Proceedings of Machine Learning Research vol 336:1–36, 2026 39th Annual Conference on Learning Theory

Distribution-Free Sequential Prediction with Abstentions

Jialin Yu YUJL1@GATECH.EDU and **Moïse Blanchard** MBLANCHARD41@GATECH.EDU
Georgia Institute of Technology, Atlanta, GA, USA

Editors: Steve Hanneke and Tor Lattimore

Abstract

We study a sequential prediction problem in which an adversary is allowed to inject arbitrarily many adversarial instances in a stream of i.i.d. instances, but at each round, the learner may also *abstain* from making a prediction without incurring any penalty if the instance was indeed corrupted. This semi-adversarial setting naturally sits between the classical stochastic case with i.i.d. instances for which function classes with finite VC dimension are learnable; and the adversarial case with arbitrary instances, known to be significantly more restrictive. For this problem, Goel et al. (2023) showed that, if the learner knows the distribution μ of clean samples in advance, learning can be achieved for all VC classes without restrictions on adversary corruptions. This is, however, a strong assumption in both theory and practice: a natural question is whether similar learning guarantees can be achieved without prior distributional knowledge, as is standard in classical learning frameworks (e.g., PAC learning or asymptotic consistency) and other non-i.i.d. models (e.g., smoothed online learning). We therefore focus on the distribution-free setting where μ is *unknown* and propose an algorithm ABSTAINBOOST based on a boosting procedure of weak learners, which guarantees sublinear error for general VC classes in *distribution-free* abstention learning for oblivious adversaries. These algorithms also enjoy similar guarantees for adaptive adversaries, for structured function classes including linear classifiers. These results are complemented with corresponding lower bounds, which reveal an interesting polynomial trade-off between misclassification error and number of erroneous abstentions.

Keywords: Statistical learning theory, Online learning, Abstention learning, Distribution-free, Adaptive and oblivious adversaries

1. Introduction

We study the classical online prediction problem in which a learner sequentially observes an instance $x_t \in \mathcal{X}$ and makes a prediction about a value y_t before observing the true label. The learner's objective is to minimize the error of its predictions $\hat{y}_t$ compared to the true value y_t. The problem has been extensively studied under two extreme scenarios: the *stochastic* setting where instances $x_1, \ldots, x_T$ are assumed to be independently and identically distributed (i.i.d.) usually from an unknown distribution μ to the learner, and the general *adversarial* setting where no assumptions are made on the instance generation process. In the stochastic setting, also known as PAC learning (Vapnik and Chervonenkis, 1974; Valiant, 1984), simple algorithms can provably learn the best prediction function within function classes $\mathcal{F}$ of finite Vapnik-Chervonenkis (VC) dimension (Vapnik and Chervonenkis, 1971, 1974; Valiant, 1984), a combinatorial measure of complexity that is well controlled for many practical hypothesis classes, including linear classifiers. On the other hand, learning is severely limited for adversarial data. Specifically, learning in the adversarial setting requires the function class to have finite so-called Littlestone dimension (Littlestone, 1988; Ben-David et al., 2009), which does not hold even for threshold function classes in $[0, 1]$.[1]

1. The threshold function class is defined as $\mathcal{F} = \{x \in [0, 1] \mapsto \mathbb{1}[x \leq x_0] : x_0 \in [0, 1]\}$.

This significant gap has sparked substantial interest in intermediate learning models which aim to relax the i.i.d. assumption while preserving strong learning guarantees. This includes distributionally-constrained adversaries (Rakhlin et al., 2011; Blanchard and Kpotufe, 2025) and specifically the smooth adversarial model which imposes density ratio constraints on data distributions (Rakhlin et al., 2011; Haghtalab et al., 2024; Block et al., 2022). As an alternative model for non-i.i.d. data, another line of work focused on algorithms robust to *adversarial corruptions* in which an adversary may corrupt training (Valiant, 1985; Kearns and Li, 1993; Bshouty et al., 2002; Awasthi et al., 2017; Gao et al., 2021; Hanneke et al., 2022; Balcan et al., 2022; Shafahi et al., 2018; Blum et al., 2021) or test data (Feige et al., 2018; Attias et al., 2019; Montasser et al., 2019, 2020b, 2021, 2022, 2020a). A central challenge with adversarial corruptions is that achieving strong predictive performance typically requires imposing substantial restrictions on the adversary, such as limiting corruptions to a small fraction of the data or to specific classes of perturbations. To circumvent these limitations, allowing the learner to *abstain* on difficult or out-of-training instances has emerged as a fruitful workaround (Chow, 1970; Bartlett and Wegkamp, 2008; Goldwasser et al., 2020; Kalai and Kanade, 2021; Cortes et al., 2016b). In practice, an abstention from the machine learning model could trigger a conservative default policy, for instance in the context of content moderation systems, or prompt human review/intervention in medical diagnoses or autonomous driving applications. In high-risk applications such as these, a cautious abstention is often significantly more desirable and cost-effective than a prediction error, and enables the model to selectively predict on instances for which it has high confidence—an approach sometimes referred to as "reliable learning" (Kanade and Thaler, 2014; Kalai and Kanade, 2021).

Distribution-free sequential learning with abstentions. We consider the sequential prediction model with abstentions introduced by Goel et al. (2023) in which an adversary may inject arbitrary corruptions within the clean sequence of i.i.d. instances from a distribution μ. In particular, the adversary may corrupt half (or all) of the data instances $x_1, \ldots, x_T$. On the other hand, the learner may abstain on instances which they suspect may be corrupted. In turn, when the learner abstains, it may either receive as feedback an optimal response policy (Goel et al., 2023) or not (Cortes et al., 2016a,b, 2018), depending on the model. For simplicity, we focus on the *realizable binary classification* case where all instances are labeled by a fixed in-class function $f^\star : \mathcal{X} \to \{0, 1\}$. In other terms, we focus on a *clean-label* model (Shafahi et al., 2018; Blum et al., 2021), where corrupted instances are still correctly labeled. Nevertheless, in general, providing clean labels for corrupted samples doesn't always help learning (Larsen et al., 2026). In this context, the learner aims to always make correct predictions whenever it does not abstain, thereby ensuring reliability, while simultaneously avoiding abstentions on clean instances. Crucially, the learner incurs no penalty for abstaining on corrupted instances; in light of the fundamental limitations for adversarial learning even for realizable data (Littlestone, 1988), this is essentially necessary to achieve meaningful learning guarantees when the data may contain a large fraction of corrupted examples.

In this abstention sequential learning framework, Goel et al. (2023) propose learning algorithms that guarantee both low misclassification error and few abstentions on clean instances for any function class $\mathcal{F}$ with finite VC dimension, under the core assumption that the clean sample distribution μ is *known* to the learner a priori. While these results provide strong theoretical evidence for the benefit of abstentions in the presence of adversarial attacks, the knowledge of μ is an important practical limitation. Indeed, assuming access to the underlying data distribution significantly departs from classical results in the PAC learning and consistency literature, in which simple algorithms such as Empirical Risk Minimization (ERM) or nearest-neighbor-based algorithms achieve desired learn-

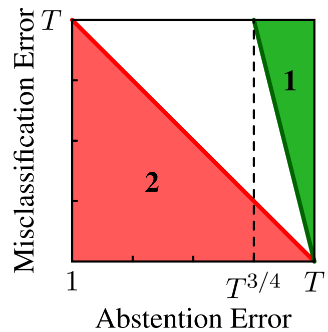

Figure 1: Tradeoffs between misclassification error and abstention error for either (1) oblivious adversaries or (2) adaptive adversaries and function classes with finite reduction dimension (see Definition 6), including linear classifiers. The green region **1** is achievable (Theorems 3 and 7), while the red region **2** is not (Theorem 4). The plot displayed in log-scale.

ing guarantees without any prior distributional knowledge. Accordingly, recent work has aimed to remove distributional assumptions in other beyond-i.i.d. learning models, including smooth adversaries (Block et al., 2024; Blanchard, 2025). For the abstention learning problem itself, Goel et al. (2023) show as a proof of concept that prior knowledge of the distribution μ is not required for certain function classes: namely axis-aligned boxes and those with VC dimension 1. This naturally leads to the following fundamental question for abstention sequential learning:

Is learning with abstentions possible for general function classes (with finite VC dimension) without prior knowledge of the data distribution μ?

Our contributions. Throughout, we refer to the number of erroneous abstentions (on non-corrupted instances) as the *abstention error*. Our results reveal a trade-off between the misclassification error and the abstention error of learning algorithms. Towards characterizing the achievable misclassification/abstention error landscape for distribution-free sequential prediction with abstentions, we make the following contributions. Our key findings are summarized in Fig. 1.

We propose a family of algorithms ABSTAINBOOST for $\alpha \in [0, 1/4]$ which guarantee $\mathcal{O}(T^{4\alpha})$ misclassification error and $\tilde{\mathcal{O}}(d^2 T^{1-\alpha})$ abstention error for any function class with VC dimension d, against *oblivious adversaries*.

Additionally, we show that ABSTAINBOOST also achieves similar learning guarantees against *adaptive adversaries* when the function class has certain additional structure (finite so-called reduction-dimension, see Definition 6). For instance, for linear classifiers in $\mathbb{R}^p$, ABSTAINBOOST achieves $\tilde{\mathcal{O}}(T^{4\alpha})$ misclassification error and $\tilde{\mathcal{O}}(p^{6.5} T^{1-\alpha})$ abstention error. This also recovers learnability against adaptive adversaries for VC-1 classes and the class of axis-aligned rectangles, provided by Goel et al. (2023) with tailored algorithms, albeit with worse learning guarantees.

Last, we complement these results with an oblivious lower bound tradeoff between misclassification and abstention. Specifically, we show that for any $\alpha \in [0, 1]$, learning algorithms must make either $\Omega(T^\alpha)$ misclassification error or $\Omega(T^{1-\alpha})$ abstention error against some oblivious adversary.

2. Preliminaries

Abstention learning setup. Throughout this paper, we denote by $\mathcal{X}$ the instance domain and let $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ be a class of measurable functions. We consider the sequential online learning problem with adversarial corruptions, where the learner may additionally abstain as introduced in Goel et al. (2023). Formally, the learning protocol proceeds as follows. Initially, an adversary fixes a distribution μ over $\mathcal{X}$ together with labeling function $f^\star \in \mathcal{F}$. The learner does not observe $f^\star$ nor μ. Then, at each iteration $t \in [T]$:

1. The adversary first decides whether or not to corrupt this round by selecting $c_t \in \{0,1\}$.
2. If $c_t = 1$, the adversary selects an arbitrary instance $x_t \in \mathcal{X}$. Otherwise, if $c_t = 0$, the new instance is sampled $x_t \sim \mu$ independently from the past.
3. The learner observes x_t and outputs $\hat{y}_t \in \{0,1,\perp\}$, where $\perp$ denotes an abstention.
4. Finally, the learner observes the (clean) label $y_t = f^\star(x_t)$.

We emphasize that this setup corresponds to a realizable setting: the values y_t are always consistent with the fixed function $f^\star \in \mathcal{F}$ even on corrupted rounds. If the learner makes a prediction $\hat{y}_t \in \{0,1\}$, it incurs a misclassification cost if it makes a mistake: $\hat{y}_t \neq y_t$. Conversely, if the learner abstains $\hat{y}_t = \perp$, it incurs an abstention cost if the round was not corrupted: $c_t = 0$. We therefore focus on the misclassification error and abstention error on all T iterations, defined by

$$\textsc{MisErr} := \sum_{t=1}^{T} \mathbb{1}[\hat{y}_t \notin \{y_t, \perp\}] \quad \text{and} \quad \textsc{AbsErr} := \sum_{t=1}^{T} \mathbb{1}[c_t = 0 \wedge \hat{y}_t = \perp] \,.$$

In words, the learner aims to make few mistakes on the rounds where it predicts to minimize its misclassification error, while not abstaining too often on non-adversarial rounds to minimize its abstention error. Importantly, by construction, the learner is allowed to abstain on adversarially injected inputs for free. This allows for an arbitrary number of corruption rounds (e.g., $T/2$ corruptions) without incurring a linear penalty in the objective, as opposed to the rejection model in Chow (1970), where there is a small but fixed cost on every abstention.

Finally, we note that the learner observes y_t even when it chooses to abstain. This is natural if we view the abstention action not merely as a null output, but as a trigger for expert consultation. One can also consider an alternative model where the learner only observes the label when it makes a prediction, which is studied under a different abstention learning framework that incurs cost for every abstention (Cortes et al., 2016a, 2018).

We consider two different adversarial models: *oblivious* and *adaptive* adversaries defined below.

Definition 1 (Oblivious and adaptive adversaries) *An adversary is* oblivious *if it fixes in advance corruption times $\{t : c_t = 1\}$ together with their corrupted instances $(x_t)_{t:c_t=1}$, without adapting to the learner's output or clean instance samples. An adversary is* adaptive *if its decision at time $t \in [T]$ may depend on all previous learner's outputs and clean instance samples.*

Complexity notion for the function class. In the sequential problem without abstentions—when the instances are i.i.d., known as the PAC learning setting—learnable function classes exactly correspond to those with finite VC dimension (Vapnik and Chervonenkis, 1971, 1974; Valiant, 1984).

Definition 2 (VC dimension) *Let $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ be a function class on $\mathcal{X}$. We say that $\mathcal{F}$* shatters *a set of points $x_1, ..., x_m \in \mathcal{X}$ if for any choice of labels $\epsilon \in \{0,1\}^m$ there exists $f \in \mathcal{F}$ such that $f(x_i) = \epsilon_i$ for all $i \in [m]$. The VC dimension of $\mathcal{F}$ is the size of the largest shattered set.*

Further notation. Next, we introduce the notion of *shattering probability*, which is a central quantity appearing in the algorithm developed by Goel et al. (2023) in the case when the distribution μ of uncorrupted samples is known. Formally, the k*-shattering probability* of a given function class $\mathcal{F}$ corresponds to the probability of k i.i.d. points sampled from μ being shattered:

$$\rho_k(\mathcal{F}, \mu) := \mathbb{P}_{x_1,\ldots,x_k \overset{iid}{\sim} \mu}[\{x_1, \ldots, x_k\} \text{ is shattered by } \mathcal{F}].$$

For instance, $\rho_1(\mathcal{F}, \mu)$ is the probability of an uncorrupted sample $x \sim \mu$ that can be labeled by both 0 and 1 by functions in $\mathcal{F}$. Equivalently, $\rho_1(\mathcal{F}, \mu)$ is the probability of this sample falling in the *disagreement region* $D(\mathcal{F}) := \{x \in \mathcal{X} : \exists f, g \in \mathcal{F}, f(x) \neq g(x)\}$. Our algorithm uses observed data points to learn and reduce the hypothesis class. Given a dataset $A = \{(x_i, y_i)\}_i$, define the *reduction* of $\mathcal{F}$ to A as $\mathcal{F} \cap A := \{f \in \mathcal{F} : \forall i,\ f(x_i) = y_i\}$. When A consists of a single labeled example (x, y), we write $\mathcal{F}^{x \to y}$ as shorthand for $\mathcal{F} \cap \{(x, y)\}$. Last, by convention $\min \emptyset = +\infty$.

3. Main Results

Our main result for oblivious adversaries is that sublinear misclassification and abstention errors can be achieved by an algorithm ABSTAINBOOST , which combines weak learners using a boosting abstention strategy. Its description is deferred to Section 3.1.

Theorem 3 *Let $\mathcal{F}$ be a function class with VC dimension d and a horizon $T \geq 1$. Then, for any $\alpha \in [0, 1/4]$, there is a choice of parameters for* ABSTAINBOOST *that achieves the following learning guarantee against any oblivious adversary:*

$$\text{MISERR} \lesssim T^{4\alpha} \quad \textit{and} \quad \mathbb{E}[\text{ABSERR}] \lesssim d^2 \log^{7/4}(T) \cdot T^{1-\alpha}.$$

In particular, the family of algorithms ABSTAINBOOST provide a polynomial tradeoff between the abstention and misclassification errors, as depicted in the green region in Fig. 1. Naturally, one may wonder if such a polynomial tradeoff is necessary—can we achieve poly$(d \log T)$ misclassification error with $T^{1-\Omega(1)}$ abstention error, or inversely? To this end, we complement Theorem 3 with the following lower bound tradeoff between abstention and misclassification errors, which shows that polynomial tradeoffs are necessary even for learning function classes with VC dimension one.

Theorem 4 *There is a VC-1 function class $\mathcal{F}$ such that the following holds. For any $T \geq 1$ and any learner, for any $\alpha \in [0, 1]$, there exists an oblivious adversary on $\mathcal{F}$ for which either*

$$\mathbb{E}[\text{MISERR}] \geq T^{\alpha}/32 \quad \textit{or} \quad \mathbb{E}[\text{ABSERR}] \geq T^{1-\alpha}/2.$$

This lower bound tradeoff is depicted in the red region within Fig. 1. While this leaves some polynomial gap between the guarantees of ABSTAINBOOST from Theorem 3, for VC-1 classes, this gives a complete characterization of the misclassification/abstention error landscape:

Remark 5 *For the special case of VC-1 function classes, Goel et al. (2023) provided a family of algorithms tuned for some parameter $\alpha \in [0,1]$*[2] *that achieves $\mathbb{E}[\textsc{MisErr}] \leq 2T^{\alpha}$ and $\mathbb{E}[\textsc{AbsErr}] \leq T^{1-\alpha}\log T$. Together with Theorem 4, this gives a tight characterization of achievable misclassification and abstention error for axis-aligned rectangles (see Appendix C for a definition) or VC-1 classes.*

Last, we turn to the case of adaptive adversaries. At the high-level, this setting is significantly more challenging since the adversary may interfere with statistical estimations carried by the learner. Nevertheless, we show that ABSTAINBOOST can still achieve learning guarantees for useful function classes with additional structure. Specifically, in addition to having finite VC dimension, we will focus on function classes with finite so-called *reduction dimension*, defined as follows:

Definition 6 *Let $l \geq 1$. We say that a function class $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ has l-reduction dimension D_l if for any $n \geq D_l$, and any $x_1, \ldots, x_n \in \mathcal{X}$:*

$$\left|\left\{(\mathcal{F} \cap A)|_{\{x_1,\ldots,x_n\}} : A \subseteq \mathcal{X} \times \{0,1\}, |A| \leq l\right\}\right| \leq n^{D_l}.$$

In words, a function class $\mathcal{F}$ has finite l-reduction dimension if the number of possible function classes that can be obtained by reducing $\mathcal{F}$ to functions consistent with l labeled datapoints grows polynomially with the number n of test points. This polynomial growth is reminiscent of the polynomial growth $\mathcal{O}(n^d)$ of the number of different labelings induced by a function class with VC dimension d on n test points, known as Sauer-Shelah's lemma (Sauer, 1972; Shelah, 1972).

Of particular interest, linear classifiers in $\mathbb{R}^d$—$\mathcal{F}_{\text{lin}}^d := \{x \in \mathbb{R}^d \mapsto \mathbb{1}[a^\top x \geq b] : a \in \mathbb{R}^d, b \in \mathbb{R}\}$—have finite reduction dimension: we show in Lemma 15 that they have l-reduction dimension $D_l = \mathcal{O}(d^2 l)$. This result uses links between oriented matroids and algebraic geometry together with Warren's theorem (Warren, 1968), which bounds the number of topological components defined by polynomial equations. We refer to Appendix C for further examples.

For function classes which have finite $\mathcal{O}(d^2 \log T)$-reduction dimension, we show that ABSTAINBOOST essentially achieves the same learning guarantee as in Theorem 3 for adaptive adversaries.

Theorem 7 *Let $\mathcal{F}$ be a function class with VC dimension d and $\lceil 5d^2 \log T \rceil$-reduction dimension D given a horizon $T \geq 1$. Then, for any $\alpha \in [0, 1/4]$, there is a choice of parameters for* ABSTAINBOOST *that achieves the following learning guarantees against any adaptive adversary:*

$$\textsc{MisErr} \lesssim T^{4\alpha} \quad \textit{and} \quad \mathbb{E}[\textsc{AbsErr}] \lesssim d^{1/2}(D \log D + \log T)^{3/2} \log(T) \cdot T^{1-\alpha}.$$

This shows that abstention learning is still possible for adaptive adversaries for structured VC function classes, with the same algorithmic ideas as for the oblivious case. We believe this can provide a useful starting point towards adaptively learning all VC classes:

Open question. *Can we achieve poly$(d)T^{1-\Omega(1)}$ misclassification and abstention error for any function class with VC dimension d, against adaptive adversaries?*

2. using their Algorithm 2 with parameter $T^{1-\alpha}$

3.1. Overview of our algorithmic approach

When the distribution μ is known, the algorithm developed by Goel et al. (2023) crucially relies on the knowledge of shattering probabilities which serve as a form of potential to decide when to abstain and which predictions to make—we refer to Section 4 for an overview of their algorithm. Hence, the main challenge for designing algorithms without prior knowledge of μ is to estimate these shattering probabilities efficiently using only the observed sequence of instances, which may contain a large fraction of corruptions. To do so, we intuitively reduce the problem to the simpler setting in which the learner has access to a few clean samples from μ, following two main steps.

Construction of weak learners. For each subset $\mathcal{T} \subseteq [T]$ of N times, we construct a weak learner which uses the instances $\mathcal{S} = \{x_t : t \in \mathcal{T}\}$ to estimate shattering probabilities. Specifically, this weak learner runs a variant of the algorithm from Goel et al. (2023) which uses shattering probability estimators based on the samples in $\mathcal{S}$ only. For convenience, we also include as parameter for the weak learner, labelings $(z_t)_{t\in\mathcal{T}}$ for all times in $\mathcal{T}$. Together, this yields a weak learner $\text{WL}(\mathcal{T}, z)$; we defer its detailed description to Algorithm 1 in Section 4.

Importantly, when $\mathcal{T}$ contains only uncorrupted iterations, these instances in $\mathcal{S}$ exactly correspond to N i.i.d. samples from μ. This essentially corresponds to the case when the learner is given N uncorrupted samples from μ a priori, for which we obtain the following learning guarantee.

Theorem 8 (Simplified Theorem 20) *Let $\mathcal{F}$ be a function class of VC dimension d and $\mathcal{O}(d^2 \log \frac{1}{\epsilon})$-reduction dimension $D \in \mathbb{N} \cup \{\infty\}$. Fix $\epsilon \in (0, 1)$. Let $\mathcal{T}$ denote the (potentially random) set of first $\tilde{\Omega}(d^2/\epsilon)$ (resp. $\tilde{\Omega}(D \log(D) d^2/\epsilon)$) non-corrupted rounds if the adversary is oblivious (resp. adaptive). Define $z := (y_t)_{t\in\mathcal{T}}$. Then, for some choice of parameters of the weak learner*

$$\textsc{MisErr}(\mathcal{T}, z) \lesssim d^2 \log(1/\epsilon) \quad \textit{and} \quad \mathbb{E}[\textsc{AbsErr}(\mathcal{T}, z)] \lesssim \epsilon T,$$

where $\textsc{MisErr}(\mathcal{T}, z)$ and $\textsc{AbsErr}(\mathcal{T}, z)$ respectively denote the misclassification and abstention error of the weak learner WL$(\mathcal{T}, z)$ as defined in Algorithm 1.

Boosting strategy to combine weak learners. We then use a BOOSTING strategy to combine the recommendations of all such weak learners $\text{WL}(\mathcal{T}, z)$ and achieve misclassification and abstention errors comparable to that of the best weak learner. We defer its detailed description to Algorithm 3 within Section 5. In fact, the boosting procedure does not require any specific property of the weak learners; we therefore present its guarantee in a general abstention learning framework with experts, of independent interest. Consider the abstention online setup with L experts in which at iteration $t \in [T]$: (1) An adversary adaptively selects the recommendation $y_{i,t} \in \{0, 1, \perp\}$ for each expert $i \in [m]$ together with the true value $y_t \in \{0, 1\}$. (2) Next, the learner observes the recommendations $y_{i,t}$ for $i \in [L]$ then selects $\hat{y}_t \in \{0, 1, \perp\}$. (3) At the end of the round, the true value y_t is revealed to the learner. We have the following guarantee on the BOOSTING procedure for this learning problem.

Theorem 9 *Fix parameters $M \geq 0$ and $s_{\max} \in [T]$. Then, for L experts, $\textsc{Boosting}_{s_{\max},M}$ achieves the following guarantee against any (adaptive) adversary:*

$$\textsc{MisErr} \leq c_0 \frac{(M \log L + \log T) T \log L}{s_{\max}} \quad \textit{and} \quad \textsc{AbsErr} \leq s_{\max} \lceil \log L \rceil + \min_{i\in[L]: M_i < M} A_i,$$

for some universal constant $c_0 > 0$, where M_i and A_i respectively denote the misclassification and abstention error of expert $i \in [L]$.

We note that usual learning-with-expert techniques do not readily apply since the boosting procedure needs to *simultaneously* ensure low excess abstention error and excess misclassification compared to good weak learners. Since Theorem 8 precisely ensures the existence of a weak learner with less than $M = \mathcal{O}(d^2 \log T)$ misclassification error and low abstention error, running BOOSTING on all weak learners and adjusting parameters yields the desired learning guarantees for Theorem 3 and Theorem 7. For completeness, the final algorithm ABSTAINBOOST is given in Algorithm 4.

4. Abstention learning with few uncorrupted samples

Algorithmic overview of Goel et al. (2023). To gain intuition on the construction of weak learners, we first review the algorithm in Goel et al. (2023), detailed in Algorithm 5, which crucially relies on prior knowledge of μ to estimate shattering probabilities $\rho_k(\cdot, \mu)$. Throughout, this section only, we denote by $\mathcal{F}_t := \mathcal{F} \cap \{(x_s, y_s) : s < t\}$ the reduced function class given prior datapoints at iteration $t \geq 1$. In particular, $\mathcal{F}_t$ is non-increasing in t and hence k-shattering probabilities $\rho_k(\mathcal{F}_t, \mu)$ are also non-increasing in t for all $k \in [d]$.

The first observation is that whenever $\rho_1(\mathcal{F}_t, \mu) \leq T^{-1}$, the algorithm only makes a prediction when the instance is fully disambiguated $x_t \notin D(\mathcal{F}_t)$, incurring no misclassification errors. Indeed, this results in a probability of a bad abstention at t of exactly $\rho_1(\mathcal{F}_t, \mu) \leq T^{-1}$ and hence a negligible overall abstention error. The goal of the algorithm is therefore to decrease the 1-shattering probability when it makes a misclassification error, while also ensuring low abstention error. This is in fact possible if the 2-shattering probability is small, say $\rho_2(\mathcal{F}_t, \mu) \leq T^{-2}$. More generally, they show the following bound for any $k \geq 1$ (Goel et al., 2023, Lemma 4.2):

$$\mathbb{P}_{x \sim \mu}\left[\min\left\{\rho_k(\mathcal{F}^{x \to 0}, \mu), \rho_k(\mathcal{F}^{x \to 1}, \mu)\right\} \geq 0.6\rho_k(\mathcal{F}, \mu)\right] \leq 5\frac{\rho_{k+1}(\mathcal{F}, \mu)}{\rho_k(\mathcal{F}, \mu)}. \tag{1}$$

In particular, provided that $\rho_{k+1}(\mathcal{F}, \mu)/\rho_k(\mathcal{F}, \mu) \leq T^{-1}$, the algorithm may abstain whenever the instance x_t falls into the bad event from Eq. (1) without incurring significant abstention cost, and otherwise predict the value $\hat{y}_t = y \in \{0, 1\}$ yielding the larger value for $\rho_k(\mathcal{F}^{x_t \to y}, \mu)$.This ensures a decrease of the k-shattering probability by a factor 0.6 whenever a misclassification error occurs, leading to the following structure for the μ-dependent algorithm of Goel et al. (2023). Since $\rho_{d+1}(\mathcal{F}, \mu) = 0$ ($\mathcal{F}$ has VC dimension d) we can first decrease the d-shattering probability until $\rho_d(\mathcal{F}_t, \mu) \leq T^{-d}$. Then, we can decrease the $(d-1)$-shattering probability until $\rho_{d-1}(\mathcal{F}_t, \mu) \leq T^{-(d-1)}$; and continue successively until we decrease $\rho_1(\mathcal{F}_t, \mu) \leq T^{-1}$.

Construction of the weak learner. We now turn to our main setting of interest when the distribution μ is unknown. As discussed within Section 3.1, we may consider in this part that the learner has access to a small set of clean i.i.d. samples from μ, which can be used to estimate the shattering probabilities required by the μ-dependent algorithm.[3] The main difficulty is that it requires estimating shattering probabilities of order $T^{-\Omega(d)}$, i.e., exponentially small in the VC dimension; while we may have at most T samples from μ even if no input was corrupted.

3. The algorithm would have similar learning guarantees if instead of knowing μ, the learner has access to estimates of shattering probabilities accurate up to say a (1 ± 0.05) factor.

First, for any finite subset $S \subseteq \mathcal{X}$—which we think of having i.i.d. samples from μ—the k-shattering probability of a function class $\mathcal{F}$ has the natural following U-statistic estimator

$$\hat{\rho}_k^S(\mathcal{F}) := \binom{|S|}{k}^{-1} \sum_{S' \subseteq S, |S'|=k} \mathbb{1}[S' \text{ is shattered by } \mathcal{F}].$$

For convenience, we denote by $\sigma_k^N(\mathcal{F}, \mu)$ the standard deviation of the random variable $\hat{\rho}_k^S(\mathcal{F})$ when S contains N i.i.d. samples from μ. To achieve Gaussian failure probabilities, we consider the median of these estimators: for a collection of subsets $\mathcal{S} = \{S_i, i \in [m]\}$, we define the k-shattering median estimator $\rho_k^{\mathcal{S}}(\mathcal{F}) := \text{Median}(\hat{\rho}_k^{S_1}(\mathcal{F}), \ldots, \hat{\rho}_k^{S_m}(\mathcal{F}))$.

The central idea behind the construction of our weak learners is the following. The standard deviation $\sigma_k^N(\mathcal{F}, \mu)$ when S contains N i.i.d. samples form μ, can be bounded as a function of the l-shattering probabilities of $\mathcal{F}$ for $l < k$ (see Eq. (5) in the appendix). Specifically, even though the S contain at most $N \ll T$ samples, the estimator $\rho_k^{\mathcal{S}}(\mathcal{F})$ may have exponential precision provided that l-shattering probabilities for $l < k$ have an exponential decay.

Lemma 10 *Fix $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$, a distribution μ on $\mathcal{X}$, $k \geq 1$, and $N \geq k^2$. Let $\eta = \frac{k^2}{N}$. Suppose that for some $c \geq 1$, we have $\rho_l(\mathcal{F}, \mu) \leq c \cdot \eta^l$ for all $l < k$. Then, $\sigma_k^N(\mathcal{F}, \mu) < \sqrt{3c \cdot \eta^k \rho_k(\mathcal{F}, \mu)}$.*

For instance, taking $c = 1$ within Lemma 10, if $\rho_l(\mathcal{F}, \mu) \leq \eta^l$ for all $l < k$, then with high probability the estimator $\rho_k^{\mathcal{S}}(\mathcal{F})$, where $\mathcal{S}$ is a collection of $\mathcal{O}(\log T)$ subsets of N i.i.d. samples from μ, accurately estimates the k-shattering probability up to a (1 ± 0.05) factor whenever $\rho_k(\mathcal{F}, \mu) \gtrsim \eta^k$.

This motivates the following strategy. At iteration t, instead of aiming to decrease the k_t-shattering probability according to a decreasing sequence for k_t as in Goel et al. (2023), we select

$$k_t := \max\{k' \in [d] : \forall l \in [k'], \rho_l^{\mathcal{S}}(\mathcal{F}_t) > \epsilon \cdot \rho_{l-1}^{\mathcal{S}}(\mathcal{F}_t)\}. \tag{2}$$

Intuitively, this ensures that the conditions required for Lemma 10 hold for $k = k_t - 1$ and $c = \rho_{k_t-1}(\mathcal{F}_t, \mu)$. This essentially shows that the estimated k_t-shattering probabilities $\rho_{k_t}(\tilde{\mathcal{F}}, \mu)$ for $\tilde{\mathcal{F}} \in \{\mathcal{F}_t, \mathcal{F}_t^{x_t \to 0}, \mathcal{F}_t^{x_t \to 1}\}$ are accurate as long as $\rho_{k_t}(\mathcal{F}_t, \mu) \gtrsim \frac{k^2}{N} \cdot \rho_{k_t-1}(\mathcal{F}_t, \mu)$, providing the desired guarantee for shattering probability estimates. In addition, by construction of k_t, we have $\frac{\rho_{k_t+1}^{\mathcal{S}}(\mathcal{F}_t)}{\rho_{k_t}^{\mathcal{S}}(\mathcal{F}_t)} \leq \epsilon$. Provided that these estimates are accurate, this shows that the probability of a bad abstention at time t is bounded by $\mathcal{O}(\epsilon)$ from Eq. (1), yielding an overall abstention error $\mathcal{O}(\epsilon T)$.

This yields our final weak learner $\text{WL}(\mathcal{T}, z)$. Before iteration $\max \mathcal{T}$, it considers that only rounds in $\mathcal{T}$ are uncorrupted. After then, it uses the samples $\{x_{t'} : t' \in \mathcal{T}\}$ to estimate k_t-shattering probabilities where k_t is defined in Eq. (2). The weak learner is summarized in Algorithm 1. As a remark, Algorithm 1 accommodates for two types of updates for the restricted function classes $\mathcal{F}_t$: setting the parameter `update` $=$ `always` corresponds to the update policy discussed here. The other alternative `update` $=$ `restricted` will only be used for adaptive adversaries.

Oblivious vs. adaptive adversaries. In the above exposition, we omitted a major technicality: the guarantee from Lemma 10 on shattering probability estimates for a function class $\tilde{\mathcal{F}}$ applies to i.i.d. samples independent from $\tilde{\mathcal{F}}$. However, Algorithm 1 requires estimating shattering probabilities of various function classes $\{\mathcal{F}_t, \mathcal{F}_t^{x_t \to 0}, \mathcal{F}_t^{x_t \to 1} : t \in (\max \mathcal{T}, T]\}$. For *oblivious* adversaries, we can check that estimates are always accurate by applying a simple union bound: the function classes considered are independent from the i.i.d. samples $\mathcal{S}$ of μ used by the learner.

Algorithm 1: Weak Learner WL$(\mathcal{T}, z)$

Input: precision $\epsilon \in [0,1]$, subset of times $\mathcal{T}$ and labels $z = (z_t)_{t\in\mathcal{T}}$, number of subsets m, update policy $\texttt{update} \in \{\texttt{always}, \texttt{restricted}\}$

1 **for** $t \leq \max \mathcal{T}$ **do** If $t \in \mathcal{T}$, predict $\hat{y}_t = z_t$, otherwise abstain $\hat{y}_t = \perp$;
2 Let $\mathcal{S} = \{S_i\}_{i=1}^m$ be a partition of $\{x_{t'} : t' \in \mathcal{T}\}$ such that $|S_1| = |S_2| = \cdots = |S_m| = \lfloor |\mathcal{T}|/m \rfloor$, discarding any remaining samples if necessary. Initialize $\mathcal{F}_t \leftarrow \mathcal{F}$.
3 **for** $t = 1 + \max \mathcal{T}, \ldots, T$ *and* $\rho_1^{\mathcal{S}}(\mathcal{F}_t) > \epsilon$ **do**
4 Observe x_t. Let $k = \max\{k' \in [d] : \forall l \in [k'], \rho_l^{\mathcal{S}}(\mathcal{F}_t) > \epsilon \cdot \rho_{l-1}^{\mathcal{S}}(\mathcal{F}_t)\}$. Note that $\rho_0^{\mathcal{S}}(\mathcal{F}_t) = 1$.
5 **if** $\min_{y\in\{0,1\}} \rho_k^{\mathcal{S}}(\mathcal{F}_t^{x_t\to y}) \geq 0.9\rho_k^{\mathcal{S}}(\mathcal{F}_t)$ **then** Abstain $\hat{y}_t = \perp$;
6 **else** Predict $\hat{y}_t = \arg\max_{y\in\{0,1\}} \rho_k^{\mathcal{S}}(\mathcal{F}_t^{x_t\to y})$;
7 Observe response y_t
8 **if** $\mathit{update} = \mathit{always}$ **then** Update $\mathcal{F}_{t+1} = \mathcal{F}_t^{x_t\to y_t}$;
9 **else if** $\mathit{update} = \mathit{restricted}$ **then** If $\hat{y}_t \notin \{y_t, \perp\}$, update $\mathcal{F}_{t+1} = \mathcal{F}_t^{x_t\to y_t}$. Otherwise, keep $\mathcal{F}_{t+1} = \mathcal{F}_t$;
10 **end**
11 **if** $\rho_1^{\mathcal{S}}(\mathcal{F}_t) \leq \epsilon$ *and* $x_t \in D(\mathcal{F}_t)$ **then** Abstain $\hat{y}_t = \perp$;
12 **if** $\rho_1^{\mathcal{S}}(\mathcal{F}_t) \leq \epsilon$ *and* $x_t \notin D(\mathcal{F}_t)$ **then** Predict consistent label $\hat{y}_t = f(x_t)$ for any $f \in \mathcal{F}_t$;

On the other hand, *adaptive* adversaries may select instances x_t so that the class $\mathcal{F}_t^{x_t\to 0}$ is badly correlated with the samples $\mathcal{S}$ used by the weak learner. To resolve this issue, we note that the current strategy would work if the reduced function classes $\mathcal{F}_t$ were only updated when mistakes are made (see line 9 of Algorithm 1). Further, the total number of mistakes is very small: at most $l - 1 := d^2 \log(1/\epsilon)$[4]. Hence, all function classes that can potentially be encountered by the learner when running the weak learner with the update policy $\texttt{update} = \texttt{restricted}$,[5] are of the form $\mathcal{F} \cap A$ where $A \subseteq \mathcal{X} \times \{0,1\}$ is a dataset of $|A| \leq l$ points.

We then show a uniform concentration result on shattering probability estimates for all such reduced classes if the l-reduction dimension of $\mathcal{F}$ is finite. As a reminder, the l-reduction dimension precisely bounds the number of different projected function classes of the form $\mathcal{F} \cap A$ for $|A| \leq l$.

Lemma 11 *There exists a universal constant $c_0 \geq 1$ such that the following holds. Fix $k, N \geq 1$. Let $l \geq 1$ and $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ be a function class with l-reduction dimension D_l. Let $\mathcal{S} := \{S_i, i \in [m]\}$ be a collection of $m \geq c_0(D_l \log(ND_l) + \log \frac{1}{\delta})$ independent sets of N i.i.d. samples from μ. Then, with probability at least $1 - \delta$,*

$$|\rho_k^{\mathcal{S}}(\mathcal{F} \cap A) - \rho_k(\mathcal{F} \cap A, \mu)| \leq 2\sigma_k^N(\mathcal{F} \cap A, \mu), \quad A \subseteq \mathcal{X} \times \{0,1\}, |A| \leq l.$$

Lemma 11 precisely shows that shattering probabilities of all function classes that may be encountered can be uniformly accurately estimated, completing the last argument for Theorem 8.

4. Recall that shattering probabilities are reduced by a factor at each mistake
5. We use a different update policy $\texttt{update} = \texttt{always}$ for oblivious adversaries to achieve stronger misclassification and abstention dependency in the VC dimension d of the function class.

Algorithm 2: Algorithm for one layer $\text{DELETE}_{s,C}$

Input: number of experts L, minimum number of predictions C, deletion number $s \geq 0$

for $t \geq 1$ **do**
  Observe $y_{i,t}$ for all experts $i \in [L]$
  For each $i \in [L]$, if $|\{t' < t : y_{i,t'} \neq \perp\}| < s$, delete prediction $\tilde{y}_{i,t} := \perp$, otherwise $\tilde{y}_{i,t} := y_{i,t}$
  if $|\{i \in [L] : \tilde{y}_{i,t} \neq \perp\}| \geq C$ **then** Predict $\hat{y}_t := \arg\max_{y \in \{0,1\}} |\{i \in [L] : \tilde{y}_{i,t} = y\}|$;
  else Abstain $\hat{y}_t = \perp$;

5. Boosting procedure

In this section, we describe our boosting procedure for the online setup with L experts (weak learners) as described in Section 3.1. The procedure requires an error bound parameter M. Specifically, up to removing predictions of weak learners with large misclassification error, we may assume that all weak learners have at most M misclassification error—for the abstention learning problem, we will use $M = \mathcal{O}(d^2 \log T)$ since Theorem 8 guarantees the existence of a good weak learner with this misclassification error. With this property in mind, the high-level idea behind our procedure is as follows. Weak learners which do not abstain for at least $s \gg M$ rounds make at least $s - M$ correct predictions and as a result, have small mistake per prediction average rate $\leq M/(s - M)$. Therefore, assuming that we can focus on these weak learners with at least s predictions, the fraction of incorrect predictions compared to correct predictions throughout all T rounds is at most $\leq M/(s - M) \ll 1$. In turn, majority-vote-inspired algorithms will achieve low misclassification error without having to abstain frequently. Since we do not know a priori which weak learners make $s \gg M$ predictions, we instead use a deletion strategy: we ignore the first s predictions of each algorithm for some adaptively chosen value of s, which precisely removes weak learners with at most s predictions. In practice, to achieve low misclassification error, our complete boosting procedure needs to perform $\log L$ rounds of such deletions. We focus on a single layer for now.

For any parameter $s \geq 0$ and minimum number of predictions C per round, we consider the algorithm DELETE which deletes the first s predictions of each expert, then follows the majority vote $\hat{y}_t$ if at least C weak learners still make a prediction in $\{0, 1\}$, and otherwise abstains $\hat{y}_t = \perp$. This is described in Algorithm 2 and is tailored for the following structured adversaries.

Definition 12 (Structured adversaries) *We say that an adversary includes at most M mistakes if for any defined round $t \geq 1$, all weak learners $i \in [L]$ have misclassification error at most M, i.e., $|\{t' \leq t : y_{i,t} \notin \{y_t, \perp\}\}| \leq M$.*

We say that an adversary includes at most C predictions per round if for any defined round $t \geq 1$, at most C weak learners make a prediction, i.e., $|\{i \in [L] : y_{i,t} \neq \perp\}| \leq C$.

Following arguments inspired from the informal overview above, we can show that there is a choice of parameter $s \geq 0$ that achieves good learning guarantee for the DELETE algorithm. We then aggregate these algorithms using a standard learning-with-experts to essentially achieve the best misclassification error among these. This yields the $\text{AGREGATE}_{s_{\max},C}$ algorithm, simply defined as the Weighted Majority Algorithm (WMA) (Littlestone and Warmuth, 1994) over $\text{DELETE}_{s,C}$ for $s \in \{0, \ldots, s_{\max}\}$ with learning rate $1/2$. Conveniently it is also deterministic. For completeness, we include the AGREGATE algorithm in Algorithm 6 in appendix. Importantly, all considered DELETE_s

algorithms delete at most $s_{\max}$ predictions for each weak learner, which will be crucial to bound the overall abstention error. Altogether, we obtain the following guarantee.

Lemma 13 *Fix $C, s_{\max} \geq 0$ and $M \geq 1$. Consider an adversary for L weak learners which includes at most M mistakes and at most C predictions per round. Denote by $\hat{y}_t$ the selected value of* AGGREGATE$_{s_{\max},C/2}$ *at iteration t. Letting n be the final iteration, we have*

$$|\{t \in [n] : \hat{y}_t \notin \{y_t, \perp\}\}| \leq \frac{32Mn\lceil \log L \rceil}{s_{\max}+1} + 4\log(s_{\max}+1).$$

Also, for any iteration $t \in [n]$ such that $\left|\{i \in [L] : z_{i,t} \neq \perp \text{ and } \left|\{t' < t : z_{i,t'} \neq \perp\right|\} \geq s_{\max}\}\right| \geq C/2$, the algorithm makes a prediction: $\hat{y}_t \neq \perp$.

We are now ready to construct our multi-layer algorithm. To ensure that each weak learner $i \in [L]$ makes at most M mistakes, we first delete all predictions $y_{i,t} \in \{0,1\}$ from weak learners i which made M misclassification mistakes in previous rounds: $y_{i,t} \leftarrow \perp$. Next, we use AGGREGATE$_{s_{\max},L/2}$, forming the first layer. From Lemma 13, it achieves low misclassification error on all times when the number of weak learners making a prediction lay within $[L/2, L]$, after removing the first $s_{\max}$ predictions of all weak learners. On the other hand, it may abstain when less than $L/2$ weak learners make predictions. For these remaining times, we then use AGGREGATE$_{s_{\max},L/4}$, forming the second layer. This ensures low misclassification error whenever the number of weak learner predictions lies within $[L/4, L/2]$, after removal of another $s_{\max}$ predictions per weak learner in this layer. We continue this process for $\lceil \log L \rceil$ layers, which takes care of all remaining times. The final algorithm BOOSTING$_{s_{\max}}$ is summarized in Algorithm 3. Importantly, the subroutine AGGREGATE$_{s_{\max},2^{-j}L}$ is run on different subset of times $\mathcal{Q}_j$ for each layer j, which corresponds to when the number of weak learner predictions (after prediction removals) is at most $2^{1-j}L$. In Algorithm 3, we write AGGREGATE$^{\mathcal{Q}_j}_{s_{\max},2^{-j}L}$ to emphasize this point.

Combining together the learning guarantees for all AGGREGATE subroutines from Lemma 13, we obtain the desired guarantee for the BOOSTING algorithm.

Proof sketch for Theorem 9. First, to check that the algorithm is well-defined we can verify that the while loop on each iteration $t \geq 1$ terminates. Indeed, on each loop, the AGGREGATE$_{s_{\max},2^{-j}L}$ makes a prediction $\hat{y}_t^{(j)} \in \{0,1\}$ if at least $2^{-j-1}L$ learner predictions remain after deleting $s_{\max}$ predictions per weak learner (see line 11 of Algorithm 3). Roughly speaking, at each loop at least half of the predictions are deleted, ensuring that j strictly increases on each loop and hence terminates after $\lceil \log L \rceil$ rounds when no weak learner predictions remain.

Next, since the boosting procedure uses predictions of the AGGREGATE$_{s_{\max},2^{-j}L}$ algorithm when there were at most $2^{1-j}L$ weak learner predictions (see definition of j on line 7 of Algorithm 3), Lemma 13 bounds the misclassification error for each layer j. The final misclassification error is essentially the sum of those for all layers, incurring an extra factor $\log L$. For the abstention error, by construction, each layer $j \in [\lceil \log L \rceil]$ deleted at most $s_{\max}$ predictions for each weak learner (see the initialization of deletion counts in line 1 of Algorithm 3). In summary, throughout the procedure only deleted $\leq s_{\max}\lceil \log L \rceil$ predictions for weak learner with at most M misclassification error. This gives the desired abstention error bound since the BOOSTING algorithm makes a prediction whenever at least one weak learner prediction remains (see line 13 of Algorithm 3). ∎

Algorithm 3: Final boosting algorithm $\text{BOOSTING}_{s_{\max}}$

Input: number of experts L, maximum per-layer deletion $s_{\max}$, mistake tolerance M

```
For j ∈ [⌈log L⌉] initialize times Q_j ← ∅ and deletion counts s_{i,j} ← s_max for all i ∈ [L]
for t ≥ 1 do
    Observe recommendation y_{i,t} for each expert i ∈ [L]
    For i ∈ [L], if |{s < t : y_{i,s} ∉ {y_s, ⊥}}| ≥ M, delete prediction: z_{i,t} ←⊥; otherwise z_{i,t} ← y_{i,t}
    Define n_t ← |{i ∈ [L] : z_{i,t} ≠⊥}|
    while n_t > 0 do
        Let j such that n_t ∈ (2^{-j} L, 2^{1-j} L]
        Let ŷ_t^{(j)} be the value selected by AGGREGATE^{Q_j}_{s_max, 2^{-j} L} for a new iteration with weak learner
          recommendations (z_{i,t})_{i∈[L]}. Add Q_j ← Q_j ∪ {t}
        if ŷ_t^{(j)} ≠⊥ then Predict ŷ_t = ŷ_t^{(j)} and break ;
        for i ∈ [L] with z_{i,t} ≠⊥ and s_{i,j} > 0 do
            Delete prediction: z_{i,t} ←⊥ and update s_{i,j} ← s_{i,j} − 1
        Update n_t ← |{i ∈ [L] : z_{i,t} ≠⊥}|
    if n_t = 0 then Abstain ŷ_t =⊥;
```

6. Conclusion

In the sequential learning with abstention framework of Goel et al. (2023), we showed that achieving sublinear abstention and misclassification errors is possible for general VC classes without knowing μ, for oblivious adversaries. We also identified structural properties of the function class $\mathcal{F}$—finite so-called *reduction dimension*—enabling learning against adaptive adversaries, which may be of independent interest for other abstention learning models. Together with corresponding lower bounds, these results show the existence of a polynomial-form tradeoff between abstention and misclassification errors. This work naturally leaves open the following two directions.

Tight characterizations of tradeoffs between abstention and misclassification errors. At the high-level, our positive results for oblivious adversaries show that $\mathcal{O}(T^{4\alpha})$ misclassification can be achieved while ensuring $\tilde{\mathcal{O}}(T^{1-\alpha})$ abstention error, leaving a gap compared to the corresponding $\Omega(T^{\alpha})$ misclassification error lower bound. The lower bound turns out to be tight for VC-1 classes for instances, but the situation is unclear for general VC classes. We note, however, that with further parameter information (e.g., which deletion parameter $s \leq s_{\max}$ to use within DELETE procedures), slight modifications of the algorithm can yield stronger misclassification guarantees, opening potential opportunities for further improvements.

Extending results to adaptive adversaries for general VC classes. Our guarantees for adaptive adversaries are limited to function classes with finite reduction dimension (and finite VC dimension). It remains open whether one can remove such constraints and achieve successful learning for all VC classes, which leads to the following open question:

> *Can we achieve poly$(d)T^{1-\Omega(1)}$ misclassification and abstention error for any function class with VC dimension d, against adaptive adversaries?*

Acknowledgments

The authors are grateful to Abhishek Shetty for useful discussions.

References

Noga Alon. Tools from higher algebra. *Handbook of combinatorics*, 2:1749–1783, 1995.

Idan Attias, Aryeh Kontorovich, and Yishay Mansour. Improved generalization bounds for robust learning. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, 2019.

Pranjal Awasthi, Maria Florina Balcan, and Philip M. Long. The power of localization for efficiently learning linear separators with noise. *Journal of the ACM*, 63(6):1–27, 2017.

Maria-Florina Balcan, Avrim Blum, Steve Hanneke, and Dheeraj Sharma. Robustly-reliable learners under poisoning attacks. In *Proceedings of the 35th Annual Conference on Learning Theory*, 2022.

Peter L Bartlett and Marten H Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9(8), 2008.

Shai Ben-David. 2 notes on classes with vapnik-chervonenkis dimension 1. *arXiv preprint arXiv:1507.05307*, 2015.

Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *COLT*, volume 3, page 1, 2009.

Anders Björner, Michel Las Vergnas, Bernd Sturmfels, Neil White, and Gunter M. Ziegler. *Oriented matroids*. Number 46 in Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1999.

Moïse Blanchard. Agnostic smoothed online learning. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, STOC '25, page 1997–2006, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715105. doi: 10.1145/3717823.3718111. URL https://doi.org/10.1145/3717823.3718111.

Moïse Blanchard and Samory Kpotufe. Distributionally-constrained adversaries in online learning. *arXiv preprint arXiv:2506.10293*, 2025.

Adam Block, Yuval Dagan, Noah Golowich, and Alexander Rakhlin. Smoothed online learning is as easy as statistical learning. In *Conference on Learning Theory*, pages 1716–1786. PMLR, 2022.

Adam Block, Alexander Rakhlin, and Abhishek Shetty. On the performance of empirical risk minimization with smoothed data. *arXiv preprint arXiv:2402.14987*, 2024.

Avrim Blum, Steve Hanneke, Jian Qian, and Han Shao. Robust learning under clean-label attack. In *Conference on Learning Theory*, pages 591–634. PMLR, 2021.

Nader H. Bshouty, Nadav Eiron, and Eyal Kushilevitz. PAC learning with nasty noise. *Theoretical Computer Science*, 288(2):255–275, 2002.

C. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Boosting with abstention. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016a. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/7634ea65a4e6d9041cfd3f7de18e334a-Paper.pdf.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *International conference on algorithmic learning theory*, pages 67–82. Springer, 2016b.

Corinna Cortes, Giulia DeSalvo, Claudio Gentile, Mehryar Mohri, and Scott Yang. Online learning with abstention. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1059–1067. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/cortes18a.html.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Uriel Feige, Yishay Mansour, and Robert E. Schapire. Robust inference for multiclass classification. In Firdaus Janoos, Mehryar Mohri, and Karthik Sridharan, editors, *Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 368–386. PMLR, 2018.

Ji Gao, Amin Karbasi, and Mohammad Mahmoody. Learning and certification under instance-targeted poisoning. In *The Conference on Uncertainty in Artificial Intelligence*, 2021.

Surbhi Goel, Steve Hanneke, Shay Moran, and Abhishek Shetty. Adversarial resilience in sequential prediction via abstention. *Advances in Neural Information Processing Systems*, 36:8027–8047, 2023.

Shafi Goldwasser, Adam Tauman Kalai, Yael Kalai, and Omar Montasser. Beyond perturbations: Learning guarantees with arbitrary adversarial test examples. *Advances in Neural Information Processing Systems*, 33:15859–15870, 2020.

Nika Haghtalab, Tim Roughgarden, and Abhishek Shetty. Smoothed analysis with adaptive adversaries. *Journal of the ACM*, 71(3):1–34, 2024.

Steve Hanneke, Amin Karbasi, Mohammad Mahmoody, Itai Mehalel, and Shay Moran. On optimal learning under targeted data poisoning. In *Advances in Neural Information Processing Systems*, volume 36, 2022.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2282952.

Adam Kalai and Varun Kanade. Towards optimally abstaining from prediction with ood test examples. *Advances in Neural Information Processing Systems*, 34:12774–12785, 2021.

Varun Kanade and Justin Thaler. Distribution-independent reliable learning. In *Conference on Learning Theory*, pages 3–24. PMLR, 2014.

Michael Kearns and Ming Li. Learning in the presence of malicious errors. *SIAM Journal on Computing*, 22(4):807–837, 1993.

Kasper Green Larsen, Chirag Pabbaraju, and Abhishek Shetty. Learning with monotone adversarial corruptions. In Matus Telgarsky and Jonathan Ullman, editors, *Proceedings of the 37th International Conference on Algorithmic Learning Theory*, volume 313 of *Proceedings of Machine Learning Research*, pages 1–18. PMLR, 23–26 Feb 2026. URL https://proceedings.mlr.press/v313/larsen26a.html.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2(4):285–318, 1988.

Nick Littlestone and Manfred K Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

Omar Montasser, Steve Hanneke, and Nathan Srebro. Vc classes are adversarially robustly learnable, but only improperly. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2512–2530. PMLR, 2019.

Omar Montasser, Surbhi Goel, Ilias Diakonikolas, and Nathan Srebro. Efficiently learning adversarially robust halfspaces with noise. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7010–7021. PMLR, 2020a.

Omar Montasser, Steve Hanneke, and Nathan Srebro. Reducing adversarially robust learning to non-robust PAC learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, 2020b.

Omar Montasser, Steve Hanneke, and Nathan Srebro. Adversarially robust learning with unknown perturbation sets. *CoRR*, abs/2102.02145, 2021.

Omar Montasser, Steve Hanneke, and Nathan Srebro. Adversarially robust learning: A generic minimax optimal learner and characterization. In *Advances in Neural Information Processing Systems*, volume 36, 2022.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Stochastic, constrained, and smoothed adversaries. *Advances in neural information processing systems*, 24, 2011.

Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13 (1):145–147, 1972.

Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. *Advances in neural information processing systems*, 31, 2018.

Saharon Shelah. A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 41(1):247–261, 1972.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Leslie G. Valiant. Learning disjunctions of conjunctions. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence*, pages 560–566, 1985.

V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971. doi: 10.1137/1116025.

Vladimir Vapnik and Alexey Chervonenkis. Theory of pattern recognition, 1974.

Hugh E Warren. Lower bounds for approximation by nonlinear manifolds. *Transactions of the American Mathematical Society*, 133(1):167–178, 1968.

Algorithm 4: ABSTAINBOOST

Input: horizon T, precision $\epsilon \in [0,1]$, number of subsets m, subset size N, maximum per-layer deletion $s_{\max} \in [T]$, mistake tolerance M, update policy $\texttt{update} \in \{\texttt{always}, \texttt{restricted}\}$

1 Run BOOSTING with maximum per-layer deletion $s_{\max}$, mistake tolerance M, on all $L := \sum_{k \leq nM} 2^k \binom{T}{k}$ weak learners $\mathrm{WL}(\mathcal{T}, z)$ run with precision ϵ, update parameter $\texttt{update}$ and number of subsets m subsets, for $\mathcal{T} \subseteq [T]$ and $z = (z_t)_{t \in \mathcal{T}} \in \{0,1\}^{\mathcal{T}}$ with $|\mathcal{T}| \leq nM$

Appendix A. Concentration inequalities

Theorem 14 (Median estimator, e.g. Theorem 2 from Lugosi and Mendelson (2019)) *Fix $\delta \in (0,1)$ and let $X_1, \ldots, X_n$ be i.i.d. samples from a distribution with mean μ and variance σ^2, with $n \geq \lceil 8 \log(1/\delta) \rceil$. Then, with probability at least $1 - \delta$,*

$$|\mathit{Median}(X_1, \ldots, X_n) - \mu| \leq 2\sigma.$$

Appendix B. Proofs of Section 3

B.1. ABSTAINBOOST description

In this section, we detail the ABSTAINBOOST which simply performs the BOOSTING strategy over all weak learners for subsets $\mathcal{T} \subset [T]$ and labelings $(z_t)_{t \in \mathcal{T}}$ of bounded size mN for some parameters $m, N \geq 1$. The algorithm is given in Algorithm 4.

B.2. Parameter choice for our main upper bound results

We first prove our guarantee for ABSTAINBOOST against oblivious adversaries.

Proof of Theorem 3 This proof combines Theorem 20 and Theorem 9. We use the following parameters (unless mentioned otherwise): $\texttt{update} = \texttt{always}$, $\epsilon = d^2 \log^{7/4}(T) T^{-\alpha}$, $m = \lceil 8 \log(dT/\epsilon) \rceil$, $N = \lceil 2000 d^2/\epsilon \rceil$, $s_{\max} = \left\lceil d^2 \log^{3/2}(T) T^{1-2\alpha} \right\rceil$ and $M = 5d^2 \log(1/\epsilon)$. The number of weak learners is $L = \sum_{k \leq Nm} 2^k \binom{T}{k}$. Provided that $s_{\max} \leq T$, Theorem 9 implies

$$\textsc{MisErr} \lesssim M \log^2 L \cdot \frac{T}{s_{\max}} \lesssim d^4 \log^{7/2} T \cdot \frac{T^{2\alpha}}{\epsilon^2} \asymp T^{4\alpha}.$$

And by Theorem 20 (the formal version of Theorem 8), for each run, the weak learner $\mathrm{WL}(\mathcal{T}, z)$ with $(\mathcal{T}, z)$ matching the first mN times of uncorrupted samples and labels satisfies

$$\textsc{MisErr}(\mathcal{T}, z) < M \text{ and } \mathbb{E}\left[\textsc{AbsErr}(\mathcal{T}, z)\right] \leq 18\epsilon T.$$

Hence by Theorem 9, provided that $\epsilon < 1$,

$$\begin{aligned}
\mathbb{E}[\textsc{AbsErr}] &\leq s_{\max} \lceil \log L \rceil + \mathbb{E}\left[\min_{\substack{i \in [L] \\ \textsc{MisErr}(\mathcal{T}_i, z_i) < M}} \textsc{AbsErr}(\mathcal{T}_i, z_i) \right] \\
&\leq s_{\max} \lceil \log L \rceil + \mathbb{E}\left[\textsc{AbsErr}(\mathcal{T}, z)\right] \\
&\lesssim d^4 T^{1-2\alpha} \log^{7/2}(T)/\epsilon + \epsilon T \asymp d^2 \log^{7/4}(T) \cdot T^{1-\alpha}.
\end{aligned}$$

We now consider the edge cases when $s_{\max} > T$ or $\epsilon \geq 1$, both of which imply $d^2 \log^{7/4}(T) > T^\alpha$. In particular, the desired abstention error statement is vacuous. In particular, 0 misclassification error can be achieved by never making any prediction, achieved by taking $M = 0$. ■

Next, we prove our guarantee for ABSTAINBOOST against adaptive adversaries, assuming the initial function class $\mathcal{F}$ has a bounded reduction dimension.

Proof of Theorem 7 The proof is exactly the same as for Theorem 3 except with the following different parameters: we choose `update = restricted`, $\epsilon = d^{1/2}(D\log D + \log T)^{3/2}\log(T) \cdot T^{-\alpha}$, $m = \lceil c_0(D\log(D) + 8D + 3\log(d/\epsilon))\rceil$, $N = \lceil 2000d^2/\epsilon \rceil$, $M = 5d^2\log(1/\epsilon)$, and $s_{\max} = \lceil d^{-1}(D\log D + \log T)^2 \log(T) \cdot T^{1-2\alpha}\rceil$. Again, we have $L = \sum_{k\leq Nm} 2^k \binom{T}{k}$ weak learners, provided that $s_{\max} \leq T$ and $\epsilon < 1$. By our choice of ϵ, we have $\log(1/\epsilon) \leq \log T$, hence the $\lceil 5d^2\log(1/\epsilon)\rceil$- reduction dimension is even smaller than the $\lceil 5d^2\log(T)\rceil$- reduction dimension D. Applying Theorem 9 again yields MISERR $\lesssim T^{4\alpha}$. Next, the same arguments using Theorems 9 and 20 show that

$$\mathbb{E}[\text{ABSERR}] \lesssim s_{\max}\lceil \log L\rceil + \epsilon T \asymp d^{1/2}(D\log D + \log T)^{3/2}\log(T)\cdot T^{1-\alpha}.$$

We next turn to the case $s_{\max} > T$ or $\epsilon \geq 1$. Again, in both cases, the abstention bound is vacuous and we can safely choose $M = 0$ to get the desired guarantee. ■

B.3. Proof of our main lower bound result

Apart from deriving error upper bounds for distribution-free abstention learning, we also explored the lower bounds of this problem. We show that there is a trade-off between misclassification and abstention errors. Intuitively, we consider the following two adversaries: Adversary 1 doesn't do corruptions and sample instances $x_1, \ldots, x_A$ from μ, while Adversary 2 decides to corrupt at every iteration, using exactly the same sequence of samples $x_1, \ldots, x_A$ until some time k when the probability of learner making a prediction is more than $1/2$. If such time doesn't exist in $[1, A]$, then the learner will make $\Omega(A)$ abstention errors in expectation when playing against Adversary 1. Otherwise, one can define the true hypothesis to be different from the prediction of learner at time k, incurring a misclassification error. For a fixed horizon T, the tree structure of a VC-1 class ensures that we can move on to the next layer of the tree and repeat the above reasoning for at least $\lfloor T/A \rfloor$ layers, which results in $\Omega(T/A)$ misclassification errors. Our result corresponds to region 2 in Fig. 1.

Proof of Theorem 4 Fix a parameter $N = 4T$. We start by constructing a function class $\mathcal{F}$ using the following tree structure: let $\mathcal{T} := \bigcup_{t=0}^{T}[N]^t$ be a rooted full N-ary tree of depth T, that is, $x \in \mathcal{T}$ is an ancestor of y if and only if x is a prefix of y (incuding $y = x$) and denote $x \preceq y$ accordingly. For convenience, we denote by $\mathcal{T}_i := \bigcup_{t=0}^{i}[N]^t$ the subtree of depth i for $i \in [T]$. We then consider the function class $\mathcal{F} : \mathcal{T} \to \{0,1\}$ containing all indicator paths from the root to a node of the tree as follows:

$$\mathcal{F} = \{\mathbb{1}[\cdot \preceq x_0] : x_0 \in \mathcal{T}\}.$$

One can easily check that $\mathcal{F}$ has VC dimension one due to its tree structure—the tree structure known to be a characterization of VC-1 classes (Ben-David, 2015).

We now fix a learning algorithm alg on $\mathcal{F}$ such that there is a constant $A \in [1/2, T/2]$ such that for any oblivious adversary on $\mathcal{F}$, $\mathbb{E}[\text{ABSTENTIONERROR}] \leq A$. Next, let $i_{\max} := \lfloor T/(2A) \rfloor$.

In the following, we recursively construct for each layer $i \leq i_{\max}$ a sequence $(x_t)_{t\leq k_i}$ within $\mathcal{T}_i$ together with $x_i^\star \in [N]^i$ such that $k_i \leq 2A \cdot i$ and run on the realizable sequence $(x_t, y_t = \mathbb{1}[x_t \preceq x_i^\star])_{t\leq k_i}$, the algorithm alg makes the following misclassification errors:

$$\forall j \leq i, \quad \mathbb{P}[\hat{y}_{k_j} = 1 - y_{k_j}] \geq \frac{1}{8},$$

where $\hat{y}_t$ denotes the prediction of alg at time t. We let $k_0 = 0$ and $x_0^\star$ be the root so that this trivially holds for $i = 0$.

Fix $i \in [i_{\max}]$ and suppose that the construction is complete for layer $i-1$. Let μ_i be the uniform distribution over the N children of $x_{i-1}^\star$. As a note, μ_i is a distribution over $[N]^i$. We now complete the sequence $(x_t)_{t\leq k_{i-1}}$ using i.i.d. samples from μ_i. Denote by $(\tilde{x}_t)_{t\in[T]}$ the corresponding sequence as well as the responses $\tilde{y}_t = \mathbb{1}[\tilde{x}_t \preceq x_{i-1}^\star]$. Note that the data $(\tilde{x}_t, \tilde{y}_t)_{t\in[T]}$ is realizable and can be viewed as containing k_{i-1} corrupted samples followed by $T - k_{i-1}$ non-corrupted samples from μ_i. Hence, when run on this realizable sequence, we must have

$$\sum_{t=k_{i-1}}^{T} \mathbb{P}[\hat{y}_t = \perp] \leq A.$$

In particular, there exists $k_i > k_{i-1}$ with $k_i \leq k_{i-1} + 2A$ such that $\mathbb{P}[\hat{y}_{k_i} = \perp] \leq 1/2$. Additionally, since the samples x_t for $t \in (k_{i-1}, k_i]$ are i.i.d. sampled from μ_i which is uniform over N elements, we have

$$\mathbb{P}[\exists t \in (k_{i-1}, k_i), \tilde{x}_{k_i} = \tilde{x}_t] \leq \sum_{t=k_{i-1}+1}^{k_i - 1} \mathbb{P}[\tilde{x}_{k_i} = \tilde{x}_t] \leq \frac{T}{N} = \frac{1}{4}.$$

As a result, introducing the $\mathcal{E}_i := \{\forall t \in (k_{i-1}, k_i), \tilde{x}_{k_i} \neq \tilde{x}_t\}$, we have

$$\mathbb{P}[\{\hat{y}_{k_i} \neq \perp\} \cap \mathcal{E}_i] \geq \frac{1}{2} - \frac{1}{4} = \frac{1}{4}.$$

By the law of total probability, we can therefore fix elements $x_t \in [N]^i$ for $t \in (k_{i-1}, k_i]$ such that (1) $x_{k_i} \notin \{x_t, t \in (k_{i-1}, k_i)\}$ and (2) alg run on the data $(x_t, \mathbb{1}[x_t \preceq x_{i-1}^\star])_{t<k_i}$ and tested on x_{k_i} satisfies $\mathbb{P}[\hat{y}_{k_i} \neq \perp] \geq 1/4$. This completes the construction of $(x_t)_{t\leq k_i}$. Next, if $\mathbb{P}[\hat{y}_{k_i} = 0] \geq 1/8$ we pose $x_i^\star := x_{k_i}$. Otherwise, we define $x_i^\star$ to be any child of $x_{i-1}^\star$ which does not belong to $\{x_t, t \in (k_{i-1}, k_i]\}$. Note that this is possible since $x_{i-1}^\star$ has $N > T$ children. We then pose $y_{k_i} = \mathbb{1}[x_{k_i} \preceq x_i^\star]$. By construction, when alg is run on $(x_t, y_t)_{t\leq k_i}$, we obtained

$$\mathbb{P}[\hat{y}_{k_i} = 1 - y_{k_i}] \geq \frac{1}{8}. \tag{3}$$

It only remains to check that for all $t \in [k_i]$ one has

$$y_t = \mathbb{1}[x_t \preceq x_i^\star]. \tag{4}$$

By definition, this holds for $t = k_i$. Since $x_{i-1}^\star$ is the parent of $x_i^\star$, the functions $\mathbb{1}[\cdot \preceq x_{i-1}^\star]$ and $\mathbb{1}[\cdot \preceq x_i^\star]$ coincide on $\mathcal{T}_i$ and hence, Eq. (4) directly holds for all $t \leq k_{i-1}$. Finally, for any $t \in (k_{i-1}, k_i)$, by construction we have $x_t \neq x_i^\star$ and as a result, $y_t = \mathbb{1}[x_t \preceq x_{i-1}^\star] = 0 = \mathbb{1}[x_t \preceq x_i^\star]$ (recall that $x_t \neq x_i^\star$ both have depth i while $x_{i-1}^\star$ has depth $i-1$).

This ends the induction and the construction of the realizable sequence $(x_t, y_t)_{t\leq k_{i_{\max}}}$. We can complete it to define $(x_t, y_t = \mathbb{1}[x_t \preceq x^\star_{i_{\max}}])_{t\leq T}$ arbitrarily. Note that this is valid since $k_{i_{\max}} \leq 2Ai_{\max} \leq T$. By construction, on this data, alg has the following misclassification error:

$$\mathbb{E}[\textsc{MisErr}] \geq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[\hat{y}_t = 1 - y_t]\right] \geq \sum_{i=1}^{i_{\max}} \mathbb{P}[\hat{y}_{k_i} \neq y_{k_i}] = \frac{i_{\max}}{8} \geq \frac{1}{32}\frac{T}{A},$$

where in the last inequality we used the fact that $A \leq T/2$. This ends the proof of the desired misclassification error bound. Note that this construction can be embedded for all $T \geq 1$ within the function class defined by a full $\mathbb{N}$-ary tree of infinite depth (which still has VC dimension 1). This ends the proof. ■

Appendix C. Reduction-dimension bounds for various hypothesis classes

In this section, we argue that many classical and practical VC function classes have small reduction dimension, making it a reasonable complexity measure for many hypothesis classes (in fact, we are not aware of "natural" VC function classes that have infinite reduction dimensions). Below, we give some examples of reduction-dimension computations for linear classifiers, VC-1 classes, axis-aligned rectangles, and subsets of bounded size.

Linear classifiers. We start with linear separators in dimension d:

$$\mathcal{F}^d_{\text{lin}} := \left\{x \in \mathbb{R}^d \mapsto \mathbb{1}[a^\top x \geq b] : a \in \mathbb{R}^d, b \in \mathbb{R}\right\},$$

which have VC dimension $d + 1$. In the following, we show that linear separators have bounded reduction dimension. The proof uses oriented matroid arguments to translate the combinatorial counting problem into an algebraic problem. Intuitively, for the linear class $\mathcal{F}^d_{\text{lin}}$, the number of restricted function classes appearing in the definition of the reduction dimension, corresponds to the number of oriented matroids given by different choice of A's (the matroids are simply the signs patterns for the linear functions on the set of considered points). We next use a standard result, which states that these matroids are characterized by sign of the determinants on any $(d+1)$ points among these. Therefore, it suffices to count the number of topological components given by signs of polynomials (the determinants), which can be studied using Warren's theorem (Warren, 1968, Theorem 2), a classical result in algebraic geometry which bounds the number of topological components defined by polynomial equations.

Lemma 15 *Let $d \geq 1$. Then, for any $l \geq 1$, $\mathcal{F}^d_{\text{lin}}$ has l-reduction dimension at most $c_1 d^2 l$ for some universal constant $c_1 > 0$.*

Proof We refer to Björner et al. (1999) for a detailed exposition of oriented matroids. We will only use their standard properties. For any vector configuration $X = (x_1, \ldots, x_n)$ in $\mathbb{R}^{d+1}$ we consider the associated oriented matroid $M_X := ([n], \mathcal{L}_X)$ where $\mathcal{L}_X$ is the collection of all covectors of X generated by linear functionals:

$$\mathcal{L}_X = \left\{(\text{sgn}(y^\top x_i))_{i\in[n]} : y \in \mathbb{R}^d\right\}.$$

Here, $\operatorname{sgn} : \mathbb{R} \to \{+, -, 0\}$ is the sign function. Since the vectors lie in $\mathbb{R}^{d+1}$, the rank r of the matroid M_X is at most $d+1$ (the rank of a matroid is the unique length of maximal chains in $\mathcal{L}_X$, for the partial order given by the inclusion of supports, minus one). Additionally, the chirotope associated with the vector configuration X is given by

$$\chi_X : (i_1, \ldots, i_r) \in [n]^r \mapsto \operatorname{sgn}(\det(x_{i_1}, \ldots, x_{i_r})) \in \{+, -, 0\}.$$

Importantly, the matroid M_X is characterized by its chirotope χ_X (by default with the determinant we fixed a global sign orientation), e.g. see (Björner et al., 1999, Theorem 3.5.5 or Corollary 3.5.12). Crucially, this implies that given collection of all signs of determinants $\operatorname{sgn}(\det(x_{i_1}, \ldots, x_{i_{d+1}})) \in \{+, -, 0\}$ between any $d+1$ vectors of the configuration, we can recover the collection of their covectors $\mathcal{L}_X$.

We are now ready to bound the l-reduction dimension of linear separators. For convenience, for any point $z \in \mathbb{R}^d$ we will denote by $\tilde{z} = (z^\top, 1)^\top \in \mathbb{R}^{d+1}$ its homogeneisation. Note that the projection of linear separators onto a set of points can be obtained from the collection of covectors of its homogeneized vector configuration (the covectors store the information $\{+, -, 0\}$ while linear separators either merge $+$ and 0, or $-$ and 0). Further, for set of $l' \leq l$ datapoints $A = \{(a_i, y_i) : i \in [l']\}$, given the projection of linear separators onto $\{x_1, \ldots, x_n, a_1, \ldots, a_{l'}\}$ we can recover $(\mathcal{F}_{\text{lin}} \cap A)|_{\{x_1,\ldots,x_n\}}$, simply by focusing on projections whose value on a_i equals y_i for all $i \in [l']$. Altogether, this shows that

$$\begin{aligned}
&\left|\left\{(\mathcal{F}_{\text{lin}}^d \cap A)|_{\{x_1,\ldots,x_n\}} : A \subseteq \mathcal{X} \times \{0,1\}, |A| \leq l\right\}\right| \\
&\leq \sum_{l'=0}^{l} 2^{l'} \left|\left\{M_{(\tilde{x}_1,\ldots,\tilde{x}_n,\tilde{a}_1,\ldots,\tilde{a}_{l'})} : a_1, \ldots, a_{l'} \in \mathbb{R}^d\right\}\right| \\
&= \sum_{l'=0}^{l} 2^{l'} \left|\left\{(i_1, \ldots, i_{d+1}) \in [n+l']^{d+1} \mapsto \operatorname{sgn}(\det(z_{i_1}, \ldots, z_{i_{d+1}}))\right.\right. \\
&\qquad\qquad \left.\left. : Z = (\tilde{x}_1, \ldots, \tilde{x}_n, \tilde{a}_1, \ldots, \tilde{a}_{l'}),\ a_1, \ldots, a_{l'} \in \mathbb{R}^d\right\}\right|.
\end{aligned}$$

For a fixed $l' \in [l]$, denote by $N_{l'}$ the l'-th term within the sum. Note that for any indices $i_1, \ldots, i_{d+1} \in [n+l']$, the function $\det(z_{i_1}, \ldots, z_{i_{d+1}})$ is a polynomial in the variable $(a_1, \ldots, a_{l'}) \in (\mathbb{R}^d)^{l'}$ of degree at most $d+1$. Therefore, $N_{l'}$ exactly corresponds to the number of sign patterns for all equations of the type $\det(z_{i_1}, \ldots, z_{i_{d+1}}) = 0$ in $(\mathbb{R}^d)^{l'}$. We can then use a variant of Warren's theorem (Alon, 1995, Proposition 5.5) which bounds this number by $(8eMK/D)^D$, where M is the number of polynomials over D variables and K is their maximum degree, if $2M \geq D$. In our context, this gives

$$N_{l'} \leq \left(8e\frac{d+1}{dl'}\binom{n+l'}{d+1}\right)^{dl'}.$$

Plugging this into the previous bound shows that

$$\left|(\mathcal{F}_{\text{lin}}^d \cap A)|_{\{x_1,\ldots,x_n\}} : A \subseteq \mathcal{X} \times \{0,1\}, |A| \leq l\right| \leq 2^l (n+l)^{c_1 d^2 l},$$

for some constant $c_1 > 0$. In turn, this shows that $D_l \leq c_2 d^2 l$ for some universal constant $c_2 > 0$. ■

Function classes with VC dimension 1. Next, we check that VC-1 classes have bounded reduction dimension, using their convenient tree representation Ben-David (2015).

Proposition 16 *Let $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ be a function class with VC dimension 1. Then, for any $l \geq 1$, $\mathcal{F}$ has l-reduction dimension $l + \mathcal{O}(1)$.*

Proof Fix a function class $\mathcal{F}$ of VC dimension 1. From (Ben-David, 2015, Theorem 4) without loss of generality, we may consider that we are given a tree ordering $\preceq$ on $\mathcal{X}$ (that is, a partial order on $\mathcal{X}$ such that every initial segment $I_x = \{y \in \mathcal{X} : y \preceq x\}$ is linearly ordered) such that all functions $f \in \mathcal{F}$ is an initial segment with respect to $\preceq$, i.e., for any $x \preceq y \in \mathcal{X}$, if $f(y) = 1$ then $f(x) = 1$.

Consider any dataset $A \subseteq \mathcal{X} \times \{0,1\}$. Let B be the dataset obtained by deleting datapoints $(x,0) \in A$ such that (1) there exists $(z,0) \in A$ with $z \prec x$ or (2) there exists $(z,1) \in A$ with $z \not\prec x$; as well as deleting datapoints $(x,1) \in A$ such that there exists $(z,1) \in A$ with $x \prec z$. Note that $\mathcal{F} \cap A = \mathcal{F} \cap B$.

Fix test points $S = \{x_1, \ldots, x_n\} \in \mathcal{X}$. We focus on the tree ordering $\preceq$ restricted to ancestors of S, that is, $T := \{x \in \mathcal{X} : x \preceq y, y \in S\}$. Then, note that when restricted to S, the function class $\mathcal{F} \cap B|_S$ is equivalent to replacing each datapoint $(x,y) \in B$ with any $(\tilde{x}, y)$ where $\tilde{x}$ has the same $\preceq$ comparisons in terms of $\preceq$ with all datapoints in $S \cup \{z : (z,y) \in B\}$. We recall that by construction, all $(x,0) \neq (z,0) \in B$ are such that x and z are incomparable. Further, there is at most one datapoint $(x,1) \in B$ with label 1 and in that case, all other datapoints $(z,0)$ in B are such that $x \prec z$. Hence, we may choose one such representative for each possible $\preceq$-comparisons with S for all datapoints in B with label 0. Note that the number of such possible $\preceq$-comparisons is bounded by the number of possible nodes in a rooted tree with leaves within S: it is at most $2|S| = 2n$. For the potential label 1 we additionally need to respect the fact that $x \prec z$ for all $(z,0) \in B$: up to duplicating representatives this gives $4n$ possible choices. In summary,

$$|\{(\mathcal{F} \cap A)|_S : A \subseteq \mathcal{X} \times \{0,1\}, |A| \leq l\}| \leq 1 + \sum_{r \leq l} ((2n)^r + (2n)^{r-1} \cdot 4n) \lesssim n^l,$$

where the first term 1 counts the empty function class if the dataset is not realizable, the second (resp. last) term corresponds to datasets B containing only 0 labels (resp. at least one 1 label). This ends the proof. ■

Axis-aligned rectangles. We now turn to axis-aligned rectangles in $\mathbb{R}^d$ defined as follows:

$$\mathcal{R}^d := \left\{ B_{a,b} : x \in \mathbb{R}^d \mapsto \mathbb{1}[a \leq x \leq b] : a, b \in \mathbb{R}^d \right\},$$

also have finite reduction dimension. Here, inequalities between vectors are meant component-wise. That is, for $a, b \in \mathbb{R}^d$, $a \leq b$ when $a_i \leq b_i$ for all $i \in [d]$. We recall that axis-aligned rectangles have VC dimension $2d$.

Proposition 17 *For any $l \geq 1$, $\mathcal{R}^d$ has l-reduction dimension at most $4d + 1$.*

Proof Note that for any dataset $A \subseteq \mathbb{R}^d \times \{0,1\}$, the reduced function class $\mathcal{R}^d \cap A$ can be reduced to separate constraints on each coordinate $i \in [d]$ on the range allowed for the interval $[a_i, b_i]$ for functions $f_{a,b} \in \mathcal{R} \cap A$:

$$\mathcal{R} \cap A = \{f_{a,b} : \forall i \in [d], a_i \in (z_i^{(1)}, z_i^{(2)}], b_i \in [z_i^{(3)}, z_i^{(4)})\},$$

where $z_i^{(1)} = \max\{z_i : (z,0) \in A, \exists(\tilde{z},1) \in A, z_i < \tilde{z}_i\}$, $z_i^{(2)} = \min\{z_i : (z,1) \in A, \exists(\tilde{z},0) \in A, \tilde{z}_i < z_i\}$, and similarly for $z_i^{(3)}, z_i^{(4)}$, with the convention $\max \emptyset = -\infty$ and $\min \emptyset = +\infty$.

Hence, given n test points $S = \{x^1, \ldots, x^n\} \subseteq \mathbb{R}^d$, the projection $\mathcal{R} \cap A|_S$ only depends on the relative ordering of $z_i^{(1)} < z_i^{(2)} \leq z_i^{(3)} < z_i^{(4)}$ compared to $\{x_i^1, \ldots, x_i^n\}$, for each $i \in [n]$. This gives at most $(n+1)^4$ choices for each $i \in [n]$. In summary,

$$|\{(\mathcal{R}^d \cap A)|_S : A \subseteq \mathcal{X} \times \{0,1\}, |A| \leq l\}| \leq 1 + (n+1)^{4d},$$

where the additional 1 comes from the case when A is not realizable. Hence, $\mathcal{R}^d$ has l-reduction at most $4d+1$. ■

Subsets of bounded size. Another classical function class example is that of subsets of size at most d: for any instance space $\mathcal{X}$, we define

$$\mathcal{S}_{\mathcal{X}}^d = \{x \in \mathcal{X} \mapsto \mathbb{1}[x \in S] : S \subseteq \mathcal{X}, |S| \leq d\},$$

which has VC dimension d by construction, provided that $|\mathcal{X}| \geq d$. We can easily check that these have low reduction dimension.

Proposition 18 *For any $\mathcal{X}$ and $l \geq 1$, $\mathcal{S}_{\mathcal{X}}^d$ has l-reduction dimension $l + \mathcal{O}(\log d)$.*

Proof The main observation is that for any dataset A, $\mathcal{S}_{\mathcal{X}}^d \cap A$ fixes the value of the function on $A_x := \{x : (x,y) \in A\}$ and on $\mathcal{X} \setminus S$, it exactly corresponds to the function class $\mathcal{S}_{\mathcal{X} \setminus A_x}^{d-r}$ where $r = |\{(x,1) \in A\}|$ is the number of ones in the dataset (since r ones have been fixed, there are at most $d-r$ remaining). Hence, projected on test points $S = \{x_1, \ldots, x_n\}$, the possible l-reduced classes correspond to fixing the value on at most l points in $S \cap A_x$, and on the rest of the test points, the reduced class is characterized by a single number $d - r \leq d - p$ where $p = |\{(x,1) : x \in S \cap A_x\}|$ is the number of ones on the l datapoints coinciding with test points in S. Hence,

$$|\{(\mathcal{S}_X^d \cap A)|_S : A \subseteq \mathcal{X} \times \{0,1\}, |A| \leq l\}| \leq 1 + \binom{n}{l} \sum_{p=0}^{d} \binom{l}{p} (d-p+1) \leq 1 + edn^l$$

where the additional 1 corresponds to the case when A is not realizable. This ends the proof. ■

Appendix D. Proofs of Section 4

We detail the algorithm from Goel et al. (2023) here, which inspires our construction of weak learners. The intuitions behind this algorithm are detailed in Section 4.

Next, we start by proving Lemma 10, which controls the variance of estimation.

Lemma 10 *Fix $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$, a distribution μ on $\mathcal{X}$, $k \geq 1$, and $N \geq k^2$. Let $\eta = \frac{k^2}{N}$. Suppose that for some $c \geq 1$, we have $\rho_l(\mathcal{F}, \mu) \leq c \cdot \eta^l$ for all $l < k$. Then, $\sigma_k^N(\mathcal{F}, \mu) < \sqrt{3c \cdot \eta^k \rho_k(\mathcal{F}, \mu)}$.*

Algorithm 5: Algorithm for abstention online learning with knowledge of the distribution μ

Initialize $k \leftarrow d$ and $\mathcal{F}_1 = \mathcal{F}$
for $t = 1, \ldots, T$ *and while* $k > 1$ **do**
    Receive x_t
    if $\min\left\{\rho_k(\mathcal{F}_t^{x_t \to 1}, \mu),\ \rho_k(\mathcal{F}_t^{x_t \to 0}, \mu)\right\} \geq 0.6\, \rho_k(\mathcal{F}_t, \mu)$ **then** output $\hat{y}_t = \perp$;
    else predict $\hat{y}_t = \arg\max_{y \in \{0,1\}} \{\rho_k(\mathcal{F}_t^{x_t \to y}, \mu)\}$;
    Upon receiving label y_t, define $\mathcal{F}_{t+1} = \mathcal{F}_t^{x_t \to y_t}$
    if $\rho_k(\mathcal{F}_{t+1}, \mu) \leq T^{-k}$ **then** Set $k \leftarrow k - 1$;
if $k = 1$ *and* $x_t \in D(\mathcal{F}_t)$ **then** $\hat{y}_t = \perp$;
if $k = 1$ *and* $x_t \notin D(\mathcal{F}_t)$ **then** Predict the (unique) consistent label for $\hat{x}_t$;

Proof Recall that

$$\hat{\rho}_k^S(\mathcal{F}) := \frac{1}{\binom{N}{k}} \sum_{S' \subseteq S, |S'| = k} 1[S' \text{ is shattered by } \mathcal{F}].$$

First consider

$$\begin{aligned}
&\mathbb{P}\left(S_1, S_2 \text{ both shattered by } \mathcal{F}\right) \\
&= \mathbb{P}\left(S_1 \text{ is shattered by } \mathcal{F} \,|S_2 \text{ is shattered by } \mathcal{F}\right) \cdot \mathbb{P}\left(S_2 \text{ is shattered by } \mathcal{F}\right) \\
&\leq \mathbb{P}\left(S_1 \backslash S_2 \text{ is shattered by } \mathcal{F}\right) \cdot \mathbb{P}\left(S_2 \text{ is shattered by } \mathcal{F}\right)
\end{aligned}$$

where the last inequality uses the independence of "$S_1 \backslash S_2$ is shattered by $\mathcal{F}$" and "S_2 is shattered by $\mathcal{F}$". Recall that $\rho_r(\mathcal{F}, \mu)$ denotes the probability of a set with cardinality r being shattered, and since we are fixing $\mathcal{F}$ and μ here we can write ρ_r as a short cut for $\rho_r(\mathcal{F}, \mu)$. Given $N \geq 2k$, we have

$$\begin{aligned}
\mathbb{E}\left[\hat{\rho}_k^S(\mathcal{F})^2\right] &= \binom{N}{k}^{-2} \sum_{r=0}^{k} \sum_{\substack{S_1, S_2 \subseteq S \\ |S_1| = |S_2| = k \\ |S_1 \cap S_2| = r}} \mathbb{P}\big(S_1, S_2 \text{ are both shattered by } \mathcal{F}\big). \\
&= \binom{N}{k}^{-2} \sum_{r=0}^{k} \sum_{\substack{S_1, S_2 \subseteq S \\ |S_1| = |S_2| = k \\ |S_1 \cap S_2| = r}} \rho_k \rho_{k-r} \\
&= \binom{N}{k}^{-2} \sum_{r=0}^{k} \binom{N}{k} \binom{k}{r} \binom{N-k}{k-r} \rho_k \rho_{k-r} \\
&= \rho_k \left(\sum_{r=0}^{k} \frac{\binom{k}{r}\binom{N-k}{k-r}}{\binom{N}{k}} \rho_{k-r} \right).
\end{aligned}$$

Therefore

$$\text{Var}\left[\hat{\rho}_k^S(\mathcal{F})\right] \leq \rho_k \left(\sum_{r=1}^{k} \frac{\binom{k}{r}\binom{N-k}{k-r}}{\binom{N}{k}} \left(\rho_{k-r} - \rho_k\right) \right). \tag{5}$$

Consider a hypergeometric distribution $X \sim \text{Hypergeometric}(N, k, k)$ then

$$\text{Var}\left[\hat{\rho}_k^S(\mathcal{F})\right] \leq \rho_k \mathbb{E}\left[\rho_{k-X} \mathbb{1}(X \geq 1)\right].$$

For simplicity, write $\eta := k^2/N$. By assumption, $\rho_l \le c\eta^{l-(k-1)}$ for all $l \le k-1$. Therefore,

$$\mathrm{Var}\left[\hat{\rho}_k^S(\mathcal{F})\right] \le \rho_k \mathbb{E}\left[c\eta^{1-X}\right]. \tag{6}$$

Let $R \sim \mathrm{Binom}(k, \frac{k}{N})$. A standard comparison inequality for sampling with/without replacement (Hoeffding, 1963, Theorem 4) implies that $\mathbb{E}[\eta^{-X}] \le \mathbb{E}[\eta^{-R}]$. And hence

$$\mathbb{E}[\eta^{-X}] \le (1 - \frac{k}{N} + \frac{k}{N\eta})^k \le \exp\left(\left(\tfrac{1}{\eta} - 1\right)\frac{k^2}{N}\right). \tag{7}$$

Combining Eq. (6) and Eq. (7) we have

$$\mathrm{Var}\left[\hat{\rho}_k^S(\mathcal{F})\right] \le \rho_k c\eta \cdot \exp\left(\left(\tfrac{1}{\eta} - 1\right)\frac{k^2}{N}\right) < 3c\eta\rho_k.$$

This ends the proof. ■

The following lemma guarantees accurate estimation of shattering probabilities up to a (1 ± 0.2) factor, which can be obtained as an immediate result of Lemma 10 under the assumption that estimation error $\left|\rho_k^{\mathcal{S}}(\mathcal{F}) - \rho_k(\mathcal{F}, \mu)\right|$ is bounded by standard deviation $\sigma_k^N(\mathcal{F}, \mu)$. Nevertheless, it's useful to state it separately as it will be repeatedly used in the proof of Theorem 20.

Lemma 19 *Let $\mathcal{F} : \mathcal{X} \to \{0, 1\}$ and μ a distribution on $\mathcal{X}$ and let $\mathcal{S} = \{S_i\}_{i\in[m]}$ where each S_i contains N iid samples from μ. Let $2000d^2/N \le \epsilon < 1$. Assume that there is an integer $k \in [d]$ and constant c' such that for $0 \le l \le k-1$,*

$$\rho_l(\mathcal{F}, \mu) \le 1.2c'\epsilon^{l-(k-1)}.$$

Then, given

$$\left|\rho_k^{\mathcal{S}}(\mathcal{F}) - \rho_k(\mathcal{F}, \mu)\right| \le 2\sigma_k^N(\mathcal{F}, \mu) \tag{8}$$

we have

$$\left|\rho_k^{\mathcal{S}}(\mathcal{F}) - \rho_k(\mathcal{F}, \mu)\right| < 0.1\sqrt{c'\epsilon\rho_k(\mathcal{F}, \mu)}$$

which implies that, whenever $\rho_k^{\mathcal{S}}(\mathcal{F}) \ge c'\epsilon$, we have $\rho_k(\mathcal{F}, \mu) \in \rho_k^{\mathcal{S}}(\mathcal{F}) \cdot [0.8, 1.2]$ as well.

Proof Since $k^2/N \le \epsilon$, we have $\rho_l \le 1.2c'(k^2/N)^{l-(k-1)}$ for $0 \le l \le k-1$, by Lemma 10,

$$\mathrm{Var}\left[\hat{\rho}_k^{S_1}(\mathcal{F})\right] < 3.6(k^2/N)c'\rho_k(\mathcal{F}, \mu).$$

With condition (8), we have

$$\begin{aligned}\left|\rho_k^{\mathcal{S}}(\mathcal{F}) - \rho_k(\mathcal{F}, \mu)\right| &\le 2\sqrt{3.6(k^2/N)c'\rho_k(\mathcal{F}, \mu)} \\ &\le 2\sqrt{3.6\epsilon c'\rho_k(\mathcal{F}, \mu)/2000} \\ &< 0.1\sqrt{c'\epsilon\rho_k(\mathcal{F}, \mu)}.\end{aligned}$$

Therefore, whenever $\rho_k^{\mathcal{S}}(\mathcal{F}) \ge c'\epsilon$, we have

$$\left|\rho_k^{\mathcal{S}}(\mathcal{F}) - \rho_k(\mathcal{F}, \mu)\right| < 0.1\sqrt{\rho_k^{\mathcal{S}}(\mathcal{F})\rho_k(\mathcal{F}, \mu)},$$

which implies that $\rho_k(\mathcal{F},\mu) \in \rho_k^S(\mathcal{F}) \cdot [0.8, 1.2]$ as well. ■

In the adaptive case, the condition Eq. (8) does not naturally hold—as discussed in Section 4, the function class $\mathcal{F}$ can be badly correlated with $\mathcal{S}$. We next prove Lemma 11, which resolves this issue by showing universal concentration over all possible function classes of the form $\mathcal{F} \cap A$ where $\mathcal{F}$ is the initial hypothesis class and A can be any subset of less than l data points in $\mathcal{X} \times \{0,1\}$.

Proof of Lemma 11 The first step of the proof involves lifting the problem from $\mathcal{X}$ to $\mathcal{X}_s := \{S \subseteq \mathcal{X}, |S| = s\}$ then applying standard uniform concentration bounds on a carefully chosen VC class $\mathcal{G} : \mathcal{X}_s \to \{0,1\}$. First, we denote by μ_s the distribution of $\{x_i, i \in [s]\}$ where $(x_i)_{i\in[s]} \overset{iid}{\sim} \mu$ are i.i.d. samples from μ. We then define the function class $\mathcal{G} : \mathcal{X}_s \to \{0,1\}$ as the collection of all functions of the form

$$\phi_{A,I} : S \in \mathcal{X}_s \mapsto \mathbb{1}_I[\hat\rho_k^S(\mathcal{F} \cap A)],$$

for any subset $A \subseteq \mathcal{X} \times \{0,1\}$ with $|A| \leq l$, and any interval $I \subseteq \mathbb{R}_+$.

We now bound the VC dimension of this function class $\mathcal{G}$. Fix any test sets $S_1,\ldots,S_m \in \mathcal{X}_s$ for $m \geq D_l/s$. Note that given values $z_1,\ldots,z_m \in \mathbb{R}_+$, the set of possible projections of functions $\mathbb{1}_I[\cdot]$ for closed intervals $I \subseteq \mathbb{R}_+$ onto $\{z_1,\ldots,z_m\}$ is upper bounded by $\binom{m+1}{2}+1$—we can order these elements by increasing order then decide of the start and end point of those which belong to the interval I. In particular,

$$\begin{aligned}
\left|\mathcal{G}|_{\{S_1,\ldots,S_m\}}\right| &\leq \left|\left\{(\hat\rho_k^{S_i}(\mathcal{F}\cap A))_{i\in[m]} : A \subseteq \mathcal{X}\times\{0,1\}, |A|\leq l\right\}\right| \cdot m(m+1) \\
&\overset{(i)}{\leq} |\{(\mathcal{F}\cap A)|_{S_1\cup\ldots\cup S_m} : A \subseteq \mathcal{X}\times\{0,1\}, |A|\leq l\}| \cdot m(m+1) \\
&\overset{(ii)}{\leq} (ms)^{D_l} \cdot m(m+1).
\end{aligned}$$

In (i) we noted that given the projection of the function class $\mathcal{F} \cap A$ onto $S_1 \cup \ldots \cup S_m$, we can construct all estimates of the form $\hat\rho_k^{S_1}(\mathcal{F}\cap A),\ldots,\hat\rho_k^{S_m}(\mathcal{F}\cap A)$. Indeed, $\hat\rho_k^S(\mathcal{F})$ only evaluates the function class $\mathcal{F}$ on points in S. In (ii) we used the definition of the restriction dimension together with $ms \geq D_l$. Recall that if $\{S_1,\ldots,S_m\}$ are shattered by $\mathcal{G}$, then $\left|\mathcal{G}|_{\{S_1,\ldots,S_m\}}\right| = 2^m$. Together with the previous bound, this shows that $\mathcal{G}$ has VC dimension at most $c_0 D_l \log(sD_l)$ for some $c_0 \geq 1$.

We can then apply standard VC uniform concentration bounds on $\mathcal{G}$, e.g. (Devroye et al., 2013, Theorem 12.5), which shows that there is a universal constant c_1 such that the following holds for any $m \geq c_1(D_l \log(sD_l) + \log\frac{1}{\delta})$. Let $S_1,\ldots,S_m \overset{iid}{\sim} \mu_s$, with probability at least $1-\delta$, for any subset $A \subseteq \mathcal{X}\times\{0,1\}$ with $|A| \leq l$ and interval $I \subseteq \mathbb{R}_+$,

$$\left|\frac{1}{m}\left|\left\{i \in [m] : \hat\rho_k^{S_i}(\mathcal{F}\cap A) \in I\right\}\right| - \mathbb{P}_{S\sim\mu_s}[\hat\rho_k^{S_i}(\mathcal{F}\cap A) \in I]\right| \leq \frac{1}{8}. \tag{9}$$

Now note that for any $A \subseteq \mathcal{X}\times\{0,1\}$, by Chebyshev's inequality we have

$$\mathbb{P}_{S\sim\mu_s}\left(|\hat\rho_k^S(\mathcal{F}\cap A) - \rho_k(\mathcal{F}\cap A,\mu)| \geq 2\sigma_k^s(\mathcal{F}\cap A,\mu)\right) \leq \frac{1}{4}.$$

In particular, for the interval $I_A := \rho_k(\mathcal{F}\cap A,\mu) + 2\sigma_k^s(\mathcal{F}\cap A,\mu)\cdot[-1,1]$, this precisely shows $\mathbb{P}_{S\sim\mu_s}[\hat\rho_k^{S_i}(\mathcal{F}\cap A) \in I_A] \geq \frac{3}{4}$. Together with Eq. (9), we obtained with probability at least $1-\delta$, for any set A of at most l datapoints,

$$\left|\left\{i \in [m] : |\hat{\rho}_k^{S_i}(\mathcal{F} \cap A) - \rho_k(\mathcal{F} \cap A, \mu)| \leq 2\sigma_k^s(\mathcal{F} \cap A, \mu)\right\}\right| > \frac{m}{2}.$$

In turn, this implies that the median value of $\hat{\rho}_k^{S_i}(\mathcal{F} \cap A)$ for $i \in [m]$ belongs to the desired interval:

$$|\rho_k^{\mathcal{S}}(\mathcal{F} \cap A) - \rho_k(\mathcal{F} \cap A, \mu)| \leq 2\sigma_k^s(\mathcal{F} \cap A, \mu).$$

This ends the proof. ■

We are now ready to prove the main guarantee for the weak learner WL$(\mathcal{T}, z)$ with appropriate parameters, whose formal and complete version is given below.

Theorem 20 *There exists a universal constant $c_0 > 0$ such that the following holds. Fix $\epsilon \in (0,1)$. Let $\mathcal{F} \subseteq \{0,1\}^{\mathcal{X}}$ be a function class of VC dimension d and fix an adversary. Let $N = \lceil 2000d^2/\epsilon \rceil$. Let $\mathcal{T}$ denote the (potentially random) set of first mN non-corrupted rounds and define $z := (y_t)_{t\in\mathcal{T}}$.*

If (1) the adversary is oblivious and `update` $=$ `always`*, $m \geq 8\log(dT/\epsilon)$, or (2) the adversary is adaptive and* `update` $=$ `restricted`*, $m \geq c_0(D\log(D) + 8D + 3\log(d/\epsilon))$ where D is the $\lceil 5d^2 \log \frac{1}{\epsilon} \rceil$-reduction dimension of $\mathcal{F}$, then*

$$\text{MISERR}(\mathcal{T}, z) < 5d^2 \log \frac{1}{\epsilon} \quad \textit{and} \quad \mathbb{E}[\text{ABSERR}(\mathcal{T}, z)] \leq 18\epsilon T,$$

where $\text{MISERR}(\mathcal{T}, z)$ and $\text{ABSERR}(\mathcal{T}, z)$ respectively denote the misclassification and abstention error of the weak learner WL$(\mathcal{T}, z)$ as defined in Algorithm 1.

Proof Before analyzing the miclassification and abstention error, we make a few remarks. First, note that for the considered subset $\mathcal{T}$, the samples $(x_t)_{t\in\mathcal{T}}$ are i.i.d. sampled from μ. Further, by construction, since these are the first non-corrupted times and the vector z contains their correct labels, in the interval of time $[\max \mathcal{T}]$ the learner WL$(\mathcal{T}, z)$ makes no classification nor abstention mistakes. As a result, it suffices to focus on the period of time $(\max \mathcal{T}, T]$.

For each $t \in \{\max \mathcal{T} + 1, \ldots, T\}$, we define the set of function classes encountered by the learner via

$$\mathcal{G} := \left\{\mathcal{F}_t, \mathcal{F}_t^{x_t \to 0}, \mathcal{F}_t^{x_t \to 1} : t \in (\max \mathcal{T}, T]\right\}.$$

Note that these are random function classes. Next, we define the event

$$\mathcal{E} := \left\{\forall \tilde{\mathcal{F}} \in \mathcal{G}, k \in [d], \quad \left|\rho_k^{\mathcal{S}}(\tilde{\mathcal{F}}) - \rho_k(\tilde{\mathcal{F}}, \mu)\right| \leq 2\sigma_k^N(\tilde{\mathcal{F}})\right\}. \tag{10}$$

For now, let us assume $\mathbb{P}(\mathcal{E}^c) \leq 3\epsilon$. We will verify this assumption later for both oblivious and adaptive adversaries, after analyzing the misclassification and abstention error.

Misclassification error. First, we analyze the misclassification error. By construction of weak learners, misclassification errors only occur at times $t \in [T]$ when $\rho_1^{\mathcal{S}}(\mathcal{F}_t) > \epsilon$ and $\min_{y\in\{0,1\}} \rho_{k_t}^{\mathcal{S}}(\mathcal{F}_t^{x_t \to y}) < 0.9\rho_{k_t}^{\mathcal{S}}(\mathcal{F}_t)$, where k_t is defined as in line 4 of Algorithm 1. Thus, since we predict with the label corresponding to $\arg\max_{y\in\{0,1\}} \rho_{k_t}^{\mathcal{S}}(\mathcal{F}_t^{x_t \to y})$ and a mistake occurred, we have

$$\rho_{k_t}^{\mathcal{S}}(\mathcal{F}_{t+1}) = \rho_{k_t}^{\mathcal{S}}(\mathcal{F}_t^{x_t \to y_t}) < 0.9\rho_{k_t}^{\mathcal{S}}(\mathcal{F}_t).$$

Meanwhile, notice that combining $\rho_1^S(\mathcal{F}_t) > \epsilon$ and $\forall l \in [k_t], \rho_l^S(\mathcal{F}_t) > \epsilon \cdot \rho_{l-1}^S(\mathcal{F}_t)$, we have

$$\rho_{k_t}^S(\mathcal{F}_t) > \epsilon \cdot \epsilon^{k_t - 1} = \epsilon^{k_t}.$$

For each fixed $k \in [d]$, $\rho_k(\mathcal{F}_t)$ is non-increasing since the $\mathcal{F}_t$ are non-increasing throughout the learning procedure. In particular, $\rho_k^S(\mathcal{F}_t)$ can be decreased by a factor $9/10$ at most $k \log_{\frac{10}{9}} \left(\frac{1}{\epsilon}\right)$ times. Hence

$$\textsc{MisErr}(\mathcal{T}, z) \leq \sum_{k=1}^{d} k \log_{\frac{10}{9}} \left(\frac{1}{\epsilon}\right) < 5d^2 \log\left(\frac{1}{\epsilon}\right).$$

Note that this result on misclassification error holds *regardless of* event $\mathcal{E}$.

Abstention error. Consider a time $t \in [\max \mathcal{T} + 1, T]$ for which k_t was defined, that is $\rho_1^S(\mathcal{F}_t) > \epsilon$. We start by proving by induction that under event $\mathcal{E}$,

$$\rho_{k'}(\mathcal{F}_t, \mu) \in \rho_{k'}^S(\mathcal{F}_t) \cdot [0.8, 1.2], \quad \forall 0 \leq k' \leq k_t. \tag{11}$$

Note that the base case $k' = 0$ is true because $\rho_0(\mathcal{F}_t, \mu) = \rho_0^S(\mathcal{F}_t) = 1$. Suppose that for some $k' \in [k_t]$ this holds for all $0 \leq l \leq k' - 1$. By construction of the weak learner, we have

$$\rho_l^S(\mathcal{F}_t) \leq \rho_{k'-1}^S(\mathcal{F}_t)\epsilon^{l-(k'-1)}, \quad 0 \leq l \leq k' - 1.$$

Together with the induction hypothesis, this implies

$$\rho_l(\mathcal{F}_t, \mu) \leq 1.2\rho_l^S(\mathcal{F}_t) \leq 1.2\rho_{k'-1}^S(\mathcal{F}_t)\epsilon^{l-(k'-1)}, \quad 0 \leq l \leq k' - 1. \tag{12}$$

Moreover, the event $\mathcal{E}$ implies that

$$\left|\rho_{k'}^S(\mathcal{F}_t) - \rho_{k'}(\mathcal{F}_t, \mu)\right| \leq 2\sigma_{k'}^N(\mathcal{F}_t). \tag{13}$$

Also, recall that by construction, we have $\rho_{k'}^S(\mathcal{F}_t) > \epsilon\rho_{k'-1}^S(\mathcal{F}_t)$. Note that this also holds when $k' = 1$ since by construction $\rho_1^S(\mathcal{F}_t) \geq \epsilon$ and $\rho_0^S(\mathcal{F}_t) = 1$. Altogether, we can now apply Lemma 19 with the parameter $c' = \rho_{k'-1}^S(\mathcal{F}_t)$, which gives the desired induction

$$\rho_{k'}(\mathcal{F}_t, \mu) \in \rho_{k'}^S(\mathcal{F}_t) \cdot [0.8, 1.2].$$

This ends the proof of Eq. (11).

Using the same arguments, we can check that Eqs. (12) and (13) still hold for $k' = k_t + 1$. Hence, applying Lemma 19 gives

$$\begin{aligned}\rho_{k_t+1}(\mathcal{F}_t, \mu) &< \rho_{k_t+1}^S(\mathcal{F}_t) + 0.1\sqrt{\epsilon\rho_{k_t}^S(\mathcal{F}_t)\rho_{k_t+1}(\mathcal{F}_t, \mu)} \\ &\overset{(i)}{\leq} \epsilon\rho_{k_t}^S(\mathcal{F}_t) + 0.1\sqrt{\epsilon\rho_{k_t}^S(\mathcal{F}_t)\rho_{k_t+1}(\mathcal{F}_t, \mu)},\end{aligned}$$

where in (i) we used the definition of k_t in line 4 of Algorithm 1: either $k_t = d$ in which case we already have $\rho_{k_t+1}^S(\mathcal{F}_t) = 0$ or $k_t < d$ in which case we must have $\rho_{k_t+1}^S(\mathcal{F}_t) \leq \epsilon\rho_{k_t}^S(\mathcal{F}_t)$. Consequently,

$$\rho_{k_t+1}(\mathcal{F}_t, \mu) \leq 1.2\epsilon\rho_{k_t}^S(\mathcal{F}_t). \tag{14}$$

Next, we aim to show that for $y \in \{0,1\}$, given $\rho^{\mathcal{S}}_{k_t}(\mathcal{F}_t^{x_t\to y}) \geq 0.9\rho^{\mathcal{S}}_{k_t}(\mathcal{F}_t)$, we have

$$\rho_{k_t}(\mathcal{F}_t^{x_t\to y},\mu) \in \rho^{\mathcal{S}}_{k_t}(\mathcal{F}_t^{x_t\to y})\cdot[0.8,1.2]. \tag{15}$$

To see this, first note that for any $l \in [0,k_t]$ we have

$$\rho_l(\mathcal{F}_t^{x_t\to y},\mu) \leq \rho_l(\mathcal{F}_t,\mu) \leq 1.2\rho^{\mathcal{S}}_l(\mathcal{F}_t) \leq 1.2(\rho^{\mathcal{S}}_{k_t}(\mathcal{F}_t)/\epsilon)\cdot\epsilon^{l-(k_t-1)}.$$

Next, since the event $\mathcal{E}$ holds,

$$\left|\rho^{\mathcal{S}}_{k_t}(\mathcal{F}_t^{x_t\to y}) - \rho_{k_t}(\mathcal{F}_t^{x_t\to y},\mu)\right| \leq 2\sigma^N_{k_t}(\mathcal{F}_t^{x_t\to y}).$$

By Lemma 19 with $c' = \rho^{\mathcal{S}}_{k_t}(\mathcal{F}_t)/\epsilon$,

$$\begin{aligned}\left|\rho^{\mathcal{S}}_{k_t}(\mathcal{F}_t^{x_t\to y}) - \rho_{k_t}(\mathcal{F}_t^{x_t\to y},\mu)\right| &< 0.1\sqrt{\rho^{\mathcal{S}}_{k_t}(\mathcal{F}_t)\rho_{k_t}(\mathcal{F}_t^{x_t\to y},\mu)}\\ &\leq 0.1\sqrt{\frac{10}{9}\rho^{\mathcal{S}}_{k_t}(\mathcal{F}_t^{x_t\to y})\rho_{k_t}(\mathcal{F}_t^{x_t\to y},\mu)}\end{aligned}$$

which implies $\rho_{k_t}(\mathcal{F}_t^{x_t\to y},\mu) \in \rho^{\mathcal{S}}_{k_t}(\mathcal{F}_t^{x_t\to y})\cdot[0.8,1.2]$, ending the proof of Eq. (15).

We are now ready to bound the abstention error. There are two possible cases for an abstention error: When $\rho^{\mathcal{S}}_1(\mathcal{F}_t) \leq \epsilon$, since $x_t \notin D(\mathcal{F}_t)$ implies $\hat{y}_t = y_t$, an error at time t can only happen when $x_t \sim \mu$ and $x_t \in D(\mathcal{F}_t)$, as in line 11 of Algorithm 1. When $\rho^{\mathcal{S}}_1(\mathcal{F}_t) > \epsilon$, an abstention error occurs if and only if the adversary has decided not to corrupt at time t (i.e. $c_t = 0$) yet the learner still abstains due to $\min_{y\in\{0,1\}}\rho^{\mathcal{S}}_{k_t}(\mathcal{F}_t^{x_t\to y}) \geq 0.9\rho^{\mathcal{S}}_{k_t}(\mathcal{F}_t)$. Let $\mathcal{H}_t$ denote the history before time t, hence $c_t \in \mathcal{H}_t$. Combining the two cases, the probability of an abstention error at time t is bounded by

$$\begin{aligned}P_t &:= \mathbb{P}(x_t \in D(\mathcal{F}_t), \rho^{\mathcal{S}}_1(\mathcal{F}_t) \leq \epsilon, c_t = 0)\\ &\qquad + \mathbb{P}(\min_{y\in\{0,1\}}\rho^{\mathcal{S}}_{k_t}(\mathcal{F}_t^{x_t\to y}) \geq 0.9\rho^{\mathcal{S}}_{k_t}(\mathcal{F}_t), \rho^{\mathcal{S}}_1(\mathcal{F}_t) > \epsilon, c_t = 0)\\ &\leq \mathbb{P}(x_t \in D(\mathcal{F}_t), \rho^{\mathcal{S}}_1(\mathcal{F}_t) \leq \epsilon, \mathcal{E}|c_t = 0) + \mathbb{P}(\min_{y\in\{0,1\}}\rho^{\mathcal{S}}_{k_t}(\mathcal{F}_t^{x_t\to y}) \geq 0.9\rho^{\mathcal{S}}_{k_t}(\mathcal{F}_t), \mathcal{E}, c_t = 0) + 2\mathbb{P}(\mathcal{E}^c)\end{aligned} \tag{16}$$

Applying Lemma 19 with $k = 1$ and $c' = 1$ yields that, under event $\mathcal{E}$,

$$\left|\rho^{\mathcal{S}}_1(\mathcal{F}_t) - \rho_1(\mathcal{F}_t,\mu)\right| < 0.1\sqrt{\epsilon\rho_1(\mathcal{F}_t,\mu)}.$$

Given $\rho^{\mathcal{S}}_1(\mathcal{F}_t) \leq \epsilon$, we must have $\rho_1(\mathcal{F}_t,\mu) \leq 1.2\epsilon$. Otherwise, if $\rho_1(\mathcal{F}_t,\mu) > 1.2\epsilon$,

$$\frac{5}{6} > \frac{\rho^{\mathcal{S}}_1(\mathcal{F}_t)}{\rho_1(\mathcal{F}_t,\mu)} > 1 - 0.1\sqrt{\frac{\epsilon}{\rho_1(\mathcal{F}_t,\mu)}} \geq 1 - \frac{0.1}{\sqrt{1.2}} \approx 0.91,$$

which is clearly a contradiction. Therefore,

$$\begin{aligned}\mathbb{P}(x_t \in D(\mathcal{F}_t), \rho^{\mathcal{S}}_1(\mathcal{F}_t) \leq \epsilon, \mathcal{E}|c_t = 0) &\leq \mathbb{P}(x_t \in D(\mathcal{F}_t), \rho_1(\mathcal{F}_t,\mu) \leq 1.2\epsilon|c_t = 0)\\ &\leq \mathbb{P}(x_t \in D(\mathcal{F}_t)\,|\rho_1(\mathcal{F}_t,\mu) \leq 1.2\epsilon, c_t = 0)\\ &\leq 1.2\epsilon.\end{aligned} \tag{17}$$

where the last inequality is because the distribution of x_t given $c_t = 0$ is independent of $\mathcal{H}_t$.

Now we turn to the second term in Eq. (16). For $y \in \{0,1\}$, since $\rho^{\mathcal{S}}_{k_t}(\mathcal{F}_t^{x_t \to y}) \geq 0.9\rho^{\mathcal{S}}_{k_t}(\mathcal{F}_t)$, from the previous discussions we have that under $\mathcal{E}$,

$$\rho_{k_t}(\mathcal{F}_t^{x_t \to y}, \mu) \in \rho^{\mathcal{S}}_{k_t}(\mathcal{F}_t^{x_t \to y}) \cdot [0.8, 1.2],$$

and by Eq. (11), $\rho_{k_t}(\mathcal{F}_t, \mu) \leq 1.2\rho^{\mathcal{S}}_{k_t}(\mathcal{F}_t)$. Hence event $\left\{\min_{y\in\{0,1\}} \rho^{\mathcal{S}}_{k_t}(\mathcal{F}_t^{x_t \to y}) \geq 0.9\rho^{\mathcal{S}}_{k_t}(\mathcal{F}_t)\right\} \cap \mathcal{E}$ implies

$$\min_{y\in\{0,1\}} \rho_{k_t}(\mathcal{F}_t^{x_t \to y}, \mu) \geq 0.8 \min_{y\in\{0,1\}} \rho^{\mathcal{S}}_{k_t}(\mathcal{F}_t^{x_t \to y}) \geq 0.72\rho^{\mathcal{S}}_{k_t}(\mathcal{F}_t) \geq 0.6\rho_{k_t}(\mathcal{F}_t, \mu).$$

Consequently,

$$\begin{aligned}
&\mathbb{P}(\min_{y\in\{0,1\}} \rho^{\mathcal{S}}_{k_t}(\mathcal{F}_t^{x_t \to y}) \geq 0.9\rho^{\mathcal{S}}_{k_t}(\mathcal{F}_t), \mathcal{E}, c_t = 0) \\
&\leq \mathbb{P}\left(\min_{y\in\{0,1\}} \rho_{k_t}(\mathcal{F}_t^{x_t \to y}, \mu) \geq 0.6\rho_{k_t}(\mathcal{F}_t, \mu), c_t = 0\right) \\
&\leq \mathbb{P}_{x\sim\mu}\left(\min_{y\in\{0,1\}} \rho_{k_t}(\mathcal{F}_t^{x \to y}, \mu) \geq 0.6\rho_{k_t}(\mathcal{F}_t, \mu)\right).
\end{aligned}$$

Note that here $\mathbb{P}_{x\sim\mu}$ is taking expectation with respect to both x and $\mathcal{H}_t$, with the distribution of $x \sim \mu$ being independent of $\mathcal{H}_t$. Further, from Goel et al. (2023, Lemma 4.2), or Eq. (1), we have

$$\mathbb{P}_{x\sim\mu}\left(\min_{y\in\{0,1\}} \rho_{k_t}(\mathcal{F}_t^{x \to y}, \mu) \geq 0.6\rho_{k_t}(\mathcal{F}_t, \mu) \,|\, \mathcal{H}_t\right) \leq \min\left\{\frac{5\rho_{k_t+1}(\mathcal{F}_t)}{\rho_{k_t}(\mathcal{F}_t)}, 1\right\}.$$

Note, that condition on $\mathcal{H}_t$, the only randomness in the above expression comes from x. Combining Eq. (17) with the last two inequalities yields

$$\begin{aligned}
P_t &\leq 1.2\epsilon + \mathbb{E}\left[\min\left\{\frac{5\rho_{k_t+1}(\mathcal{F}_t)}{\rho_{k_t}(\mathcal{F}_t)}, 1\right\}\right] + 2\mathbb{P}(\mathcal{E}^c) \\
&\leq 1.2\epsilon + \mathbb{E}\left[\frac{5\rho_{k_t+1}(\mathcal{F}_t)}{\rho_{k_t}(\mathcal{F}_t)}\mathbb{1}(\mathcal{E})\right] + 3\mathbb{P}(\mathcal{E}^c) \\
&\leq 1.2\epsilon + \mathbb{E}\left[\frac{7.5\rho^{\mathcal{S}}_{k_t+1}(\mathcal{F}_t, \mu)}{\rho^{\mathcal{S}}_{k_t}(\mathcal{F}_t)}\right] + 3\mathbb{P}(\mathcal{E}^c) \leq 17.7\epsilon.
\end{aligned}$$

where the third inequality follows from Eqs. (11) and (14), and the last inequality follows by $\mathbb{P}(\mathcal{E}^c) \leq 3\epsilon$ and the definition of k_t. Consequently,

$$\mathbb{E}[\textsc{AbsErr}(\mathcal{T}, z)] \leq \sum_{t=1}^{T} P_t \leq 17.7\epsilon T.$$

Finally, we need to verify the assumption that $\mathbb{P}(\mathcal{E}^c) \leq 3\epsilon$ for oblivious and adaptive adversaries. Formally, our goal is to show that, with at most 3ϵ probability of failure,

$$\forall \tilde{\mathcal{F}} \in \mathcal{G}, k \in [d], \quad \left|\rho^{\mathcal{S}}_k(\tilde{\mathcal{F}}) - \rho_k(\tilde{\mathcal{F}}, \mu)\right| \leq 2\sigma^N_k(\tilde{\mathcal{F}}). \tag{18}$$

We will separately show Eq. (18) for oblivious and adaptive adversaries.

Verifying $\mathbb{P}(\mathcal{E}^c) \leq 3\epsilon$ for oblivious adversaries. When the adversary is oblivious, we may consider without loss of generality that the sequence of corrupted times as well as the corrupted samples are deterministic. Also, since in oblivious setting we let the algorithm update $\mathcal{F}_t$ at every time $t > \max \mathcal{T}$, we have $\mathcal{F}_t = \mathcal{F} \cap \{(x_\tau, y_\tau)\}_{\tau=\max \mathcal{T}+1}^{t-1}$. Apart from the fixed corrupted samples, the sequence $(x_t)_{t=\max \mathcal{T}+1}^{T}$ is independent of $(x_t)_{t\in\mathcal{T}}$ and as a result, $\mathcal{G}$ is independent of $(x_t)_{t\in\mathcal{T}}$. Also recall that by definition, for any given function class $\tilde{\mathcal{F}}$ and $k \in [d]$, the estimator $\hat{\rho}_k^{S_1}(\tilde{\mathcal{F}})$ is unbiaised for $\rho_k(\tilde{\mathcal{F}}, \mu)$. Hence, using concentration bounds on the median in Theorem 14 with $\delta = \epsilon/dT$ and the union bound yields

$$\mathbb{P}(\mathcal{E}^c) = \mathbb{E}\left[\mathbb{P}(\mathcal{E}^c \mid \mathcal{G})\right] \leq \frac{\epsilon}{dT}|\mathcal{G}| \leq 3\epsilon.$$

Verifying $\mathbb{P}(\mathcal{E}^c) \leq \epsilon$ for adaptive adversaries. For adaptive adversaries, the argument is more complex. This is because the samples $\mathcal{S}$ used for estimation (which are basically samples in $(x_t)_{t\in\mathcal{T}}$) are potentially dependent on the sequence $(x_t)_{t=\max \mathcal{T}+1}^{T}$, hence we can no longer assume samples in $\mathcal{S}$ are independent $\mathcal{G}$. Despite not having this convenience, we can utilize Lemma 11 which shows universal concentration over all possible function classes of the form $\mathcal{F} \cap A$ where $\mathcal{F}$ is the initial hypothesis class and A can be any subset of less than l data points in $\mathcal{X} \times \{0,1\}$.

For a run of $\text{WL}(\mathcal{T}, z)$, define the set of times

$$\mathcal{M} := \{t \in (\max \mathcal{T}, T] : \text{WL}(\mathcal{T}, z) \text{ makes a mistake at time } t\}.$$

Since in adaptive setting we only let the algorithm update $\mathcal{F}_t$ when there is a misclassification error, each function class $\tilde{\mathcal{F}} \in \mathcal{G}$ is of the form $\mathcal{F} \cap A$ for some $A = \{(x_\tau, y_\tau) : \tau \in \mathcal{M}, \tau < t\}$, or $A = \{(x_\tau, y_\tau) : \tau \in \mathcal{M}, \tau < t\} \cap (x_t, z)$ for $z \in \{0,1\}$. Either way, $|A| \leq |\mathcal{M}| + 1$.

From the analysis of Misclassification error, $|\mathcal{M}| < 5d^2 \log(1/\epsilon)$ almost surely. Therefore, let $N = \lceil 2000d^2/\epsilon \rceil$, $\delta = \epsilon/d$ and $l = \lceil 5d^2 \log(1/\epsilon) \rceil$ in Lemma 11. Then, one can check the condition on m for Lemma 11 is satisfied:

$$\lceil c_0(D \log(ND) + \log(1/\delta)) \rceil \leq \lceil c_0(D \log(D) + 8D + 3\log(d/\epsilon)) \rceil \leq m.$$

Hence for each $k \in [d]$, with probabilty at least $1 - \epsilon/d$,

$$|\rho_k^{\mathcal{S}}(\mathcal{F} \cap A) - \rho_k(\mathcal{F} \cap A, \mu)| \leq 2\sigma_k^N(\mathcal{F} \cap A, \mu), \quad A \subseteq \mathcal{X} \times \{0,1\},\ |A| \leq \left\lceil 5d^2 \log(1/\epsilon) \right\rceil.$$

In particular, every function class in $\mathcal{G}$ takes the form $\mathcal{F} \cap A$ for some set of data points A with $|A| \leq \lceil 5d^2 \log(1/\epsilon) \rceil$. Hence

$$\forall \tilde{\mathcal{F}} \in \mathcal{G}, \quad \left|\rho_k^{\mathcal{S}}(\tilde{\mathcal{F}}) - \rho_k(\tilde{\mathcal{F}}, \mu)\right| \leq 2\sigma_k^N(\tilde{\mathcal{F}}).$$

Union bounding over all $k \in [d]$ finishes the proof. ■

Appendix E. Proofs of Section 5

We start by proving the learning guarantee for the DELETE algorithm which will be a subroutine within the complete boosting procedure.

Lemma 21 *Fix $C \geq 0$ and $M \geq 1$. Consider an adversary for L weak learners which includes at most M mistakes and at most C predictions per round. Denote by $\hat{y}_t(s) \in \{0, 1, \perp\}$ the value selected by* $\text{DELETE}_{s,\frac{C}{2}}$ *at iteration t. Then, for any $s_{\max} \geq 1$ and defined iteration $n \geq 1$, there exists $s \in \{0, \ldots, s_{\max}\}$ such that*

$$|\{t \in [n] : \hat{y}_t(s) \notin \{y_t, \perp\}\}| \leq \frac{8Mn \lceil \log L \rceil}{s_{\max} + 1}.$$

Proof For each weak learner $i \in [L]$ we denote by $r_i = |\{t \in [n] : z_{i,t} \neq \perp\}|$ its total number of predictions. Note that if $r_i = 0$ for all $i \in [L]$ then weak learners always abstain and hence the desired result is immediate: the deletion algorithms also always abstain. We suppose this is not the case from now. We fix a deletion parameter $s \geq 0$ and denote by $N(s) = |\{t \in [n] : \hat{y}_t(s) \notin \{y_t, \perp\}\}|$ the number of mistakes of $\text{DELETE}_{s,\frac{C}{2}}$. We consider the random variable

$$r := r_k \quad \text{where} \quad k \sim \text{Unif}(\{i \in [L] : r_i > 0\})$$

is sampled uniformly among weak learners with at least one prediction. Next, for any $l \geq 1$ we define q_l to be the 2^{-l}-quantile of r, that is, $q_l \in \mathbb{N}$ and

$$\mathbb{P}[r \geq q_l] \geq 2^{-l} > \mathbb{P}[r > q_l].$$

We also pose $q_0 = 0$. Note that since there are at most L rows, all quantiles q_l are equal for $l \geq \log L$ to $r_{\max} := \max_{i \in [L]} r_i$. We first consider the case when $s \geq r_{\max}$. In that case, after deletions, we have $\tilde{z}_{a,b} = \perp$ for all $(a, b) \in [n] \times [L]$ hence $N(s) = 0$. On the other hand, for $l \in [\lceil \log L \rceil]$, note that for any integer $s \in [q_{l-1}, q_l)$, we have

$$\mathbb{E}[r - s \mid r > s] \geq (q_l - s)\mathbb{P}[r \geq q_l \mid r > s] \geq (q_l - s)\frac{\mathbb{P}[r \geq q_l]}{\mathbb{P}[r > q_{l-1}]} \geq \frac{q_l - s}{2}.$$

In particular, this shows that

$$\begin{aligned} Cn &\overset{(i)}{\geq} \sum_{t \in [n], i \in [L]} \mathbb{1}[\tilde{z}_{i,t} \neq \perp] = \mathbb{E}[r - s \mid r > s] \cdot |\{i \in [L] : \exists t \in [n], \tilde{z}_{i,t} \neq \perp\}| \\ &\overset{(ii)}{\geq} \frac{q_l - s}{2M} \sum_{i \in [L]} |\{t \in [n] : \tilde{z}_{i,t} \notin \{y_t, \perp\}\}| \\ &\overset{(iii)}{\geq} \frac{q_l - s}{2M} \cdot \frac{C}{4} N(s) \end{aligned}$$

In (i) we used the fact that at each round, at most C weak learners make predictions. In (ii), we used the previous lower bound on $\mathbb{E}[r - s \mid r > s]$ as well as the assumption that each weak learner makes at most M mistakes. In (iii) we used the fact that by construction, $\text{DELETE}_{s,C/2}$ follows the majority vote if at least $C/2$ weak learners make a prediction. Hence, if $\hat{y}_t(s) \notin \{y_t, \perp\}$ then at least $C/4$ weak learners must have made a mistake. Altogether, we showed that

$$N(s) \leq \frac{8Mn}{q_l - s}.$$

Algorithm 6: Aggregate algorithm for one layer AGGREGATE$_{s_{\max},C}$

Input: number of experts L, minimum number of predictions C, maximum deletion $s_{\max} \geq 0$

Initialize weights $w_{s,1} = 1$ for all $s \in \{0, \ldots, s_{\max}\}$
for $t \geq 1$ **do**
  For $y \in \{0, 1, \perp\}$ let $W_t(y) := \sum_{s \leq s_{\max}: \hat{y}_t(s) = y} w_{s,t}$ where $\hat{y}_t(s)$ is the selection of DELETE$_{s_t,C}$
  Select $\hat{y}_t = \arg\max_{y \in \{0,1,\perp\}} W_t(y)$
  For $s \in \{0, \ldots, s_{\max}\}$, if $\hat{y}_t(s) \notin \{y_t, \perp\}$, update $w_{s,t+1} = \frac{w_{s,t}}{2}$; otherwise keep $w_{s,t+1} = w_{s,t}$

We recall that there are at most $\lceil \log L \rceil$ values of form q_l for $l \in [\lceil \log L \rceil]$. In particular, either $s_{\max} + 1 \leq \lceil \log L \rceil$ in which case the desired bound is immediate since $N(s) \leq n$. Or $s_{\max} + 1 > \lceil \log L \rceil$ hence by the pigeonhole principle, either $s = r_{\max} \leq s_{\max}$ or there exists $s = q_{l-1}$ for $l \in [\lceil \log L \rceil]$ such that $q_l - s \geq (s_{\max} + 1) / \lceil \log L \rceil$. In all cases, this proves the existence of $s \in \{0, \ldots, s_{\max}\}$ such that

$$N(s) \leq \frac{8Mn \lceil \log L \rceil}{s_{\max} + 1}.$$

This ends the proof. ■

Then, we present the full algorithm of AGGREGATE omitted in Section 5, which combines the predictions of the DELETE$_{s,C}$ algorithms for $s \in \{0, \ldots, s_{\max}\}$.

We next prove the misclassification error of the AGGREGATE algorithm using standard misclassification guarantees for WMA (Littlestone and Warmuth, 1994).

Proof of Lemma 13 The misclassification error bound for AGGREGATE$_{s_{\max},C/2}$ is essentially an application of standard bounds for WMA, e.g. Littlestone and Warmuth (1994) (the proof directly generalizes to multiclass case), for our parameters we can use $M_T(\text{WMA}) \leq 4(M^\star + \log N)$ for where $M_T(\text{WMA})$ (resp. $M^\star$) is the number of mistakes of WMA (resp. of the best expert) for N experts. Hence, the misclassification error of the WMA aggregates satisfies

$$|\{t \in [n] : \hat{y}_t \notin \{y_t, \perp\}\}| \leq 4 \min_{s \leq s_{\max}} |\{t \in [n] : \hat{y}_t(s) \notin \{y_t, \perp\}\}| + 4 \log(s_{\max} + 1).$$

Furthering the right-hand side using Lemma 21 gives the desired misclassification error bound.

Next, by construction, each algorithm DELETE$_{s,C/2}$ for $s \in \{0, \ldots, s_{\max}\}$, deletes at most the first $s_{\max}$ predictions for each weak learner. Hence for any iteration $t \in [n]$, the prediction of any weak learner $i \in [L]$ such that $|\{t' < t : z_{i,t'} \neq \perp\}| \geq s_{\max}$ will not be deleted for any algorithm DELETE$_{s,C/2}$ with $s \leq s_{\max}$. In particular, if at least $C/2$ such weak learners make a prediction then all algorithms DELETE$_{s,C/2}$ for $s \leq s_{\max}$ make a prediction. Hence, AGGREGATE$_{s_{\max},C/2}$ also makes a prediction. ■

We are now ready to prove the main guarantee for our boosting procedure BOOSTING.

Proof of Theorem 9 Fix the parameters $\epsilon, s_{\max}, M$. As in the algorithm, we denote by $y_{t,i}$ the prediction of the weak learner $i \in [L]$ at time $t \in [T]$ for the sequence of instances generated by the adversary.

Correctness of the algorithm. We start by checking that Algorithm 3 is well-defined, specifically, that the while loop in lines 6-13 at each round $t \geq 1$ terminates. Indeed, fix any iteration of this loop, and let $j \in [\lceil \log L \rceil]$ be the index such that $n_t := |\{i \in [L] : z_{i,t} \neq \perp\}| \in (2^{-j}L, 2^{1-j}L]$. From Lemma 13, the layer-j subroutine $\text{AGGREGATE}^{\mathcal{Q}_j}_{s_{\max}, 2^{-j}L}$ makes a prediction $\hat{y}_t^{(j)} \neq \perp$ whenever at least $2^{-j}L$ weak learners $i \in [L]$ make a prediction $z_{i,t} \neq \perp$ and have made at least $s_{\max}$ predictions in previous inputs to the layer-j subroutine. Additionally, note that the quantity $s_{\max} - s_{i,j}$ precisely counts the number of predictions of weak learner i in previous inputs to this subroutine. Hence, after running lines 10-12, the prediction $z_{i,t}$ of weak learner $i \in [L]$ is only kept if it made at least $s_{\max}$ predictions in previous inputs to the layer-j subroutine. For clarity, denote by $\tilde{z}_{i,t}$ the updated value of $z_{t,i}$ after this operation. Altogether, this shows that if $\hat{y}_t^{(j)} = \perp$, then

$$\tilde{n}_t := |\{i \in [L] : \tilde{z}_{i,t} \neq \perp\}| < 2^{-j}L.$$

In turn, this shows that the variable j is strictly increasing at each loop iteration, and hence must terminate when either a prediction is made or $n_t = 0$.

Misclassification error. In line 4 of Algorithm 3, we delete predictions of all weak learners with at least M mistakes. To avoid confusions, we denote by $z_{i,t}^{(0)}$ the corresponding updated recommendation of weak learner $i \in [L]$ at iteration $t \in [n]$:

$$z_{i,t}^{(0)} := \begin{cases} \perp & \text{if } |\{s < t : y_{i,s} \notin \{y_s, \perp\}\}| \geq M, \\ y_{i,t} & \text{otherwise.} \end{cases}$$

By construction, for each weak learner $i \in [L]$, the recommendations $z_{i,t}^{(0)}$ for $t \in [T]$ contain at most M mistakes.

For any layer $j = 1, \ldots, \lceil \log L \rceil$, we denote by $\mathcal{Q}_{j,T}$ the final value of $\mathcal{Q}_j$ at the end of the algorithm, that is, $\mathcal{Q}_j$ corresponds to the set of times on which we ran the subroutine $\text{AGGREGATE}_{s_{\max}, 2^{-j}L}$. Next, for any $t \in \mathcal{Q}_{j,T}$, denote by $(z_{i,t}^{(j)})_{i \in [L]}$ the weak learner predictions input to this subroutine at iteration t. Note that the layer-j weak learner recommendations $z_{i,t}^{(j)}$ for $i \in [L]$ and $t \in \mathcal{Q}_{j,T}$ are obtained from their counterpart $z_{i,t}^{(0)}$ by deleting some predictions: $z_{i,t}^{(j)} \in \{z_{i,t}^{(0)}, \perp\}$. Indeed, throughout Algorithm 3, the only updates of the quantities $z_{t,i}$ are deletions, see line 11. In turn this shows that the input fed to the subroutine $\text{AGGREGATE}_{s_{\max}, 2^{-j}L}$ contain at most M mistakes, as per Definition 12. By construction, these inputs also have at most $2^{1-j}L$ predictions per round—see line 7 of Algorithm 3—as per Definition 12. Hence, Lemma 13 bounds the misclassification error of the layer-j subroutine by

$$\sum t \in \mathcal{Q}_{j,T} \mathbb{1}[\hat{y}_t^{(j)} \notin \{y_t, \perp\}] \leq \frac{32M|\mathcal{Q}_{j,T}| \lceil \log L \rceil}{s_{\max} + 1} + 4 \log(s_{\max} + 1). \tag{19}$$

Since at each iteration $t \in [T]$ we follow the prediction of one of the subroutines or abstain, we can bound the total misclassification error by the sum of misclassification error of the subroutines:

$$\begin{aligned} \sum_{t=1}^{T} \mathbb{1}[\hat{y}_t \notin \{y_t, \perp\}] &\leq \sum_{j=1}^{\lceil \log L \rceil} \sum_{t \in \mathcal{Q}_{j,T}} \mathbb{1}[\hat{y}_t^{(j)} \notin \{y_t, \perp\}] \\ &\leq \frac{32MT \lceil \log L \rceil^2}{s_{\max} + 1} + 4 \log(s_{\max} + 1) \lceil \log L \rceil, \end{aligned}$$

where in the last inequality we used Eq. (19) and the fact that $\mathcal{Q}_{j,T} \subseteq [T]$ for all layers j. Using $s_{\max} \leq T$ gives the desired bound for the misclassification error of Algorithm 3.

Abstention error. By construction, the algorithm abstains $\hat{y}_t = 0$ only if at the end of while loop in lines 6-13, all weak learner updated recommendations are abstentions: $n_t = |\{i \in [L] : z_{i,t} \neq \perp\}| = 0$. We recall that these updated recommendations are obtained from $z_{i,t}^{(0)}$ by potentially deleting predictions in line 11. Note, however, that there are at most $s_{\max}$ deletions for each layer j throughout the complete procedure, as depicted by the counts $s_{i,j}$. Formally, if we denote by $z_{i,t}$ its value at the end of the while loop, for each $i \in [L]$ we have

$$|\{t \in [T] : \hat{y}_t = \perp \text{ and } z_{i,t}^{(0)} \neq \perp\}| \leq |\{t \in [T] : z_{i,t} = \perp \text{ and } z_{i,t}^{(0)} \neq \perp\}| \leq s_{\max} \lceil \log L \rceil .$$

Additionally, if weak learner i makes strictly less than M mistakes, then line 4 of Algorithm 3 never deletes its predictions and hence $y_{t,i} = z_{i,t}^{(0)}$. Together with the previous equation this implies

$$\textsc{AbsErr} \leq \textsc{AbsErr}(i) + s_{\max} \lceil \log L \rceil ,$$

where $\textsc{AbsErr}(i)$ denotes the abstention error of weak learner i. This ends the proof. ■