Agnostic Online Learning with Reliable Abstention

Omar Montasser
Department of Statistics and Data Science
Yale University
omar.montasser@yale.edu

Abstract

We study online learning in the adversarial injection model [GHMS23], where each instance is either drawn i.i.d. from an unknown distribution or is an arbitrary adversarial injection. The learner never knows which rounds are adversarial, but may abstain without penalty on adversarial injections. The learner's goal is to make few total misclassifications and few abstentions on i.i.d. instances. This model lies between classical statistical learning, governed by the VC dimension, and adversarial online learning, governed by the Littlestone dimension. Unlike prior work, which has largely focused on the realizable setting, we consider an agnostic formulation that allows arbitrary labels. We present an online algorithm that guarantees sublinear excess misclassification error and sublinear abstention error, against oblivious and adaptive adversaries, for function classes that admit bounded bracketing number witnessed by brackets of bounded VC dimension. Our algorithm is distribution-independent and does *not* require prior knowledge of the distribution. As corollaries, we obtain new learning guarantees for halfspaces (linear classifiers) and richer classes such as feed-forward neural networks with threshold activations.

1 Introduction

A stark gap arises in learning theory when considering online (or sequential) classification between the fully-stochastic and fully-adversarial worlds. In the stochastic world, where the online data sequence is independent and identically distributed (i.i.d.) from some (unknown) distribution D, it is well-known that a function class $\mathcal{H}$ is learnable if and only if its VC dimension is finite [VC71, LM12]. In the adversarial world, where the online data sequence is adversarially and adaptively chosen, a function class $\mathcal{H}$ is learnable if and only if its Littlestone dimension is finite [Lit87, BPS09]. The stark gap between these two worlds is exemplified by arguably one of the simplest function classes: thresholds on the unit interval, which has VC dimension 1 but *infinite* Littlestone dimension.

This gap continues to inspire and motivate research that explores intermediate relaxations between the two worlds, with the aim of obtaining strong learning guarantees (as in the stochastic world) but under more general conditions (as in the adversarial world). For example, this includes research on smoothed online learning [RST11, HRS24, BDGR22], distributionally-constrained adversaries [BK25, BSR26], transductive online learning [BKM97, HMS23], and introducing relaxed benchmarks to compete with [MSZ25].

The focus of this work is on one particularly relevant model introduced by Goel et al. [GHMS23] for beyond worst-case online classification. This model considers online classification of i.i.d. sequences interspersed with arbitrary adversarial injections where the learner is allowed to *abstain*. Specifically, in each time step $t = 1, \ldots, T$, the adversary decides whether: (a) x_t is drawn i.i.d. from an (unknown) distribution D, or (b) x_t is an adversarial injection chosen arbitrarily from $\mathcal{X}$. The adversary's decisions are never revealed to the learner. Importantly, the learner is allowed to *abstain* on adversarial injections, but is penalized for abstaining on i.i.d. instances. The goal of the learner is to

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

make as few misclassifications as possible, and as few abstentions as possible on the i.i.d. subsequence,

$$\mathrm{MissErr} := \sum_{t=1}^{T} \mathbb{1}\left[\hat{y}_t \in \{0,1\} \text{ and } \hat{y}_t \neq y_t\right] \text{ and } \mathrm{AbsErr} := \sum_{t=1}^{T} \mathbb{1}\left[\hat{y}_t = \perp \text{ and } x_t \text{ is i.i.d.}\right]. \quad (1)$$

Prior work on this problem has mainly focused on the *realizable* setting where the ground-truth labels of the i.i.d. sequence *and* the adversarial injections are consistent with some target function $h_\star \in \mathcal{H}$, i.e., $y_t = h_\star(x_t)$ [GHMS23, YB26, EG26]. In this work, we consider the more challenging *agnostic* setting of this problem, and by that we mean that the labels of the i.i.d. sequence and the adversarial injections are *arbitrarily* chosen by the adversary and need not be consistent with some $h_\star \in \mathcal{H}$. The central question that we study is:

What learning guarantees can we hope to achieve for agnostic *online classification with abstention under adversarial injections?*

As is standard in agnostic online learning [BPS09], a sensible objective is to require the learner to achieve low misclassification error (MissErr) relative to the best classifier $h_\star \in \mathcal{H}$,

$$\mathrm{MissErr} - \mathrm{OPT} = \sum_{t=1}^{T} \mathbb{1}\left[\hat{y}_t \in \{0,1\} \text{ and } \hat{y}_t \neq y_t\right] - \min_{h \in \mathcal{H}} \sum_{t=1}^{T} \mathbb{1}[h(x_t) \neq y_t], \quad (2)$$

while simultaneously requiring low abstention error (AbsErr) as defined in (1). In fact, we will also consider competing with a *stricter* benchmark, $\mathrm{OPT}_{\mathrm{cen}}$, which is the best-in-class error on the (subset) rounds the learner did *not* abstain (see Remark 1),

$$\mathrm{OPT}_{\mathrm{cen}} := \min_{h \in \mathcal{H}} \sum_{t=1}^{T} \mathbb{1}[\hat{y}_t \neq \perp] \cdot \mathbb{1}[h(x_t) \neq y_t]. \quad (3)$$

Before presenting our contributions, we put in context some of the results in the realizable setting ($\mathrm{OPT} = 0$). It is important to distinguish between two cases: (i) the learner has prior knowledge of distribution D (distribution-dependent), and (ii) the learner has *no* prior knowledge of D (distribution-independent). With prior knowledge of D, Goel et al. [GHMS23] proposed a learning algorithm that achieves a total misclassification error of $O(d^2 \log(1/\varepsilon))$ and abstention error of $O(\varepsilon T)$, where d denotes the VC dimension of $\mathcal{H}$. More recently, in the more challenging case of *no* prior knowledge of D, Yu and Blanchard [YB26] proposed a learning algorithm that achieves misclassification error of $O(T^{3\alpha})$ and abstention error of $O(d^2 \log^{5/3}(T) \cdot T^{1-\alpha})$ for any $\alpha \in [0, 1/3]$ against *oblivious* adversaries. Furthermore, Yu and Blanchard [YB26] extended those guarantees to *adaptive* adversaries for function classes satisfying additional structure that they introduce (finite reduction dimension) which includes halfspaces. Another important result is that, without prior knowledge of D (i.e., distribution-independent), polynomial tradeoffs between misclassification and abstention errors are *unavoidable* even for oblivious adversaries [EG26, YB26]. Specifically, there are function classes with VC dimension 1 such that any algorithm must incur either $\Omega(\sqrt{T})$ misclassification error or $\Omega(\sqrt{T})$ abstention error [EG26, YB26]. Beyond the realizable setting, no learning guarantees were known prior to this work except perhaps for the special case of thresholds under random classification noise [Hei25].

1.1 Our Contributions

In this paper, we study the problem of online classification of i.i.d. sequences interspersed with arbitrary adversarial injections introduced by [GHMS23], with a particular focus on obtaining *agnostic* guarantees making no assumptions on how the labels are generated.

Results. Our formal results are presented in Section 3 and summarized in Table 1. We provide an overview below:

- In Theorem 1, we present an algorithm (Algorithm 1) that achieves sublinear misclassification and abstention errors in the *agnostic* and realizable settings, against oblivious *and* adaptive adversaries, for function classes $\mathcal{H}$ that admit bounded bracketing number witnessed by brackets of bounded VC dimension. Our algorithm is distribution-independent and does *not* require prior knowledge of the underlying distribution.

- In Theorem 2, we extend quantitatively similar learning guarantees to a *censored* variant of the problem where the learner observes feedback labels only in the rounds where it does *not* abstain and is required to compete with the best-in-class on these rounds.
- In Corollary 1, we instantiate our generic guarantees to establish new and improved results for the class of halfspaces in $\mathbb{R}^d$. In Corollary 2, we establish new learning guarantees for richer function classes such as intersections of halfspaces and feed-forward neural networks. Along the way, we establish a new generic structural result bounding the ε-bracketing number of compositions of classes (Lemma 2) which may be of independent interest.

Result	Setting	Misclassification	Abstention
Theorem 1	agnostic	$\sqrt{T^{1+\alpha}\log N}$	$T^{1-\alpha} + d_{\mathcal{B}}\log^2 T$
	realizable	$T^{\alpha}\log N$	$T^{1-\alpha} + d_{\mathcal{B}}\log^2 T$
Theorem 2 (censored)	agnostic	$\sqrt{T^{1+\alpha}\log N}$	$T^{1-\alpha} + d_{\mathcal{B}}^2\log^4 T$
	realizable	$T^{\alpha}\log N$	$T^{1-\alpha} + d_{\mathcal{B}}\log^2 T$

Table 1: Summary of our guarantees, for any $\alpha \in [0,1]$ and against any adversary, up to constant factors. In the agnostic setting, we bound $\mathbb{E}[\mathrm{MissErr} - \mathrm{OPT}]$ (Theorem 1) and $\mathbb{E}[\mathrm{MissErr} - \mathrm{OPT}_{\mathrm{cen}}]$ (Theorem 2), and in the realizable setting, we bound $\mathbb{E}[\mathrm{MissErr}]$. Here $N = \mathcal{N}_{[\,]}(\mathcal{H}, \mathcal{B}, 1/2T)$ is the bracketing number of $\mathcal{H}$ at scale $1/2T$ using brackets from $\mathcal{B}$, and $d_{\mathcal{B}}$ is the VC dimension of the bracket class $\mathcal{B}$ (see Definition 1 and Definition 2). For halfspaces in $\mathbb{R}^d$, $\log N = O(d\log(dT))$ and $d_{\mathcal{B}} = O(d^2\log d)$ (Corollary 1). For example, $\alpha = 1/3$ gives $T^{2/3}$ rates for both errors in the agnostic setting, and $\alpha = 1/2$ gives $\sqrt{T}$ rates in the realizable setting, up to logarithmic and class-dependent factors.

Techniques. Our algorithmic approach departs significantly from the earlier approaches of [GHMS23, YB26, EG26], and is based on the concept of *brackets* and *bracketing number* (see Definition 1 and Definition 2). A brief background and history on bracketing number is deferred to the beginning of Section 3.1. We remark below a few features of our bracketing-based approach.

- The agnostic setting presents new technical challenges that do not seem amenable to previous approaches, while our bracketing-based approach overcomes those challenges. A more detailed discussion is presented in Section 4.
- Earlier results against *adaptive* adversaries were known only in the realizable setting and under different structural conditions: finite reduction dimension [YB26], and inference dimension and certificate dimension [EG26]. Our bracketing-based approach provides new techniques to handle adaptive adversaries, which enables us to extend new learning guarantees against adaptive adversaries beyond halfspaces to richer classes such as feed-forward neural networks in the realizable *and* agnostic settings.
- We highlight that as part of our bracketing-based approach we provide a technique to construct valid brackets from adversarially contaminated sequences. This avoids an *exponential* complexity of "guessing" the hidden i.i.d. subsequence [YB26].

1.2 Related Literature

Beyond worst-case online learning. A broad line of work seeks intermediate models of online learning between the fully-stochastic and fully-adversarial settings. For example, smoothed online learning restricts an adversary by requiring each instance to be generated from a distribution with bounded density ratio relative to a fixed base measure [HRS24, BDGR22]. Recent work has also removed or weakened the assumption that the base measure is known [BRS24, Bla25]. More recently, the smoothed online learning framework has been extended to capture more general distributionally-constrained adversaries [BK25, BSR26]. A different perspective which keeps the adversary unrestricted but instead considers relaxed benchmarks to compete with has also been studied [MSZ25]. Another classical intermediate model is *transductive* online learning [BKM97, KK05], where the adversary fixes the instances $x_1, \ldots, x_T$ in advance and reveals them, unlabeled, to the learner, who then predicts their labels online. In the realizable case, the optimal mistake bound for transductive online learning exhibits a trichotomy governed by the VC and Littlestone dimensions

[HMS23], and its gap to standard online learning was recently tightly quantified [CHMS25]. One of the main consequences that follow from these different intermediate online learning models is stronger learning guarantees, bypassing the worst-case Littlestone dimension lower bounds and establishing positive guarantees for function classes with bounded VC dimension.

Abstention and selective classification. Learning with an abstention option has a long history, beginning with the error–abstention tradeoff of Chow [Cho70], and later formalized in several selective classification and learning-with-rejection frameworks (for statistical batch settings) [RS88, Kiv95, FMS04, HW06, EYW10, CDM16]. For online learning, abstention with fully adversarial sequences has also been studied [LLWS11, SZB10], but it is impossible to learn even intervals on the line in this setting. One notable framework that is close to the adversarial injection model considered in this work, is PQ-learning [GKKM20]. This framework considers a transductive setting where a learning algorithm is presented with labeled training examples from a distribution P and arbitrary unlabeled test examples and is required to classify test examples with low error and low abstention rate under P. The adversarial injection model [GHMS23] can be thought of as the online version of PQ-learning, since the test examples are revealed one-by-one (online) as opposed to all-at-once (transductive). Another closely related model is testable learning with distribution shift (TDS) [KSV24], where the learner is allowed to abstain if the test set is not "close" to the training distribution; recent work shows that this coarse-grained model is actually equivalent to PQ-learning [PKSV26]. Also related is relatively smart learning [DP26], a batch framework in which a learner competes with the best semi-supervised guarantee certifiable from unlabeled data.

Adversarially robust learning. The motivation for abstaining on suspicious inputs is related to robustness to adversarial examples [SZS+14, BCM+13, GSS15], where the goal is to learn a predictor that is robustly correct on adversarial perturbations of i.i.d. test examples. There has been a series of works establishing theoretical guarantees for adversarially robust learning [e.g., MHS19, APU23, Mon23]. Note that this line of work typically assumes "perturbed" test examples (e.g., perturbations of bounded norm), while the adversarial injection model studied in this work considers *arbitrary* test examples with no restrictions. There are also works focusing on theoretical guarantees against adversarial attacks on the training data [e.g., BHQS21, BBHS22].

2 Preliminaries and Background

Problem Setup. We denote by $\mathcal{X}$ the instance space, and by $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ the benchmark function class that we would like to compete with. We consider the problem of online classification in the adversarial injection model introduced in [GHMS23]. Formally, the learning protocol proceeds as follows. Initially, an adversary (or nature) fixes a distribution D over $\mathcal{X}$. The learner may or may not know the distribution D. Then, at each time step $t \in [T] := \{1, \ldots, T\}$:

1. The adversary decides whether or not to inject an adversarial instance by choosing $c_t \in \{1, 0\}$.
2. If $c_t = 0$, then $x_t \overset{\text{iid}}{\sim} D$ independently of the past, while if $c_t = 1$, then adversary chooses an arbitrary $x_t \in \mathcal{X}$. Afterwards, the adversary chooses an arbitrary label $y_t \in \{0, 1\}$ for x_t.
3. The learner observes x_t and outputs $\hat{y}_t \in \{0, 1, \perp\}$, where $\perp$ denotes an abstention.
4. The learner observes the label y_t.

We emphasize that in this work we consider the more general *agnostic* setting where the labels y_t chosen by the Adversary are *arbitrary*, and need not be consistent with some ground-truth function $h_\star \in \mathcal{H}$ as was considered in prior works [GHMS23, YB26, EG26].

If the learner makes a prediction $\hat{y}_t \in \{0, 1\}$ and $\hat{y}_t \neq y_t$, then this is counted as a misclassification mistake. If the learner abstains $\hat{y}_t = \perp$ and x_t is iid from D (i.e., $c_t = 0$ and there is no adversarial injection), then this is counted as an abstention mistake. Then, the goal for the learner is to make few misclassifications relative to the best-in-class predictor $h_\star \in \mathcal{H}$,

$$\text{MissErr} - \text{OPT} = \sum_{t=1}^{T} \mathbb{1}\left[\hat{y}_t \in \{0,1\} \text{ and } \hat{y}_t \neq y_t\right] - \min_{h \in \mathcal{H}} \sum_{t=1}^{T} \mathbb{1}[h(x_t) \neq y_t]. \tag{4}$$

Simultaneously, the learner should make as few abstentions as possible on the iid examples drawn from D, i.e., the timesteps with no adversarial injections,

$$\mathrm{AbsErr} := \sum_{t=1}^{T} \mathbb{1}\left[c_t = 0 \text{ and } \hat{y}_t = \perp\right]. \tag{5}$$

Remark 1 (Censored Variant)**.** As discussed in [GHMS23, YB26], we will also consider a *censored* variant of this protocol where the learner observes a feedback label y_t only when it does *not* abstain, i.e., $\hat{y}_t \neq \perp$. In particular, for the agnostic setting we consider in this work, we will additionally require the learner to compete with a *stricter* benchmark

$$\mathrm{OPT}_{\mathrm{cen}} := \min_{h \in \mathcal{H}} \sum_{t=1}^{T} \mathbb{1}[\hat{y}_t \neq \perp] \cdot \mathbb{1}[h(x_t) \neq y_t]. \tag{6}$$

Note that $\mathrm{OPT}_{\mathrm{cen}} \leq \mathrm{OPT}$, since it only considers the best-in-class in a subset of the rounds.

Remark 2 (Adaptive vs. Oblivious)**.** As discussed in [YB26], an oblivious adversary chooses in advance the corruption times $\{t : c_t = 1\}$ with their corresponding adversarial injections $(x_t)_{t:c_t=1}$, while an adaptive adversary's decision (and choice) of corruption at time t may depend on the entire history of interaction with the learner before time t. The techniques we present will handle both adversaries.

Brackets. A central concept and object that will be important for our algorithm and analysis is the notion of *brackets* and *bracketing number*. For our purposes, a *bracket* is a pair $b = (b^-, b^+)$ of binary functions $b^-, b^+ : \mathcal{X} \to \{0, 1\}$, such that $b^-(x) \leq b^+(x)$ for all $x \in \mathcal{X}$. We say that b brackets $h \in \mathcal{H}$ if $b^-(x) \leq h(x) \leq b^+(x)$ for every $x \in \mathcal{X}$.

Definition 1 (Bracketing Number)**.** Let $\mathcal{H}$ be a hypothesis class, D be a distribution over $\mathcal{X}$, and let $\varepsilon \in (0, 1)$. The *ε-bracketing number* of class $\mathcal{H}$ with respect to distribution D, denoted by $\mathcal{N}_{[\,]}(\mathcal{H}, D, \varepsilon)$, is the smallest cardinality N of a family of brackets $\mathcal{B}_{D,\varepsilon} = \{(b_i^-, b_i^+)\}_{i=1}^N$ such that for every $h \in \mathcal{H}$, there exists $i \in [N]$ satisfying

$$\forall x \in \mathcal{X} : b_i^-(x) \leq h(x) \leq b_i^+(x) \qquad \text{and} \qquad \mathbb{P}_{X \sim D}(b_i^-(X) \neq b_i^+(X)) \leq \varepsilon.$$

We call such a family an *ε-bracketing* of $\mathcal{H}$ with respect to D.

Another concept that will be important for us is the complexity of the bracket class.

Definition 2 (Complexity of Bracket Class)**.** We say that a class $\mathcal{B} \subseteq \{0, 1\}^{\mathcal{X}}$ brackets $\mathcal{H}$ if for every distribution D over $\mathcal{X}$ and every $\varepsilon \in (0, 1)$, there exists an ε-bracketing $\mathcal{B}_{D,\varepsilon} \subseteq \mathcal{B} \times \mathcal{B}$ of $\mathcal{H}$ with respect to D. Note that this definition requires all binary functions from which brackets for $\mathcal{H}$ are constructed to belong to $\mathcal{B}$. We denote by $\mathcal{N}_{[\,]}(\mathcal{H}, \mathcal{B}, \varepsilon)$ the *worst-case* ε-bracketing number of $\mathcal{H}$ (using brackets from $\mathcal{B}$) over all distributions,

$$\mathcal{N}_{[\,]}(\mathcal{H}, \mathcal{B}, \varepsilon) := \sup_D \inf \left\{|\mathcal{B}_{D,\varepsilon}| : \mathcal{B}_{D,\varepsilon} \subseteq \mathcal{B} \times \mathcal{B} \text{ and } \mathcal{B}_{D,\varepsilon} \text{ is an } \varepsilon\text{-bracketing of } \mathcal{H} \text{ wrt } D\right\}. \tag{7}$$

Finally, we denote by $d_{\mathcal{B}}$ the VC dimension of the bracket class $\mathcal{B}$.

3 Main Results

Our main result is an algorithm that achieves sublinear misclassification and abstention errors in the *agnostic* setting, against oblivious *and* adaptive adversaries, for function classes $\mathcal{H}$ that admit bounded bracketing number using brackets of bounded VC dimension. Our algorithm is distribution-independent and requires *no* prior knowledge of the distribution.

Theorem 1. *Let $\mathcal{H}$ be a hypothesis class, and let $\mathcal{B}$ be a bracket class for $\mathcal{H}$ with VC dimension $d_{\mathcal{B}}$. Let $T \geq 1$ be a time horizon. Then, for any parameter $\alpha \in [0, 1]$, Algorithm 1 achieves the following guarantee against* any *adversary and any distribution D over $\mathcal{X}$:*

$$\mathbb{E}[\mathrm{MissErr} - \mathrm{OPT}] \lesssim \sqrt{T^{1+\alpha} \log\left(\mathcal{N}_{[\,]}(\mathcal{H}, \mathcal{B}, 1/2T)\right)} \quad \textit{and} \quad \mathbb{E}[\mathrm{AbsErr}] \lesssim T^{1-\alpha} + d_{\mathcal{B}} \log^2(T)\,.$$

Moreover, in the realizable setting where $\mathrm{OPT} = 0$, Algorithm 1 satisfies

$$\mathbb{E}[\mathrm{MissErr}] \lesssim T^{\alpha} \log\left(\mathcal{N}_{[\,]}(\mathcal{H}, \mathcal{B}, 1/2T)\right) \quad \textit{and} \quad \mathbb{E}[\mathrm{AbsErr}] \lesssim T^{1-\alpha} + d_{\mathcal{B}} \log^2(T)\,.$$

Algorithm 1: Online Classification and Abstention via Brackets

Input: horizon T, round length τ, benchmark class $\mathcal{H}$, and bracket class $\mathcal{B}$ for $\mathcal{H}$.

Let $R \leftarrow \lceil T/\tau \rceil$.

for $r = 1 \wedge t = 1, \ldots, \tau$ **do**

  Receive x_t from adversary, output an abstention $\hat{y}_t = \perp$.

for $r = 2, \ldots, R$ **do**

  Let $I_r = \{(r-1)\tau + 1, \ldots, \min(r\tau, T)\}$.

  Let $S_r = (x_1, \ldots, x_{(r-1)\tau})$, and $n_r = |S_r|$.

  Let $\mathcal{B}_r$ be a collection of brackets for $\mathcal{H}$ on S_r at scale $\varepsilon = 1/(2n_r)$.

  Initialize p_1 to be a uniform distribution over the set of experts $\{E_b : b \in \mathcal{B}_r\}$, each expert E_b is defined as follows

$$E_b(x) := b^-(x) \text{ if } b^-(x) = b^+(x), \text{ otherwise } \perp .$$

  for $t \in I_r$ **do**

    Receive x_t from adversary.

    Sample a random expert $E_b \sim p_t$ and predict $\hat{y}_t = E_b(x_t) \in \{0, 1, \perp\}$.

    Upon receiving the feedback label y_t, update the mixture p_t:

$$p_{t+1}(E_b) = p_t(E_b) \exp\left(-\eta \mathbb{1}\{E_b(x_t) = 1 - y_t\}\right) / Z_t,$$

    where Z_t is a normalization constant.

We provide an overview of our algorithmic approach, technical challenges, and proof strategy in Section 4. The full proof of Theorem 1 is deferred to Appendix A. Observe that this result exhibits a polynomial tradeoff between misclassification and abstention error rates. Such a polynomial tradeoff is unavoidable even in the realizable case (when $\mathrm{OPT} = 0$), as shown in [EG26, YB26]. Specifically, there are function classes with VC dimension 1 such that any algorithm must incur either $\Omega(\sqrt{T})$ misclassification error or $\Omega(\sqrt{T})$ abstention error [EG26, YB26].

Next, we present learning guarantees in the *censored* variant of the problem (see Remark 1) where the learner receives label feedback y_t only in the rounds where it does *not* abstain ($\hat{y}_t \neq \perp$). Recall that here we require the learner to compete with a *stricter* benchmark, $\mathrm{OPT}_{\mathrm{cen}}$ in (6), which considers the best-in-class only in the rounds where the learner did not abstain.

Theorem 2. *Let $\mathcal{H}$ be a hypothesis class, and let $\mathcal{B}$ be a bracket class for $\mathcal{H}$ with VC dimension $d_{\mathcal{B}}$. Let $T \geq 1$ be a time horizon. Then, for any parameter $\alpha \in [0, 1]$, Algorithm 2 achieves the following guarantee against* any *adversary and any distribution D over $\mathcal{X}$:*

$$\mathbb{E}[\mathrm{MissErr} - \mathrm{OPT}_{\mathrm{cen}}] \lesssim \sqrt{T^{1+\alpha} \log\left(\mathcal{N}_{[\,]}(\mathcal{H}, \mathcal{B}, 1/2T)\right)} \quad \textit{and} \quad \mathbb{E}[\mathrm{AbsErr}] \lesssim T^{1-\alpha} + d_{\mathcal{B}}^2 \log^4(T) .$$

Moreover, in the realizable setting where $\mathrm{OPT}_{\mathrm{cen}} = 0$, Algorithm 2 satisfies

$$\mathbb{E}[\mathrm{MissErr}] \lesssim T^{\alpha} \log\left(\mathcal{N}_{[\,]}(\mathcal{H}, \mathcal{B}, 1/2T)\right) \quad \textit{and} \quad \mathbb{E}[\mathrm{AbsErr}] \lesssim T^{1-\alpha} + d_{\mathcal{B}} \log^2(T) .$$

Note that in the realizable setting, we recover the same guarantee as in the original problem (up to constants), and in the agnostic setting, the only difference is the lower-order term $d_{\mathcal{B}}^2 \log^4(T)$, instead of $d_{\mathcal{B}} \log^2(T)$, in the abstention error bound. Algorithm 2 is a slight variation of Algorithm 1, and the proof only requires a few changes to the proof of Theorem 1. We defer those details to Appendix B.

Remark 3 (First-Order Rates)**.** We remark that because our algorithm is based on Multiplicative Weights, we immediately obtain improved first-order regret bounds of the form

$$\mathbb{E}[\mathrm{MissErr} - \mathrm{OPT}] \lesssim \sqrt{T^{\alpha}\, \mathbb{E}[\mathrm{OPT}]\, \log\left(\mathcal{N}_{[\,]}(\mathcal{H}, \mathcal{B}, 1/2T)\right)} + T^{\alpha} \log\left(\mathcal{N}_{[\,]}(\mathcal{H}, \mathcal{B}, 1/2T)\right) ,$$

while the abstention rate remains the same as in Theorem 1.

3.1 Refined Guarantees for Some Natural Function Classes

Given the new generic upper bounds of Theorem 1 in terms of the bracketing number of a class $\mathcal{H}$ and the VC dimension of the bracketing class $\mathcal{B}$, a natural question to ask here is: which function classes

$\mathcal{H}$ admit bounded bracketing numbers? At first sight, one might hope that the bracketing number behaves much like the covering number: once a class has finite VC dimension d, the ε-covering number scales as $(1/\varepsilon)^{O(d)}$ [Hau95]. The story, however, is more subtle. While it has been shown that finite VC dimension does indeed imply *finite* ε-bracketing number under every distribution [AN10, AN12], there are classes with VC dimension 2 where the ε-bracketing number can depend arbitrarily badly on $1/\varepsilon$ [Mal12, vH13, BKMS21], ruling out any universal ε-bracketing bound which scales as $(1/\varepsilon)^{O(d)}$. This is not the end of the story. It turns out that for many natural geometric function classes, the ε-bracketing number indeed scales as $(1/\varepsilon)^{O(d)}$. This includes halfspaces (linear classifiers), polynomial threshold functions, and semi-algebraic classes [BKMS21, DGM22].

We begin with refined guarantees for the class of halfspaces in $\mathbb{R}^d$, and after that we present a new generic structural result bounding the ε-bracketing number of compositions of classes which enables us to extend our learning guarantees to richer function classes such as intersections of halfspaces and feed-forward neural networks with threshold activations.

Lemma 1 (Theorem 16 and Lemma 18 of [BKMS21])**.** *Let $\mathcal{H} = \{x \mapsto \mathbb{1}\{\langle w, x\rangle \le b\} \mid w \in \mathbb{R}^d, b \in \mathbb{R}\}$ be the class of halfspaces in $\mathbb{R}^d$, and let $\mathcal{B}$ be the class of intersections and unions of at most $d+2$ halfspaces from $\mathcal{H}$. Then, for any $\varepsilon \in (0,1)$, $\mathcal{N}_{[\,]}(\mathcal{H}, \mathcal{B}, \varepsilon) \le (d/\varepsilon)^{O(d)}$.*

Combining Theorem 1 with Lemma 1 establishes new guarantees for halfspaces in the realizable and agnostic settings as stated in the following corollary.

Corollary 1 (Halfspaces)**.** *Let $\mathcal{H}$ be halfspaces in $\mathbb{R}^d$. Let $T \ge 1$ be a time horizon. Then, for any parameter $\alpha \in [0,1]$, Algorithm 1 achieves the following guarantee against* any *adversary:*

$$\mathbb{E}[\mathrm{MissErr} - \mathrm{OPT}] \lesssim \sqrt{T^{1+\alpha} \cdot d \log(2dT)} \text{ and } \mathbb{E}[\mathrm{AbsErr}] \lesssim T^{1-\alpha} + d^2 \log(d) \log^2(T).$$

Moreover, in the realizable setting where $\mathrm{OPT} = 0$, Algorithm 1 satisfies

$$\mathbb{E}[\mathrm{MissErr}] \lesssim T^{\alpha} \cdot d \log(2dT) \quad \text{and} \quad \mathbb{E}[\mathrm{AbsErr}] \lesssim T^{1-\alpha} + d^2 \log(d) \log^2(T).$$

In particular, for adaptive adversaries in the realizable setting, this result improves on the previous bounds of [YB26] which established an $O(T^{3\alpha})$ misclassification error and an $\tilde{O}(d^{4.67} T^{1-\alpha})$ abstention error. Observe that Corollary 1 immediately implies results for more complex function classes such as k-degree polynomial threshold functions in $\mathbb{R}^d$, and more generally any function class with small sign-rank (i.e., embeddable in halfspaces).

Beyond halfspaces and towards establishing learning guarantees for more complex function classes, we present below a new generic structural result bounding the ε-bracketing number of compositions of classes. This result can be viewed as analogous to classical results bounding the VC dimension of compositions of classes [BEHW89, AGHM23], and extends earlier results bounding the bracketing number of unions and intersections of classes [HRS20].

Lemma 2 (Bracketing Number of Compositions)**.** *Let $\mathcal{H}_i \subseteq \{0,1\}^{\mathcal{X}}$ be classes of binary functions for $i \in [k]$, and let $\mathcal{F}$ be a binary function class over $\{0,1\}^k$ with VC dimension $d_{\mathcal{F}}$. Consider the composition of $\mathcal{F}$ with $\mathcal{H}_1, \ldots, \mathcal{H}_k$,*

$$\mathcal{G} := \left\{ x \mapsto f\big(h_1(x), \ldots, h_k(x)\big) : \ f \in \mathcal{F}, \ h_i \in \mathcal{H}_i \right\}.$$

Then, for every distribution D over $\mathcal{X}$ and for every $\varepsilon_1, \ldots, \varepsilon_k \in [0,1]$,

$$\mathcal{N}_{[\,]}(\mathcal{G}, D, \varepsilon_1 + \cdots + \varepsilon_k) \le e\, 2^{k d_{\mathcal{F}}} \prod_{i=1}^{k} \mathcal{N}_{[\,]}(\mathcal{H}_i, D, \varepsilon_i).$$

The proof of Lemma 2 is deferred to Appendix D. We highlight an interesting aspect of Lemma 2 which is that the bracketing of composition $\mathcal{G}$ depends on the bracketing number of $\mathcal{H}_1, \ldots, \mathcal{H}_k$ and the VC dimension of $\mathcal{F}$, and does *not* depend on the bracketing number of $\mathcal{F}$. Furthermore, as we discuss in detail in Section 4, it will be important for our purposes to be able to construct brackets from i.i.d. samples (that may contain adversarial injections), and we note here that the construction used in the proof of Lemma 2 preserves this property. That is, if bracketings for $\mathcal{H}_i$ can be constructed for (adversarially contaminated) i.i.d. samples, then a bracketing for $\mathcal{G}$ can be constructed from (adversarially contaminated) i.i.d. samples (see Remark 4).

Combining Lemmas 1 and 2 with Theorem 1 yields new learning guarantees for complex function classes such as intersections of halfspaces and feed-forward neural networks with threshold activations.

Corollary 2 (Compositions of Halfspaces). *Let $\mathcal{G} = \{f(h_1, \ldots, h_k) : f \in \mathcal{F}, h_1, \ldots, h_k \in \mathcal{H}\}$, where $\mathcal{H}$ is the class of halfspaces in $\mathbb{R}^d$ and $\mathcal{F}$ is an* arbitrary *function class with VC dimension $d_{\mathcal{F}}$. Let $T \geq 1$ be a time horizon. Then, for any parameter $\alpha \in [0, 1]$, Algorithm 1 achieves the following guarantee against* any *adversary and any distribution D over $\mathcal{X}$:*

$$\mathbb{E}[\text{MissErr}-\text{OPT}] \lesssim \sqrt{T^{1+\alpha} \cdot k(d_{\mathcal{F}} + d\log(2dkT))} \text{ and } \mathbb{E}[\text{AbsErr}] \lesssim T^{1-\alpha} + kd^2 \log(d) \log^2(T).$$

Moreover, in the realizable setting where $\text{OPT} = 0$, Algorithm 1 satisfies

$$\mathbb{E}[\text{MissErr}] \lesssim T^{\alpha} \cdot k(d_{\mathcal{F}} + d\log(2dkT)) \quad \text{and} \quad \mathbb{E}[\text{AbsErr}] \lesssim T^{1-\alpha} + kd^2 \log(d) \log^2(T).$$

Guarantees for the censored variation. Similar to how we established Corollary 1 and Corollary 2, we obtain similar quantitative guarantees for the censored variant of the problem (described in Remark 2) by invoking our generic censored guarantee (Theorem 2) and combining it with Lemmas 1 and 2. The formal statements are deferred to Appendix B.

4 Overview of Our Algorithmic Approach

Warmup: with prior knowledge of the marginal distribution D. Consider the simpler setting where we know a priori the distribution D over $\mathcal{X}$ generating the i.i.d. sequence. In this case, our algorithmic approach starts with constructing a finite ε-bracketing $\mathcal{B}_{D,\varepsilon}$ for the class $\mathcal{H}$ with respect to distribution D, where $|\mathcal{B}_{D,\varepsilon}| = \mathcal{N}_{[\,]}(\mathcal{H}, D, \varepsilon)$. Given a bracket $b = (b^-, b^+) \in \mathcal{B}_{D,\varepsilon}$, we associate with it an abstaining expert

$$E_b(x) := b^-(x) \text{ if } b^-(x) = b^+(x), \text{ otherwise } \perp .$$

By the bracketing definition (Definition 1), observe that for every $h \in \mathcal{H}$ there exists a bracket $b = (b^-, b^+) \in \mathcal{B}_{D,\varepsilon}$ for h satisfying

$$\forall x \in \mathcal{X} : b^-(x) \leq h(x) \leq b^+(x) \qquad \text{and} \qquad \mathbb{P}_{X\sim D}(b^-(X) \neq b^+(X)) \leq \varepsilon.$$

This implies the following *key* property: if b brackets h, then the abstaining expert E_b never makes more misclassifications than h, it either predicts exactly as h or abstains (see Lemma 4). At the same time, on i.i.d. samples from D, the abstention probability of E_b is exactly the disagreement mass of the bracket, which is at most ε. Therefore, the best expert competes with the best comparator $h \in \mathcal{H}$ on misclassification loss, while *every* expert abstains rarely on i.i.d. samples. Thus, running Multiplicative Weights over the finite family of abstaining experts $\{E_b : b \in \mathcal{B}_{D,\varepsilon}\}$ would guarantee

$$\mathbb{E}[\text{MissErr} - \text{OPT}] \lesssim \sqrt{T \cdot \log \mathcal{N}_{[\,]}(\mathcal{H}, D, \varepsilon)} \quad \text{and} \quad \mathbb{E}[\text{AbsErr}] \leq \sum_{t=1}^{T} \mathbb{P}(\hat{y}_t = \perp | c_t = 0) \leq \varepsilon T.$$

For completeness, we present a formal claim and proof of this simpler setting in Appendix C.

Challenge: with *no* prior knowledge of marginal distribution D. The actual problem is harder because the learning algorithm does *not* know the distribution D, and moreover does not know which observed examples are i.i.d. draws from D and which are adversarial injections. Thus, we cannot simply precompute a bracketing $\mathcal{B}_{D,\varepsilon}$, nor can we first identify the i.i.d. subsequence and then construct brackets based on this subsequence. The central challenge is therefore:

How to construct valid brackets directly from sequences that are contaminated with adversarial injections, without first identifying which points are iid?

We first discuss a tempting strategy that fails in the agnostic setting. In the realizable setting, Yu and Blanchard [YB26] provide a boosting framework which runs exponentially many weak learners on "guessed" i.i.d. subsequences. Each weak learner treats its given subsequence as i.i.d. and proceeds to run a sample-based online learning algorithm for prediction with abstention. The booster guarantees an abstention error close to that of the best weak learner up to an additive term, but pays a multiplicative overhead in misclassification error. A natural adaptation here is therefore to build bracket-experts from every candidate i.i.d. subsequence of the observed history and feed them to the same booster. The difficulty is that our goal in the agnostic setting is to compete with the best-in-class

(OPT) up to an *additive* regret term, whereas the boosting strategy of Yu and Blanchard [YB26] would incur a prohibitively large multiplicative factor in front of OPT.

We take a different approach that avoids the complexity of guessing exponentially many possibilities. Specifically, our algorithm does *not* need to guess i.i.d. subsequences of the history, and instead directly constructs brackets from the entire history, even though it may contain adversarial injections. Recall that Algorithm 1 has access to a class $\mathcal{B}$ that brackets $\mathcal{H}$ (as in Definition 2). It partitions the time horizon into blocks of length τ, and at the beginning of each block r it uses the entire (possibly contaminated) history S_r to construct a fresh bracketing family $\mathcal{B}_r$ for $\mathcal{H}$ at empirical scale $\varepsilon = 1/(2n_r)$, where $n_r = |S_r|$. It then runs a new Multiplicative Weights instance over the induced abstaining experts $\{E_b : b \in \mathcal{B}_r\}$ during this block.

The misclassification analysis is then relatively standard: on each block r, there is a bracket $b \in \mathcal{B}_r$ which brackets the best-in-class comparator $h^\star$, and so the corresponding expert E_b either abstains or predicts exactly as $h^\star$ (see Lemma 4). Thus, invoking the regret guarantee of Multiplicative Weights (Lemma 3) on each block and summing over blocks yields

$$\mathbb{E}[\mathrm{MissErr} - \mathrm{OPT}] \lesssim \sum_{r=2}^{R} \sqrt{\tau \log |\mathcal{B}_r|} \le \sqrt{\frac{T^2}{\tau} \log \mathcal{N}_{[\,]}(\mathcal{H}, \mathcal{B}, 1/2T)}.$$

The more challenging part is the abstention analysis. How can brackets constructed from i.i.d. samples have low disagreement mass? And, how can this carry through under adversarial injections? Recall here that all brackets $b = (b^-, b^+)$ belong to a class $\mathcal{B}$ of finite VC dimension $d_{\mathcal{B}}$ (Definition 2). It follows that the disagreement class $\{x \mapsto \mathbb{1}[b(x) \neq b'(x)] : b, b' \in \mathcal{B}\}$ has finite VC dimension $O(d_{\mathcal{B}})$ [BEHW89]. Therefore, with enough i.i.d. samples, any bracket $b = (b^-, b^+)$ with low disagreement on the empirical sample will have low disagreement on the underlying distribution. In Lemma 6, we appeal to standard uniform convergence bounds to argue that with high probability over the draw of m iid samples from D,

$$\forall b^-, b^+ \in \mathcal{B},\ \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{b^-(x) \neq b^+(x)\} = 0 \qquad \Longrightarrow \qquad \mathbb{P}_D(b^-(X) \neq b^+(X)) \lesssim \frac{d_{\mathcal{B}} \log(m)}{m}.$$

In a *key* Lemma 7, we prove that this remains true even when the adversary, after seeing the m iid samples, decides to inject additional $n - m$ adversarial points. Specifically, we show that if the empirical disagreement of a bracket $b = (b^-, b^+) \in \mathcal{B}_r$ on the full observed sequence S_r is at most $1/(2n_r)$, then in fact the disagreement is zero on the i.i.d. subsequence hidden inside S_r, and therefore the probability that an expert E_b abstains under distribution D is small.

Given that each block r in Algorithm 1 has length τ, there can be at most 2τ abstentions on i.i.d. instances before Algorithm 1 collects at least τ i.i.d. instances. After that, for blocks r in which the number of i.i.d. instances m_r available in S_r is at least τ, every bracket in $\mathcal{B}_r$ abstains on a fresh clean example with probability $\tilde{O}(d_{\mathcal{B}}/m_r)$, which sums to only a logarithmic contribution over later blocks. Choosing $\tau = \lceil T^{1-\alpha} \rceil$ yields the tradeoff of Theorem 1. The full formal proof is deferred to Appendix A. Notably, Algorithm 1 need not detect when its brackets have low abstention probability: more i.i.d. instances simply yield better brackets, while more adversarial injections can be safely handled by abstaining without penalty.

5 Discussion

In this work, we present new algorithms and learning guarantees for online classification with abstention under adversarial injections [GHMS23], in the agnostic and realizable settings, against oblivious and adaptive adversaries, for function classes that admit bounded bracketing numbers witnessed by brackets of bounded VC dimension. We discuss below a few directions for future research.

VC Dimension vs. Bracketing Number. Our approach fundamentally relies on the concept of brackets and is limited to function classes with bounded bracketing number. A natural technical question that arises in our work is: if a function class $\mathcal{H}$ with VC dimension d has finite ε-bracketing number satisfying $\mathcal{N}_{[\,]}(\mathcal{H}, D, \varepsilon) \le (1/\varepsilon)^{O(d)}$, then does there exist a class $\mathcal{B}$ with VC dimension $\mathrm{poly}(d)$ that brackets $\mathcal{H}$ (as in Definition 2)? In other words, whenever small bracketing number is

possible, is it possible to achieve with brackets with low complexity? This question does not appear to be studied in the literature, and it would strengthen recent results [DGM22] bounding the bracketing number of several natural function classes if resolved positively.

Guarantees for VC classes against adaptive adversaries. While our results establish new guarantees against adaptive adversaries in both the realizable and agnostic settings for function classes with bounded bracketing numbers (including halfspaces and feedforward nets), as discussed earlier, bracketing numbers do not behave "nicely" for every VC class. Therefore, our approach does not fully resolve the open question of whether one can achieve sublinear $\mathrm{poly}(d)T^{1-\Omega(1)}$ misclassification and abstention error rates against adaptive adversaries for every function class of VC dimension d [GHMS23, YB26].

Computational aspects. The runtime of our bracketing-based algorithm is linear in the bracketing number which for halfspaces in $\mathbb{R}^d$ scales as $n^{O(d)}$ where n is the size of the (contaminated) sequence. We remark that this is a huge improvement over previous approaches which require calculating involved shattering statistics [GHMS23] and running exponentially many algorithmic instantiations [YB26]. It would be interesting to explore more efficient algorithms for halfspaces. More broadly, it would also be interesting to think about oracle-efficient algorithms with access to reasonable oracles (e.g., ERM).

References

[AGHM23] Noga Alon, Alon Gonen, Elad Hazan, and Shay Moran. Boosting simple learners. *TheoretiCS*, 2, 2023. URL: `https://doi.org/10.46298/theoretics.23.8`, `doi:10.46298/THEORETICS.23.8`.

[AN10] Terrence M. Adams and Andrew B. Nobel. Uniform convergence of Vapnik-Chervonenkis classes under ergodic sampling. *The Annals of Probability*, 38(4):1345–1367, 2010. `doi:10.1214/09-AOP511`.

[AN12] Terrence M. Adams and Andrew B. Nobel. Uniform approximation of Vapnik-Chervonenkis classes. *Bernoulli*, 18(4):1405–1420, 2012. `doi:10.3150/11-BEJ379`.

[APU23] Hassan Ashtiani, Vinayak Pathak, and Ruth Urner. Adversarially robust learning with tolerance. In Shipra Agrawal and Francesco Orabona, editors, *International Conference on Algorithmic Learning Theory, February 20-23, 2023, Singapore*, volume 201 of *Proceedings of Machine Learning Research*, pages 115–135. PMLR, 2023. URL: `https://proceedings.mlr.press/v201/ashtiani23a.html`.

[BBHS22] Maria-Florina Balcan, Avrim Blum, Steve Hanneke, and Dravyansh Sharma. Robustly-reliable learners under poisoning attacks. In *Conference on Learning Theory*, Proceedings of Machine Learning Research, pages 4498–4534. PMLR, 2022. URL: `https://proceedings.mlr.press/v178/balcan22a.html`.

[BCM+13] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

[BDGR22] Adam Block, Yuval Dagan, Noah Golowich, and Alexander Rakhlin. Smoothed online learning is as easy as statistical learning. In Po-Ling Loh and Maxim Raginsky, editors, *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 1716–1786. PMLR, 2022. URL: `https://proceedings.mlr.press/v178/block22a.html`.

[BEHW89] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. ACM*, 36(4):929–965, 1989. `doi:10.1145/76359.76371`.

[BHQS21] Avrim Blum, Steve Hanneke, Jian Qian, and Han Shao. Robust learning under clean-label attack. In *Conference on Learning Theory, COLT 2021*, Proceedings of Machine Learning Research, pages 591–634. PMLR, 2021. URL: http://proceedings.mlr.press/v134/blum21a.html.

[BK25] Moïse Blanchard and Samory Kpotufe. Distributionally-constrained adversaries in online learning. *CoRR*, abs/2506.10293, 2025. URL: https://doi.org/10.48550/arXiv.2506.10293, arXiv:2506.10293, doi:10.48550/ARXIV.2506.10293.

[BKM97] Shai Ben-David, Eyal Kushilevitz, and Yishay Mansour. Online learning versus offline learning. *Machine Learning*, 29(1):45–63, 1997. doi:10.1023/A:1007465907571.

[BKMS21] Mark Braverman, Gillat Kol, Shay Moran, and Raghuvansh R. Saxena. Near optimal distributed learning of halfspaces with two parties. In *Proceedings of the 34th Annual Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 724–758, 2021.

[Bla25] Moïse Blanchard. Agnostic smoothed online learning. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pages 1997–2006. ACM, 2025. doi:10.1145/3717823.3718111.

[BPS09] Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009. URL: http://www.cs.mcgill.ca/%7Ecolt2009/papers/032.pdf#page=1.

[BRS24] Adam Block, Alexander Rakhlin, and Abhishek Shetty. On the performance of empirical risk minimization with smoothed data. In Shipra Agrawal and Aaron Roth, editors, *The Thirty Seventh Annual Conference on Learning Theory, June 30 - July 3, 2023, Edmonton, Canada*, volume 247 of *Proceedings of Machine Learning Research*, pages 596–629. PMLR, 2024. URL: https://proceedings.mlr.press/v247/block24a.html.

[BSR26] Moïse Blanchard, Abhishek Shetty, and Alexander Rakhlin. Characterizing online and private learnability under distributional constraints via generalized smoothness. *CoRR*, abs/2602.20585, 2026. URL: https://doi.org/10.48550/arXiv.2602.20585, arXiv:2602.20585, doi:10.48550/ARXIV.2602.20585.

[CBL06] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[CDM16] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *Algorithmic Learning Theory*, volume 9925 of *Lecture Notes in Computer Science*, pages 67–82. Springer, 2016. doi:10.1007/978-3-319-46379-7_5.

[CHMS25] Zachary Chase, Steve Hanneke, Shay Moran, and Jonathan Shafer. Optimal mistake bounds for transductive online learning. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025*, 2025. URL: https://arxiv.org/abs/2512.12567.

[Cho70] C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi:10.1109/TIT.1970.1054406.

[DGM22] Kunal Dutta, Arijit Ghosh, and Shay Moran. Uniform brackets, containers, and combinatorial Macbeath regions. In *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*, volume 215 of *LIPIcs*, pages 59:1–59:10. Schloss Dagstuhl - Leibniz-Zentrum f"ur Informatik, 2022. URL: https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ITCS.2022.59, doi:10.4230/LIPIcs.ITCS.2022.59.

[DP26] Shaddin Dughmi and Alireza F. Pour. Relatively smart: A new approach for instance-optimal learning. In Steve Hanneke and Tor Lattimore, editors, *The Thirty Ninth Annual Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pages 2107–2144. PMLR, 2026. URL: https://proceedings.mlr.press/v336/dughmi26a.html.

[EG26] Ezra Edelman and Surbhi Goel. Reliable abstention under adversarial injections: Lower bounds and new upper bounds. In *Advances in Neural Information Processing Systems 39: Annual Conference on Neural Information Processing Systems 2026, NeurIPS 2026*, 2026. URL: https://arxiv.org/abs/2602.20111.

[EYW10] Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010. URL: https://jmlr.org/papers/v11/el-yaniv10a.html.

[FMS04] Yoav Freund, Yishay Mansour, and Robert E. Schapire. Generalization bounds for averaged classifiers. *The Annals of Statistics*, 32(4):1698 – 1722, 2004. doi:10.1214/009053604000000058.

[GHMS23] Surbhi Goel, Steve Hanneke, Shay Moran, and Abhishek Shetty. Adversarial resilience in sequential prediction via abstention. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10–16, 2023*, 2023. URL: https://proceedings.neurips.cc/paper_files/paper/2023/hash/1967f962c7c2083618236d80eeb9d1ac-Abstract-Conference.html.

[GKKM20] Shafi Goldwasser, Adam Tauman Kalai, Yael Kalai, and Omar Montasser. Beyond perturbations: Learning guarantees with arbitrary adversarial test examples. In *Advances in Neural Information Processing Systems 33*, 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/b6c8cf4c587f2ead0c08955ee6e2502b-Abstract.html.

[GSS15] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL: http://arxiv.org/abs/1412.6572.

[Hau95] David Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995. doi:10.1016/0097-3165(95)90052-7.

[Hei25] Carolin Heinzler. Adversarial resilience against clean-label attacks in realizable and noisy settings. *CoRR*, abs/2504.13966, 2025. URL: https://doi.org/10.48550/arXiv.2504.13966, arXiv:2504.13966, doi:10.48550/ARXIV.2504.13966.

[HMS23] Steve Hanneke, Shay Moran, and Jonathan Shafer. A trichotomy for transductive online learning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10–16, 2023*, 2023. URL: https://proceedings.neurips.cc/paper_files/paper/2023/hash/3e32af2df2cd13dfbcbe6e8d38111068-Abstract.html.

[HRS20] Nika Haghtalab, Tim Roughgarden, and Abhishek Shetty. Smoothed analysis of online and differentially private learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[HRS24] Nika Haghtalab, Tim Roughgarden, and Abhishek Shetty. Smoothed analysis with adaptive adversaries. *J. ACM*, 71(3):19, 2024. doi:10.1145/3656638.

[HW06] Radu Herbei and Marten H. Wegkamp. Classification with reject option. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 34(4):709–721, 2006. doi:10.1002/cjs.5550340410.

[Kiv95] Jyrki Kivinen. Learning reliably and with one-sided error. *Math. Syst. Theory*, 28(2):141–172, 1995. doi:10.1007/BF01191474.

[KK05] Sham M. Kakade and Adam Tauman Kalai. From batch to transductive online learning. In *Advances in Neural Information Processing Systems 18, NIPS 2005, Vancouver, British Columbia, Canada, December 5–8, 2005*, pages 611–618, 2005. URL: https://proceedings.neurips.cc/paper_files/paper/2005/hash/17693c91d9204b7a7646284bb3adb603-Abstract.html.

[KSV24] Adam R. Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. Testable learning with distribution shift. In *The Thirty Seventh Annual Conference on Learning Theory*, Proceedings of Machine Learning Research, pages 2887–2943. PMLR, 2024. URL: https://proceedings.mlr.press/v247/klivans24a.html.

[Lit87] Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Mach. Learn.*, 2(4):285–318, 1987. doi:10.1007/BF00116827.

[LLWS11] Lihong Li, Michael L. Littman, Thomas J. Walsh, and Alexander L. Strehl. Knows what it knows: a framework for self-aware learning. *Mach. Learn.*, 82(3):399–443, 2011. URL: https://doi.org/10.1007/s10994-010-5225-4, doi:10.1007/S10994-010-5225-4.

[LM12] Alessandro Lazaric and Rémi Munos. Learning with stochastic inputs and adversarial outputs. *J. Comput. Syst. Sci.*, 78(5):1516–1537, 2012. URL: https://doi.org/10.1016/j.jcss.2011.12.027, doi:10.1016/J.JCSS.2011.12.027.

[Mal12] Yu. V. Malykhin. Bracketing entropy and VC-dimension. *Mathematical Notes*, 91(6):800–807, 2012. doi:10.1134/S0001434612050264.

[MHS19] Omar Montasser, Steve Hanneke, and Nathan Srebro. VC classes are adversarially robustly learnable, but only improperly. In *Conference on Learning Theory, COLT 2019*, Proceedings of Machine Learning Research, pages 2512–2530. PMLR, 2019. URL: http://proceedings.mlr.press/v99/montasser19a.html.

[Mon23] Omar Montasser. Theoretical foundations of adversarially robust learning. *CoRR*, abs/2306.07723, 2023. URL: https://doi.org/10.48550/arXiv.2306.07723, arXiv:2306.07723, doi:10.48550/ARXIV.2306.07723.

[MSZ25] Omar Montasser, Abhishek Shetty, and Nikita Zhivotovskiy. Beyond worst-case online classification: Vc-based regret bounds for relaxed benchmarks. In Nika Haghtalab and Ankur Moitra, editors, *The Thirty Eighth Annual Conference on Learning Theory*, Proceedings of Machine Learning Research, pages 4168–4202. PMLR, 2025. URL: https://proceedings.mlr.press/v291/montasser25a.html.

[PKSV26] Shyamal Patel, Adam R. Klivans, Konstantinos Stavropoulos, and Arsen Vasilyan. Equivalence of coarse and fine-grained models for learning with distribution shift. In Steve Hanneke and Tor Lattimore, editors, *The Thirty Ninth Annual Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pages 4022–4049. PMLR, 2026. URL: https://proceedings.mlr.press/v336/patel26a.html.

[RS88] Ronald L. Rivest and Robert H. Sloan. Learning complicated concepts reliably and usefully. In Howard E. Shrobe, Tom M. Mitchell, and Reid G. Smith, editors, *Proceedings of the 7th National Conference on Artificial Intelligence*, pages 635–640. AAAI Press / The MIT Press, 1988. URL: http://www.aaai.org/Library/AAAI/1988/aaai88-113.php.

[RST11] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Stochastic, constrained, and smoothed adversaries. In *Advances in Neural Information Processing Systems 24*, pages 1764–1772, 2011. URL: https://proceedings.neurips.cc/paper/2011/hash/692f93be8c7a41525c0baf2076aecfb4-Abstract.html.

[SZB10] Amin Sayedi, Morteza Zadimoghaddam, and Avrim Blum. Trading off mistakes and don't-know predictions. In *Advances in Neural Information Processing Systems 23*, pages 2092–2100. Curran Associates, Inc., 2010. URL: https://proceedings.neurips.cc/paper/2010/hash/286674e3082feb7e5afb92777e48821f-Abstract.html.

[SZS+14] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL: http://arxiv.org/abs/1312.6199.

[VC71] 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.

[vH13] Ramon van Handel. The universal Glivenko-Cantelli property. *Probability Theory and Related Fields*, 155(3–4):911–934, 2013. doi:10.1007/s00440-012-0416-5.

[YB26] Jialin Yu and Moïse Blanchard. Distribution-free sequential prediction with abstentions. In Steve Hanneke and Tor Lattimore, editors, *The Thirty Ninth Annual Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pages 6976–7011. PMLR, 2026. URL: https://proceedings.mlr.press/v336/yu26a.html.

A Proof of Theorem 1

In this section, we provide a full proof of Theorem 1. We split the proof into two main parts: (a) misclassification error analysis, and (b) abstention error analysis.

A.1 Misclassification Error Analysis

We begin with stating a standard Lemma on the regret guarantee of Multiplicative Weights.

Lemma 3 (See, e.g. Theorem 2.2 and Theorem 2.4 in [CBL06])**.** *Given a finite set of experts $F = \{f_1, \ldots, f_N\}$ and an arbitrary sequence of loss functions $\ell_1, \ldots, \ell_T : F \to [0,1]$, running the Multiplicative Weights algorithm using experts F with parameter $\eta > 0$ produces a sequence of mixture strategies $p_1, \ldots, p_T$ which satisfy*

$$\sum_{t=1}^{T} \mathbb{E}_{f\sim p_t} \ell_t(f) - L^\star \le \frac{\log N}{\eta} + \frac{\eta T}{8}, \quad \text{where } L^\star := \min_{f\in F} \sum_{t=1}^{T} \ell_t(f).$$

In particular, choosing $\eta = \sqrt{8\log N/T}$ guarantees $\sum_{t=1}^{T} \mathbb{E}_{f\sim p_t} \ell_t(f) - L^\star \le \sqrt{(T/2)\log N}$. Moreover, if $L^\star = 0$, choosing $\eta \ge \log(2)$, guarantees $\sum_{t=1}^{T} \mathbb{E}_{f\sim p_t} \ell_t(f) \le 2\log N$. These guarantees hold for every sequence of loss functions, including sequences chosen adaptively (e.g., by an adaptive adversary), since each p_t is a deterministic function of $\ell_1, \ldots, \ell_{t-1}$.

The following *key* Lemma states that an expert E_b based on a bracket b for a predictor h never makes more misclassifications than h.

Lemma 4. *Let $h \in \mathcal{H}$, let $b = (b^-, b^+)$ bracket h, and let (x, y) be any labeled example. Then*

$$\mathbb{1}\big[E_b(x) \in \{0,1\} \text{ and } E_b(x) \ne y\big] \le \mathbb{1}\big[h(x) \ne y\big].$$

Proof. If $E_b(x) = 1$, then $b^-(x) = 1$, hence $h(x) = 1$ because $b^- \le h$. Thus the left-hand side equals $\mathbb{1}[y = 0] = \mathbb{1}[h(x) \ne y]$. If $E_b(x) = 0$, then $b^+(x) = 0$, hence $h(x) = 0$ because $h \le b^+$. Thus the left-hand side equals $\mathbb{1}[y = 1] = \mathbb{1}[h(x) \ne y]$. If $E_b(x) = \perp$, the left-hand side is 0. □

We are now ready to bound the misclassification error.

Lemma 5 (Misclassification Error)**.** *For any distribution D over $\mathcal{X}$ and any (possibly adaptive) adversary, Algorithm 1 satisfies*

$$\mathbb{E}\left[\text{MissErr} - \text{OPT}\right] \le \sqrt{(T^2/2\tau) \cdot \log \mathcal{N}_{[\,]}(\mathcal{H}, \mathcal{B}, 1/2T)}.$$

Moreover, in the realizable case where $\text{OPT} = 0$,

$$\mathbb{E}\left[\text{MissErr}\right] \le 2(T/\tau) \cdot \log \mathcal{N}_{[\,]}(\mathcal{H}, \mathcal{B}, 1/2T).$$

Proof. Recall that Algorithm 1 divides the time horizon $\{1, \ldots, T\}$ into blocks of length τ, with a total of $R = \lceil T/\tau \rceil$ blocks. For each block $r \in \{1, \ldots, R\}$, recall that $I_r = \{(r-1)\tau + 1, \ldots, \min(r\tau, T)\}$ is the set of timesteps in block r.

For each timestep t, let H_{t-1} denote the history of the interaction before time t (including the adversary's choices and Algorithm 1's random choices and predictions), and let H^+_{t-1} denote H_{t-1} together with (c_t, x_t, y_t), i.e., the history after the adversary has selected the current example and its label, but before Algorithm 1 randomizes at time t. For each block $r \in \{2, \ldots, R\}$, timestep $t \in I_r$, and bracket $b \in \mathcal{B}_r$, let $\ell_t(E_b) := \mathbb{1}\{E_b(x_t) = 1 - y_t\}$ denote the misclassification loss of expert E_b at time t. Observe that the family $\mathcal{B}_r$ and the mixture p_t are determined by H_{t-1}, and that Algorithm 1 samples $E_b \sim p_t$ using fresh randomness that is independent of H^+_{t-1}. Therefore,

$$\mathbb{E}\left[\mathbb{1}\big[\hat{y}_t \in \{0,1\} \text{ and } \hat{y}_t \ne y_t\big] \,\middle|\, \mathsf{H}^+_{t-1}\right] = \mathbb{E}_{E_b\sim p_t} \ell_t(E_b).$$

Since Algorithm 1 abstains on every timestep of block 1, by the law of total expectation,

$$\mathbb{E}\left[\text{MissErr}\right] = \mathbb{E}\left[\sum_{r=2}^{R} \sum_{t\in I_r} \mathbb{E}_{E_b\sim p_t} \ell_t(E_b)\right].$$

We will bound the quantity inside the expectation for every realization of the interaction between the adversary and Algorithm 1.

Fix such a realization, and let $h_\star \in \arg\min_{h\in\mathcal{H}} \sum_{t=1}^{T} \mathbb{1}\left[h(x_t) \neq y_t\right]$ for the realized sequence $(x_t, y_t)_{t=1}^{T}$. For each block $r \in \{2, \ldots, R\}$, the family $\mathcal{B}_r$ contains some bracket $b_{r,\star} = (b^-_{r,\star}, b^+_{r,\star})$ such that

$$b^-_{r,\star}(x) \leq h_\star(x) \leq b^+_{r,\star}(x) \;\; \forall x \in \mathcal{X}.$$

By invoking Lemma 4, it follows that the total misclassification error of expert $E_{b_{r,\star}}$ is bounded from above by the total misclassification error of $h_\star$,

$$\sum_{t\in I_r} \ell_t(E_{b_{r,\star}}) \leq \sum_{t\in I_r} \mathbb{1}\left[h_\star(x_t) \neq y_t\right].$$

Within block r, the mixtures $(p_t)_{t\in I_r}$ are the iterates of Multiplicative Weights on the loss functions $(\ell_t)_{t\in I_r}$, and the regret guarantee of Multiplicative Weights (Lemma 3) holds for every sequence of loss functions, in particular for the realized one, even though it may have been chosen adaptively. Combining this with the inequality above, and recalling that $|I_r| \leq \tau$, gives

$$\sum_{t\in I_r} \mathbb{E}_{E_b\sim p_t} \ell_t(E_b) \leq \sum_{t\in I_r} \mathbb{1}\left[h_\star(x_t) \neq y_t\right] + \sqrt{\frac{\tau}{2}\log|\mathcal{B}_r|}.$$

Summing over $r = 2, \ldots, R$ yields

$$\begin{aligned}\sum_{r=2}^{R}\sum_{t\in I_r} \mathbb{E}_{E_b\sim p_t} \ell_t(E_b) &\leq \sum_{r=2}^{R}\sum_{t\in I_r} \mathbb{1}\left[h_\star(x_t) \neq y_t\right] + \sum_{r=2}^{R}\sqrt{\frac{\tau}{2}\log|\mathcal{B}_r|} \\ &\leq \mathrm{OPT} + \sqrt{\frac{T^2}{2\tau}\log \mathcal{N}_{[\,]}(\mathcal{H}, \mathcal{B}, 1/2T)}.\end{aligned}$$

Since this holds for every realization of the interaction, taking expectations proves the first inequality. The proof of the second inequality is identical, but we invoke the realizable guarantee of Multiplicative Weights instead (Lemma 3), which applies since $\mathrm{OPT} = 0$ implies that $\sum_{t\in I_r} \ell_t(E_{b_{r,\star}}) = 0$ for every block r. □

A.2 Abstention Error Analysis

We now turn to bounding the abstention error of Algorithm 1. We begin with stating a standard uniform convergence guarantee for VC classes.

Lemma 6 (Theorem A2.1 and Proposition A2.1 of [BEHW89])**.** *Let $\mathcal{G} \subseteq \{0,1\}^{\mathcal{X}}$ be a binary class with finite VC dimension $d_{\mathcal{G}} < \infty$. Then, for any $\delta \in (0,1), m \geq 1$, and any distribution D over $\mathcal{X}$, with probability at least $1-\delta$ over $X_1, \ldots, X_m \overset{\text{iid}}{\sim} D$,*

$$\forall g \in \mathcal{G}: \;\; \frac{1}{m}\sum_{i=1}^{m} g(X_i) = 0 \qquad \Longrightarrow \qquad \mathbb{P}_D(g(X) = 1) \leq \frac{2}{m}\left(d_{\mathcal{G}}\log\left(\frac{2em}{d_{\mathcal{G}}}\right) + \log\left(\frac{2}{\delta}\right)\right).$$

We are now ready to state and prove a *key* Lemma showing that if we are constructing brackets for $\mathcal{H}$ using a bracketing class $\mathcal{B}$ with bounded VC dimension (see Definition 2), then it is possible to construct such brackets from i.i.d. samples from a distribution D even after an adversary contaminates the samples with arbitrary adversarial injections. Note that the adversarial injections may be chosen *after seeing the entire* i.i.d. *sample*, and they may depend arbitrarily on it.

Lemma 7 (Constructing Brackets with Adversarially Contaminated Samples)**.** *Fix a bracketing class $\mathcal{B} \subseteq \{0,1\}^{\mathcal{X}}$, a distribution D on $\mathcal{X}$, an integer $m \geq 1$ and $\delta \in (0,1)$. With probability at least $1-\delta$ over $Z_1, \ldots, Z_m \overset{\text{iid}}{\sim} D$:*

for every $n \geq m$, every sequence $S = (x_1, \ldots, x_n)$ which contains $Z_1, \ldots, Z_m$, and every $b^-, b^+ \in \mathcal{B}$,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{b^-(x_i) \neq b^+(x_i)\} \leq \frac{1}{2n} \Longrightarrow \mathbb{P}_D(b^-(Z) \neq b^+(Z)) \leq \frac{2}{m}\left(d_{\mathcal{B}}\log\left(\frac{2em}{d_{\mathcal{B}}}\right) + \log\left(\frac{2}{\delta}\right)\right).$$

Proof. Consider the class of parities over $\mathcal{B}$: $\{x \mapsto \mathbb{1}\{b(x) \neq b'(x)\} \mid b, b' \in \mathcal{B}\}$. By classical results [BEHW89], it follows that the VC dimension of this class is at most $O(d_{\mathcal{B}})$. Let A denote the event over $Z_1, \ldots, Z_m \overset{\text{iid}}{\sim} D$ that for every $b^-, b^+ \in \mathcal{B}$,

$$\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{b^-(Z_i) \neq b^+(Z_i)\} = 0 \Longrightarrow \mathbb{P}_D(b^-(Z) \neq b^+(Z)) \leq \frac{2}{m}\left(d_{\mathcal{B}} \log\left(\frac{2em}{d_{\mathcal{B}}}\right) + \log\left(\frac{2}{\delta}\right)\right).$$

By invoking Lemma 6, it follows that $\mathbb{P}(A) \geq 1 - \delta$. Observe now that for any $Z_1, \ldots, Z_m \overset{\text{iid}}{\sim} D$ and any sequence $S \in \mathcal{X}^n$ which contains $Z_1, \ldots, Z_m$, for every $b^-, b^+ \in \mathcal{B}$,

$$\frac{1}{n}\sum_{x \in S} \mathbb{1}\{b^-(x) \neq b^+(x)\} \leq \frac{1}{2n} \Longrightarrow \frac{1}{n}\sum_{x \in S} \mathbb{1}\{b^-(x) \neq b^+(x)\} = 0 \Longrightarrow \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{b^-(Z_i) \neq b^+(Z_i)\} = 0.$$

Therefore, it follows that event A implies the target event A' over $Z_1, \ldots, Z_m \overset{\text{iid}}{\sim} D$:

$\forall n \geq m, \forall S \in \mathcal{X}^n$ such that $\forall_{1 \leq i \leq m} Z_i \in S, \forall b^-, b^+ \in \mathcal{B}$,

$$\frac{1}{n}\sum_{x \in S} \mathbb{1}\{b^-(x) \neq b^+(x)\} \leq \frac{1}{2n} \Longrightarrow \mathbb{P}_D(b^-(Z) \neq b^+(Z)) \leq \frac{2}{m}\left(d_{\mathcal{B}} \log\left(\frac{2em}{d_{\mathcal{B}}}\right) + \log\left(\frac{2}{\delta}\right)\right).$$

Thus, $\mathbb{P}(A') \geq 1 - \delta$. This argument is deterministic once the clean sample is fixed, and therefore it holds simultaneously for all realized contamination sequences, regardless of how they were selected based on $(Z_1, \ldots, Z_m)$. □

We are now ready to proceed with bounding the abstention error of Algorithm 1.

Lemma 8 (Abstention Error)**.** *Algorithm 1 satisfies*

$$\mathbb{E}[\text{AbsErr}] \leq 2\tau + O\left(d_{\mathcal{B}} \log(T)\right) \log\left(\frac{T}{\tau}\right).$$

Proof. Recall that Algorithm 1 divides the time horizon $\{1, \ldots, T\}$ into blocks of length τ, with a total of $R = \lceil T/\tau \rceil$ blocks. For each block $r \in \{1, \ldots, R\}$, recalling that $I_r = \{(r-1)\tau + 1, \ldots, \min(r\tau, T)\}$ is the set of timesteps in block r, let

$$m_r := \sum_{t < (r-1)\tau + 1} \mathbb{1}[c_t = 0], \qquad M_r := \sum_{t \in I_r} \mathbb{1}[c_t = 0].$$

So m_r is the number of clean iid examples before the start of block r, and M_r is the number of clean iid examples appearing during block r.

An equivalent view of interaction. We first describe a helpful way of thinking about the interaction between Algorithm 1 and the adversary. Let $Z_1, Z_2, \ldots \overset{\text{iid}}{\sim} D$ be an auxiliary sequence generated in advance independently of the adversary. At time t, let $N_{t-1} := \sum_{s<t} \mathbb{1}\{c_s = 0\}$. If the adversary chooses $c_t = 0$, set $x_t := Z_{N_{t-1}+1}$; if it chooses $c_t = 1$, then the adversary chooses an arbitrary adversarial injection $x_t \in \mathcal{X}$. Because the adversary decides $c_t \in \{0, 1\}$ before seeing $Z_{N_{t-1}+1}$, it holds that $Z_{N_{t-1}+1} \sim D$ and is independent of the past, so this construction has the same law as the original interaction. Moreover, almost surely, for every t, the clean examples among $x_1, \ldots, x_t$, listed chronologically, are exactly $Z_1, \ldots, Z_{N_t}$. In particular, for every block r, the history S_r contains $Z_1, \ldots, Z_{m_r}$ as a subsequence. It therefore suffices to prove the claimed bound under this view.

We now proceed with the analysis of the abstention error, noting that throughout the proof, expectations are over the draws of iid examples $Z_1, Z_2, \ldots, \overset{\text{iid}}{\sim} D$ (as in the view described above), the adversary's possible randomness, and Algorithm 1's internal randomness.

Let r_τ be the first block where the number of iid examples available before the start of the block is at least τ

$$r_\tau := \min\{r \in \{2, \ldots, R+1\} : m_r \geq \tau\},$$

with the convention $r_\tau = R + 1$ if no such block exists.

We will analyze the expected AbsErr on iid examples by splitting it into two parts: (a) expected AbsErr for all blocks $r < r_\tau$ (before collecting τ iid examples), and (b) expected AbsErr for all blocks $r \geq r_\tau$ (after collecting at least τ iid examples).

$$\mathbb{E}\left[\mathrm{AbsErr}\right] = \mathbb{E}\left[\sum_{r=1}^{R}\sum_{t\in I_r}\mathbb{1}\{c_t = 0 \ \wedge \ \hat{y}_t = \perp\}\right] \tag{8}$$

$$= \mathbb{E}\left[\sum_{r<r_\tau}\sum_{t\in I_r}\mathbb{1}\{c_t = 0 \ \wedge \ \hat{y}_t = \perp\} + \sum_{r\geq r_\tau}\sum_{t\in I_r}\mathbb{1}\{c_t = 0 \ \wedge \ \hat{y}_t = \perp\}\right]. \tag{9}$$

Abstention error before collecting τ iid examples. The total number of clean iid examples in blocks $1, \ldots, r_\tau - 1$ equals m_{r_τ}. If $r_\tau = R + 1$, then $m_{r_\tau} < \tau$ by definition. Otherwise, by definition of r_τ, note that $m_{r_\tau - 1} < \tau$, and because each block has length τ, the number of iid examples that appeared in block $r_\tau - 1$ can be at most τ, $M_{r_\tau - 1} \leq \tau$. It follows then that $m_{r_\tau} = m_{r_\tau - 1} + M_{r_\tau - 1} < 2\tau$. Therefore, the maximum number of abstentions on clean iid examples that happen in blocks $1, \ldots, r_\tau - 1$ is at most $2\tau - 1$:

$$\mathbb{E}\left[\sum_{r<r_\tau}\sum_{t\in I_r}\mathbb{1}\{c_t = 0 \ \wedge \ \hat{y}_t = \perp\}\right] \leq 2\tau - 1.$$

Abstention error after collecting at least τ iid examples. For the remainder of the analysis, we will need to introduce probability events that capture whether the (random) iid examples collected can be used to construct valid brackets with low abstention rate. Specifically, for each $j \in \{1, \ldots, T\}$, let A_j be the event in Lemma 7 for the clean iid examples $Z_1, \ldots, Z_j$ with confidence parameter $\delta_j = 1/T^2$. On event A_j, every adversarially contaminated sequence $S \in \mathcal{X}^n$ containing the clean examples $Z_1, \ldots, Z_j$ and every $b^-, b^+ \in \mathcal{B}$ satisfy

$$\frac{1}{n}\sum_{x\in S}\mathbb{1}\left\{b^-(x) \neq b^+(x)\right\} \leq \frac{1}{2n} \quad \Longrightarrow \quad \mathbb{P}_D(b^-(Z) \neq b^+(Z)) \leq \frac{O\left(d_{\mathcal{B}}\log(T)\right)}{j}. \tag{10}$$

Now, recall that in each round $r \in \{2, \ldots, R\}$, Algorithm 1 works by constructing brackets $\mathcal{B}_r$ based on the entire sequence of examples S_r available so far. However, we do not know a priori the number of clean examples $m_r \in \{1, \ldots, T\}$ that are contained in S_r since it depends on the interaction between the adversary and Algorithm 1 before the start of block r. So, we do not know apriori which events A_j for $j \in \{1, \ldots, T\}$ will be relevant for the analysis of Algorithm 1. Hence, we will consider the intersection event

$$A := \bigcap_{j=1}^{T} A_j,$$

where all events $A_1, \ldots, A_T$ hold simultaneously. By a union bound,

$$\mathbb{P}(\bar{A}) \leq \sum_{j=1}^{T}\mathbb{P}(\bar{A}_j) \leq \sum_{j=1}^{T}\frac{1}{T^2} = \frac{1}{T}.$$

When event A does not hold, the total abstention error on clean examples is at most T, therefore, by law of total expectation,

$$\mathbb{E}\left[\sum_{r\geq r_\tau}\sum_{t\in I_r}\mathbb{1}\{c_t = 0 \ \wedge \ \hat{y}_t = \perp\}\right] \leq \mathbb{E}\left[\mathbb{1}\{A\}\sum_{r\geq r_\tau}\sum_{t\in I_r}\mathbb{1}\{c_t = 0 \ \wedge \ \hat{y}_t = \perp\}\right] + \mathbb{P}(\bar{A})T \tag{11}$$

$$\leq \mathbb{E}\left[\mathbb{1}\{A\}\sum_{r\geq r_\tau}\sum_{t\in I_r}\mathbb{1}\{c_t = 0 \ \wedge \ \hat{y}_t = \perp\}\right] + 1. \tag{12}$$

Thus, it remains to complete the analysis of abstention error when event A holds. Fix a block $r \in \{r_\tau, \dots, R\}$. By definition of r_τ, the number of iid examples available before the start of block r, $m_r \geq \tau$. Note that the set of collected examples so far S_r contains $Z_1, \dots, Z_{m_r}$ as a subset. By the construction of the brackets $\mathcal{B}_r$, every bracket $b = (b^-, b^+) \in \mathcal{B}_r$ satisfies

$$\frac{1}{n_r} \sum_{x \in S_r} \mathbb{1}\left\{b^-(x) \neq b^+(x)\right\} \leq \frac{1}{2n_r}.$$

Therefore, under event A_{m_r}, (10) implies

$$\mathbb{P}_{Z \sim D}(E_b(Z) = \perp) \leq \frac{O\left(d_{\mathcal{B}} \log(T)\right)}{m_r} \quad \text{for every } b \in \mathcal{B}_r.$$

So, we can now use this property to argue that the expected number of abstentions on iid examples in block r will be low because all experts have a low probability of abstaining.

Fix a timestep $t \in I_r$ in block r, and condition on the history of the interaction up to time $t-1$. Observe that the mixture p_t over experts in $\mathcal{B}_r$ maintained by Algorithm 1 is already fixed. If $c_t = 0$, then $x_t = Z_{N_{t-1}+1}$ and $Z_{N_{t-1}+1} \sim D$ independent of the history so far, thus conditioning on A_{m_r},

$$\mathbb{P}\left(\hat{y}_t = \perp \,|c_t = 0, A_{m_r}\right) = \sum_{b \in \mathcal{B}_r} p_t(E_b)\mathbb{P}(E_b = \perp) \leq \sum_{b \in \mathcal{B}_r} \left(p_t(E_b) \frac{O\left(d_{\mathcal{B}} \log(T)\right)}{m_r}\right) = \frac{O\left(d_{\mathcal{B}} \log(T)\right)}{m_r}.$$

Summing over $t \in I_r$ gives

$$\mathbb{E}\left[\mathbb{1}\{A_{m_r}\} \sum_{t \in I_r} \mathbb{1}\{c_t = 0 \ \wedge \ \hat{y}_t = \perp\}\right] \leq \mathbb{E}\left[M_r \frac{O\left(d_{\mathcal{B}} \log(T)\right)}{m_r}\right].$$

Since $A \subseteq A_{m_r}$, summing over all blocks $r \geq r_\tau$ yields

$$\mathbb{E}\left[\mathbb{1}\{A\} \sum_{r=r_\tau}^{R} \sum_{t \in I_r} \mathbb{1}\{c_t = 0 \ \wedge \ \hat{y}_t = \perp\}\right] \leq \mathbb{E}\left[\sum_{r=r_\tau}^{R} \frac{M_r}{m_r}\right] \cdot O\left(d_{\mathcal{B}} \log(T)\right).$$

For every late block $r \geq r_\tau$, we have $m_r \geq \tau$ and $M_r \leq \tau$, so $0 \leq M_r/m_r \leq 1$. Using $u \leq 2\log(1+u)$ for $u \in [0,1]$ and $m_{r+1} = m_r + M_r$,

$$\frac{M_r}{m_r} \leq 2\log\left(1 + \frac{M_r}{m_r}\right) = 2\log\left(\frac{m_{r+1}}{m_r}\right).$$

Consequently, summing over $r = r_\tau, \dots, R$ telescopes to

$$\sum_{r=r_\tau}^{R} \frac{M_r}{m_r} \leq 2 \sum_{r=r_\tau}^{R} \log\left(\frac{m_{r+1}}{m_r}\right) \leq 2\log\left(\frac{m_{R+1}}{m_{r_\tau}}\right) \leq 2\log\left(\frac{T}{\tau}\right).$$

Putting everything together,

$$\mathbb{E}\left[\text{AbsErr}\right] = \mathbb{E}\left[\sum_{r=1}^{R} \sum_{t \in I_r} \mathbb{1}\{c_t = 0 \ \wedge \ \hat{y}_t = \perp\}\right] \tag{13}$$

$$\leq 2\tau - 1 + \mathbb{E}\left[\mathbb{1}\{A\} \sum_{r=r_\tau}^{R} \sum_{t \in I_r} \mathbb{1}\{c_t = 0 \ \wedge \ \hat{y}_t = \perp\}\right] + 1 \tag{14}$$

$$\leq 2\tau + O\left(d_{\mathcal{B}} \log(T)\right) \log\left(\frac{T}{\tau}\right). \tag{15}$$

This concludes the proof. □

Proof of Theorem 1. The proof follows from Lemma 5 and Lemma 8, by choosing $\tau = \lceil T^{1-\alpha} \rceil$. □

B Censored Feedback and Competing with a Refined OPT

In this section, we focus on a variant of the problem where the learner is not provided with feedback labels y_t in the rounds or timesteps it chooses to abstain $\hat{y}_t = \perp$ (see Remark 1). Importantly, in the agnostic setting, we will require the learner to compete with the best-in-class only in the rounds where the learner did not abstain, which we recall here

$$\mathrm{OPT}_{\mathrm{cen}} = \min_{h \in \mathcal{H}} \sum_{t=1}^{T} \mathbb{1}[\hat{y}_t \neq \perp] \cdot \mathbb{1}[h(x_t) \neq y_t].$$

Note that $\mathrm{OPT}_{\mathrm{cen}}$ is a more refined benchmark than OPT since $\mathrm{OPT}_{\mathrm{cen}} \leq \mathrm{OPT}$. Throughout this section, we write $N := \mathcal{N}_{[\,]}(\mathcal{H}, \mathcal{B}, 1/2T)$, and assume that $N \geq 3$, so that $\log N \geq 1$. We begin by restating Theorem 2.

Theorem 2 (Restated)**.** *Let $\mathcal{H}$ be a hypothesis class, and let $\mathcal{B}$ be a bracket class for $\mathcal{H}$ with VC dimension $d_{\mathcal{B}}$. Let $T \geq 1$ be a time horizon. Then, for any parameter $\alpha \in [0, 1]$, Algorithm 2 achieves the following guarantee against* any *adversary and any distribution D over $\mathcal{X}$:*

$$\mathbb{E}[\mathrm{MissErr} - \mathrm{OPT}_{\mathrm{cen}}] \leq 4\sqrt{T^{1+\alpha} \log N} \quad \textit{and} \quad \mathbb{E}[\mathrm{AbsErr}] \leq 3T^{1-\alpha} + O\left(d_{\mathcal{B}}^2 \log^4 T\right).$$

Moreover, in the realizable setting where $\mathrm{OPT}_{\mathrm{cen}} = 0$, Algorithm 2 satisfies

$$\mathbb{E}[\mathrm{MissErr}] \leq 4T^{\alpha} \log N \quad \textit{and} \quad \mathbb{E}[\mathrm{AbsErr}] \leq 2T^{1-\alpha} + O\left(d_{\mathcal{B}} \log^2 T\right).$$

Algorithm. Algorithm 2 has the same block structure and uses the same bracketing families $\mathcal{B}_r$ and experts $\{E_b : b \in \mathcal{B}_r\}$ as Algorithm 1. It only replaces the inner loop over the timesteps $t \in I_r$ of each block $r \in \{2, \ldots, R\}$. At each timestep, it computes the mass a_t of the experts that abstain on x_t. If $a_t > \rho$, it abstains and therefore receives no label. Otherwise, it predicts using an expert sampled from the experts that do not abstain on x_t, and updates its mixture as in Algorithm 1. The realizable and agnostic settings only differ in the choice of the parameters ρ and η (see the proof of Theorem 2 below).

Algorithm 2: Censored variant of Algorithm 1 (inner loop of block r)

Input: abstention threshold $\rho \in (0, 1/2]$ and learning rate $\eta > 0$.

for $t \in I_r$ **do**

  Receive x_t from the adversary, and compute $a_t = \sum_{b \in \mathcal{B}_r} p_t(E_b) \mathbb{1}\{E_b(x_t) = \perp\}$.

  if $a_t > \rho$ **then**

    Output an abstention $\hat{y}_t = \perp$, receive no label, and set $p_{t+1} = p_t$.

  else

    Sample a random expert $E_b \sim p_t(\cdot \mid E_b(x_t) \neq \perp)$ and predict $\hat{y}_t = E_b(x_t) \in \{0, 1\}$.

    Upon receiving the feedback label y_t, update the mixture p_t:

$$p_{t+1}(E_b) = p_t(E_b) \exp\left(-\eta \mathbb{1}\{E_b(x_t) = 1 - y_t\}\right) / Z_t,$$

    where Z_t is a normalization constant.

Differences from the uncensored setting. Let $J_r := \{t \in I_r : a_t \leq \rho\}$ denote the (random) set of rounds in block r where Algorithm 2 does not abstain, and let $\varepsilon_r := O(d_{\mathcal{B}} \log T)/m_r$ denote the bound, established in the proof of Lemma 8, on the probability that any expert in block r abstains on an i.i.d. example. The following table summarizes the differences between the two algorithms and their analyses.

	Algorithm 1 (Theorem 1)	Algorithm 2 (Theorem 2)
Abstains when	the sampled expert abstains	$a_t > \rho$
Labels observed	in every round	only in the rounds in J_r
Mixture updated on	every round of block r	only the rounds in J_r
Benchmark	OPT	$\mathrm{OPT}_{\mathrm{cen}}$
Misclassification	Lemma 5	Lemma 9: extra factor $1/(1-\rho)$
Abstention	Lemma 8: $\mathbb{E}[a_t] \leq \varepsilon_r$	Lemma 10: $\mathbb{P}(a_t > \rho) \leq \varepsilon_r/\rho$

The extra factor $1/(1-\rho)$ in the misclassification error bound arises because Algorithm 2 must predict even when some experts abstain, and the extra factor $1/\rho$ in the abstention error bound arises because it abstains based on the threshold ρ rather than on the prediction of the sampled expert.

Lemma 9 (Misclassification Error under Censored Feedback)**.** *For any distribution D over $\mathcal{X}$, any (possibly adaptive) adversary, and any $\rho \in (0, 1/2]$, Algorithm 2 with $\eta = \sqrt{8 \log |\mathcal{B}_r|/\tau}$ in each block r satisfies*

$$\mathbb{E}\left[\text{MissErr}\right] \leq \frac{1}{1-\rho}\left(\mathbb{E}\left[\text{OPT}_{\text{cen}}\right] + \sqrt{(T^2/2\tau)\cdot \log N}\right).$$

Moreover, in the realizable case where $\text{OPT}_{\text{cen}} = 0$, Algorithm 2 with $\eta \geq \log(2)$ satisfies $\mathbb{E}\left[\text{MissErr}\right] \leq \frac{1}{1-\rho}\cdot 2(T/\tau)\cdot \log N$.

Proof. We follow the proof of Lemma 5, and use its notation H_{t-1}, H^{+}_{t-1}, and $\ell_t(E_b) = \mathbb{1}\{E_b(x_t) = 1 - y_t\}$. Since the mixture p_t is determined by H_{t-1}, the quantity a_t and the decision whether to abstain are determined by H^{+}_{t-1}, and $\hat{y}_t \neq \perp$ if and only if $t \in J_r$, for every block $r \geq 2$ and $t \in I_r$.

The first difference from the proof of Lemma 5 is that, when it does not abstain, Algorithm 2 samples its expert from $p_t(\cdot \mid E_b(x_t) \neq \perp)$. Since abstaining experts incur no loss, for every $t \in J_r$,

$$\mathbb{P}\left(\hat{y}_t = 1 - y_t \mid \mathsf{H}^{+}_{t-1}\right) = \frac{1}{1-a_t}\sum_{b\in\mathcal{B}_r} p_t(E_b)\,\ell_t(E_b) \leq \frac{1}{1-\rho}\,\mathbb{E}_{E_b\sim p_t}\,\ell_t(E_b).$$

Since Algorithm 2 abstains on every timestep of block 1, by the law of total expectation,

$$\mathbb{E}\left[\text{MissErr}\right] \leq \frac{1}{1-\rho}\,\mathbb{E}\left[\sum_{r=2}^{R}\sum_{t\in J_r}\mathbb{E}_{E_b\sim p_t}\,\ell_t(E_b)\right],$$

and we bound the quantity inside the expectation for every realization of the interaction.

The second difference is that Algorithm 2 only updates its mixture on the rounds in J_r, and so the mixtures $(p_t)_{t\in J_r}$ are the iterates of Multiplicative Weights on the loss functions $(\ell_t)_{t\in J_r}$. Since $|J_r| \leq \tau$, Lemma 3 with $\eta = \sqrt{8\log|\mathcal{B}_r|/\tau}$ bounds the regret on J_r by $\frac{\log|\mathcal{B}_r|}{\eta} + \frac{\eta|J_r|}{8} \leq \sqrt{(\tau/2)\log|\mathcal{B}_r|}$, even though J_r is selected adaptively. Let $h_\star$ be a minimizer in the definition of OPT_{cen} for the realized interaction. As in the proof of Lemma 5, applying Lemma 4 to a bracket of $h_\star$ in $\mathcal{B}_r$, and using $|\mathcal{B}_r| \leq N$, gives

$$\sum_{t\in J_r}\mathbb{E}_{E_b\sim p_t}\,\ell_t(E_b) \leq \sum_{t\in J_r}\mathbb{1}\left[h_\star(x_t) \neq y_t\right] + \sqrt{\frac{\tau}{2}\log N}.$$

Since the rounds where Algorithm 2 does not abstain are exactly $\bigcup_{r\geq 2} J_r$, summing over $r = 2,\ldots,R$ bounds the quantity inside the expectation by $\text{OPT}_{\text{cen}} + \sqrt{(T^2/2\tau)\log N}$, which proves the first claim. For the second claim, $\text{OPT}_{\text{cen}} = 0$ implies that the bracket of $h_\star$ incurs no loss on J_r, and we invoke the realizable guarantee of Lemma 3 instead. □

Lemma 10 (Abstention Error under Censored Feedback)**.** *For any $\rho \in (0,1)$, Algorithm 2 satisfies*

$$\mathbb{E}\left[\text{AbsErr}\right] \leq 2\tau + O\left(\frac{d_{\mathcal{B}}\log T}{\rho}\right)\log\left(\frac{T}{\tau}\right).$$

Proof. The proof is identical to that of Lemma 8, except for the bound on the probability of abstaining on an i.i.d. example in a block $r \geq r_\tau$: Algorithm 1 abstains with probability a_t, whereas Algorithm 2 abstains exactly when $a_t > \rho$. Fix a block $r \geq r_\tau$ and a timestep $t \in I_r$. As in the proof of Lemma 8, on the event A_{m_r}, conditioned on H_{t-1} and on $c_t = 0$, the example x_t is a fresh draw from D on which every expert E_b with $b \in \mathcal{B}_r$ abstains with probability at most ε_r, and therefore $\mathbb{E}[a_t \mid \mathsf{H}_{t-1}, c_t = 0] \leq \varepsilon_r$. By Markov's inequality,

$$\mathbb{P}\left(\hat{y}_t = \perp \mid \mathsf{H}_{t-1}, c_t = 0\right) = \mathbb{P}\left(a_t > \rho \mid \mathsf{H}_{t-1}, c_t = 0\right) \leq \frac{\varepsilon_r}{\rho}.$$

Following the rest of the proof of Lemma 8 with this additional factor of $1/\rho$ proves the claim. □

Proof of Theorem 2. Let $\tau = \lceil T^{1-\alpha} \rceil$. In the agnostic setting, we run Algorithm 2 with $\eta = \sqrt{8 \log |\mathcal{B}_r| / \tau}$ in each block r, and $\rho = \min\left\{1/2, \sqrt{\log N / T^{1-\alpha}}\right\}$. Then $\rho T \le \sqrt{T^{1+\alpha} \log N}$: this holds with equality if $\rho < 1/2$, and otherwise $\log N \ge T^{1-\alpha}/4$, and so $\rho T = T/2 \le \sqrt{T^{1+\alpha} \log N}$. Since $\mathrm{OPT}_{\mathrm{cen}} \le T$ and $\frac{1}{1-\rho} \le 2$, Lemma 9 gives

$$\begin{aligned}\mathbb{E}\left[\mathrm{MissErr} - \mathrm{OPT}_{\mathrm{cen}}\right] &\le \frac{\rho}{1-\rho}\mathbb{E}\left[\mathrm{OPT}_{\mathrm{cen}}\right] + \frac{1}{1-\rho}\sqrt{\frac{T^2}{2\tau}\log N} \\ &\le 2\rho T + \sqrt{2T^{1+\alpha}\log N} \le 4\sqrt{T^{1+\alpha}\log N}.\end{aligned}$$

Moreover, $\frac{1}{\rho} \le 2 + \sqrt{T^{1-\alpha}/\log N} \le 2 + \sqrt{T^{1-\alpha}}$, since $\log N \ge 1$. Therefore, Lemma 10, $\log(T/\tau) \le \log T$, and the AM-GM inequality give

$$\mathbb{E}\left[\mathrm{AbsErr}\right] \le 2\tau + O\left(d_{\mathcal{B}} \log^2 T\right) \cdot \left(2 + \sqrt{T^{1-\alpha}}\right) \le 3T^{1-\alpha} + O\left(d_{\mathcal{B}}^2 \log^4 T\right).$$

In the realizable setting, we run Algorithm 2 with $\eta = \log 2$ and $\rho = 1/2$. Then, Lemma 9 gives $\mathbb{E}\left[\mathrm{MissErr}\right] \le 4(T/\tau)\log N \le 4T^{\alpha}\log N$, and Lemma 10 gives $\mathbb{E}\left[\mathrm{AbsErr}\right] \le 2T^{1-\alpha} + O\left(d_{\mathcal{B}} \log^2 T\right)$. □

B.1 Censored Guarantees for Some Natural Function Classes

Corollary 3 (Halfspaces)**.** *Let $\mathcal{H}$ be halfspaces in $\mathbb{R}^d$. Let $T \ge 1$ be a time horizon. Then, for any parameter $\alpha \in [0,1]$, Algorithm 2 achieves the following guarantee against* any *adversary:*

$$\mathbb{E}[\mathrm{MissErr} - \mathrm{OPT}_{\mathrm{cen}}] \lesssim \sqrt{T^{1+\alpha} \cdot d\log(2dT)} \text{ and } \mathbb{E}[\mathrm{AbsErr}] \lesssim T^{1-\alpha} + d^4 \log^2(d) \log^4(T).$$

Moreover, in the realizable setting where $\mathrm{OPT}_{\mathrm{cen}} = 0$, Algorithm 2 satisfies

$$\mathbb{E}[\mathrm{MissErr}] \lesssim T^{\alpha} \cdot d\log(2dT) \text{ and } \mathbb{E}[\mathrm{AbsErr}] \le 2T^{1-\alpha} + O(d^2 \log(d) \log^2(T)).$$

Corollary 4 (Compositions of Halfspaces)**.** *Let $\mathcal{G} = \{f(h_1, \ldots, h_k) : f \in \mathcal{F}, h_1, \ldots, h_k \in \mathcal{H}\}$, where $\mathcal{H}$ is the class of halfspaces in $\mathbb{R}^d$ and $\mathcal{F}$ is an* arbitrary *function class with VC dimension $d_{\mathcal{F}}$. Let $T \ge 1$ be a time horizon. Then, for any parameter $\alpha \in [0,1]$, Algorithm 2 achieves the following guarantee against* any *adversary:*

$$\begin{aligned}\mathbb{E}[\mathrm{MissErr} - \mathrm{OPT}_{\mathrm{cen}}] &\lesssim \sqrt{T^{1+\alpha} \cdot k(d_{\mathcal{F}} + d\log(2dkT))}, \text{ and} \\ \mathbb{E}[\mathrm{AbsErr}] &\lesssim T^{1-\alpha} + k^2 d^4 \log^2(d) \log^4(T).\end{aligned}$$

Moreover, in the realizable setting where $\mathrm{OPT}_{\mathrm{cen}} = 0$, Algorithm 2 satisfies

$$\mathbb{E}[\mathrm{MissErr}] \lesssim T^{\alpha} \cdot k(d_{\mathcal{F}} + d\log(2dkT)) \text{ and } \mathbb{E}[\mathrm{AbsErr}] \le 2T^{1-\alpha} + O(kd^2 \log(d) \log^2(T)).$$

C Warmup: Known Marginal Distribution

For completeness, we formalize here the simple known-distribution warmup discussed in the technical overview.

Algorithm 3: Known-distribution learning via fixed bracket experts

Input: horizon T, hypothesis class $\mathcal{H}$, distribution D over $\mathcal{X}$, and scale $\varepsilon \in (0,1)$.

Let $\mathcal{B}_{D,\varepsilon}$ be an ε-bracketing family for $\mathcal{H}$ with respect to D such that $|\mathcal{B}_{D,\varepsilon}| = \mathcal{N}_{[\,]}(\mathcal{H}, D, \varepsilon)$.

For each bracket $b = (b^-, b^+) \in \mathcal{B}_{D,\varepsilon}$, define an abstaining expert

$$E_b(x) := b^-(x) \text{ if } b^-(x) = b^+(x), \text{ otherwise } \perp.$$

Initialize p_1 to be a uniform distribution over $\{E_b : b \in \mathcal{B}_{D,\varepsilon}\}$.

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

  Receive x_t from the adversary.

  Sample a random expert $E_b \sim p_t$ and predict $\hat{y}_t = E_b(x_t) \in \{0, 1, \perp\}$.

  Upon receiving the label y_t, update the mixture:

$$p_{t+1}(E_b) = p_t(E_b) \exp\left(-\eta \mathbb{1}\{E_b(x_t) = 1 - y_t\}\right) / Z_t,$$

  where Z_t is a normalization constant.

Theorem 3. *Let $\mathcal{H}$ be a hypothesis class, let D be a distribution on $\mathcal{X}$, and let $\varepsilon \in (0,1)$. Then, against any (possibly adaptive) adversary, Algorithm 3 with parameter $\eta = \sqrt{8 \log \mathcal{N}_{[\,]}(\mathcal{H}, D, \varepsilon)/T}$ guarantees*

$$\mathbb{E}[\mathrm{MissErr} - \mathrm{OPT}] \le \sqrt{(T/2) \log \mathcal{N}_{[\,]}(\mathcal{H}, D, \varepsilon)} \qquad \textit{and} \qquad \mathbb{E}[\mathrm{AbsErr}] \le \varepsilon T.$$

Proof. Write $\mathcal{B} := \mathcal{B}_{D,\varepsilon}$ and $N := |\mathcal{B}| = \mathcal{N}_{[\,]}(\mathcal{H}, D, \varepsilon)$. We use the notation H_{t-1}, H^+_{t-1}, and $\ell_t(E_b) = \mathbb{1}\{E_b(x_t) = 1 - y_t\}$ from the proof of Lemma 5, with Algorithm 3 in place of Algorithm 1. Since $\mathcal{B}$ is fixed in advance, the mixture p_t is determined by H_{t-1}, and at time t Algorithm 3 samples $E_b \sim p_t$ using fresh randomness that is independent of H^+_{t-1}.

Misclassification error. As in the proof of Lemma 5, the conditional expected misclassification error at time t given H^+_{t-1} is $\mathbb{E}_{E_b \sim p_t} \ell_t(E_b)$, and therefore, by the law of total expectation,

$$\mathbb{E}[\mathrm{MissErr}] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}_{E_b \sim p_t} \ell_t(E_b)\right].$$

We bound the quantity inside the expectation for every realization of the interaction. Fix such a realization, and let $h_\star \in \arg\min_{h \in \mathcal{H}} \sum_{t=1}^{T} \mathbb{1}\{h(x_t) \neq y_t\}$ for the realized sequence. Since $\mathcal{B}$ is an ε-bracketing of $\mathcal{H}$, it contains a bracket $b_\star$ for $h_\star$, and Lemma 4 gives $\ell_t(E_{b_\star}) \le \mathbb{1}\{h_\star(x_t) \neq y_t\}$ for every t. Combining this with the regret guarantee of Multiplicative Weights (Lemma 3), which holds for every sequence of loss functions, gives

$$\sum_{t=1}^{T} \mathbb{E}_{E_b \sim p_t} \ell_t(E_b) \le \sum_{t=1}^{T} \ell_t(E_{b_\star}) + \sqrt{(T/2) \log N} \le \mathrm{OPT} + \sqrt{(T/2) \log N}.$$

Taking expectations proves the bound on $\mathbb{E}[\mathrm{MissErr} - \mathrm{OPT}]$.

Abstention error. Since $\mathcal{B}$ has minimum size, every bracket $b = (b^-, b^+) \in \mathcal{B}$ has disagreement mass $\mathbb{P}_{X \sim D}(b^-(X) \neq b^+(X)) \le \varepsilon$ (otherwise it could be removed from $\mathcal{B}$). Moreover, $E_b(x) = \perp$ exactly when $b^-(x) \neq b^+(x)$. Fix a round t. Conditioned on H_{t-1} and on $c_t = 0$, the example x_t is a fresh draw from D that is independent of the expert $E_b \sim p_t$ sampled by Algorithm 3, and therefore

$$\mathbb{P}\left(\hat{y}_t = \perp \middle|\, \mathsf{H}_{t-1}, c_t = 0\right) = \sum_{b \in \mathcal{B}} p_t(E_b)\, \mathbb{P}_{X \sim D}\left(b^-(X) \neq b^+(X)\right) \le \varepsilon.$$

Hence, $\mathbb{P}(c_t = 0 \text{ and } \hat{y}_t = \perp \mid \mathsf{H}_{t-1}) \le \varepsilon\, \mathbb{P}(c_t = 0 \mid \mathsf{H}_{t-1})$. Summing over t and taking expectations gives

$$\mathbb{E}[\mathrm{AbsErr}] = \sum_{t=1}^{T} \mathbb{P}\left(c_t = 0 \text{ and } \hat{y}_t = \perp\right) \le \varepsilon\, \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\{c_t = 0\}\right] \le \varepsilon T. \qquad \square$$

D Bracketing Number of Compositions

Lemma 2 (Bracketing Number of Compositions)**.** *Let $\mathcal{H}_i \subseteq \{0,1\}^{\mathcal{X}}$ be classes of binary functions for $i \in [k]$, and let $\mathcal{F}$ be a binary function class over $\{0,1\}^k$ with VC dimension $d_{\mathcal{F}}$. Consider the composition of $\mathcal{F}$ with $\mathcal{H}_1, \ldots, \mathcal{H}_k$,*

$$\mathcal{G} := \left\{x \mapsto f\big(h_1(x), \ldots, h_k(x)\big) : \; f \in \mathcal{F},\; h_i \in \mathcal{H}_i\right\}.$$

Then, for every distribution D over $\mathcal{X}$ and for every $\varepsilon_1, \ldots, \varepsilon_k \in [0,1]$,

$$\mathcal{N}_{[\,]}(\mathcal{G}, D, \varepsilon_1 + \cdots + \varepsilon_k) \le e\, 2^{k d_{\mathcal{F}}} \prod_{i=1}^{k} \mathcal{N}_{[\,]}(\mathcal{H}_i, D, \varepsilon_i).$$

Remark 4. The construction used below in the proof of Lemma 2 satisfies the following additional property. For any integer $m \ge 1$, if for each $i \in [k]$ a family of brackets for $\mathcal{H}_i$ can be constructed from m i.i.d. samples from D (that may be adversarially contaminated as in Lemma 7) with scale $\varepsilon_i(m, \delta_i)$ with probability at least $1 - \delta_i$, then a family of brackets for the composition $\mathcal{G}$ can be

constructed from m i.i.d. samples from D (that may be adversarially contaminated as in Lemma 7) with scale $\sum_i \varepsilon_i(m, \delta_i)$ with probability at least $1 - \sum_i \delta_i$.

For example, when each $\mathcal{H}_i$ is halfspaces in $\mathbb{R}^d$, the bracket for $\mathcal{G}$ will have scale proportional to $k \cdot \frac{O(d^2 \log(d) \log(m) + \log(1/\delta))}{m}$ implying a final abstention error of Algorithm 1 with an additive term proportional to $O(kd^2 \log(d) \log^2(T))$ which can be smaller than the VC dimension of $\mathcal{G}$ (or its underlying bracketing class).

Proof of Lemma 2. If $\mathcal{N}_{[\,]}(\mathcal{H}_i, D, \varepsilon_i) = \infty$ for some $i \in [k]$, there is nothing to prove, so assume that all of these bracketing numbers are finite. For each $i \in [k]$, let $\mathcal{B}_i$ be an ε_i-bracketing of $\mathcal{H}_i$ with respect to D of size $|\mathcal{B}_i| = \mathcal{N}_{[\,]}(\mathcal{H}_i, D, \varepsilon_i)$.

Construction. For a tuple of brackets $b = (b_1, \ldots, b_k) \in \mathcal{B}_1 \times \cdots \times \mathcal{B}_k$ with $b_i = (b_i^-, b_i^+)$ and a point $x \in \mathcal{X}$, let

$$Q_b(x) := \left\{ z \in \{0,1\}^k : \; b_i^-(x) \le z_i \le b_i^+(x) \text{ for all } i \in [k] \right\}$$

be the set of label vectors that are consistent with the brackets $b_1, \ldots, b_k$ at x. Since $b_i^-(x) \le b_i^+(x)$ for every $i \in [k]$, the set $Q_b(x)$ is nonempty. For each $f \in \mathcal{F}$, define

$$g_{b,f}^-(x) := \min_{z \in Q_b(x)} f(z) \qquad \text{and} \qquad g_{b,f}^+(x) := \max_{z \in Q_b(x)} f(z).$$

Clearly $g_{b,f}^-(x) \le g_{b,f}^+(x)$ for every $x \in \mathcal{X}$, so $(g_{b,f}^-, g_{b,f}^+)$ is a bracket. We will show that

$$\mathcal{B}_{\mathcal{G}} := \left\{ (g_{b,f}^-, g_{b,f}^+) : \; b \in \mathcal{B}_1 \times \cdots \times \mathcal{B}_k, \; f \in \mathcal{F} \right\}$$

is an $(\varepsilon_1 + \cdots + \varepsilon_k)$-bracketing of $\mathcal{G}$ with respect to D.

Bracketing. Fix $g = f(h_1, \ldots, h_k) \in \mathcal{G}$ with $f \in \mathcal{F}$ and $h_i \in \mathcal{H}_i$. For each $i \in [k]$, let $b_i = (b_i^-, b_i^+) \in \mathcal{B}_i$ be a bracket of h_i with $\mathbb{P}_{X \sim D}(b_i^-(X) \ne b_i^+(X)) \le \varepsilon_i$, and let $b = (b_1, \ldots, b_k)$. For every $x \in \mathcal{X}$, we have $b_i^-(x) \le h_i(x) \le b_i^+(x)$ for all $i \in [k]$, that is, $(h_1(x), \ldots, h_k(x)) \in Q_b(x)$, and therefore

$$g_{b,f}^-(x) \le f\big(h_1(x), \ldots, h_k(x)\big) = g(x) \le g_{b,f}^+(x).$$

Hence, $(g_{b,f}^-, g_{b,f}^+)$ brackets g. Moreover, if $b_i^-(x) = b_i^+(x)$ for all $i \in [k]$, then $Q_b(x)$ is a singleton, and so $g_{b,f}^-(x) = g_{b,f}^+(x)$. Thus,

$$\left\{ x \in \mathcal{X} : \; g_{b,f}^-(x) \ne g_{b,f}^+(x) \right\} \subseteq \bigcup_{i=1}^k \left\{ x \in \mathcal{X} : \; b_i^-(x) \ne b_i^+(x) \right\},$$

and by a union bound,

$$\mathbb{P}_{X \sim D}\left(g_{b,f}^-(X) \ne g_{b,f}^+(X) \right) \le \sum_{i=1}^k \mathbb{P}_{X \sim D}\left(b_i^-(X) \ne b_i^+(X) \right) \le \sum_{i=1}^k \varepsilon_i.$$

Counting. Since $\mathcal{F}$ is a class of functions over the domain $\{0,1\}^k$ of size 2^k, the Sauer–Shelah–Perles lemma gives $|\mathcal{F}| \le \sum_{j=0}^{d_{\mathcal{F}}} \binom{2^k}{j} \le \left(e 2^k / d_{\mathcal{F}}\right)^{d_{\mathcal{F}}} \le e\, 2^{k d_{\mathcal{F}}}$ when $d_{\mathcal{F}} \ge 1$, where the last inequality uses $(e/d)^d \le e$ for every $d \ge 1$ (and $|\mathcal{F}| \le 1$ when $d_{\mathcal{F}} = 0$). Therefore,

$$\mathcal{N}_{[\,]}(\mathcal{G}, D, \varepsilon_1 + \cdots + \varepsilon_k) \le |\mathcal{B}_{\mathcal{G}}| \le |\mathcal{F}| \prod_{i=1}^k |\mathcal{B}_i| \le e\, 2^{k d_{\mathcal{F}}} \prod_{i=1}^k \mathcal{N}_{[\,]}(\mathcal{H}_i, D, \varepsilon_i). \qquad \square$$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction are formally stated in the main paper and the appendices, and they apply to the problem setting that we formally introduced. They accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The theoretical results apply to the specific problem settings (which we formally defined) and under specific conditions (which appear in the formal theorem statements). In addition, limitations of approach proposed are discussed in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The assumptions are clearly stated as part of the problem formulation or in the formal theorem statements. All the theoretical results have proofs, in either the main paper or the appendices.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [N/A]

Justification: The paper does not include experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: The paper does not include experiments requiring code.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work is purely theoretical and there is no immediate societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper doesn't involve data or models and hence poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.