write $w' = wB_1$ for $B_1 = \max_T |w'_T| = \exp(O(d \log d)$. We have

$$P(c_i \cdot z) = w' \cdot c_i^{\otimes d} = B_1 \binom{t}{d} w \cdot \mathbf{v}_i.$$

Hence if $|c_i \cdot z| \leq d/2$, then the quantity above equals $\mathrm{Xor}_{c_i}(z) \in \{\pm 1\}$, else it takes on values in $\mathbb{R}$.

By our choice of $d = \Theta(\sqrt{t \log t})$, we have

$$\log\left(B_1 \binom{t}{d}\right) = \log B_1 + O(\log(d \log t)) = O(d \log d) \leq t^{1/2+o(1)}.$$

By our choice of $L := \exp(\log^c k)$, we have

$$\begin{aligned}
\log L &= (\log k)^c \\
&= (d \log n)^{c+o(1)} && \text{(by (11))} \\
&= d^{c+o(1)} (\log n)^{c+o(1)} \\
&= t^{c/2+o(1)} (\log n)^{s(1-c+\varepsilon)/2+o(1)} && \text{(by } d = \Theta(\sqrt{t \log t}) \text{ and } s = 2c/(1-c+\varepsilon)) \\
&= t^{c/2+o(1)} t^{(1-c+\varepsilon)/2+o(1)} && \text{(by } t = \lfloor \log^s n \rfloor) \\
&= t^{1/2+\varepsilon/2+o(1)}. && (12)
\end{aligned}$$

Therefore, for sufficiently large n,

$$L \geq a B_1 \binom{t}{d},$$

for some constant a so that $\tanh(a) \geq 1 - \alpha$, where we use our choice of constant $\alpha := 1/4 - \eta/2 > 0$.

Consider the function $g(\mathbf{v}) = \tanh(L' w \cdot \mathbf{v}) \in \Sigma_L$. We can find $a' \geq a$ such that $L = a' B_1 \binom{t}{d}$. We have for each $i \in [m]$,

$$\begin{aligned}
g(\mathbf{v}_i) &= \tanh(Lw \cdot \mathbf{v}_i) \\
&= \tanh\left(a' B_1 w \cdot \begin{pmatrix} t \\ \leq d \end{pmatrix} \mathbf{v}_i\right) \\
&= \tanh(a' w' \cdot c_i^{\otimes d}) \\
&= \tanh(a' p(c_i \cdot z)).
\end{aligned}$$

Therefore, $g(\mathbf{v}_i) \geq 1 - \alpha$ if $p(c_i \cdot z) = 1$, and $g(\mathbf{v}_i) \leq -1 + \alpha$ if $p(c_i \cdot z) = -1$. Moreover, it is clear that $g(\mathbf{v}_i) \in [-1, 1]$ always holds.

Recall our definition $\alpha := 1/4 - \eta/2 > 0$. If $\mathrm{val}(I) \geq 1 - \eta$, then by taking the function g derived from z such that $\mathrm{val}(z, I) \geq 1 - \eta = 1/2 + 2\alpha$, excluding the at most αm clauses that fail to satisfy $|c_i \cdot z| \leq d/2$, we get $g \in \Sigma_L$ such that

$$\frac{1}{m} \sum_{i=1}^{m} g(\mathbf{v}_i) b_i \geq (1/2 + \alpha) \times (1 - \alpha) + (1/2 - \alpha) \times (-1) \geq \alpha.$$

Thus based on the methodology of (Daniely et al., 2014) (see e.g. Theorem 2.1 in (Daniely, 2016)), we can apply our weak agnostic learning algorithm A to efficiently distinguish the case with $\mathrm{val}(I) \geq 1 - \eta$ and the case with uniformly random clauses with success probability at least $3/4$, solving the t-XorSat refutation problem.

As long as the running time of algorithm A is bounded by $k^{\log(L)^{1-\varepsilon}}$ for some $\varepsilon > 0$, the running time for t-XorSat refutation is bounded by

$$k^{O(\log(L)^{1-\varepsilon})} \leq n^{O(d\log(L)^{1-\varepsilon})}.$$

By (12), we have

$$d\log(L)^{1-\varepsilon} = t^{1/2+o(1)}t^{(1-\varepsilon)(1/2+\varepsilon/2+o(1))} = t^{1-\varepsilon^2/2+o(1)}.$$

Thus the running time for t-XorSat refutation is bounded by $n^{o(t)}$, as desired. ■

Appendix H. Proofs from Section E

H.1. Proof of Theorem 32

We break the proof in a sequence of lemmas, starting with simplifying the objective function.

Lemma 57 *Let $w_0 = \mathbf{E}_{\mathcal{D}}[z\varphi_{\mathbf{v}}]$. Then $w_0 \in B_\Gamma(s)$ and for any $w \in \Gamma$ we have*

$$\mathop{\mathbf{E}}_{(\mathbf{v},z)\sim\mathcal{D}}[w(\mathbf{v})z] = \langle w, w_0\rangle_\Gamma. \tag{13}$$

Proof For any $w \in \Gamma$ we can write the correlation objective as

$$\mathop{\mathbf{E}}_{(\mathbf{v},z)\sim\mathcal{D}}[w(\mathbf{v})z] = \mathbf{E}[\langle w, \varphi_{\mathbf{v}}\rangle_\Gamma z] = \langle w, \mathbf{E}[z\varphi_{\mathbf{v}}]\rangle_\Gamma = \langle w, w_0\rangle_\Gamma.$$

To bound its norm, observe that

$$\|w_0\|_\Gamma = \|\mathop{\mathbf{E}}_{\mathcal{D}}[z\varphi_{\mathbf{v}}]\|_\Gamma \leq \max_{\mathbf{v}\in\Delta_k, z\in\{\pm1\}} \|z\varphi_{\mathbf{v}}\|_\Gamma \leq s.$$

■

Next we show that we can approximate w_0 uniformly from samples by the function

$$\tilde{w}_0 = \frac{1}{n}\sum_{i=1}^n z_i\varphi_{\mathbf{v}_i}.$$

Lemma 58 *For any $\delta \in (0, 1/2)$, for some $n_0 = O(r^2s^2\alpha^{-2}\log(1/\delta))$, for any $n \geq n_0$ and any $(\mathbf{v}_1, z_1), \ldots, (\mathbf{v}_n, z_n)$ drawn i.i.d. from $\mathcal{D}$, with probability at least $1-\delta$,*

$$\|\tilde{w}_0 - w_0\|_\Gamma \leq \frac{\alpha}{3r}. \tag{14}$$

The proof uses McDiarmid's inequality (see e.g. Lemma 26.4 of (Shalev-Shwartz and Ben-David, 2014)).

Proof We can write

$$\|\tilde{w}_0 - w_0\|_\Gamma = \left\|\sum_{i=1}^n \frac{z_i\varphi_{\mathbf{v}_i}}{n} - w_0\right\|_\Gamma = \frac{1}{n}\left\|\sum_{i=1}^n (z_i\varphi_{\mathbf{v}_i} - w_0)\right\|_\Gamma$$

Since each term $z_i\varphi_{\mathbf{v}_i} - w_0$ has expectation 0, and the terms are independent, for $i \neq j$

$$\mathbf{E}[\langle z_i\varphi_{\mathbf{v}_i} - w_0, z_j\varphi_{\mathbf{v}_j} - w_0\rangle_\Gamma] = 0$$

Hence we can bound

$$\begin{aligned}\mathbf{E}[\|\tilde{w}_0 - w_0\|_\Gamma^2] &= \frac{1}{n^2}\,\mathbf{E}\left[\sum_{i=1}^n \|z_i\varphi_{\mathbf{v}_i} - w_0\|_\Gamma^2 + \sum_{i\neq j}\langle z_i\varphi_{\mathbf{v}_i} - w_0, \mathbf{z}_j\varphi_{\mathbf{v}_j} - w_0\rangle\right] \\ &= \frac{1}{n}\,\mathbf{E}[\|z_1\varphi_{\mathbf{v}_1} - w_0\|_\Gamma^2] \\ &\leq 4s^2/n \leq (\alpha/(6r))^2.\end{aligned}$$

by our choice of n. By convexity,

$$\mathbf{E}[\|\tilde{w}_0 - w_0\|_\Gamma] \leq \frac{\alpha}{6r}. \tag{15}$$

Note that each i.i.d. term $z_i\varphi_{\mathbf{v}_i}$ in the definition of $\tilde{w}_0$ has norm $\|z_i\varphi_{\mathbf{v}_i}\|_\Gamma \leq s$, so by McDiarmid's inequality, with probability at least $1-\delta$,

$$\left|\,\|\tilde{w}_0 - w_0\|_\Gamma - \mathbf{E}[\|\tilde{w}_0 - w_0\|_\Gamma]\right| \leq \alpha/(6r). \tag{16}$$

Combining this with Equation (15) gives the desired claim. ■

We need the following simple helper lemma to finish proving Theorem 32:

Lemma 59 *Let $w, \tilde{w}$ be elements of a Hilbert space Γ. If $\tilde{w} \neq 0$, define $\bar{w} = \tilde{w}/\|\tilde{w}\|_\Gamma$. If $\tilde{w} = 0$, define $\bar{w}$ to be an arbitrary element of $B_\Gamma(1)$. Then*

$$\langle w, \bar{w}\rangle \geq \|\tilde{w}\|_\Gamma - \|w - \tilde{w}\|_\Gamma \geq \|w\|_\Gamma - 2\|w - \tilde{w}\|_\Gamma.$$

Proof We have

$$\langle w, \bar{w}\rangle \geq \langle \tilde{w}, \bar{w}\rangle - \|w - \tilde{w}\|_\Gamma = \|\tilde{w}\|_\Gamma - \|w - \tilde{w}\|_\Gamma \geq \|w\|_\Gamma - 2\|w - \tilde{w}\|_\Gamma.$$

■

Proof [Proof of Theorem 32] In the weak agnostic learning task, we assume that there exists $w \in B_\Gamma(r)$ so that

$$\mathop{\mathbf{E}}_{(\mathbf{v},\mathbf{z})\sim\mathcal{D}}[w(\mathbf{v})z] \geq \alpha.$$

Under this assumption, Theorem 57 tells us that $\langle w_0, w\rangle_\Gamma \geq \alpha$. Since $w \in B_\Gamma(r)$, we have $\|w\|_\Gamma \leq r$, and by the Cauchy-Schwarz inequality, $r\,\|w_0\|_\Gamma \geq \|w_0\|_\Gamma\|w\|_\Gamma \geq \langle w_0, w\rangle_\Gamma$. Therefore, we can assume that $\|w_0\|_\Gamma \geq \alpha/r$ in the weak agnostic learning task.

Theorem 58 ensures that (14) holds with probability at least $1-\delta$. As long as (14) holds, by Theorem 59 we have

$$\left\langle w_0, \frac{\tilde{w}_0}{\|\tilde{w}_0\|_\Gamma}\right\rangle \geq \|w_0\|_\Gamma - 2\|w_0 - \tilde{w}\|_\Gamma \geq \frac{\alpha}{3r}. \tag{17}$$

The output w_2 of Algorithm 1 can be expressed as

$$w_2 = \frac{\tilde{w}_0}{s \, \|\tilde{w}_0\|_\Gamma}. \tag{18}$$

Combining (17) and (18), we know that with probability at least $1 - \delta$,

$$\langle w_0, w_2 \rangle \geq \frac{\alpha}{3rs}.$$

By Theorem 57, the inequality above implies the weak learning guarantee, namely, $\mathbf{E}[w_2(\mathbf{v})z] \geq \alpha/(3rs)$. Finally, it is clear that $\|w_2\|_\Gamma \leq 1/s$, so $w_2 \in B_\Gamma(1/s)$, as desired. ■

H.2. Proof of Theorem 33

Consider a distribution $\mathcal{D}$ of $(\mathbf{v}, \mathbf{y}) \in \Delta_k \times \mathcal{E}_k$, and define $\mathbf{z} := \mathbf{y} - \mathbf{v}$. Define $w_0^{(j)} := \mathbf{E}[\mathbf{z}^{(j)} \varphi_{\mathbf{v}}]$, where $\mathbf{z}^{(j)}$ is the j-th coordinate of $\mathbf{z}$. For n i.i.d. data points $(\mathbf{v}_1, \mathbf{y}_1), \ldots, (\mathbf{v}_n, \mathbf{y}_n)$, define $\mathbf{z}_i := \mathbf{y}_i - \mathbf{v}_i$. Define $\tilde{w}_0^{(j)} := \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i^{(j)} \varphi_{\mathbf{v}_i}$.

Lemma 60 *When $n \geq Ckr^2 s^2 \varepsilon^{-2} \log(1/\delta)$, with probability at least $1 - \delta$,*

$$\sum_{j=1}^k \|\tilde{w}_0^{(j)} - w_0^{(j)}\|_\Gamma \leq \alpha/(3r). \tag{19}$$

Proof We first show that

$$\sum_{j=1}^k \mathbf{E} \, \|\tilde{w}_0^{(j)} - w_0^{(j)}\|_\Gamma \leq \alpha/(6r).$$

For every j,

$$\begin{aligned}
\mathbf{E}[\|\tilde{w}_0^{(j)} - w_0^{(j)}\|_\Gamma^2] &= \mathbf{E}\left[\left\| \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i^{(j)} \varphi_{\mathbf{v}_i} - w_0^{(j)} \right\|_\Gamma^2\right] \\
&= \frac{1}{n^2} \sum_{i=1}^n \mathbf{E} \, \|\mathbf{z}_i^{(j)} \varphi_{\mathbf{v}_i} - w_0^{(j)}\|_\Gamma^2 + \frac{1}{n^2} \sum_{i \neq i'} \langle \mathbf{z}_i^{(j)} \varphi_{\mathbf{v}_i} - w_0^{(j)}, \mathbf{z}_{i'}^{(j)} \varphi_{\mathbf{v}_{i'}} - w_0^{(j)} \rangle \\
&= \frac{1}{n^2} \sum_{i=1}^n \mathbf{E} \, \|\mathbf{z}_i^{(j)} \varphi_{\mathbf{v}_i} - w_0^{(j)}\|_\Gamma^2 \\
&= \frac{1}{n} \mathbf{E}[\|\mathbf{z}_1^{(j)} \varphi_{\mathbf{v}_1} - w_0^{(j)}\|_\Gamma^2] \\
&\leq \frac{1}{n} \mathbf{E}[\|\mathbf{z}_1^{(j)} \varphi_{\mathbf{v}_1}\|_\Gamma^2] \\
&\leq \frac{s^2}{n} \mathbf{E}[(\mathbf{z}_1^{(j)})^2]
\end{aligned}$$

By Cauchy-Schwarz,

$$\sum_{j=1}^k \sqrt{\mathbf{E}[(\mathbf{z}_1^{(j)})^2]} \leq \sqrt{k \sum_{j=1}^k \mathbf{E}[(\mathbf{z}_1^{(j)})^2]} \leq \sqrt{k \sum_{j=1}^k \mathbf{E} \, |\mathbf{z}_1^{(j)}|} \leq \sqrt{2k}.$$

Therefore,

$$\sum_{j=1}^{k} \mathbf{E}\,\|\tilde{w}_0^{(j)} - w_0^{(j)}\|_\Gamma \le \sum_{j=1}^{k} \sqrt{\mathbf{E}[\|\tilde{w}_0^{(j)} - w_0^{(j)}\|_\Gamma^2]} \le \frac{s\sqrt{2k}}{\sqrt{n}} \le \alpha/(6r).$$

Finally, we apply McDiarmid's inequality to the following function of $(\mathbf{z}_1, \mathbf{v}_1), \ldots, (\mathbf{z}_n, \mathbf{v}_n)$:

$$\sum_{j=1}^{k} \|\tilde{w}_0^{(j)} - w_0^{(j)}\|_\Gamma = \sum_{j=1}^{k} \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{z}_i^{(j)} \varphi_{\mathbf{v}_i} - w_0^{(j)} \right\|_\Gamma$$

and get that with probability at least $1 - \delta$,

$$\sum_{j=1}^{k} \|\tilde{w}_0^{(j)} - w_0^{(j)}\|_\Gamma \le \sum_{j=1}^{k} \mathbf{E}\,\|\tilde{w}_0^{(j)} - w_0^{(j)}\|_\Gamma + \alpha/(6r) \le \alpha/(3r).$$

■

Proof [Proof of Theorem 33] In the auditing task, we assume that there exists $w \in B_\Gamma(r)^k$ such that

$$\mathbf{E}[\langle \mathbf{y} - \mathbf{v}, w(\mathbf{v}) \rangle] \ge \alpha.$$

Using our definition of $\mathbf{z} := \mathbf{y} - \mathbf{v}$ and $w_0^{(\ell)} := \mathbf{E}[\mathbf{z}^{(\ell)} \varphi_{\mathbf{v}}]$, by Theorem 57 we have

$$\sum_{\ell=1}^{k} \|w_0^{(\ell)}\|_\Gamma \ge \frac{1}{r} \sum_{\ell=1}^{k} \langle w_0^{(\ell)}, w^{(\ell)} \rangle_\Gamma = \frac{1}{r} \sum_{\ell=1}^{k} \mathbf{E}[\mathbf{z}^{(j)} w^{(j)}(\mathbf{v})] = \frac{1}{r}\,\mathbf{E}[\langle \mathbf{y} - \mathbf{v}, w(\mathbf{v}) \rangle] \ge \alpha/r.$$

Theorem 60 ensures that (19) holds with probability at least $1 - \delta$. Define $\bar{w}_0 := \tilde{w}_0 / \|\tilde{w}_0\|_\Gamma$ if $\tilde{w}_0 \neq 0$, and define $\bar{w}_0 := 0$ if $\tilde{w}_0 = 0$. As long as (19) holds, by Theorem 59,

$$\sum_{\ell=1}^{k} \langle w_0^{(\ell)}, \bar{w}_0^{(\ell)} \rangle \ge \sum_{\ell=1}^{k} \|w_0^{(\ell)}\|_\Gamma - 2 \sum_{\ell=1}^{k} \|\tilde{w}_0^{(\ell)} - w_0^{(\ell)}\|_\Gamma \ge \alpha/(3r).$$

In Algorithm 2, we have $w_2^{(\ell)} = \bar{w}_0^{(\ell)}/s$, and thus the inequality above implies

$$\sum_{\ell=1}^{k} \langle w_0^{(\ell)}, w_2^{(\ell)} \rangle \ge \alpha/(3rs).$$

Therefore by Theorem 57, with probability at least $1 - \delta$,

$$\mathbf{E}[\langle \mathbf{y} - \mathbf{v}, w_2(\mathbf{v}) \rangle] = \sum_{\ell=1}^{k} \mathbf{E}[\mathbf{z}^{(\ell)} w_2^{(\ell)}(\mathbf{v})] = \sum_{\ell=1}^{k} \langle w_0^{(\ell)}, w_2^{(\ell)} \rangle \ge \alpha/(3rs).$$

This proves that the output w_2 of Algorithm 2 satisfies the requirement of the auditing task. Finally, it is clear that each $w_2^{(\ell)}$ has norm $\|w_2^{(\ell)}\|_\Gamma \le 1/s$, so $w_2 \in B_\Gamma(1/s)^k$, as desired. ■