Fast Rates for Nonparametric Online Learning: From Realizability to Learning in Games

Constantinos Daskalakis* Noah Golowich†

April 13, 2022

Abstract

We study fast rates of convergence in the setting of nonparametric online regression, namely where regret is defined with respect to an arbitrary function class which has bounded complexity. Our contributions are two-fold:

- In the realizable setting of nonparametric online regression with the absolute loss, we propose a randomized proper learning algorithm which gets a near-optimal cumulative loss in terms of the sequential fat-shattering dimension of the hypothesis class. In the setting of online classification with a class of Littlestone dimension d, our bound reduces to $d \cdot \operatorname{poly}\log T$. This result answers a question as to whether proper learners could achieve near-optimal cumulative loss; previously, even for online classification, the best known cumulative loss was $\tilde{O}(\sqrt{dT})$. Further, for the real-valued (regression) setting, a cumulative loss bound with near-optimal scaling on sequential fat-shattering dimension was not even known for *improper* learners, prior to this work.
- Using the above result, we exhibit an independent learning algorithm for general-sum binary games of Littlestone dimension d, for which each player achieves regret $\tilde{O}(d^{3/4} \cdot T^{1/4})$. This result generalizes analogous results of Syrgkanis et al. (2015) who showed that in finite games the optimal regret can be accelerated from $O(\sqrt{T})$ in the adversarial setting to $O(T^{1/4})$ in the game setting.

To establish the above results, we introduce several new techniques, including: a hierarchical aggregation rule to achieve the optimal cumulative loss for real-valued classes, a multi-scale extension of the proper online realizable learner of Hanneke et al. (2021), an approach to show that the output of such nonparametric learning algorithms is stable, and a proof that the minimax theorem holds in all online learnable games.

*MIT CSAIL. costis@csail.mit.edu. Supported by NSF Awards CCF-1901292, DMS-2022448 and DMS-2134108, by a Simons Investigator Award, by the Simons Collaboration on the Theory of Algorithmic Fairness, by a DSTA grant, and by the DOE PhILMs project (No. DE-AC05-76RL01830).

†MIT CSAIL. nzg@mit.edu. Supported by a Fannie & John Hertz Foundation Fellowship and an NSF Graduate Fellowship.

Contents

1 Introduction

The success of deep learning has increased the importance of studying the learnability of nonparametric and high-dimensional models across all areas within learning theory and its applications. In this paper our goal is to advance our understanding of learning such models in two prominent settings, online learning and games.

In the classical setting of online learning [CBL06, SS11], a learner observes a sequence of labeled examples (x_t, y_t), generated adaptively by an adversary, and, at each round $t \geq 1$, is asked to make a prediction $f_t(x_t)$ about the true label y_t, by choosing a *hypothesis* f_t that depends only on the history of previous examples. A common goal is to minimize *regret*: for a loss function $\ell(\hat{y}, y)$ giving the penalty for predicting $\hat{y}$ when the true label is y, and a class $\mathcal{F}$ of hypotheses, the regret is the difference between the learner's total prediction loss, $\sum_{t=1}^{T} \ell(f_t(x_t), y_t)$, and the best possible loss in hindsight the learner could have obtained by choosing a single $f^\star \in \mathcal{F}$ over all T rounds.

Online learning has been studied for various instantiations of $\mathcal{F}$ and ℓ as well as various constraints on the learner and the adversary, drawing its importance from its versatility and intimate connections to other learning settings. Indeed, given the adversarial nature of the sequence of examples (x_t, y_t), online learning generalizes supervised learning, where these pairs are i.i.d., while beautiful connections have been forged between online learning and private learning [BLM20, ALMM19], contextual bandits [FR20], reinforcement learning [DYM21], adversarial sampling [ABED+21], learning of quantum states [ACH+19], and learning in games [RS13, SALS15]. In particular, online learning is a central primitive whose study unlocks understanding in many other learning-theoretic settings.

The starting point for our work is that, while the optimal no-regret algorithms are very well understood when the hypothesis class $\mathcal{F}$ is finite, low-dimensional, or parametric, our understanding of the optimal regret bounds and the algorithms achieving them is much more limited for nonparametric classes. For example, while a celebrated paper by Littlestone [Lit88] determines the optimal regret bound of online classification in the *realizable setting*, namely when $f^\star$ achieves 0 loss, his algorithm is not *proper*, namely the hypotheses f_t may not belong to the model class $\mathcal{F}$ (see also [HLM21, Ang88]); the optimal regret for proper realizable learning (by a randomized algorithm) remains elusive. In the non-Boolean setting (i.e., of regression) much less is known.

The contributions of our work, overviewed in the next section, are two-fold. First, we answer several outstanding questions, obtaining near-optimal regret bounds for proper online learning (for both classification and regression) in the realizable setting. Second, we use our new results to advance our understanding of learning in games in the nonparametric setting, which has become increasingly important due to the applications of adversarial training in robust learning, generative adversarial networks, and multi-agent reinforcement learning. Here, our results are the first to obtain fast rates for regret of independent learning in two-player zero-sum nonparametric games and more generally in multi-player general-sum nonparametric games.

1.1 Model and overview of results

We consider the standard setting of online learning with absolute loss: two agents, a *learner* and an *adversary*, interact over a total of T rounds, for some $T \in \mathbb{N}$. The learner and adversary are given at the onset a set $\mathcal{X}$ and a set $\mathcal{F}$ consisting of $[0,1]$-valued functions on $\mathcal{X}$, known as *hypotheses*. In the setting of *proper online learning*, the players perform the following for each round $1 \leq t \leq T$: the learner chooses a hypothesis $f_t \in \mathcal{F}$ (which may be random), and the adversary

picks $(x_t, y_t) \in \mathcal{X} \times [0,1]$ (which may be random) denoting a *feature* x_t together with its *label* y_t. Then the example (x_t, y_t) is revealed to the learner, who suffers loss $|f_t(x_t) - y_t|$. We allow the adversary to be *adaptive*, meaning that it can choose each example (x_t, y_t) based on the history of moves $f_1, \dots, f_{t-1}$ and $(x_1, y_1), \dots, (x_{t-1}, y_{t-1})$. In general, the goal of the learner is to minimize its expected *regret*, namely

$$\mathrm{Reg}_T := \mathbb{E}\left[\sum_{t=1}^{T} |f_t(x_t) - y_t| - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} |f(x_t) - y_t|\right]. \tag{1}$$

In this paper we are concerned with the case when the function class $\mathcal{F}$ is *nonparametric* in nature, meaning that it is infinite or extremeley large; thus regret guarantees in terms of $\log|\mathcal{F}|$ are insufficient, and we instead aim for guarantees in terms of combinatorial complexity measures of $\mathcal{F}$.[1]

Near-optimal cumulative loss for the realizable setting. A fundamental setting in which online learning is studied is the *realizable setting*, which means that the adversary is constrained to choose the sequence (x_t, y_t), $1 \le t \le T$ so that there is some $f^\star \in \mathcal{F}$ so that $f^\star(x_t) = y_t$ for all t. In this case, the regret of the learner (1) reduces to $\mathbb{E}\left[\sum_{t=1}^{T} |f_t(x_t) - y_t|\right]$, namely the total expected error made by the learner over all T rounds; we call this quantity the *cumulative loss* of the learner.

The study of the optimal cumulative loss for an online learner dates back to the seminal work of Littlestone [Lit88], who showed that in the case of *online classification* (namely, where hypotheses $f \in \mathcal{F}$ map to $\{0,1\}$, and $y_t \in \{0,1\}$ for all t), the optimal cumulative loss (also known as the *mistake bound* in the binary setting) for a hypothesis class $\mathcal{F}$ is given by a combinatorial parameter of $\mathcal{F}$ known as the *Littlestone dimension*, denoted $\mathrm{Ldim}(\mathcal{F}) \in \mathbb{N}$. One limitation of the result of [Lit88] is that this mistake bound was only shown for an *improper* learner, meaning that the learner's hypotheses f_t may not belong to the class $\mathcal{F}$. In many settings, such as the setting of learning in games discussed below, an improper learning algorithm is insufficient to solve the task at hand: for instance, for learning in games, the hypothesis class $\mathcal{F}$ denotes the set of actions available to the learning agent, who must choose a valid action at each time step. However, there are many settings in which improper learners have better statistical or computational properties than proper learners (such as [HLM15, HM16, Han16, HK16, FKL+18, Ang88, DSS14, MHS19]; see [HLM21] for further examples). It is therefore natural to ask whether there is a similar proper-improper gap in the setting of online realizable classification: *is there a near-optimal (randomized) proper learner for online classification in the realizable setting*?

We further consider the generalization of the above question to the setting of *regression*, i.e., the general case where hypotheses $f \in \mathcal{F}$ and labels y_t are real-valued. In this case the natural generalization of the Littlestone dimension is the *sequential fat-shattering dimension* (Definition 2.1). Surprisingly, prior work has not characterized the optimal cumulative loss for real-valued hypothesis classes in terms of the sequential fat-shattering dimension, even for improper learners.

[1] Note that our informal usage of the term "nonparametric" differs slightly from some other instances in the literature, such as [RST17, FK18], in which it is used to refer specifically to classes with inverse-polynomial growth in the empirical entropy numbers. While we consider, for example, generic binary classes of finite Littlestone dimension to be nonparametric, works such as [RST17, FK18] would not do so.

Therefore, our fully general question for the realizable setting is the following:

What is the optimal cumulative loss (in terms of sequential fat-shattering dimension) for realizable online regression? Can it be achieved by a (randomized) proper algorithm? ($\star$)

There appears to be some confusion in the literature regarding the latter part (proper learnability) of the question ($\star$), even for the special case of binary classification: [HLM21] states (without proof) that "unlike the realizable setting, in the agnostic setting nearly optimal randomized proper learners can exist."[2] We show that a randomized proper learner can obtain a cumulative loss bound in the realizable setting that is off from the optimal bound (of $\mathrm{Ldim}(\mathcal{F})$) by only a $\operatorname{poly}\log T$ factor.[3] Furthermore, we can extend our upper bound to the more general setting of online regression:

Theorem 1.1 (Informal version of Theorem 5.15)**.** *There is a randomized proper learner that achieves cumulative loss of* $O\left(\inf_{\alpha\in[0,1]}\left\{\alpha T + \int_\alpha^1 \mathrm{sfat}_\delta(\mathcal{F})d\delta\right\}\right)\cdot\operatorname{poly}\log T$ *in the realizable setting. In the special case of online classification, this bound becomes* $O(\mathrm{Ldim}(\mathcal{F}))\cdot\operatorname{poly}\log T$*.*

We remark that randomization is necessary for proper realizable learning: there are trivial classes, such as the class of point functions on an infinite domain, which have Littlestone dimension 1 but for which any deterministic proper learner cannot achieve any finite cumulative loss bound. Nevertheless, we show in Proposition 4.1 that there is a deterministic *improper* learner that achieves the cumulative loss bound of Theorem 1.1.

As alluded to above, we further show that the cumulative loss bound of Theorem 1.1 is optimal (up to a $\operatorname{poly}\log T$ factor) among any bound that depends only on sequential fat-shattering dimension:

Proposition 1.2 (Lower bound)**.** *For any non-increasing function* $s:[0,1]\to\mathbb{Z}_{\geq 0}$ *and* $T\in\mathbb{N}$*, there is some function class* $\mathcal{F}$ *so that* $\mathrm{sfat}_\alpha(\mathcal{F})\leq s(\alpha)$ *for all* $\alpha\in[0,1]$*, but for which any algorithm (not necessarily proper) has cumulative loss at least*

$$\Omega\left(\frac{1}{\log T}\cdot\inf_{\alpha\in[1/T,1]}\left\{\alpha T+\int_\alpha^1 s(\eta)d\eta\right\}\right) \tag{2}$$

We remark that unlike in the case of classification, the matching bound of Theorem 1.1 and Proposition 1.2 for online regression is not *instance optimal*: there may be some classes $\mathcal{F}$ for which an algorithm can achieve a cumulative loss much smaller than (2).[4] We leave the question of determining a quantity that characterizes the optimal cumulative loss in an instance-dependent manner to future work. Nevertheless, we believe that the bound of Theorem 1.1 is of interest for the following reasons: first, in [BDR21], under a mild growth condition, sequential fat-shattering dimension is shown to characterize the minimax regret in the agnostic (non-realizable) setting, meaning it is natural to ask what its relationship to the optimal cumulative loss is in the related realizable setting; second, the result of Theorem 1.1, for the real-valued setting (regression), is a crucial component in the proof of our result for learning in games (Theorem 1.4 below), *even for binary-valued games* (at a high level, this is the case because players can randomize their actions).

[2] See the bottom of page 4 in [HLM21].

[3] Whether such a $\operatorname{poly}\log T$ factor can be removed remains an open question.

[4] This could be the case, for instance, if for each $x\in\mathcal{X}$, the values $\{f(x):f\in\mathcal{F}\}$ are all distinct. Thus, after seeing a single example $(x_1, f^\star(x_1))$, the algorithm knows the identity of $f^\star$ and can predict correctly at all rounds $t>1$.

A stable proper learner and applications. Next, we describe some applications of Theorem 1.1, culminating in our result giving fast rates for learning in games in a nonparametric setting (Theorem 1.4). First, it is necessary to describe a strengthening of Theorem 1.1, namely that the cumulative loss bound of Theorem 1.1 holds for a learner that produces *stable* predictions. Traditionally, *stability* of the predictions produced by an online learner, in the sense that the predictions do not change much from round to round, has been a hallmark of online learning algorithms. In the finite-dimensional setting, such stability is classically achieved via the use of an appropriate regularizer [BT03, CBL06], but can also be obtained from the use of more unorthodox methods such as follow-the-perturbed-leader [KV05]. Further, such stability has inspired connections between online learning and other areas in learning theory, such as differentially private learning [BLM20], and the study of generalization in deep neural networks [HRS15].

Despite the recent growth of work on online learning in nonparametric settings, we are not aware of any results establishing stability of the predictions. Moreover, many of the techniques in nonparametric settings, such as the nonconstructive approach that proceeds via application of a minimax theorem together with symmetrization [RST15a, RST15b], seem fundamentally unable to establish such stability bounds (see Section 1.2). Proposition 1.3 below address this deficiency of existing work; to state the result, we introduce the following notation. For a hypothesis class $\mathcal{F}$, we denote the set of finite-support distributions on $\mathcal{F}$ by $\Delta^\circ(\mathcal{F})$; elements of $\Delta^\circ(\mathcal{F})$ will typically be denoted with bars, e.g., $\bar{f} \in \Delta^\circ(\mathcal{F})$. For $\bar{f}_1, \bar{f}_2 \in \Delta^\circ(\mathcal{F})$, let $\frac{1}{2}\left\|\bar{f}_1 - \bar{f}_2\right\|_1$ denote the total variation distance between $\bar{f}_1, \bar{f}_2$, which is well-defined by the finite-supportedness of $\bar{f}_1, \bar{f}_2$. We remark that the proper randomized learner of Theorem 1.1 outputs, for each round t, a hypothesis distributed according to a finite-support distribution $\bar{f}_t \in \Delta^\circ(\mathcal{F})$.

Proposition 1.3 (Stability; informal version of Theorem 5.15)**.** *Fix any $\eta > 0$. The proper randomized learner of Theorem 1.1, which chooses $\bar{f}_t \in \Delta^\circ(\mathcal{F})$ for each round t, may be modified to satisfy $\left\|\bar{f}_t - \bar{f}_{t+1}\right\|_1 \le \eta$ for all t, at the cost of a cumulative loss of $\frac{\operatorname{poly}\log T}{\eta} \cdot O\left(\inf_{\alpha \in [0,1]}\left\{\alpha T + \int_\alpha^1 \operatorname{sfat}_\delta(\mathcal{F}) d\delta\right\}\right)$.*

While Proposition 1.3, which applies only to the realizable setting, is of some interest in its own right, we believe it is most notable for its applications: broadly speaking, we use Proposition 1.3 to establish that many guarantees of online learning in the finite-dimensional *non-realizable* (i.e., agnostic) setting that make use of stability extend to the nonparametric case as well.

Application: fast rates for learning in games. An extensive line of work over the last decade (starting with [DDK11]; see Section 1.2) has shown that minimax $\Omega(\sqrt{T})$ lower bounds on regret can be circumvented if multiple agents implement learning algorithms from a particular family in the context of repeatedly playing a (finite) game. In Theorem 1.4 below, we show that such results hold true in the nonparametric setting as well. To state Theorem 1.4, we introduce the following preliminaries: we consider general-sum games with K players, who have action sets $\mathcal{F}_1, \dots, \mathcal{F}_K$. For simplicity, we restrict our attention to binary-valued games, namely where each player k's payoff function is of the form $\ell_k : \mathcal{F}_1 \times \cdots \times \mathcal{F}_K \to \{0, 1\}$. We only assume that for each player k, the class $\mathcal{F}_k^{\ell_k} := \{f_{-k} \mapsto \ell_k(f_k, f_{-k}) : f_k \in \mathcal{F}_k\}$ has finite Littlestone dimension.[5]

[5]Some assumption on the game is necessary to guarantee existence of Nash equilibria: there is a 2-player zero-sum game, "Guess the larger number" (GTLN), which has infinite Littlestone dimension, but which has no ϵ-approximate Nash equilibrium for any $\epsilon < 1$ (see [HLM21]). A necessary and sufficient condition for a binary-valued game and all its subgames to contain approximate Nash equilibria is that the game not contain an embedded copy of GTLN; we leave it as an interesting future direction to extend our results in some form to such games.

In the setting of *independent learning algorithms for repeated game playing* [DDK11, RS13], the following procedure occurs over T rounds: for each round $t \le T$, each player $k \in [K]$ plays a (finite-support) distribution over actions, denoted $\bar{f}_k^t \in \Delta^\circ(\mathcal{F}_k)$. Then each player k suffers loss $\mathbb{E}_{(f_1,\dots,f_K)\sim(\bar{f}_1^t,\dots,\bar{f}_K^t)}[\ell_k(f_1,\dots,f_K)]$. Further, each player k observes the function mapping each of its actions $f_k \in \mathcal{F}_k$ to its expected loss (under $\bar{f}_1^t,\dots,\bar{f}_{k-1}^t,\bar{f}_{k+1}^t,\dots,\bar{f}_K^t$) had it played f_k; it uses this information to adapt its play in future rounds. In the setting of independent learning algorithms in *finite* normal form games, the foundational work of [SALS15] showed that if each player implements the algorithm Optimistic Exponential Weights (also known as Optimistic Hedge), then each player k can achieve regret $O(\log^{3/4}|\mathcal{F}_k| \cdot \sqrt{K} \cdot T^{1/4})$. This result has since been improved multiple times, culminating in [DFG21] which obtains regret $O(K \cdot \mathrm{poly}(\log|\mathcal{F}_k|, \log T))$. Our main result for independent learning in games is an analogue of the result of [SALS15] for the nonparametric setting:

Theorem 1.4 (Informal version of Theorem 7.2)**.** *There is an independent learning algorithm (`Optimistic SOA-Experts`, Algorithm 3), so that the following holds. Fix a game G of finite Littlestone dimension, as above. If the players repeatedly play G with each player using `Optimistic SOA-Experts`, then each player k suffers regret $\tilde{O}(\mathrm{Ldim}(\mathcal{F}_k^{\ell_k})^{3/4} \cdot \sqrt{K} \cdot T^{1/4})$.*

1.2 Related work

The present paper lies at the confluence of many distinct lines of work on both statistical (i.i.d.) and adversarial (online) learning, as well as game theory, which we summarize below.

Fast rates in online & offline learning. Our two main results, Theorems 1.1 and 1.4, both beat a $\Omega(\sqrt{T})$ lower bound on regret for many hypothesis classes of interest, such as classes of finite Littlestone dimension, by making additional assumptions about the adversary. A multitude of such results on *fast rates* has been established, for both offline and online problems, over the past two decades. In finite-dimensional online settings, fast rates (in many cases, on the order of $\log T$) can be obtained if the loss function has special structure, such as if it is exp-concave [CBL06], or more generally satisfies a *mixability* [Vov95, HKW98, Vov01] or *stochastic mixability* [vEGM+15] condition. Such results have been extended to the nonparametric setting for several special cases of exp-concave losses, including the square loss [RS14a] and log loss [RST15b, BFR20, FKL+18]. Unlike our results, these works often allow for a (arbitrary) non-realizable adversary. In the case of a realizable adversary but for the harder case of absolute loss, the halving algorithm [SS11] obtains logarithmic regret for finite classes $\mathcal{F}$ in the case of binary classification; it is generalized by the Standard Optimal Algorithm (SOA) of [Lit88] for the infinite case.

A similarly extensive line of work has pursued fast rates in the offline setting (i.e., where the examples (x_t, y_t), $1 \le t \le T$, are i.i.d. according to some distribution). It has long been known that in the realizable setting for binary classification, excess risk of $\tilde{O}(\mathrm{VCdim}(\mathcal{F})/T)$ is achievable[6] by a proper learning algorithm [VK06, BEHW89], such as empirical risk minimization. This bound has been improved by logarithmic factors several times [HLW94, Sim15, Han16]. This work was generalized to the real-valued (regression) setting in [Men02], in which an analogue of Theorem 1.1 for the offline realizable setting was established. If the (non-sequential) α-fat-shattering dimension grows as α^{-p}, $p \in (0,2)$, the rate obtained by [Men02, Theorem 4.1] for the *offline* setting is

[6] In the offline case, statistical rates are usually normalized by the number of samples T; we follow this convention, noting that in the online case we do not normalize by T.

$\tilde{O}(T^{-2/(2+p)})$, whereas if the sequential fat-shattering dimension grows as α^{-p}, the (normalized) rate of Theorem 1.1 for the *online* setting is $\tilde{O}(T^{-\min\{1,1/p\}})$. While Proposition 1.2 shows that the bound of Theorem 1.1 is best possible, it is unclear if this is the case for the offline rates of [Men02]; we note, though, that [Men02] conjectured that their rates were best-possible in the offline settting.

Local Rademacher complexities and fast rates. Extending the techniques of [Men02], several works in the offline setting introduced *local Rademacher complexities* [Kol11, KP04, BKP04, BBM05, Men14, RST17]; these works derive fast rates, which are often data-dependent in nature, under a wider spectrum of assumptions generalizing realizability. In particular, the rates are generally phrased in terms of a fixed point of the modulus of continuity of the local Rademacher complexities around an optimal hypothesis. These results on local Rademacher complexities generalized and unified many previous papers (such as [SST12]) which showed that, as in the online setting, fast rates are attainable in the offline setting under additional restrictions on the loss function, such as smoothness. As such, it would be of great interest to have a similarly powerful theory of local Rademacher complexities in the online setting. Initial steps toward this objective were made in [RSS12], but the notion of local sequential Rademacher complexity from [RSS12] seems quite limited in nature, as it does not recover most of the existing nonparametric results on fast online rates mentioned above, as well as our own results. In the online setting, the effect of localization can be obtained in some special cases, such as learning with square loss, by using *offset Rademacher complexities* [RS14a, LRS15]; extending such techniques to our setting of realizability with absolute loss is an interesting open problem.

Fast rates for learning in games. A parallel line of work proving fast rates for regret of online learning assumes that the adversary for the online learning algorithm is itself the output of another online learner, in the context of repeated game-playing. The seminal result in this direction was that of [DDK11], which described an algorithm for learning with d experts that achieves the minimax regret of $O(\sqrt{T \log d})$ for a (worst-case) adversary, but which obtains regret $\mathrm{poly}(\log T, \log d)$ when it plays against itself in a two-player zero-sum game with d actions per player. Similar results have since been shown for various other algorithms in two-player, zero-sum games [HAM21, RS13]. For the more challenging case of multi-player general-sum games, [SALS15] showed that when all players use any algorithm from the family of Optimistic Mirror Descent (OMD) algorithms, each player has regret $O(T^{1/4} \cdot \log^{3/4} d)$. This was subsequently improved by [CP20] who showed a regret bound of $O(T^{1/6} \cdot \log^{5/6} d)$ for each player when there are only 2 players and both use Optimistic Hedge (a special case of OMD), and then by [DFG21] which obtained a near-optimal regret bound of $O(\log d \cdot \log^4 T)$ for any number of players under Optimistic Hedge. The techniques used to achieve these results have been successfully extended to achieve fast rates in various other settings, including learning in games with bandit feedback [WL18, BLLW19, WLA20] and learning in extensive-form games [FKS19]. All existing works in this direction are parametric in nature, considering a finite expert (i.e., hypothesis) class. Our Theorem 1.4 is the first to consider such results in the nonparametric setting.

Relation to existing work on constrained adversaries. Several of our results, such as Theorem 1.4 and our stable path-length regret bound in Theorem 6.1 (which is used to prove Theorem 1.4) may be seen as showing that minimax regret lower bounds can be broken if certain constraints

are placed on the adversary. The work [RST11] develops a version of sequential Rademacher complexity to characterize the optimal rates for online learning with constrained adversaries in a general setting. It is shown in [RST11, Proposition 13] that this generic technique recovers the path-length regret bound of [SALS15] for learning with d experts. This result for finite hypothesis classes is *not* extended in [RST11] to the more general nonparametric setting of Theorem 6.1, though such an extension appears possible in principle, if an appropriate Bernstein-type uniform convergence lemma for trees could be shown. However, there is a more significant limitation of the framework of [RST11], which is that this framework is nonconstructive in nature and thus the implied learning algorithm is not shown to be stable in the sense of Proposition 1.3. If the learner is not stable, then if used in a game, it does not produce stable losses for the other agents and thus it is impossible to use the framework of [RST11] to derive Theorem 1.4.[7]

2 Preliminaries

Notation We use the following generic notation. Given a sequence $X_1, \dots, X_n$ (e.g., of sets, or elements of sets), we will denote it by $X_{1:n}$. For a set $\mathcal{S}$, let $\Delta^\circ(\mathcal{S})$ denote the set of finite support measures on $\mathcal{S}$. For $n \in \mathbb{N}$, let $[n] = \{1, 2, \dots, n\}$. For integers $\lambda \in \mathbb{Z}$, we set $\alpha_\lambda := 2^{-\lambda}$ to denote the various scales our algorithms will operate at; typically (but not always) we will have $\lambda \geq 1$. For a function $f : \mathcal{X} \to \mathbb{R}$, let $\|f\|_{\infty,\mathcal{X}} = \sup_{x\in\mathcal{X}} |f(x)|$, and for finite-support distributions $\bar{f}_1, \bar{f}_2 \in \Delta^\circ(\mathcal{F})$, let $\left\|\bar{f}_1 - \bar{f}_2\right\|_1$ denote twice their total variation distance.

2.1 Online learning: combinatorial quantities

We introduce some notation and definitions regarding the setting of online nonparametric regression. Consider sets $\mathcal{X}, \mathcal{Y}$ and let $\mathcal{Y}^\mathcal{X}$ denote the set of $\mathcal{Y}$-valued functions on $\mathcal{X}$; usually we will have either $\mathcal{Y} = \{0, 1\}$ or $\mathcal{Y} = [0, 1]$. Throughout the paper we will assume that $\mathcal{F} \subset [0, 1]^\mathcal{X}$ is a known hypothesis class. We first define the sequential fat-shattering dimension, which is a combinatorial quantity that characterizes online learnability in the real-valued setting. To do so, we review some notation (from [RST15a]) regarding binary trees. For a set $\mathcal{Z}$, a *$\mathcal{Z}$-valued tree* $\mathbf{z}$ is a complete rooted binary tree each of whose nodes are labeled by an element of $\mathcal{Z}$. Let d denote the depth of the tree. For each $1 \leq t \leq d$, we identify the 2^{t-1} nodes of $\mathbf{z}$ at depth t with the sequences $\epsilon_{1:t-1} = (\epsilon_1, \dots, \epsilon_{t-1}) \in \{-1, 1\}^{t-1}$; the value ϵ_i determines whether one must take the left or right child at the ith step on the path from the root of $\mathbf{z}$ to the given vertex. We denote the label of the vertex $\epsilon_{1:t-1}$ by $\mathbf{z}_t(\epsilon_{1:t-1})$, so that $\mathbf{z}_t$ is a function mapping $\{-1, 1\}^{t-1} \to \mathcal{Z}$. The data of the tree $\mathbf{z}$ consists of the d-tuple $(\mathbf{z}_1, \dots, \mathbf{z}_d)$. Nodes at the final level, namely of the form $(\epsilon_1, \dots, \epsilon_d)$, are called *leaves*.

Definition 2.1 (Sequential fat-shattering dimension)**.** For a class $\mathcal{F} \subset [0, 1]^\mathcal{X}$ and $\alpha > 0$, its *α-sequential fat-shattering dimension*, denoted $\mathrm{sfat}_\alpha(\mathcal{F})$, is the largest positive integer d so that there is a complete $\mathcal{X}$-valued binary tree $\mathbf{x}$ and a complete $[0, 1]$-valued binary tree $\mathbf{s}$, both of depth d,

[7] One might hope that the constructive relaxation-based approach of [RSS12] would allow one to use the framework of [RST11] to produce stable learners. While this is the case in some finite-dimensional settings (see Section 10 of [RSS12]), this strategy fails in the general nonparametric setting since the basic Meta-algorithm of [RSS12] requires the computation of a fixed point each iteration, which may be non-stable. Our analysis runs into a similar challenge involving a per-round fixed-point operation, but we are able to overcome it using the particular structure of our algorithm, and this technique does not appear to extend to the setting of [RSS12]; see Section 3.

so that for all $k_{1:d} \in \{-1,1\}^d$, there is some $f \in \mathcal{F}$ so that $k_t \cdot (f(\mathbf{x}_t(k_{1:t-1})) - \mathbf{s}_t(k_{1:t-1})) \geq \alpha/2$ for all $t \in [d]$. In such a case, the class $\mathcal{F}$ is said to *α-shatter* the tree $\mathbf{x}$, as *witnessed* by $\mathbf{s}$.

To work with the sequential fat-shattering dimension at a given scale α, it is often useful to discretize the class $\mathcal{F}$ in the sense given in the below definition:

Definition 2.2 (Scale-sensitive restrictions)**.** Fix a class $\mathcal{F} \subset [0,1]^{\mathcal{X}}$ and $\alpha \in (0,1)$. Fix $(x,y) \in \mathcal{X} \times [0,1]$. We define the *$\alpha$-restriction of $\mathcal{F}$ to (x,y)*, denoted $\mathcal{F}|^{\alpha}_{(x,y)}$, to be the set:

$$\mathcal{F}|^{\alpha}_{(x,y)} := \{f \in \mathcal{F} : \lfloor y/\alpha \rfloor = \lfloor f(x)/\alpha \rfloor\}.$$

Equivalently, we have that $\mathcal{F}|^{\alpha}_{(x,y)} = \{f \in \mathcal{F} : f(x) \in [j\alpha, (j+1)\alpha)\}$, where $j = \lfloor y/\alpha \rfloor$.

In the case where $\mathcal{F}$ is $\{0,1\}$-valued (in which the sequential fat-shattering dimension reduces to the Littlestone dimension), the well-known *standard optimal algorithm (SOA)* [Lit88] gives the optimal cumulative loss (i.e., mistake bound) in the realizable setting. The SOA is an improper learning algorithm, but the hypotheses it outputs nevertheless have a certain structure which will prove useful in our setting as well; Definition 2.3 below generalizes such "SOA hypotheses" to the real-valued setting.

Definition 2.3 (SOA hypothesis)**.** Fix a class $\mathcal{F} \subset [0,1]^{\mathcal{X}}$ and a parameter $\alpha \in (0,1)$. For each $x \in \mathcal{X}$, set $s_x := \mathrm{sfat}_\alpha(\mathcal{F})$, and for $0 \leq j < \lfloor 1/\alpha \rfloor + 1$, set $s_{x,j} := \mathrm{sfat}_\alpha\left(\mathcal{F}|^{\alpha}_{(x,j\alpha)}\right)$.

The *SOA hypothesis for $\mathcal{F}$ at scale α*, denoted $\mathrm{SOA}(\mathcal{F},\alpha) \in [0,1]^{\mathcal{X}}$, is defined as follows. Fix any $x \in \mathcal{X}$, and let $j^\star$ to be chosen as small as possible so that $s_{x,j^\star} \geq s_{x,j}$ for all $0 \leq j < \lfloor 1/\alpha \rfloor + 1$. Then set $\mathrm{SOA}(\mathcal{F},\alpha)(x) := (j^\star+1)\alpha$.

Roughly speaking, the SOA hypothesis discretizes $\mathcal{F}$ using the scale parameter α and maps each x into the bucket such that the restricted class has maximum sequential fat-shattering dimension. Finally we introduce the notion of dual classes: for $\mathcal{F} \subset [0,1]^{\mathcal{X}}$, its *dual class*, denoted $\mathcal{F}^\star \subset [0,1]^{\mathcal{F}}$, is the class $\{f \mapsto f(x) \ : \ x \in \mathcal{X}\}$; thus $\mathcal{F}^\star$ is in bijection with $\mathcal{X}$.

The binary case: Littlestone classes. Finally, we mention the specialization of the above concepts to the binary-valued case, namely when $\mathcal{F} \subset \{0,1\}^{\mathcal{X}}$. Here, $\mathrm{sfat}_\alpha(\mathcal{F})$ is constant as a function of $\alpha \in [0,1]$, and this constant value is called the *Littlestone dimension* of $\mathcal{F}$, denoted $\mathrm{Ldim}(\mathcal{F})$. The scale parameter α in Definitions 2.2 and 2.3 is unnecessary, so restrictions are denoted $\mathcal{F}|_{(x,y)}$ and the SOA hypothesis is denoted by $\mathrm{SOA}(\mathcal{F})$.

3 Overview of techniques

The main technical innovation in our paper is a stable proper learning algorithm in the realizable setting, `Multi-scale Proper Learner` (Algorithm 2), which obtains the guarantees of Theorem 1.1 and Proposition 1.3 (stated formally in Theorem 5.15). As mentioned previously, the guarantee of Theorem 1.1 is new even for an improper learning algorithm in the setting of realizable regression. We therefore begin by describing a simple improper learning algorithm, `Multi-scale Improper Learner` (Algorithm 1), which obtains the optimal cumulative loss of $O\left(\min_\alpha \left\{\alpha T + \int_0^1 \mathrm{sfat}_\eta(\mathcal{F}) d\eta\right\}\right)$, as stated in Proposition 4.1.

3.1 A multi-scale improper learner

The starting point for the algorithm `Multi-scale Improper Learner` is the following simple algorithm which generalizes the Standard Optimal Algorithm of [Lit88]: fix some $\alpha > 0$ at the beginning of the learning procedure, and set $\mathcal{F}^1 = \mathcal{F}$. For each $t \geq 1$, predict the hypothesis $\mathrm{SOA}(\mathcal{F}^t, \alpha)$. After observing each example (x_t, y_t), if it is the case that $|\mathrm{SOA}(\mathcal{F}^t, \alpha)(x_t) - y_t| > \alpha$, then set $\mathcal{F}^{t+1} \leftarrow \mathcal{F}^t|^{\alpha}_{(x_t,y_t)}$, and otherwise set $\mathcal{F}^{t+1} \leftarrow \mathcal{F}^t$. It is straightforward to show (see Lemma A.2) that if $|\mathrm{SOA}(\mathcal{F}, \alpha)(x_t) - y_t| > \alpha$, then $\mathrm{sfat}_\alpha(\mathcal{F}^t|^{\alpha}_{(x_t,y_t)}) < \mathrm{sfat}_\alpha(\mathcal{F}^t)$, meaning that for each round t at which this algorithm makes a mistake larger than α, we have $\mathrm{sfat}_\alpha(\mathcal{F}^{t+1}) < \mathrm{sfat}_\alpha(\mathcal{F}^t)$. Thus the cumulative loss for the algorithm is at most $\alpha T + \mathrm{sfat}_\alpha(\mathcal{F})$. Even if we optimize over α, thus obtaining a cumulative loss of $\min_{\alpha \in [0,1]}\{\alpha T + \mathrm{sfat}_\alpha(\mathcal{F})\}$, we still do not get close to the optimal cumulative loss. For instance, if $\mathrm{sfat}_\alpha(\mathcal{F}) = \Theta(\alpha^{-p})$ for some $p \in (0,1)$, then the bound of Proposition 4.1 is constant in the horizon T, whereas $\min_\alpha\{\alpha T + \alpha^{-p}\} = \Theta(T^{p/(1+p)})$.

The key to obtaining better rates is to understand how to aggregate the predictions of SOA hypotheses at multiple scales α. This is similar in spirit to the technique of *chaining* [Dud78], which can be used to bound excess risk with an integral of (empirical) entropies by constructing a multi-scale cover. In our setting, though, it is the actual *predictions* of an algorithm which we wish to aggregate over multiple scales, and doing so appears to be quite different from chaining *covers* at multiple scales.

In `Multi-scale Improper Learner`, we address this challenge as follows: for an appropriate parameter $\Lambda \leq \log T$, we maintain a total of Λ subclasses of $\mathcal{F}$, denoted $\mathcal{F}_1, \ldots, \mathcal{F}_\Lambda$, at each round t. Letting $\alpha_\lambda = 2^{-\lambda}$ for each $\lambda \in [\Lambda]$, each subclass $\mathcal{F}_\lambda$ is updated in response to the examples (x_t, y_t) as described above, for the scale α_λ. For each round t, and each possible point $x_t \in \mathcal{X}$, the Λ subclasses each produce a prediction, $\mathrm{SOA}(\mathcal{F}_\lambda, \alpha_\lambda)(x_t) \in [0,1]$, for $\lambda \in [\Lambda]$. The main difficulty one faces is: *which of these Λ options should be chosen as the algorithm's prediction for x_t?*

`Multi-scale Improper Learner` answers this question using a simple aggregation rule we call the *hierarchical aggregation rule* (see Definitions 4.1 and 4.2). For any given point x_t, this rule chooses a single prediction out of the Λ elements $g_\lambda := \mathrm{SOA}(\mathcal{F}_\lambda, \alpha_\lambda)(x_t)$, $1 \leq \lambda \leq \Lambda$, as follows: it chooses $g_{\bar{\lambda}}$, where $\bar{\lambda} \geq 1$ is as small as possible so that $|g_{\bar{\lambda}} - g_{\bar{\lambda}+1}| > 2\alpha_{\bar{\lambda}}$ (if no such $\bar{\lambda} < \Lambda$ exists, set $\bar{\lambda} = \Lambda$). This aggregation rule satisfies the following key property (see Lemma 4.2): fix any choice of the true label y_t for the point x_t, and set $\delta_t := |y_t - \bar{g}_\lambda|$ to be the algorithm's error. Then there is some $\lambda' \in [\Lambda]$ so that $\alpha_{\lambda'} \geq \Omega(\delta_t)$ and $|\mathrm{SOA}(\mathcal{F}_{\lambda'}, \alpha_{\lambda'})(x_t) - y_t| > \alpha_{\lambda'}$; the proof of this fact requires some delicate case-work. Thus, the potential function $\sum_{\lambda=1}^{\Lambda} \alpha_\lambda \cdot \mathrm{sfat}_{\alpha_\lambda}(\mathcal{F}_\lambda)$ decreases by $\Omega(\delta_t)$ at each round t, which allows us to bound the total error over all T rounds by the integral $\alpha_\Lambda T + \int_{\alpha_\Lambda}^{1} \mathrm{sfat}_\eta(\mathcal{F}) d\eta$.

3.2 Obtaining the optimal cumulative loss for a proper learner

We proceed to describe our proper learner (`Multi-scale Proper Learner`, Algorithm 2) which obtains the same cumulative loss (up to a $\mathrm{poly}\log T$ factor) as `Multi-scale Improper Learner`. At a high level, `Multi-scale Proper Learner` uses the constructive framework of [HLM21] to "make proper" our improper learning algorithm. However, the algorithm and its analysis is not merely a case of generalizing that of [HLM21], which only treated the case of classification, to the real-valued (regression) setting. Rather, as mentioned in Section 1.1, our proper learner improves quantitatively upon the state of the art even in the special case of classification: we manage to obtain a poly-logarithmic (in T) cumulative loss for general Littlestone classes, and the previously

best known bound was $O(\sqrt{T})$ [BDPSS09, RST15a, HLM21].

Thus, we begin by describing how we can obtain an improved cumulative loss bound for a (randomized) proper learning algorithm for binary classification. At a high level, we build off the approach of [HLM21]: roughly speaking, this approach maintains a multiset $\mathcal{T}$ of subclasses $\mathcal{F}^i$ of $\mathcal{F}$, each accompanied by a weight $w^i \geq 0$. At each iteration, it considers the distribution Q over the hypotheses $\mathrm{SOA}(\mathcal{F}^i)$ weighted according to the values w^i, and tries to find a finite-support distribution $\bar{f}$ over hypotheses in $\mathcal{F}$, whose expectation is close to that of Q. If it can find such a distribution $\bar{f} \in \Delta^\circ(\mathcal{F})$, it uses $\bar{f}_t := \bar{f}$ as its output on the next iteration t. If such $\bar{f}$ does not exist, an application of the minimax theorem implies the existence of a sequence of elements (x_j, y_j) in $\mathcal{X} \times \{0,1\}$, such that, when we replace the $\mathcal{F}^i$ by the restrictions $\mathcal{F}^i|_{(x_j,y_j)}$ for all i and j, a certain potential function of $\mathcal{T}$ is decreased by an appreciable amount. This potential function can only decrease a bounded number of times, which implies that we must eventually come to a point at which a desired $\bar{f}$ can be found.

The main limitation of the above approach that precludes a $\operatorname{poly}\log T$ cumulative loss bound is the notion of closeness of the weighted average (improper) hypothesis h to the randomized (proper) hypothesis $\bar{f} \in \Delta^\circ(\mathcal{F})$. In [HLM21], a certain fixed scale α was chosen, and it was shown that we can find $\bar{f}$ so that for all (x, y) satisfying $\mathbb{E}_{h\sim Q}[|h(x) - y|] < \alpha$, then $\mathbb{E}_{f\sim\bar{f}}[|\bar{f}(x) - y|] \leq O(\alpha)$. This approach leads to cumulative loss of $\alpha T + \tilde{O}(\mathrm{Ldim}(\mathcal{F})/\alpha)$, which is never less than $O(\sqrt{\mathrm{Ldim}(\mathcal{F}) \cdot T})$. To improve upon this bound, we have to find $\bar{f}$ so that for *all* scales $\alpha \in [1/T, 1]$, if $\mathbb{E}_{h\sim Q}[|h(x)-y|] < \alpha$, then $\mathbb{E}_{f\sim\bar{f}}[|f(x)-y|] \leq O(\alpha)$ (step 2a of Algorithm 2). In the case that there does not exist a $\bar{f}$ satisfying this *stronger* condition, then when we apply the minimax theorem, we end with a sequence in $\mathcal{X} \times \{0,1\}$ satisfying a *weaker* condition (Lemma 5.8). Via a careful analysis of the potential function alluded to above, it turns out that this weaker condition is still sufficient to ensure a decrease in the potential (Lemma 5.10).

Furthermore, because of the multi-scale nature of this argument, our application of the minimax theorem is to a general real-valued function class, even in the case when $\mathcal{F}$ is binary-valued. Of course, it is necessary to prove that the minimax theorem actually holds in such settings. We show that it is sufficient for our needs to establish that the minimax theorem holds in general for real-valued classes which are online learnable (i.e., have sequential fat-shattering dimension finite at all scales). This fact, in turn, is proven in Section 8.

3.3 A multi-scale proper learner for regression

The proof of Theorem 1.1 (obtained by `Multi-scale Proper Learner`, Algorithm 2) follows, roughly speaking, by combining the hierarchical aggregation of SOA hypotheses (from the improper learner for realizable regression) with the insights from the previous section needed to obtain the optimal cumulative loss for binary classes. In particular, the weighted average hypothesis h formed each round from the previous section is replaced by a weighted average of hierarchically aggregated SOA hypotheses in the sense of Definition 4.2; the SOA hypotheses to be aggregated are collected in a data structure we call a *weighted subclass collection* (Definition 5.1). The resulting algorithm is "doubly multi-scale" in the following sense: we need to use multiple scales in the sense described in the previous paragraph to characterize the closeness of h and $\bar{f}$, but we also need multiple scales to deal with the growth of $\mathrm{sfat}_\alpha(\mathcal{F})$ as $\alpha \to 0$. This creates additional technical challenges; see Section 5 for details.

3.4 Making the proper learner stable

Next we address the stability property of `Multi-scale Proper Learner`, namely the proof of Proposition 1.3. We begin with the case of improper learning for binary classification, in which case the Standard Optimal Algorithm simply outputs the hypothesis $\mathrm{SOA}(\mathcal{F}^t)$ at each round t, and updates $\mathcal{F}^{t+1} \leftarrow \mathcal{F}^t|_{(x_t,y_t)}$ if it incorrectly predicts (x_t, y_t) (and otherwise sets $\mathcal{F}^{t+1} \leftarrow \mathcal{F}^t$). The key insight that allows a stable improper learner here is that $\mathcal{F}^t$ is only updated in the event of a mistake[8], and there are only $\mathrm{Ldim}(\mathcal{F})$ mistakes overall. Thus, for any $\eta > 0$, if we instead output the uniform distribution over the past $1/\eta$ hypotheses, $\mathrm{SOA}(\mathcal{F}^t), \ldots, \mathrm{SOA}(\mathcal{F}^{t-(1/\eta)})$, each original mistake will incur at most $1/\eta$ new ones, leading to a cumulative loss of $\mathrm{Ldim}(\mathcal{F})/\eta$. Further, the total variation distance between consecutive averages of $1/\eta$ hypotheses is at most 2η. Thus, for *improper learning for classification*, we immediately obtain the guarantee of Proposition 1.3. To obtain the same cumulative loss for a proper learner (and in the regression setting), we essentially pass the above insight into the machinery described in the previous sections. In particular, we show that due to the fact that we only make updates to subclasses $\mathcal{F}^i$ of the weighted subclass collection $\mathcal{T}$ when $\mathcal{F}^i$ makes a mistake, the collection $\mathcal{T}$ changes slowly (Lemmas 5.7 and 5.11), which allows us to show that averaging over a window of $1/\eta$ rounds only degrades the cumulative loss by a factor of $1/\eta$.

3.5 Application: fast rates for learning in games

At last we can overview the proof of Theorem 1.4, which leans heavily on our stable proper learner (Theorem 1.1 and Proposition 1.3). The main technical component of Theorem 1.4 is a *path-length* regret bound for a stable proper learner (Theorem 6.1), which shows (for a stable learning algorithm) that if consecutive losses fed by the adversary are close, then we can obtain improved regret (i.e., beating $O(\sqrt{T})$). At a high level, the idea of the proof of Theorem 6.1 is to use the "SOA-experts" technique of [BDPSS09, RST15a][9] which uses the existence of an online cover of bounded size for the hypothesis class $\mathcal{F}$ for any data sequence $x_1, x_2, \ldots, x_T$.[10] Each element of this online cover is interpreted as an expert, which runs an instance of our proper realizeable learner (`Multi-scale Proper Learner`). Typically one uses an online experts algorithm (such as exponential weights, i.e., Hedge) to learn the best expert in this cover. In order to obtain path-length regret bounds, we replace Hedge with Optimistic Hedge [RS13, SALS15] and use (as a black-box) the path-length regret bound of [SALS15]. Crucially, the stability property of the output of `Multi-scale Proper Learner` (Proposition 1.3) implies that (a) the outputs of the experts produce slowly-changing losses for the Optimistic Hedge algorithm, which is necessary to get strong path-length regret bounds, and (b) the outputs of the Optimistic Hedge are therefore slowly changing, meaning that in the game setting, other agents' losses are slowly changing. One additional challenge that occurs in the proof is that because each agent is playing randomized strategies, the function class we must work with is that which takes as input a distribution over examples $\mathcal{X}$, and thus is real-valued (even though we are in the setting of a binary game). In Lemma 7.1, we nevertheless show that its sequential fat-shattering dimension can be bounded in terms of the Littlestone dimension of the original binary-valued class, which allows us to use our results for proper realizable learning in the

[8] Some instantions of the Standard Optimal Algorithm restrict $\mathcal{F}^{t+1} \leftarrow \mathcal{F}^t|_{(x_t,y_t)}$ even if there is not a mistake at step t, though this is not necessary.

[9] For the latter reference [RST15a], see in particular the version at `https://arxiv.org/pdf/1006.1138v1.pdf`.

[10] See also the online version of the Sauer-Shelah lemma, [RS14b, Theorem 13.7].

real-valued setting. The full proof may be found in Sections 6 and 7.

4 A near-optimal improper cumulative loss bound

As a warm-up, we derive an optimal cumulative loss bound in the realizable setting for the easier case of improper learning, which remained open prior to this work. As noted previously, a cumulative loss bound of $O\left(\min_{\alpha\in[0,1]}\{\alpha T + \mathrm{sfat}_\alpha(\mathcal{F})\}\right)$ is immediate from the definition of $\mathrm{sfat}_\alpha(\cdot)$, but this regret bound is suboptimal in many cases, for instance when the sequential fat-shattering dimension exhibits growth $\mathrm{sfat}_\alpha(\mathcal{F}) \asymp \alpha^{-p}$ for some $p \in (0,1)$.

To improve upon this trivial bound, it is necessary to consider the sequential fat-shattering dimension at multiple scales α, somewhat analogously to how chaining is used to improve statistical rates in the agnostic setting. Our techniques for doing so differ substantially from chaining since rather than considering covers at different scales, we consider different hypotheses at different scales. To aggregate the predictions of the hypotheses at varying scales, we introduce *hierarchical aggregation rules* in Definition 4.1 below. First, we define the scales we will consider: for $\lambda \in \mathbb{Z}$, define $\alpha_\lambda := 2^{-\lambda}$. Throughout this section, we will fix some Λ (which ultimately will depend on the growth of $\mathrm{sfat}_\alpha(\mathcal{F})$ as $\alpha \to 0$) and consider scales $\alpha_1, \alpha_2, \ldots, \alpha_\Lambda$.

Definition 4.1 (Hierarchical aggregation)**.** For a sequence of real numbers $g_1, \ldots, g_\Lambda \in [0,1]$, we define the *hierarchical aggregation rule* $\mathrm{HAgg}(g_1, \ldots, g_\Lambda) \in [0,1]$ to be $g_{\bar\lambda}$, where $\bar\lambda$ is chosen so that for $2 \le \lambda' \le \bar\lambda$, it holds that $|g_{\lambda'} - g_{\lambda'-1}| \le 2\alpha_{\lambda'-1}$, yet $|g_{\bar\lambda} - g_{\bar\lambda+1}| > 2\alpha_{\bar\lambda}$, if such $\bar\lambda$ exists; if no such $\bar\lambda$ exists, set $\bar\lambda = \Lambda$. We will call this value of $\bar\lambda$ the *cutoff point* and denote it by $\bar\lambda = \bar\lambda(g_1, \ldots, g_\Lambda)$.

The individual hypotheses referred to above (to which a hierarchical aggregation rule is applied) will be the SOA hypotheses at differing scales (Definition 2.3), namely $\mathrm{SOA}(\mathcal{F}_\lambda, \alpha_\lambda)$, for various classes $\mathcal{F}_\lambda$. We next define the *SOA hypothesis for a sequence*:

Definition 4.2 (SOA hypotheses for sequences)**.** Given a sequence $\mathcal{F}_{1:\lambda} = (\mathcal{F}_1, \ldots, \mathcal{F}_\lambda)$ of hypothesis classes, define its *SOA hypothesis*, denoted $\mathrm{SOA}(\mathcal{F}_{1:\Lambda})$, as

$$\mathrm{SOA}(\mathcal{F}_{1:\Lambda})(x) = \mathrm{HAgg}(\mathrm{SOA}(\mathcal{F}_1, \alpha_1)(x), \mathrm{SOA}(\mathcal{F}_2, \alpha_2)(x), \ldots, \mathrm{SOA}(\mathcal{F}_\Lambda, \alpha_\Lambda)(x)).$$

We also denote the *cutoff point* for the sequence $(\mathrm{SOA}(\mathcal{F}_1, \alpha_1)(x), \ldots, \mathrm{SOA}(\mathcal{F}_\Lambda, \alpha_\Lambda)(x))$ by

$$\bar\lambda(\mathcal{F}_{1:\Lambda}, x) := \bar\lambda(\mathrm{SOA}(\mathcal{F}_1, \alpha_1)(x), \ldots, \mathrm{SOA}(\mathcal{F}_\Lambda, \alpha_\Lambda)(x)).$$

Algorithm 1, `Multi-scale Improper Learner`, presents an improper proper learner that uses the SOA hypothesis for sequences presented in Definition 4.2. The following proposition upper bounds the number of mistakes made by `Multi-scale Improper Learner`.

Proposition 4.1 (Optimal cumulative loss bound for improper learning)**.** *Suppose* $(x_t, y_t) \in \mathcal{X} \times [0,1]$ *and* $y_t = f^\star(x_t)$ *for some* $f^\star \in \mathcal{F}$ *for all* $t \in [T]$*. Then the predictions* $\hat y_t$*,* $t \in [T]$ *of* `Multi-scale Improper Learner` *(Algorithm 1) satisfy*

$$\sum_{t=1}^{T} |\hat y_t - y_t| \le C \cdot \inf_{\alpha\in[0,1]} \left\{ \alpha T + \int_\alpha^1 \mathrm{sfat}_\eta(\mathcal{F}) d\eta \right\}. \tag{3}$$

for some constant C*.*

Algorithm 1: `Multi-scale Improper Learner`

Input: Function class $\mathcal{F} \subset [0,1]^{\mathcal{X}}$, time horizon $T \in \mathbb{N}$, scale parameter $\Lambda \in \mathbb{N}$.

1. For $1 \le \lambda \le \Lambda$, initialize $\mathcal{F}_\lambda \leftarrow \mathcal{F}$.
2. For $1 \le t \le T$:
   (a) Observe x_t, and predict $\hat{y}_t := \mathrm{SOA}(\mathcal{F}_{1:\Lambda})(x_t)$.
   (b) Observe y_t, suffer loss $\delta_t := |y_t - \mathrm{SOA}(\mathcal{F}_{1:\Lambda})(t)|$.
   (c) Set λ_t, $1 \le \lambda_t \le \Lambda + 1$, to be $\Lambda + 1$ if $\delta_t \le \alpha_\Lambda$, and otherwise as small as possible so that $\delta_t > \alpha_{\lambda_t}$.
   (d) Set $\bar{\lambda}_t$, $1 \le \bar{\lambda}_t \le \Lambda$, to be the cutoff point $\bar{\lambda}_t := \bar{\lambda}(\mathcal{F}_{1:\Lambda}, x_t)$.
   (e) Update $\mathcal{F}_{\lambda'} \leftarrow \mathcal{F}_{\lambda'}|^{\alpha_{\lambda'}}_{(x_t,y_t)}$ for all $\lambda' \ge \min\{\lambda_t, \bar{\lambda}_t + 1\}$ such that $\mathrm{sfat}_{\alpha_{\lambda'}}(\mathcal{F}_{\lambda'}|^{\alpha_{\lambda'}}_{(x_t,y_t)}) < \mathrm{sfat}_{\alpha_{\lambda'}}(\mathcal{F}_{\lambda'})$.

Proof. Choose $\alpha \in [0,1]$ which minimizes $\alpha T + \int_\alpha^1 \mathrm{sfat}_\eta(\mathcal{F}) d\eta$; since we assume that $\mathrm{sfat}_c(\mathcal{F}) \ge 1$ for a constant c, we can assume that $\alpha \ge 1/T$ with the loss of a constant factor. Set $\Lambda = \lfloor \log 1/(2\alpha) \rfloor$. By bounding the integral in (3) below by the appropriate Riemann sum, it suffices to show that for some constant $C > 0$, we have

$$\sum_{t=1}^{T} |\hat{y}_t - y_t| \le T\alpha_\Lambda + C \sum_{\lambda=0}^{\Lambda} \mathrm{sfat}_{\alpha_\lambda}(\mathcal{F}) \cdot \alpha_\lambda.$$

Define, for $1 \le t \le T+1$,

$$\Phi_t(\mathcal{F}_{1:\Lambda}) := (T + 1 - t) \cdot \alpha_\Lambda + 16 \sum_{\lambda=1}^{\Lambda} \alpha_\lambda \cdot \mathrm{sfat}_{\alpha_\lambda}(\mathcal{F}_\lambda).$$

Below we will abbreviate Φ_t for the value $\Phi_t(\mathcal{F}_{1:\Lambda})$, where $\mathcal{F}_{1:\Lambda}$ is the sequence maintained by the algorithm at the beginning of round t. It is straightforward that $\Phi_1 = T\alpha_\Lambda + 16 \sum_{\lambda=1}^{\Lambda} \alpha_\lambda \cdot \mathrm{sfat}_{\alpha_\lambda}(\mathcal{F})$. Moreover, Φ_t is non-negative for all $t \le T+1$. We will show that $\Phi_t - \Phi_{t+1} \ge \delta_t$ for all t, which will imply the statement of the lemma.

Fix any $t \le T$, and let $\mathcal{F}_1, \ldots, \mathcal{F}_\Lambda$ denote the subclasses maintained by `Multi-scale Improper Learner` at the beginning of round t. We apply Lemma 4.2 for the sequence $\mathcal{F}_1, \ldots, \mathcal{F}_\Lambda$, $\delta = \delta_t$, and $(x, y) = (x_t, y_t)$. Note that the parameter λ in the statement of Lemma 4.2 is λ_t, and $\bar{\lambda}(\mathcal{F}_{1:\Lambda}, x) = \bar{\lambda}_t$. Lemma 4.2 then implies that at least one of the following holds:

- Either $\delta_t \le \alpha_\Lambda$, which implies that $\Phi_t - \Phi_{t+1} \ge \alpha_\Lambda \ge \delta_t$, as desired; or
- There is some $\lambda' \in [\Lambda]$ satisfying $\lambda' \ge \min\{\bar{\lambda}_t + 1, \lambda_t\}$ so that $|\mathrm{SOA}(\mathcal{F}_{\lambda'}, \alpha_{\lambda'})(x_t) - y_t| > \alpha_{\lambda'} \ge \delta_t/16$. By Lemma A.2, it follows that $\mathrm{sfat}_{\alpha_{\lambda'}}(\mathcal{F}_{\lambda'}|^{\alpha_{\lambda'}}_{(x_t,y_t)}) < \mathrm{sfat}_{\alpha_{\lambda'}}(\mathcal{F}_{\lambda'})$, which implies that $\Phi_t - \Phi_{t+1} \ge 16\alpha_{\lambda'} \ge \delta_t$, as desired.

In both cases, we thus get a decrease in the potential of at least δ_t, completing the proof of the proposition. ∎

Lemma 4.2 is the main technical lemma used in the proof of Proposition 4.1, used to show a decrease in the potential function therein.

Lemma 4.2. *Fix any $\Lambda \in \mathbb{N}$, any sequence of subclasses $\mathcal{F}_1, \dots, \mathcal{F}_\Lambda \subset \mathcal{F}$, and consider any $(x, y) \in \mathcal{X} \times [0, 1]$. Set $\delta := |y - \mathrm{SOA}(\mathcal{F}_{1:\Lambda})(x)|$, and define $\lambda \in \{1, 2, \dots, \Lambda\}$ to be $\Lambda + 1$ if $\delta \le \alpha_\Lambda$, and otherwise as small as possible so that $\delta > \alpha_\lambda$. Then at least one of the below holds:*

- *$\delta \le \alpha_\Lambda$ and $\bar{\lambda}(\mathcal{F}_{1:\Lambda}, x) = \Lambda$; or*
- *For some λ' satisfying $\min\{\bar{\lambda}(\mathcal{F}_{1:\Lambda}, x)+1, \lambda\} \le \lambda' \le \min\{\bar{\lambda}(\mathcal{F}_{1:\Lambda}, x)+1, \Lambda\}$, we have $|\mathrm{SOA}(\mathcal{F}_{\lambda'}, \alpha_{\lambda'})(x) - y| > \alpha_{\lambda'} \ge \delta/16$.*

Proof. Set $\bar{\lambda} = \bar{\lambda}(\mathcal{F}_{1:\Lambda}, x)$. Note that the choice of λ in the statement of the lemma ensures that $\delta \le 2\alpha_\lambda$. Further, by Definition 4.2, we have that $\mathrm{SOA}(\mathcal{F}_{\bar{\lambda}}, \alpha_{\bar{\lambda}})(x) = \mathrm{SOA}(\mathcal{F}_{1:\Lambda})(x)$.

We consider the following cases:

- First suppose that $\delta \le \alpha_\Lambda$ (i.e., $\lambda = \Lambda + 1$). If $\bar{\lambda} = \Lambda$, then we are done; otherwise, it must hold that $|\mathrm{SOA}(\mathcal{F}_{\bar{\lambda}}, \alpha_{\bar{\lambda}})(x) - \mathrm{SOA}(\mathcal{F}_{\bar{\lambda}+1}, \alpha_{\bar{\lambda}+1})(x)| > 2\alpha_{\bar{\lambda}}$. But $\delta = |y - \mathrm{SOA}(\mathcal{F}_{\bar{\lambda}}, \alpha_{\bar{\lambda}})(x)| \le \alpha_\Lambda \le \alpha_{\bar{\lambda}}$, meaning that $|\mathrm{SOA}(\mathcal{F}_{\lambda'}, \alpha_{\lambda'})(x) - y| > \alpha_{\bar{\lambda}} > \alpha_{\lambda'} \ge \delta/2$ with $\lambda' = \bar{\lambda} + 1$, thus verifying the second item in the lemma's statement.
- In the next case, suppose that $\Lambda \ge \lambda \ge \bar{\lambda} + 1$. If it is not the case that $|\mathrm{SOA}(\mathcal{F}_{\bar{\lambda}}, \alpha_{\bar{\lambda}})(x) - \mathrm{SOA}(\mathcal{F}_{\bar{\lambda}+1}, \alpha_{\bar{\lambda}+1})(x)| > 2\alpha_{\bar{\lambda}}$, then we must have $\bar{\lambda} = \Lambda$ and so $\lambda = \Lambda+1$, which contradicts our assumption of $\lambda \le \Lambda$ in this case. Otherwise, $|\mathrm{SOA}(\mathcal{F}_{\bar{\lambda}}, \alpha_{\bar{\lambda}})(x) - \mathrm{SOA}(\mathcal{F}_{\bar{\lambda}+1}, \alpha_{\bar{\lambda}+1})(x)| > 2\alpha_{\bar{\lambda}}$ holds. Moreover we have

  $$|\mathrm{SOA}(\mathcal{F}_{\bar{\lambda}}, \alpha_{\bar{\lambda}})(x) - y| = \delta \le 2\alpha_\lambda \le 2\alpha_{\bar{\lambda}+1} = \alpha_{\bar{\lambda}},$$

  Hence $|\mathrm{SOA}(\mathcal{F}_{\bar{\lambda}_t+1}, \alpha_{\bar{\lambda}+1})(x) - y| > \alpha_{\bar{\lambda}} > \alpha_{\bar{\lambda}+1} \ge \delta/2$, so in this case we may again choose $\lambda' = \bar{\lambda} + 1$.
- In the final case, $\lambda \le \bar{\lambda}$ (and the previous cases do not apply). Thus here $|y - \mathrm{SOA}(\mathcal{F}_{\bar{\lambda}}, \alpha_{\bar{\lambda}})(x)| = \delta > \alpha_\lambda \ge \alpha_{\bar{\lambda}}$. We consider two sub-cases:
  - In the event that $\lambda \ge \bar{\lambda} - 3$ (i.e., $\lambda \in \{\bar{\lambda} - 3, \bar{\lambda} - 2, \bar{\lambda} - 1, \bar{\lambda}\}$), we therefore have that $\delta \le 2\alpha_\lambda \le 16\alpha_{\bar{\lambda}}$, meaning that, with $\lambda' = \bar{\lambda}$, $|y - \mathrm{SOA}(\mathcal{F}_{\lambda'}, \alpha_{\lambda'})(x)| > \alpha_{\lambda'} \ge \delta/16$.
  - In the other subcase, we have $\lambda < \bar{\lambda} - 3$; then we have

    $$\begin{aligned}
    &|\mathrm{SOA}(\mathcal{F}_{\bar{\lambda}}, \alpha_{\bar{\lambda}})(x) - \mathrm{SOA}(\mathcal{F}_{\lambda+3}, \alpha_{\lambda+3})(x)| \\
    \le &\sum_{\lambda'=\lambda+3}^{\bar{\lambda}-1} |\mathrm{SOA}(\mathcal{F}_{\lambda'}, \alpha_{\lambda'})(x) - \mathrm{SOA}(\mathcal{F}_{\lambda'+1}, \alpha_{\lambda'+1})(x)| \\
    \le &\sum_{\lambda'=\lambda+3}^{\bar{\lambda}-1} 2\alpha_{\lambda'} \\
    \le &4\alpha_{\lambda+3} = \alpha_\lambda/2.
    \end{aligned}$$

    It follows that $|\mathrm{SOA}(\mathcal{F}_{\lambda+3}, \alpha_{\lambda+3})(x) - y| > \alpha_\lambda/2 > \alpha_{\lambda+3} \ge \delta/16$.

□

5 A near-optimal proper cumulative loss bound

Throughout this section, we consider a real-valued class $\mathcal{F} \subset [0,1]^{\mathcal{X}}$, so that $\mathrm{sfat}_\alpha(\mathcal{F}) < \infty$ for all $\alpha > 0$. Having established a cumulative loss bound in Proposition 4.1 for improper learning of $\mathcal{F}$ in the realizable setting, we turn to the more challenging case of proper learning of $\mathcal{F}$. In addition to the ideas on hierarchical aggregation used in the case of improper learner (Section 4 above), a key tool we use is a generalization of the proper online realizable learner of [HLM21]. It is necessary, however, to make substantial modifications to the algorithm of [HLM21]: for one, it only applies in the setting of binary classification, but even in that setting it provides a (significantly) suboptimal cumulative loss of $\tilde{O}(\sqrt{\mathrm{Ldim}(\mathcal{F}) \cdot T})$; our proper algorithm gives cumulative loss of $O(\mathrm{Ldim}(\mathcal{F}) \cdot \mathrm{poly}\log(T))$. Thus, we introduce new techniques to correct both of these shortcomings.

We begin by introducing some notation. Fix some scale parameter $\Lambda \in \mathbb{N}$ (ultimately, Λ will be chosen identically as in the proof of Proposition 4.1). We will consider scales $\alpha_\lambda = 2^{-\lambda}$ for $\lambda \in \mathbb{Z}$. We will primarily be considering values of λ in the set $[\Lambda] = \{1, 2, \ldots, \Lambda\}$ but occasionally will refer to α_λ for other (integral) values of λ.

Definition 5.1 (Weighted subclass collection)**.** A *weighted subclass collection* $\mathcal{T}$ is a tuple $\mathcal{T} = (\mathcal{T}_1, \ldots, \mathcal{T}_\Lambda)$, where for each $\lambda \in [\Lambda]$, $\mathcal{T}_\lambda$ is a multiset of tuples of the form $\mathcal{T}_\lambda = \{(\mathcal{G}_\lambda^1, w_\lambda^1), \ldots, (\mathcal{G}_\lambda^{|\mathcal{T}_\lambda|}, w_\lambda^{|\mathcal{T}_\lambda|})\}$, where for each $1 \le v_\lambda \le |\mathcal{T}_\lambda|$, we have $w_\lambda^{v_\lambda} \ge 0$ and $\mathcal{G}_\lambda^{v_\lambda} \subset \mathcal{F}$.

We will use the letter w to denote the collection of all $w_\lambda^{v_\lambda}$, for $\lambda \in [\Lambda]$ and $1 \le v_\lambda \le |\mathcal{T}_\lambda|$, and the letter $\mathcal{G}$ to denote the collection of all $\mathcal{G}_\lambda^{v_\lambda}$, for $\lambda \in [\Lambda]$ and $1 \le v_\lambda \le |\mathcal{T}_\lambda|$. We introduce the following notation to denote a weighted subclass collection $\mathcal{T}$: we will write $\mathcal{T} = [\mathcal{G}, w]$.[11]

In words, the weighted subclass collection $\mathcal{T}$ denotes a collection of subclasses of $\mathcal{F}$ together with non-negative weights for each scale λ; our algorithm will use a weighted aggregation of the SOA hypotheses of these subclasses, according to the weights $w_\lambda^{v_\lambda}$. For a weighted subclass collection $\mathcal{T}$, let $\mathcal{N}_\Lambda(\mathcal{T})$ denote the set of sequences $(v_1, \ldots, v_\Lambda)$, where for $\lambda \in [\Lambda]$, $1 \le v_\lambda \le |\mathcal{T}_\lambda|$ (i.e., $(\mathcal{G}_\lambda^{v_\lambda}, w^{v_\lambda}) \in \mathcal{T}_\lambda$). We will abbreviate the sequence $(v_1, \ldots, v_\Lambda)$ with the letter v, so that we have $v \in \mathcal{N}_\Lambda(\mathcal{T})$. We also abbreviate the sequence $(\mathcal{G}_1^{v_1}, \ldots, \mathcal{G}_\Lambda^{v_\Lambda})$ as $\mathcal{G}_{1:\Lambda}^v$ and the sequence $(w_1^{v_1}, \ldots, w_\Lambda^{v_\Lambda})$ as $w_{1:\Lambda}^v$. For each $v \in \mathcal{N}_\Lambda(\mathcal{T})$, we next define

$$P_w(v) := \prod_{\lambda=1}^{\Lambda} \frac{w_\lambda^{v_\lambda}}{\sum_{u_\lambda=1}^{|\mathcal{T}_\lambda|} w_\lambda^{u_\lambda}}.$$

It is evident from the above definition that $\sum_{v \in \mathcal{N}_\Lambda(\mathcal{T})} P_w(v) = 1$.

For a sequence $\mathcal{G}_{1:\Lambda}$ and $(x, y) \in \mathcal{X} \times [0,1]$, define its *truncated error at the point* (x, y) as

$$\mathrm{TErr}(\mathcal{G}_{1:\Lambda}, x, y) := \max\left\{ |\operatorname{SOA}(\mathcal{G}_{1:\Lambda})(x) - y|, \alpha_{\bar{\lambda}(\mathcal{G}_{1:\Lambda}, x)} \right\}.$$

The intuition behind the truncated error is as follows: recall that $\operatorname{SOA}(\mathcal{G}_{1:\Lambda})(x) = \operatorname{SOA}(\mathcal{G}_{1:\Lambda}, \alpha_{\bar{\lambda}(\mathcal{G}_{1:\Lambda}, x)})(x)$, meaning that $\operatorname{SOA}(\mathcal{G}_{1:\Lambda})(x)$ is, in general, only accurate up to an additive $\alpha_{\bar{\lambda}(\mathcal{G}_{1:\Lambda}, x)}$. Thus, if it happens that $|\operatorname{SOA}(\mathcal{G}_{1:\Lambda})(x) - y| \ll \alpha_{\bar{\lambda}(\mathcal{G}_{1:\Lambda}, x)}$, then this is due to "luck"; it turns out that in order to ensure that certain potential functions always decrease it is convenient to still force us to pay $\alpha_{\bar{\lambda}(\mathcal{G}_{1:\Lambda}, x)}$ in our error bounds when such "lucky" situations occur.

[11] This notation emphasizes the use of the letters $\mathcal{G}, w$ to denote the subclasses and weights, respectively, belonging to $\mathcal{T}$. If we wish to describe another weighted subclass collection, we might notate it as $\mathcal{S} = [\mathcal{H}, z]$, replacing the pairs $(\mathcal{G}_\lambda^{v_\lambda}, w_\lambda^{v_\lambda})$ with the pairs $(\mathcal{H}_\lambda^{v_\lambda}, z_\lambda^{v_\lambda})$.

Now fix a weighted subclass collection $\mathcal{T} = [\mathcal{G}, w]$; we will write

$$\mathrm{TErr}(\mathcal{T}, x, y) := \sum_{v \in \mathcal{N}_\Lambda(\mathcal{T})} P_w(v) \cdot \mathrm{TErr}(\mathcal{G}^v_{1:\Lambda}, x, y)$$

to denote the average truncated error of a element $\mathcal{G}^v_{1:\Lambda}$, drawn according to the distribution $P_w(v)$. Next define for $x \in \mathcal{X}$,

$$\mathrm{Vote}_{\mathcal{T}}(x) := \sum_{v \in \mathcal{N}_\Lambda(\mathcal{T})} P_w(v) \cdot \mathrm{SOA}(\mathcal{G}^v_{1:\Lambda})(x).$$

For $\epsilon \in [0,1]$, define

$$\mathrm{HighVote}(\mathcal{T}, \epsilon) := \{(x, \mathrm{Vote}_{\mathcal{T}}(x)) : x \in \mathcal{X},\ \mathrm{TErr}(\mathcal{T}, x, \mathrm{Vote}_{\mathcal{T}}(x)) \le \epsilon\}. \tag{4}$$

In words, $\mathrm{HighVote}(\mathcal{T}, \epsilon)$ is the set of tuples $(x, \mathrm{Vote}_{\mathcal{T}}(x))$ for which the truncated error of $\mathrm{TErr}(\mathcal{G}^v_{1:\Lambda}, x, \mathrm{Vote}_{\mathcal{T}}(x))$ is at most ϵ, when v is drawn from the distribution induced by $P_w(v)$, for $v \in \mathcal{N}_\Lambda(\mathcal{T})$. The set $\mathrm{HighVote}(\mathcal{T}, \epsilon)$ should be interpreted as the set of tuples $(x, \mathrm{Vote}_{\mathcal{T}}(x))$ about which the hypotheses $\mathrm{SOA}(\mathcal{G}^v_{1:\Lambda})$, weighted according to $v \sim P_w(\cdot)$, are "nearly unanimous" (up to error ϵ) about the label of the point x. The quantities $\mathrm{Vote}_{\mathcal{T}}(x)$, $\mathrm{HighVote}(\mathcal{T}, \epsilon)$ are generalizations of the analogous quantities defined in [HLM21] to the real-valued case.

For $f \in \mathcal{F}$, let $\delta_f \in \Delta^\circ(\mathcal{F})$ denote the point mass at f. For $N \in \mathbb{N}$ we define the class

$$\mathrm{Rand}(\mathcal{F}^N) = \left\{ \frac{1}{N} \cdot (\delta_{f_1} + \cdots + \delta_{f_n}) \ :\ f_1, \ldots, f_N \in \mathcal{F} \right\} \subset \Delta^\circ(\mathcal{F}), \tag{5}$$

to be the collection of (uniform) averages of N hypotheses in $\mathcal{F}$. We will often denote elements of $\mathrm{Rand}(\mathcal{F}^N)$ with bars, e.g., given $f_1, \ldots, f_N$, we will denote the corresponding element of $\mathrm{Rand}(\mathcal{F}^N)$ by $\bar{f}$, so that $\bar{f} = (\delta_{f_1} + \cdots + \delta_{f_N})/N$. The algorithm `Multi-scale Proper Learner` outputs elements of $\mathrm{Rand}(\mathcal{F}^N)$ for an appropriate integer N.

5.1 Some results on weighted subclass collections

We begin by proving some results on weighted subclass collections (Definition 5.1) and their relation to the notions of truncated error and Highvote defined above. Lemma 5.1 shows that the prediction made by the voting hypothesis, $\mathrm{Vote}_{\mathcal{T}}(x)$, achieves the optimal truncated error up to a constant factor (of 2).

Lemma 5.1. *Fix a weighted subclass collection $\mathcal{T} = [\mathcal{G}, w]$. For any $x \in \mathcal{X}$, it holds that*

$$\mathrm{TErr}(\mathcal{T}, x, \mathrm{Vote}_{\mathcal{T}}(x)) \le 2 \cdot \min_{y \in [0,1]} \{\mathrm{TErr}(\mathcal{T}, x, y)\}.$$

Note that the conclusion of Lemma 5.1 can be rewritten as:

$$\sum_{v \in \mathcal{N}_\Lambda(\mathcal{T})} P_w(v) \cdot \mathrm{TErr}(\mathcal{G}^v_{1:\Lambda}, x, \mathrm{Vote}_{\mathcal{T}}(x)) \le 2 \cdot \min_{y \in [0,1]} \sum_{v \in \mathcal{N}_\Lambda(\mathcal{T})} P_w(v) \cdot \mathrm{TErr}(\mathcal{G}^v_{1:\Lambda}, x, y).$$

Proof of Lemma 5.1. Set $y_0 := \arg\min_{y \in [0,1]} \sum_{v \in \mathcal{N}_\Lambda(\mathcal{T})} P_w(v) \cdot \mathrm{TErr}(\mathcal{G}^v_{1:\Lambda}, x, y)$ and $y_1 = \mathrm{Vote}_{\mathcal{T}}(x)$. We assume that $y_1 \ge y_0$ (the other case $y_1 \le y_0$ is treated in a symmetric manner). Set

$$\mathcal{S}_-(x) := \{v \in \mathcal{N}_\Lambda(\mathcal{T}) :\ \mathrm{SOA}(\mathcal{G}^v_{1:\Lambda})(x) < y_1\}$$
$$\mathcal{S}_+(x) := \{v \in \mathcal{N}_\Lambda(\mathcal{T}) :\ \mathrm{SOA}(\mathcal{G}^v_{1:\Lambda})(x) \ge y_1\}.$$

Since y_1 is the weighted mean of the quantities $\mathrm{SOA}(\mathcal{G}^v_{1:\Lambda})(x)$ (according to the weights $P_w(v)$), over $v \in \mathcal{N}_\Lambda(\mathcal{T})$, it holds that

$$\sum_{v\in\mathcal{S}_-(x)} P_w(v)\cdot|y_1-\mathrm{SOA}(\mathcal{G}^v_{1:\Lambda})(x)| = \sum_{v\in\mathcal{S}_+(x)} P_w(v)\cdot|y_1-\mathrm{SOA}(\mathcal{G}^v_{1:\Lambda})(x)|. \tag{6}$$

For $v \in \mathcal{S}_+(x)$, since $\mathrm{SOA}(\mathcal{G}^v_{1:\Lambda})(x) \geq y_1 \geq y_0$, we have that $|\,\mathrm{SOA}(\mathcal{G}^v_{1:\Lambda})(x)-y_0| \geq |\,\mathrm{SOA}(\mathcal{G}^v_{1:\Lambda})(x)-y_1|$, and thus

$$\mathrm{TErr}(\mathcal{G}^v_{1:\Lambda},x,y_0) \geq \mathrm{TErr}(\mathcal{G}^v_{1:\Lambda},x,y_1). \tag{7}$$

Hence

$$\sum_{v\in\mathcal{S}_+(x)} P_w(v)\cdot\mathrm{TErr}(\mathcal{G}^v_{1:\Lambda},x,y_1) \leq \sum_{v\in\mathcal{S}_+(x)} P_w(v)\cdot\mathrm{TErr}(\mathcal{G}^v_{1:\Lambda},x,y_0). \tag{8}$$

Note that for any $v \in \mathcal{S}_-(x)$, if $\alpha_{\bar\lambda(\mathcal{G}^v_{1:\Lambda},x)} > |\,\mathrm{SOA}(\mathcal{G}^v_{1:\Lambda})(x)-y_1|$, we again have that

$$\mathrm{TErr}(\mathcal{G}^v_{1:\Lambda},x,y_0) \geq \mathrm{TErr}(\mathcal{G}^v_{1:\Lambda},x,y_1) = \alpha_{\bar\lambda(\mathcal{G}^v_{1:\Lambda},x)},$$

since $\alpha_{\bar\lambda(\mathcal{G}^v_{1:\Lambda},x)}$ is the minimum, for all $y \in [0,1]$, of $\mathrm{TErr}(\mathcal{G}^v_{1:\Lambda},x,y)$. Thus, using (6) for the first inequality below,

$$\begin{aligned}\sum_{v\in\mathcal{S}_-(x)} P_w(v)\cdot\mathrm{TErr}(\mathcal{G}^v_{1:\Lambda},x,y_1) &\leq \sum_{v\in\mathcal{S}_+(x)} P_w(v)\cdot|\,\mathrm{SOA}(\mathcal{G}^v_{1:\Lambda})(x)-y_1| + \sum_{v\in\mathcal{S}_-(x)} P_w(v)\cdot\mathrm{TErr}(\mathcal{G}^v_{1:\Lambda},x,y_0)\\ &\leq \sum_{v\in\mathcal{S}_+(x)} P_w(v)\cdot\mathrm{TErr}(\mathcal{G}^v_{1:\Lambda},x,y_0) + \sum_{v\in\mathcal{S}_-(x)} P_w(v)\cdot\mathrm{TErr}(\mathcal{G}^v_{1:\Lambda},x,y_0) \qquad (9)\\ &= \sum_{v\in\mathcal{N}_\Lambda(\mathcal{T})} P_w(v)\cdot\mathrm{TErr}(\mathcal{G}^v_{1:\Lambda},x,y_0), \qquad (10)\end{aligned}$$

where (9) uses (7). Combining (8) and (10) gives the desired conclusion. □

Lemma 5.2 shows that if the truncated error $\mathrm{TErr}(\mathcal{T},x,y)$ is large for some $\mathcal{T}$ and (x,y), then we can get a lower bound on the weight of hypotheses $(\mathcal{G}^{v_\lambda}_\lambda, w^{v_\lambda}_\lambda)$ in the multisets $\mathcal{T}_\lambda$ for which $\mathrm{SOA}(\mathcal{G}^{v_\lambda}_\lambda,\alpha_\lambda)(x)$ is not close to y. This lemma will be used to show that if the truncated error for some example (x_t,y_t) is large, then we can update the classes $\mathcal{G}^{v_\lambda}_\lambda$ and the weights $w^{v_\lambda}_\lambda$ in a way that decreases a certain potential function.

Lemma 5.2. *Fix a weighted subclass collection $\mathcal{T}=[\mathcal{G},w]$. For any example $(x,y)\in\mathcal{X}\times[0,1]$, it holds that*

$$\alpha_\Lambda + \sum_{\lambda=1}^{\Lambda}\alpha_\lambda\cdot\frac{\sum_{v_\lambda=1}^{|\mathcal{T}_\lambda|} w^{v_\lambda}_\lambda\cdot\mathbb{1}[|\,\mathrm{SOA}(\mathcal{G}^{v_\lambda}_\lambda,\alpha_\lambda)(x)-y|>\alpha_\lambda]}{\sum_{v_\lambda=1}^{|\mathcal{T}_\lambda|} w^{v_\lambda}_\lambda} \geq \frac{1}{16}\cdot\sum_{v\in\mathcal{N}_\Lambda(\mathcal{T})} P_w(v)\cdot\mathrm{TErr}(\mathcal{G}^v_{1:\Lambda},x,y). \tag{11}$$

Proof. Consider any $v\in\mathcal{N}_\Lambda(\mathcal{T})$. We will assign the mass $P_w(v)$ to some tuple $(\mathcal{G}^{v_\lambda}_\lambda, w^{v_\lambda}_\lambda)\in\mathcal{T}_\lambda$, for some $\lambda\in[\Lambda]$ which satisfies $\alpha_\lambda+\alpha_\Lambda \geq \Omega(\mathrm{TErr}(\mathcal{G}^v_{1:\Lambda},x,y))$, in the following manner. Set $\bar\lambda_v := \bar\lambda(\mathcal{G}^v_{1:\Lambda},x)$. Then by Lemma 4.2 with $\mathcal{F}_{1:\Lambda}=\mathcal{G}^v_{1:\Lambda}=(\mathcal{G}^{v_1}_1,\ldots,\mathcal{G}^{v_\Lambda}_\Lambda)$, at least one of the below is the case:

- $|\operatorname{SOA}(\mathcal{G}^v_{1:\Lambda})(x) - y| \le \alpha_\Lambda$ and $\bar{\lambda}_v = \Lambda$; or
- For some $\lambda' \le \bar{\lambda}_v + 1$, we have $|\operatorname{SOA}(\mathcal{G}^{v_{\lambda'}}_{\lambda'}, \alpha_{\lambda'})(x) - y| > \alpha_{\lambda'} \ge |\operatorname{SOA}(\mathcal{G}^v_{1:\Lambda})(x) - y|/16$.

Set λ' to be the value guaranteed by the second item above, in the event that it holds, and $\lambda' = \perp$ otherwise. Then

$$\alpha_\Lambda + \alpha_{\lambda'} \cdot \mathbb{1}[|\operatorname{SOA}(\mathcal{G}^{v_{\lambda'}}_{\lambda'}, \alpha_{\lambda'})(x) - y| > \alpha_{\lambda'}] \ge \max\left\{\frac{1}{16} \cdot |\operatorname{SOA}(\mathcal{G}^v_{1:\Lambda})(x) - y|, \frac{\alpha_{\bar{\lambda}_v}}{2}\right\} \ge \frac{1}{16} \cdot \operatorname{TErr}(\mathcal{G}^v_{1:\Lambda}, x, y).$$

(Note that in the event $\lambda' = \perp$, we have that $\mathbb{1}[|\operatorname{SOA}(\mathcal{G}^{v_{\lambda'}}_{\lambda'}, \alpha_{\lambda'})(x) - y| > \alpha_{\lambda'}] = 0$, meaning that the left-hand side of the above expression is well-defined.) For the element v, we now assign weight $P_w(v)$ to the tuple $(\mathcal{G}^{v_{\lambda'}}_{\lambda'}, v^{v_{\lambda'}}_{\lambda'}) \in \mathcal{T}_{\lambda'}$, in the event that $\lambda' \neq \perp$ (and do not assign the weight $P_w(v)$ to any tuple in the event that $\lambda' = \perp$).

Note that total weight (taken over all $v \in \mathcal{N}_\Lambda(\mathcal{T})$) that could be assigned to any tuple $(\mathcal{G}^{v_\lambda}_\lambda, w^{v_\lambda}_\lambda) \in \mathcal{T}_\lambda$, for any $\lambda \in [\Lambda]$, is at most

$$\sum_{u \in \mathcal{N}_\Lambda(\mathcal{T}): u_\lambda = v_\lambda} P_w(u) = \frac{w^{v_\lambda}_\lambda}{\sum_{u_\lambda=1}^{|\mathcal{T}_\lambda|} w^{u_\lambda}_\lambda}.$$

Thus (11) follows. □

Lemma 5.3 shows that if some SOA hypothesis corresponding to a sequence has large error on a point (x, y), then the SOA hypothesis for the sequence must have large truncated error on (x, y).

Lemma 5.3. *Fix a sequence $\mathcal{F}_{1:\Lambda}$ and a point $(x, y) \in \mathcal{X} \times [0, 1]$. If, for some $\lambda \in [\Lambda]$, $|\operatorname{SOA}(\mathcal{F}_\lambda, \alpha_\lambda)(x) - y| > 5\alpha_\lambda$, then $\operatorname{TErr}(\mathcal{F}_{1:\Lambda}, x, y) > \alpha_\lambda$.*

Proof. Fix $\lambda \in [\Lambda]$ so that $|\operatorname{SOA}(\mathcal{F}_\lambda, \alpha_\lambda)(x) - y| > 5\alpha_\lambda$, and set $\bar{\lambda} := \bar{\lambda}(\mathcal{F}_{1:\Lambda}, x)$. Since $\operatorname{TErr}(\mathcal{F}_{1:\Lambda}, x, y) \ge \alpha_{\bar{\lambda}}$, the lemma clearly holds if $\bar{\lambda} \le \lambda$. Otherwise, we have that for all λ' satisfying $\lambda \le \lambda' < \bar{\lambda}$, $|\operatorname{SOA}(\mathcal{F}_{\lambda'}, \alpha_{\lambda'})(x) - \operatorname{SOA}(\mathcal{F}_{\lambda'+1}, \alpha_{\lambda'+1})(x)| \le 2\alpha_{\lambda'}$. Therefore

$$|\operatorname{SOA}(\mathcal{F}_{1:\Lambda})(x) - \operatorname{SOA}(\mathcal{F}_\lambda, \alpha_\lambda)(x)| \le \sum_{\lambda'=\lambda}^{\bar{\lambda}-1} |\operatorname{SOA}(\mathcal{F}_{\lambda'}, \alpha_{\lambda'})(x) - \operatorname{SOA}(\mathcal{F}_{\lambda'+1}, \alpha_{\lambda'+1})(x)| \le \sum_{\lambda'=\lambda}^{\bar{\lambda}-1} 2\alpha_{\lambda'} \le 4\alpha_\lambda,$$

and it follows that $\operatorname{TErr}(\mathcal{F}_{1:\Lambda}, x, y) \ge |y - \operatorname{SOA}(\mathcal{F}_{1:\Lambda})(x)| \ge \alpha_\lambda$. □

Notice that Lemma 5.3 is not true if the truncated error $\operatorname{TErr}(\mathcal{F}_{1:\Lambda}, x, y)$ is replaced with the absolute loss $|y - \operatorname{SOA}(\mathcal{F}_{1:\Lambda})(x)|$: it could be the case that for the given λ in the lemma statement, the cutoff point $\bar{\lambda}(\mathcal{F}_{1:\Lambda}, x)$ is much smaller than λ but $\operatorname{SOA}(\mathcal{F}_{1:\Lambda})(x) = \operatorname{SOA}(\mathcal{F}_{\bar{\lambda}}, \alpha_{\bar{\lambda}})(x)$ happens to be very close to y.

The next lemma shows that if a weighted subclass collection $\mathcal{T} = [\mathcal{G}, w]$ is "nearly unanimous" about the label y of a point x (in the sense of Highvote, defined in (4)), then most of the individual (single-scale) hypotheses $\operatorname{SOA}(\mathcal{G}^{v_\lambda}_\lambda, \alpha_\lambda)$ from $\mathcal{T}$ must predict (x, y) approximately correctly. It may be seen as a sort of converse to Lemma 5.2, which shows that if the truncated error $\operatorname{TErr}(\mathcal{T}, x, y)$ is small, then many of the SOA hypotheses at individual scales must be inaccurate on (x, y).

Lemma 5.4. *Consider any weighted subclass collection $\mathcal{T} = [\mathcal{G}, w]$, and $(x, y) \in \mathcal{X} \times [0, 1]$. If $(x, y) \in \mathrm{HighVote}(\mathcal{T}, \alpha)$ for some $\alpha \geq 0$, then for all $\lambda \in [\Lambda]$,*

$$\frac{\sum_{v_\lambda=1}^{|\mathcal{T}_\lambda|} w_\lambda^{v_\lambda} \cdot \mathbb{1}[|\,\mathrm{SOA}(\mathcal{G}_\lambda^{v_\lambda}, \alpha_\lambda)(x) - y| > 5\alpha_\lambda]}{\sum_{u_\lambda=1}^{|\mathcal{T}_\lambda|} w_\lambda^{u_\lambda}} \leq \frac{\alpha}{\alpha_\lambda}. \tag{12}$$

Proof. The conclusion of the lemma is immediate if $\alpha \geq \alpha_\lambda$, so we may assume from here on that $\alpha < \alpha_\lambda$.

By Lemma 5.3, for any $v \in \mathcal{N}_\Lambda(\mathcal{T})$ and $\lambda \in [\Lambda]$ for which $|\,\mathrm{SOA}(\mathcal{G}_\lambda^{v_\lambda}, \alpha_\lambda)(x) - y| > 5\alpha_\lambda$, it holds that $\mathrm{TErr}(\mathcal{G}_{1:\Lambda}^{v}, x, y) > \alpha_\lambda$. Thus, for any tuple $(\mathcal{G}_\lambda^{v_\lambda}, w_\lambda^{v_\lambda}) \in \mathcal{T}_\lambda$ for which $|\,\mathrm{SOA}(\mathcal{G}_\lambda^{v_\lambda}, \alpha_\lambda)(x) - y| > 5\alpha_\lambda$ there is a set $\mathcal{S}_{v_\lambda} \subset \mathcal{N}_\Lambda(\mathcal{T})$ (namely, the set of all u for which $u_\lambda = v_\lambda$) so that for all $u \in \mathcal{S}_{v_\lambda}$, $\mathrm{TErr}(\mathcal{G}_{1:\Lambda}^{u}, x, y) > \alpha_\lambda$ and so that $\sum_{u \in \mathcal{S}_{v_\lambda}} P_w(u) = \frac{w_\lambda^{v_\lambda}}{\sum_{u_\lambda=1}^{|\mathcal{T}_\lambda|} w_\lambda^{u_\lambda}}$.

That $(x, y) \in \mathrm{HighVote}(\mathcal{T}, \alpha)$ means that $\sum_{v \in \mathcal{N}_\Lambda(\mathcal{T})} P_w(v) \cdot \mathrm{TErr}(\mathcal{G}_{1:\Lambda}^{v}, x, y) \leq \alpha$. By Markov's inequality, for any λ for which $\alpha_\lambda > \alpha$,

$$\sum_{v \in \mathcal{N}_\Lambda(\mathcal{T})} P_w(v) \cdot \mathbb{1}[\mathrm{TErr}(\mathcal{G}_{1:\Lambda}^{v}, x, y) > \alpha_\lambda] \leq \frac{\alpha}{\alpha_\lambda}.$$

Thus, the mass (under $P_w(\cdot)$) of the union of all the sets $\mathcal{S}_{v_\lambda}$ is at most $\frac{\alpha}{\alpha_\lambda}$. Since the sets $\mathcal{S}_{v_\lambda}$ are pairwise disjoint, (12) follows. ∎

Lemma 5.5. *Suppose P is a distribution supported on $[0, 1]$ so that $\mathbb{E}_{Z \sim P}[Z] > \alpha$ for some $\alpha \in [2\alpha_\Lambda, 1]$. Then there is some $\lambda \in [\Lambda]$ so that $\mathbb{E}_{Z \sim P}[\mathbb{1}[Z > \alpha_\lambda]] > \frac{\alpha}{4\Lambda\alpha_\lambda}$.*

Proof. Suppose for the purpose of contradiction that for all $\lambda \in [\Lambda]$, $\mathbb{E}[\mathbb{1}[Z > \alpha_\lambda]] \leq \frac{\alpha}{4\Lambda\alpha_\lambda}$. Then

$$\begin{aligned}
\mathbb{E}[Z] \leq& \sum_{\lambda=1}^{\Lambda} \mathbb{P}[\alpha_\lambda < Z \leq \alpha_{\lambda-1}] \cdot \alpha_{\lambda-1} + \alpha_\Lambda \\
\leq& \alpha_\Lambda + \sum_{\lambda=1}^{\Lambda} \frac{\alpha\alpha_{\lambda-1}}{4\Lambda\alpha_\lambda} \\
\leq& \alpha_\Lambda + \alpha/2 \leq \alpha,
\end{aligned}$$

a contradiction. ∎

Lemma 5.6 shows that if two length-Λ sequences of real numbers have their first λ_0 positions equal, then their hierarchical aggregations (Definition 4.1) differ by only $O(\alpha_{\lambda_0})$.

Lemma 5.6. *Suppose $\Lambda \in \mathbb{N}$, and $g_1, \ldots, g_\Lambda, g'_1, \ldots, g'_\Lambda \in [0, 1]$. Suppose $\lambda_0 \leq \Lambda$ is such that for all $\lambda \leq \lambda_0$, $g_\lambda = g'_\lambda$. Then $|\,\mathrm{HAgg}(g_1, \ldots, g_\Lambda) - \mathrm{HAgg}(g'_1, \ldots, g'_\Lambda)| \leq 8\alpha_{\lambda_0}$.*

Proof. Set $\bar{\lambda} := \bar{\lambda}(g_1, \ldots, g_\Lambda)$ and $\bar{\lambda}' := \bar{\lambda}(g'_1, \ldots, g'_\Lambda)$. In particular, we have $\mathrm{HAgg}(g_1, \ldots, g_\Lambda) = g_{\bar{\lambda}}$ and $\mathrm{HAgg}(g'_1, \ldots, g'_\Lambda) = g'_{\bar{\lambda}'}$.

By symmetry, we may assume without loss of generality that $\bar{\lambda} \geq \bar{\lambda}'$. If $\bar{\lambda} = \bar{\lambda}' \leq \lambda_0$, then $\mathrm{HAgg}(g_1, \ldots, g_\Lambda) = g_{\bar{\lambda}} = g_{\bar{\lambda}'} = \mathrm{HAgg}(g'_1, \ldots, g'_\Lambda)$, and the claim of the lemma is immediate. If $\bar{\lambda} > \bar{\lambda}'$, then $\bar{\lambda}' < \Lambda$ and the definition of $\bar{\lambda}, \bar{\lambda}'$ gives that $|g'_{\bar{\lambda}'} - g'_{\bar{\lambda}'+1}| > 2\alpha_{\bar{\lambda}'} > 2\alpha_{\bar{\lambda}} \geq |g_{\bar{\lambda}'} - g_{\bar{\lambda}'+1}|$,

and in particular $|g'_{\bar{\lambda}'} - g'_{\bar{\lambda}'+1}| \neq |g_{\bar{\lambda}'} - g_{\bar{\lambda}'+1}|$. Hence $\lambda_0 \leq \bar{\lambda}'$. Thus, from here on, we may assume that $\lambda_0 \leq \bar{\lambda}' \leq \bar{\lambda}$.

Now note that

$$|g_{\bar{\lambda}} - g_{\lambda_0}| \leq \sum_{\lambda=\lambda_0}^{\bar{\lambda}-1} |g_\lambda - g_{\lambda+1}| \leq \sum_{\lambda=\lambda_0}^{\bar{\lambda}-1} 2\alpha_\lambda \leq 4\alpha_{\lambda_0}$$

and

$$|g'_{\bar{\lambda}'} - g'_{\lambda_0}| \leq \sum_{\lambda=\lambda_0}^{\bar{\lambda}'-1} |g'_\lambda - g'_{\lambda+1}| \leq \sum_{\lambda=\lambda_0}^{\bar{\lambda}'-1} 2\alpha_\lambda \leq 4\alpha_{\lambda_0}.$$

Using that $g_{\lambda_0} = g'_{\lambda_0}$, we get that $|\operatorname{HAgg}(g_1, \ldots, g_\Lambda) - \operatorname{HAgg}(g'_1, \ldots, g'_\Lambda)| = |g_{\bar{\lambda}} - g'_{\bar{\lambda}'}| \leq 8\alpha_{\lambda_0}$. □

For multisets $\mathcal{S}, \mathcal{S}'$ of tuples of the form $(\mathcal{G}^i, w^i)$ for $\mathcal{G}^i \subset \mathcal{F}$, $w^i \geq 0$, define

$$\Delta(\mathcal{S}, \mathcal{S}') = \frac{1}{2} \sum_{\mathcal{H} \subset \mathcal{F}} \left| \frac{\sum_{(\mathcal{G}^i, w^i) \in \mathcal{S}} w^i \cdot \mathbb{1}[\mathcal{G}^i = \mathcal{H}]}{\sum_{(\mathcal{G}^i, w^i) \in \mathcal{S}} w^i} - \frac{\sum_{(\mathcal{G}^{i\prime}, w^{i\prime}) \in \mathcal{S}'} w^{i\prime} \cdot \mathbb{1}[\mathcal{G}^{i\prime} = \mathcal{H}]}{\sum_{(\mathcal{G}^{i\prime}, w^{i\prime}) \in \mathcal{S}'} w^{i\prime}} \right|. \tag{13}$$

Note that $\Delta(\mathcal{S}, \mathcal{S}')$ is the total variation distance between the distributions on subclasses of $\mathcal{F}$ induced by the weights w_i and w'_i.

Lemma 5.7 shows a sensitivity-type result for the voting rule $\operatorname{Vote}_{\mathcal{T}}(\cdot)$ and for the truncated error $\operatorname{TErr}(\mathcal{T}, \cdot)$: if two weighted subclass collections $\mathcal{T}, \mathcal{T}'$ are such that their components $\mathcal{T}_\lambda, \mathcal{T}'_\lambda$ are close in the sense of (13) for each λ, then the voting rules and truncated errors for $\mathcal{T}, \mathcal{T}'$ are close.

Lemma 5.7. *Fix any $\lambda \in [\Lambda]$ and consider weighted subclass collections $\mathcal{T} = [\mathcal{G}, w], \mathcal{T}' = [\mathcal{G}', w']$. Then for any $x \in \mathcal{X}$,*

$$|\operatorname{Vote}_{\mathcal{T}}(x) - \operatorname{Vote}_{\mathcal{T}'}(x)| \leq 8 \cdot \sum_{\lambda=1}^{\Lambda} \alpha_\lambda \cdot \Delta(\mathcal{T}_\lambda, \mathcal{T}'_\lambda).$$

Moreover, for any pair $(x, y) \in \mathcal{X} \times [0, 1]$,

$$\left|\operatorname{TErr}(\mathcal{T}, x, y) - \operatorname{TErr}(\mathcal{T}', x, y)\right| \leq 8 \cdot \sum_{\lambda=1}^{\Lambda} \alpha_\lambda \cdot \Delta(\mathcal{T}_\lambda, \mathcal{T}'_\lambda). \tag{14}$$

Proof. First we define finite-support distributions Q_λ, Q'_λ, for each $\lambda \in [\Lambda]$, over the set of subclasses of $\mathcal{F}$, as follows: for $\lambda \in [\Lambda]$ and $\mathcal{H}_\lambda \subset \mathcal{F}$, define

$$Q_\lambda(\mathcal{H}_\lambda) := \frac{\sum_{v_\lambda=1}^{|\mathcal{T}_\lambda|} w_\lambda^{v_\lambda} \cdot \mathbb{1}[\mathcal{G}_\lambda^{v_\lambda} = \mathcal{H}_\lambda]}{\sum_{v_\lambda=1}^{|\mathcal{T}_\lambda|} w_\lambda^{v_\lambda}}, \qquad Q'_\lambda(\mathcal{H}_\lambda) := \frac{\sum_{v_\lambda=1}^{|\mathcal{T}'_\lambda|} w_\lambda^{v_\lambda\prime} \cdot \mathbb{1}[\mathcal{G}_\lambda^{v_\lambda\prime} = \mathcal{H}_\lambda]}{\sum_{v_\lambda=1}^{|\mathcal{T}'_\lambda|} w_\lambda^{v_\lambda\prime}}$$

By the definition (13), there is a coupling between Q_λ, Q'_λ, which we denote as $\tilde{Q}_\lambda((\mathcal{H}_\lambda, \mathcal{H}'_\lambda))$, so that

$$\sum_{(\mathcal{H}_\lambda, \mathcal{H}'_\lambda)} \tilde{Q}_\lambda((\mathcal{H}_\lambda, \mathcal{H}'_\lambda)) \cdot \mathbb{1}[\mathcal{H}_\lambda \neq \mathcal{H}'_\lambda] = \Delta(\mathcal{T}_\lambda, \mathcal{T}'_\lambda). \tag{15}$$

Next we define a distribution $\tilde{Q}$ over pairs of sequences $\mathcal{H}_{1:\Lambda}, \mathcal{H}'_{1:\Lambda}$ of subsets of $\mathcal{F}$, as follows: make independent draws $(\mathcal{H}_1, \mathcal{H}'_1) \sim \tilde{Q}_1, \ldots, (\mathcal{H}_\Lambda, \mathcal{H}'_\Lambda) \sim \tilde{Q}_\Lambda$. Now $\tilde{Q}$ is the distribution of the resulting pair $(\mathcal{H}_{1:\Lambda}, \mathcal{H}'_{1:\Lambda})$; this is equivalent (up to notational differences) to defining $\tilde{Q}$ as the product of the distributions $\tilde{Q}_1, \ldots, \tilde{Q}_\Lambda$. Note that the marginal (under $\tilde{Q}$) of any sequence $\mathcal{H}_{1:\Lambda}$ is simply $\sum_{v \in \mathcal{N}_\Lambda(\mathcal{T}): \mathcal{G}^v_{1:\Lambda} = \mathcal{H}_{1:\Lambda}} P_w(v)$; an analogous statement holds for the marginal of any sequence $\mathcal{H}'_{1:\Lambda}$.

Given two sequences $\mathcal{G}_{1:\Lambda}, \mathcal{G}'_{1:\Lambda}$ of subsets of $\mathcal{F}$, let $\lambda^\star(\mathcal{G}_{1:\Lambda}, \mathcal{G}'_{1:\Lambda})$ denote the minimum value of λ so that $\mathcal{G}_\lambda \not\equiv \mathcal{G}'_\lambda$ (and set $\lambda^\star(\mathcal{G}_{1:\Lambda}, \mathcal{G}'_{1:\Lambda}) = \infty$ if such λ does not exist). Now we can compute, for any $x \in \mathcal{X}$,

$$
\begin{aligned}
&|\operatorname{Vote}_{\mathcal{T}}(x) - \operatorname{Vote}_{\mathcal{T}'}(x)| \\
=&\left| \sum_{(\mathcal{H}_{1:\Lambda}, \mathcal{H}'_{1:\lambda})} \tilde{Q}(\mathcal{H}_{1:\Lambda}, \mathcal{H}'_{1:\Lambda}) \cdot (\operatorname{SOA}(\mathcal{H}_{1:\Lambda})(x) - \operatorname{SOA}(\mathcal{H}'_{1:\Lambda})(x)) \right| \\
\leq& \sum_{(\mathcal{H}_{1:\Lambda}, \mathcal{H}'_{1:\Lambda})} \tilde{Q}(\mathcal{H}_{1:\Lambda}, \mathcal{H}'_{1:\Lambda}) \cdot |\operatorname{SOA}(\mathcal{H}_{1:\Lambda})(x) - \operatorname{SOA}(\mathcal{H}'_{1:\Lambda})(x)| \\
\leq& \sum_{(\mathcal{H}_{1:\Lambda}, \mathcal{H}'_{1:\Lambda})} \tilde{Q}(\mathcal{H}_{1:\Lambda}, \mathcal{H}'_{1:\Lambda}) \cdot 8 \cdot \alpha_{\lambda^\star(\mathcal{H}_{1:\Lambda}, \mathcal{H}'_{1:\Lambda})} \qquad (16) \\
\leq& 8 \sum_{\lambda=1}^{\Lambda} \Delta(\mathcal{T}_\lambda, \mathcal{T}'_\lambda) \cdot \alpha_\lambda, \qquad (17)
\end{aligned}
$$

where (16) uses Lemma 5.6, and (17) uses the fact that by (15), the total mass under $\tilde{Q}$ of pairs $(\operatorname{SOA}(\mathcal{G}_{1:\Lambda}), \operatorname{SOA}(\mathcal{G}'_{1:\Lambda}))$ for which $\lambda^\star(\mathcal{G}_{1:\Lambda}, \mathcal{G}'_{1:\Lambda}) = \lambda$ (and thus $\operatorname{SOA}(\mathcal{G}_\lambda, \alpha_\lambda) \not\equiv \operatorname{SOA}(\mathcal{G}'_\lambda, \alpha_\lambda)$), is at most $\Delta(\mathcal{T}_\lambda, \mathcal{T}'_\lambda)$. To establish (14), we note that

$$
\begin{aligned}
&\left|\operatorname{TErr}(\mathcal{T}, x, y) - \operatorname{TErr}(\mathcal{T}', x, y)\right| \\
=&\left| \sum_{(\mathcal{H}_{1:\Lambda}, \mathcal{H}'_{1:\Lambda})} \tilde{Q}(\mathcal{H}_{1:\Lambda}, \mathcal{H}'_{1:\Lambda}) \cdot (\operatorname{TErr}(\mathcal{H}_{1:\Lambda}, x, y) - \operatorname{TErr}(\mathcal{H}'_{1:\Lambda}, x, y)) \right| \\
\leq& \sum_{(\mathcal{H}_{1:\Lambda}, \mathcal{H}'_{1:\Lambda})} \tilde{Q}(\mathcal{H}_{1:\Lambda}, \mathcal{H}'_{1:\Lambda}) \cdot \left| \max\{|\operatorname{SOA}(\mathcal{H}_{1:\Lambda})(x) - y|, \alpha_{\bar{\lambda}(\mathcal{H}_{1:\Lambda}, x)}\} - \max\{|\operatorname{SOA}(\mathcal{H}'_{1:\Lambda})(x) - y|, \alpha_{\bar{\lambda}(\mathcal{H}'_{1:\Lambda}, x)}\} \right| \\
\leq& \sum_{(\mathcal{H}_{1:\Lambda}, \mathcal{H}'_{1:\Lambda})} \tilde{Q}(\mathcal{H}_{1:\Lambda}, \mathcal{H}'_{1:\Lambda}) \cdot \left(\max\{|\operatorname{SOA}(\mathcal{H}_{1:\Lambda})(x) - \operatorname{SOA}(\mathcal{H}'_{1:\Lambda})(x)|, |\alpha_{\bar{\lambda}(\mathcal{H}_{1:\Lambda}, x)} - \alpha_{\bar{\lambda}(\mathcal{H}'_{1:\Lambda}, x)}|\} \right) \\
\leq& \sum_{(\mathcal{H}_{1:\Lambda}, \mathcal{H}'_{1:\Lambda})} \tilde{Q}(\mathcal{H}_{1:\Lambda}, \mathcal{H}'_{1:\Lambda}) \cdot \max\{8 \cdot \alpha_{\lambda^\star(\mathcal{H}_{1:\Lambda}, \mathcal{H}'_{1:\Lambda})}, 2 \cdot \alpha_{\lambda^\star(\mathcal{H}_{1:\Lambda}, \mathcal{H}'_{1:\Lambda})}\} \\
\leq& 8 \sum_{\lambda=1}^{\Lambda} \Delta(\mathcal{T}_\lambda, \mathcal{T}'_\lambda) \cdot \alpha_\lambda.
\end{aligned}
$$

□

5.2 The proper learning algorithm

Our proper (and stable) learning algorithm, `Multi-scale Proper Learner`, is presented in Algorithm 2. The algorithm maintains a weighted subclass collection $\mathcal{T}$; for each $n \geq 1$, let $\mathcal{T}^n =$

$(\mathcal{T}_1^n, \ldots, \mathcal{T}_\Lambda^n)$ denote this weighted subclass collection $\mathcal{T}$ directly after round n of the outer while loop. For each $n \geq 1$, let $t(n)$ denote the value of t maintained by the algorithm at the beginning of round n. If the round number n is clear, we will often drop the superscript n. The algorithm repeatedly performs the following process: at each round $n \geq 1$ of the outer while loop, it first checks if there is some randomized predictor $\bar{f}$ satisfying the condition in step 2a. If so, then it sets $\bar{f}_t = \bar{f}$, i.e., it uses this predictor $\bar{f}$ to predict the next example (x_t, y_t) (step 2(a)i); note that there $t = t(n)$, so $(x_t, y_t) = (x_{t(n)}, y_{t(n)})$. The algorithm then uses (x_t, y_t) to update $\mathcal{T}$ in step 2(a)ii, replacing various classes $\mathcal{G}_\lambda^{v_\lambda}$ in the weighted subclass collection which *incorrectly* predict (x_t, y_t) with appropriate restrictions at the scale α_λ, and downweighting the corresponding weights $w_\lambda^{v_\lambda}$.

The more challenging case is when the condition in step 2a is not satisfied; in this case, it turns out that by the minimax theorem (Lemma 5.8), we can find a dataset at each scale λ which satisfies a property (step 2b of `Multi-scale Proper Learner`) which is roughly "dual" to the property of step 2a. This property guarantees that we can perform further restrictions on the subclasses $\mathcal{G}_\lambda^{v_\lambda}$ (and decrease the weights $w_\lambda^{v_\lambda}$ accordingly); it turns out that after a bounded number steps of doing so, the property in step 2a will be satisfied, and we can process the next example (x_{t+1}, y_{t+1}).

We next introduce some further notation. For each $n \geq 1$, and $\lambda \in [\Lambda]$, let $W_{\lambda,n}$ denote the total of all weights in $\mathcal{T}_\lambda$ after round n, i.e., $W_{\lambda,n} = \sum_{(\mathcal{G}_\lambda^{v_\lambda}, w_\lambda^{v_\lambda}) \in \mathcal{T}_\lambda^n} w_\lambda^{v_\lambda}$. Further set $W_{\lambda,0} = 1$ for all $\lambda \in [\Lambda]$. The quantities $W_{\lambda,n}$, for $\lambda \in [\Lambda]$, will be used as a potential function to track the progress of `Multi-scale Proper Learner` over rounds n.

Finally, given the class $\mathcal{F} \subset [0,1]^{\mathcal{X}}$, denote by $\mathcal{F}_{\text{abs}} \subset [0,1]^{\mathcal{X} \times [0,1]}$ the "absolute loss class" of $\mathcal{F}$, namely $\{(x,y) \mapsto |f(x) - y| \ : \ f \in \mathcal{F}\}$. We will also need to consider the dual class of $\mathcal{F}_{\text{abs}}$, denoted $\mathcal{F}_{\text{abs}}^\star$:

$$\mathcal{F}_{\text{abs}}^\star := \{h : \mathcal{F} \to [0,1] : \quad \exists (x,y) \in \mathcal{X} \times [0,1], \text{ such that } h(f) = |f(x) - y| \ \forall f \in \mathcal{F}\}.$$

By Lemma A.6 with $k = 1$, $\mathcal{Z} = \mathcal{X} \times [0,1]$, and $\phi(a, (x,y)) = |a - y|$ (which is 1-Lipschitz), we have that $\text{sfat}_\alpha(\mathcal{F}_{\text{abs}}) \leq O(\text{sfat}_\alpha(\mathcal{F}) \cdot \log(1/\alpha))$ for all $\alpha > 0$. Thus $\text{sfat}_\alpha(\mathcal{F}_{\text{abs}}) < \infty$ for all $\alpha > 0$, meaning that (by Lemma 8.4), $\text{sfat}_\alpha(\mathcal{F}_{\text{abs}}^\star) < \infty$ for all $\alpha > 0$. Let $\text{fat}_\alpha(\cdot)$ denote the (non-sequential) fat-shattering dimension (see Section A for the definition). Since $\text{sfat}_\alpha(\mathcal{G}) \geq \text{fat}_\alpha(\mathcal{G})$ holds for any class $\mathcal{G}$, we get that $\text{fat}_\alpha(\mathcal{F}_{\text{abs}}) < \infty$ and $\text{fat}_\alpha(\mathcal{F}_{\text{abs}}^\star) < \infty$ for all $\alpha > 0$.

At some points in our proof we will need to use basic uniform convergence properties for the class $\mathcal{F}_{\text{abs}}, \mathcal{F}_{\text{abs}}^\star$; for this we make the following definitions:

$$V := \frac{10C_0 \cdot \text{fat}_{c_0\alpha_\Lambda/10}(\mathcal{F}_{\text{abs}}) \log(10/\alpha_\Lambda)}{\alpha_\Lambda^2}, \qquad V^\star := \frac{10C_0 \cdot \text{fat}_{c_0\alpha_\Lambda/10}(\mathcal{F}_{\text{abs}}^\star) \log(10/\alpha_\Lambda)}{\alpha_\Lambda^2}, \tag{18}$$

$$m_\lambda := \left\lceil \frac{C_1 V}{\alpha_\lambda} \right\rceil \quad \forall \lambda \in [\Lambda], \tag{19}$$

where C_0, c_0 are the constants of Theorem A.3, and $C_1 > 0$ is a sufficiently large constant. As the parameters $V, V^\star$ will not show up in our rates for regret, we do not attempt to optimize their dependence on any of the relevant parameters.

Finally, set

$$\mu_0 = \min_{\lambda \in [\Lambda]} \alpha_\lambda m_\lambda, \qquad \mu_1 = \max_{\lambda \in [\Lambda]} \alpha_\lambda m_\lambda. \tag{20}$$

As long as the constant C_1 is sufficiently large, we have that $\mu_1/\mu_0 \leq 2$.

Recall from Algorithm 2 that we set $C_\Lambda = 3072\Lambda^3$. The below lemma uses the minimax theorem to show that if the condition in step 2a of `Multi-scale Proper Learner` fails at some step, then the condition in step 2b does not fail; thus, this lemma establishes that the algorithm is well-defined, i.e., can run as claimed.

Lemma 5.8. *Suppose the condition in step 2a of `Multi-scale Proper Learner` (Algorithm 2) fails at some time step, i.e., there is no $\bar{f} \in \mathrm{Rand}(\mathcal{F}^{V^\star})$ so that for each $\lambda \in [\Lambda - 6]$,*

$$\sup_{(x,y)\in \mathrm{HighVote}(\mathcal{T},\alpha_\lambda/C_\Lambda)} \mathbb{E}_{f\sim\bar{f}}\left[|f(x)-y|\right] \le \alpha_\lambda.$$

Then for every $\lambda \in [\Lambda]$ there is a collection of tuples $(\tilde{x}_1^\lambda, \tilde{y}_1^\lambda), \ldots, (\tilde{x}_{m_\lambda}^\lambda, \tilde{y}_{m_\lambda}^\lambda) \in \mathrm{HighVote}(\mathcal{T}, \alpha_\lambda/C_\Lambda)$, so that for every $f \in \mathcal{F}$, there are some $\lambda \in [\Lambda]$, $\lambda' \in [\Lambda - 3]$ so that $\frac{1}{m_\lambda}\sum_{j=1}^{m_\lambda} \mathbb{1}[|f(\tilde{x}_j) - \tilde{y}_j| > \alpha_{\lambda'}] > \frac{\alpha_\lambda}{16\Lambda\alpha_{\lambda'}}$.

Proof. Fix some weighted subclass collection $\mathcal{T}$ so that there is no $\bar{f} \in \mathrm{Rand}(\mathcal{F}^{V^\star})$ so that for each $\lambda \in [\Lambda - 6]$, $\sup_{(x,y)\in \mathrm{HighVote}(\mathcal{T},\alpha_\lambda/C_\Lambda)} \mathbb{E}_{f\sim\bar{f}}\left[|f(x)-y|\right] \le \alpha_\lambda$. By Theorem A.3 applied to the class $\mathcal{F}^\star$ and by the definition of $V^\star$ in (18), for every finite support measure Q on $\mathcal{F}$, there is some $\lambda \in [\Lambda - 6]$ so that

$$\sup_{(x,y)\in \mathrm{HighVote}(\mathcal{T},\alpha_\lambda/C_\Lambda)} \mathbb{E}_{f\sim Q}\left[|f(x)-y|\right] > 2\alpha_\lambda/3.$$

For each $(x, y) \in \mathcal{X} \times [0, 1]$, let $\alpha(x, y)$ be the smallest value of α_λ (for $\lambda \le \Lambda - 6$) for which $(x, y) \in \mathrm{HighVote}(\mathcal{T}, \alpha_\lambda/C_\Lambda)$.[12] Hence for every finite support measure Q on $\mathcal{F}$, there is some $(x, y) \in \mathcal{X} \times [0, 1]$ so that $\mathbb{E}_{f\sim Q}[|f(x) - y|] > \alpha(x, y)/2$. Now consider the function class $\mathcal{G} \subset \mathbb{R}^{\mathcal{F}}$, defined by

$$\mathcal{G} := \left\{ f \mapsto \frac{|f(x)-y|}{\alpha(x,y)} \;:\; (x,y) \in \mathcal{X}\times[0,1] \right\}.$$

Note that there are a finite number of possible values of $\alpha(x, y)$, namely α_λ for $\Lambda - 6 \ge \lambda \ge -O(\log(C_\Lambda))$. For each such possible value of λ, define

$$\mathcal{G}_\lambda := \left\{ f \mapsto \frac{|f(x)-y|}{\alpha_\lambda} \;:\; \alpha(x,y) = \alpha_\lambda \right\}.$$

Note that $\mathcal{G} = \bigcup_{\Lambda-6\ge\lambda\ge -O(\log C_\Lambda)} \mathcal{G}_\lambda$. Further, since each $\mathcal{G}_\lambda$ is simply a subclass of $\mathcal{F}^\star_{\mathrm{abs}}$ scaled by $1/\alpha_\lambda$, it holds that $\mathrm{sfat}_\alpha(\mathcal{G}_\lambda) \le \mathrm{sfat}_{\alpha\alpha_\lambda}(\mathcal{F}^\star_{\mathrm{abs}}) < \infty$ for all $\alpha > 0$. Thus, by Corollary A.7, $\mathrm{sfat}_\alpha(\mathcal{G}) < \infty$ for all $\alpha > 0$. Then by Theorem 8.8,

$$\begin{aligned} 1/2 &\le \inf_{Q\in\Delta^\circ(\mathcal{F})} \sup_{P\in\Delta^\circ(\mathcal{X}\times[0,1])} \mathbb{E}_{f\sim Q,(x,y)\sim P}\left[\frac{|f(x)-y|}{\alpha(x,y)}\right] \\ &= \sup_{P\in\Delta^\circ(\mathcal{X}\times[0,1])} \inf_{Q\in\Delta^\circ(\mathcal{F})} \mathbb{E}_{f\sim Q,(x,y)\sim P}\left[\frac{|f(x)-y|}{\alpha(x,y)}\right]. \end{aligned}$$

Thus we may find a finite-support measure $P^\star \in \Delta^\circ(\mathcal{X} \times [0, 1])$ so that for every $f \in \mathcal{F}$,

$$\mathbb{E}_{(x,y)\sim P^\star}\left[\frac{|f(x)-y|}{\alpha(x,y)}\right] > 1/3. \tag{21}$$

[12] Note that it could be the case that $\alpha(x, y) \ge 1$, i.e., $\lambda \le 0$.

For each λ satisfying $2 \le \lambda \le \Lambda - 6$, let $P^\star_\lambda \in \Delta^\circ(\mathcal{X} \times [0,1])$ be the distribution of $(x,y) \sim P^\star$, conditioned on $\alpha(x,y) = \alpha_\lambda$. Further, for the case $\lambda = 1$, let $P^\star_1$ be the distribution of $(x,y) \sim P^\star$, conditioned on $\alpha(x,y) \le \alpha_1$. (If $P^\star\{(x,y) : \alpha(x,y) = \alpha_\lambda\} = 0$ for some λ, then let $P^\star_\lambda$ be an arbitrary finite-support distribution on $\mathcal{X} \times [0,1]$.) By Theorem A.3 with the function class given by $\mathcal{F}_{\mathrm{abs}}$ and by definition of V in (18) and of m_λ in (19) for each λ, we have the following: for each $\lambda \in [\Lambda - 6]$, there is a dataset $S^\lambda := \{(\tilde{x}^\lambda_1, y^\lambda_1), \dots, (\tilde{x}^\lambda_{m_\lambda}, \tilde{y}^\lambda_{m_\lambda})\}$ of size m_λ so that for any $f \in \mathcal{F}$ satisfying $\mathbb{E}_{(x,y)\sim P^\star_\lambda}[|f(x) - y|] > \alpha_\lambda/3$, we have $\frac{1}{m_\lambda}\sum_{j=1}^{m_\lambda} |f(\tilde{x}^\lambda_j) - \tilde{y}^\lambda_j| > \alpha_\lambda/4$.

But by (21), we see that for every $f \in \mathcal{F}$, there is some $\lambda \in [\Lambda - 6]$ so that $\mathbb{E}_{(x,y)\sim P^\star_\lambda}[|f(x) - y|] > \alpha_\lambda/3$. Hence, for any $f \in \mathcal{F}$, there is some $\lambda \in [\Lambda - 6]$ so that $\frac{1}{m_\lambda}\sum_{j=1}^{m_\lambda} |f(\tilde{x}^\lambda_j) - \tilde{y}^\lambda_j| > \alpha_\lambda/4$, which implies, by Lemma 5.5 with the value of Λ set to $\Lambda - 3$ (using that $\alpha_\lambda/4 \ge 2\alpha_{\Lambda-3}$ since $\lambda \le \Lambda - 6$), that for some $\lambda' \in [\Lambda - 3]$, $\frac{1}{m_\lambda}\sum_{j=1}^{m_\lambda} \mathbb{1}[|f(\tilde{x}^\lambda_j) - \tilde{y}^\lambda_j| > \alpha_{\lambda'}] > \frac{\alpha_\lambda}{16\Lambda\alpha_{\lambda'}}$. (Here we have used that $|f(\tilde{x}^\lambda_j) - \tilde{y}^\lambda_j| \le 1$ for all j, λ.) □

Lemma 5.9 shows that the potentials $W_{\lambda,n}$ (for $\lambda \in [\Lambda]$) decrease in each round n of `Multi-scale Proper Learner` if the condition in step 2a succeeds in round n.

Lemma 5.9. *Consider any round n in the algorithm for which the condition in step 2a of `Multi-scale Proper Learner` (Algorithm 2) holds. Let $t = t(n)$ be the value of t at step 2(a)ii. If $\delta_t > 32\alpha_\Lambda$, then it holds that for some $\lambda \in [\Lambda]$, $W_{\lambda,n} \le W_{\lambda,n-1} \cdot \left(1 - \frac{\delta_t}{64\Lambda\alpha_\lambda}\right)$. Further, for all $\lambda' \in [\Lambda]$, $W_{\lambda',n} \le W_{\lambda',n-1}$.*

Proof. Recall that $\delta_t = \mathrm{TErr}(\mathcal{T}, x_t, y_t)$, where $\mathcal{T} = [\mathcal{G}, w]$ is the weighted subclass collection maintained by `Multi-scale Proper Learner` at the beginning of round n (equivalently, at the end of round $n-1$). By Lemma 5.2, there is some $\lambda \in [\Lambda]$ so that

$$\alpha_\lambda \cdot \frac{\sum_{v_\lambda=1}^{|\mathcal{T}_\lambda|} w^{v_\lambda}_\lambda \cdot \mathbb{1}[|\operatorname{SOA}(\mathcal{G}^{v_\lambda}_\lambda, \alpha_\lambda)(x) - y| > \alpha_\lambda]}{W_{\lambda,n-1}} \ge \frac{1}{\Lambda}\left(\frac{1}{16}\delta_t - \alpha_\Lambda\right) \ge \frac{\delta_t}{32\Lambda}.$$

Since $\gamma \le 1/2$, it follows that

$$W_{\lambda,n} \le W_{\lambda,n-1} \cdot \left(1 - \frac{\delta_t}{32\Lambda\alpha_\lambda}\right) + W_{\lambda,n-1} \cdot \frac{\delta_t}{32\Lambda\alpha_\lambda} \cdot \gamma \le W_{\lambda,n-1} \cdot \left(1 - \frac{\delta_t}{64\Lambda\alpha_\lambda}\right).$$

The fact that $W_{\lambda',n} \le W_{\lambda',n-1}$ for all $\lambda' \in [\Lambda]$ is immediate. □

Complementing the previous Lemma 5.9, Lemma 5.10 shows that the potential $W_{\lambda,n}$ decreases in round n of `Multi-scale Proper Learner` if the condition in step 2a fails in round n.

Lemma 5.10. *Consider any round n for which the condition in step 2a of `Multi-scale Proper Learner` (Algorithm 2) fails. For all $\lambda' \in [\Lambda]$, it holds that*

$$W_{\lambda',n} \le W_{\lambda',n-1} \cdot \left(A + \frac{3\mu_1\Lambda}{\alpha_{\lambda'} \cdot C_\Lambda}\right).$$

Further, for any $\lambda' \in [\Lambda]$, letting $w^{v_{\lambda'},j}_{\lambda',\lambda}$ denote the weights constructed in step 2(b)iA at round n, we have

$$\sum_{(\lambda,j,b,v_{\lambda'}):\mathcal{G}^{v_{\lambda'},j,b}_{\lambda',\lambda} \neq \emptyset} w^{v_{\lambda'},j}_{\lambda',\lambda} \le \frac{3\mu_1\Lambda}{\alpha_{\lambda'} \cdot C_\Lambda} \cdot W_{\lambda',n-1}, \tag{22}$$

where the summation is over $\lambda \in [\Lambda], j \in [m_\lambda], 0 \le b \le \lfloor 1/\alpha_{\lambda'} \rfloor + 1$, and $v_{\lambda'} \in [|\mathcal{T}^{n-1}_{\lambda'}|]$.

Algorithm 2: `Multi-scale Proper Learner`

Input: Function class $\mathcal{F} \subset [0,1]^{\mathcal{X}}$, time horizon $T \in \mathbb{N}$, scale parameter $\Lambda \in \mathbb{N}$, constants $C_1, C_2 > 0$, $C_\Lambda = 3072\Lambda^3$.

1. For $\lambda \in [\Lambda]$, initialize $\mathcal{T} = [\mathcal{G}, w]$ to be the weighted subclass collection with $\mathcal{T}_\lambda = \{(\mathcal{F}, 1)\}$ for each λ (i.e., $\mathcal{G}_\lambda^1 = \mathcal{F}$ and $w_\lambda^1 = 1$ for all $\lambda \in [\Lambda]$). Set $\gamma := \frac{\alpha_\Lambda}{C_\Lambda}$, $m_\lambda := \left\lceil \frac{C_1 V}{\alpha_\lambda} \right\rceil$, $A = \frac{\mu_1}{\alpha_\Lambda}$, $t \leftarrow 1$ *(recall definitions of V in (18) and μ_1 in (20)).*

2. While $t \leq T$:

   (a) If, there is $\bar{f} \in \mathrm{Rand}(\mathcal{F}^{V^\star})$ so that, for each $\lambda \in [\Lambda - 6]$, $\sup_{(x,y) \in \mathrm{HighVote}(\mathcal{T}, \frac{\alpha_\lambda}{C_\Lambda})} \mathbb{E}_{f \sim \bar{f}} [|f(x) - y|] \leq \alpha_\lambda$: *(recall definition of $V^\star$ in (18))*

   i. Choose $\bar{f}_t = \bar{f}$. On the next example x_t, draw $f \sim \bar{f}_t$ and predict $f(x_t)$.

   ii. Receive y_t, and let $\delta_t := \mathrm{TErr}(\mathcal{T}, x_t, y_t)$.

   - For each $1 \leq \lambda \leq \Lambda$, and each $v_\lambda \in [|\mathcal{T}_\lambda|]$:

   A. If $|\mathrm{SOA}(\mathcal{G}_\lambda^{v_\lambda}, \alpha_\lambda)(x_t) - y_t| > \alpha_\lambda$, set $w_\lambda^{v_\lambda} \leftarrow \gamma \cdot w_\lambda^{v_\lambda}$.

   B. If $|\mathrm{SOA}(\mathcal{G}_\lambda^{v_\lambda}, \alpha_\lambda)(x_t) - y_t| > \alpha_\lambda$, set $\mathcal{G}_\lambda^{v_\lambda} \leftarrow \mathcal{G}_\lambda^{v_\lambda}|_{(x_t, y_t)}^{\alpha_\lambda}$.

   C. If $\mathcal{G}_\lambda^{v_\lambda} = \emptyset$, remove $(\mathcal{G}_\lambda^{v_\lambda}, w_\lambda^{v_\lambda})$ from $\mathcal{T}$.

   iii. Set $t \leftarrow t + 1$.

   (b) Else, choose $\{(\tilde{x}_1^\lambda, \tilde{y}_1^\lambda), \ldots, (\tilde{x}_{m_\lambda}^\lambda, \tilde{y}_{m_\lambda}^\lambda)\} \subset \mathrm{HighVote}(\mathcal{T}, \frac{\alpha_\lambda}{C_\Lambda})$ for all $\lambda \in [\Lambda]$ so as to satisfy the following property: for every $f \in \mathcal{F}$, there are some $\lambda \in [\Lambda], \lambda' \in [\Lambda - 3]$ so that $\frac{1}{m_\lambda} \sum_{j=1}^{m_\lambda} \mathbb{1}[|f(\tilde{x}_j^\lambda) - \tilde{y}_j^\lambda| > \alpha_{\lambda'}] > \frac{\alpha_\lambda}{16\Lambda\alpha_{\lambda'}}$:

   i. For each $\lambda, \lambda' \in [\Lambda]$:

   - For each $v_{\lambda'} \in [|\mathcal{T}_{\lambda'}|]$, and each $j \in [m_\lambda]$:

   A. If $|\mathrm{SOA}(\mathcal{G}_{\lambda'}^{v_{\lambda'}}, \alpha_{\lambda'})(\tilde{x}_j^\lambda) - \tilde{y}_j^\lambda| \leq 5\alpha_{\lambda'}$, set

$$w_{\lambda',\lambda}^{v_{\lambda'},j} \leftarrow \gamma \cdot w_{\lambda'}^{v_{\lambda'}},$$

$$\mathcal{G}_{\lambda',\lambda}^{v_{\lambda'},j,b} \leftarrow \begin{cases} \mathcal{G}_{\lambda'}^{v_{\lambda'}}|_{(\tilde{x}_j^\lambda, b\alpha_{\lambda'})}^{\alpha_{\lambda'}} & : \quad |b\alpha_{\lambda'} - \tilde{y}_j^\lambda| > 6\alpha_{\lambda'} \\ \emptyset & : \quad |b\alpha_{\lambda'} - \tilde{y}_j^\lambda| \leq 6\alpha_{\lambda'}, \end{cases} \qquad \forall 0 \leq b \leq \lfloor 1/\alpha_{\lambda'} \rfloor + 1.$$

   B. Otherwise, set

$$w_{\lambda',\lambda}^{v_{\lambda'},j} \leftarrow w_{\lambda'}^{v_{\lambda'}}$$
$$\mathcal{G}_{\lambda',\lambda}^{v_{\lambda'},j,0} \leftarrow \mathcal{G}_{\lambda'}^{v_{\lambda'}}, \qquad \mathcal{G}_{\lambda',\lambda}^{v_{\lambda'},j,b} \leftarrow \emptyset \quad \forall 1 \leq b \leq \lfloor 1/\alpha_{\lambda'} \rfloor + 1.$$

   ii. For $\lambda' \in [\Lambda]$, set

$$\mathcal{T}_{\lambda'} \leftarrow A \cdot \mathcal{T}_{\lambda'} \cup \bigcup_{\lambda \in [\Lambda]} \{(\mathcal{G}_{\lambda',\lambda}^{v_{\lambda'},j,b}, w_{\lambda',\lambda}^{v_{\lambda'},j}) : \ j \in [m_\lambda], \ b \leq \left\lfloor \frac{1}{\alpha_{\lambda'}} \right\rfloor + 1, \ v_{\lambda'} \in [|\mathcal{T}_{\lambda'}|], \ \mathcal{G}_{\lambda',\lambda}^{v_{\lambda'},j,b} \neq \emptyset\}.$$

Proof. Let $\mathcal{T} = \mathcal{T}^{n-1} = [\mathcal{G}, w]$ denote the weighted subclass collection at the end of round $n-1$ (i.e., at the beginning of round n). Consider the datasets $\{(\tilde{x}_1^\lambda, \tilde{y}_1^\lambda), \dots, (\tilde{x}_{m_\lambda}^\lambda, \tilde{y}_{m_\lambda}^\lambda)\} \subset \mathrm{HighVote}(\mathcal{T}, \alpha_\lambda / C_\Lambda)$ constructed in step 2b. Fix any $\lambda, \lambda' \in [\Lambda]$, and $j \in [m_\lambda]$. Since $(\tilde{x}_j^\lambda, \tilde{y}_j^\lambda) \in \mathrm{HighVote}(\mathcal{T}, \frac{\alpha_\lambda}{C_\Lambda})$, by Lemma 5.4 with $\alpha = \alpha_\lambda / C_\Lambda$, at most a fraction $\frac{\alpha_\lambda}{\alpha_{\lambda'} \cdot C_\Lambda}$ of the weight $w_{\lambda'}^{v_{\lambda'}}$, for $1 \le v_{\lambda'} \le |\mathcal{T}_{\lambda'}|$, satisfies $\mathbb{1}[|\,\mathrm{SOA}(\mathcal{G}_{\lambda'}^{v_{\lambda'}}, \alpha_{\lambda'})(\tilde{x}_j^\lambda) - \tilde{y}_j^\lambda| > 5\alpha_{\lambda'}$. Let $\mathcal{S}_{\lambda,0}$ be the set of such indices $v_{\lambda'}$, and let $\mathcal{S}_{\lambda,1} = [|\mathcal{T}_{\lambda'}|] \backslash \mathcal{S}_{\lambda,0}$. Thus, letting $w_{\lambda',\lambda}^{v_{\lambda'},j}$ denote the weights constructed in step 2(b)iA at round n,

$$\sum_{v_{\lambda'} \in \mathcal{S}_{\lambda,0}} w_{\lambda',\lambda}^{v_{\lambda'},j} + \frac{2}{\alpha_{\lambda'}} \sum_{v_{\lambda'} \in \mathcal{S}_{\lambda,1}} w_{\lambda',\lambda}^{v_{\lambda'},j} \le \left(\frac{\alpha_\lambda}{\alpha_{\lambda'} \cdot C_\Lambda} + \frac{2\gamma}{\alpha_{\lambda'}} \right) \cdot W_{\lambda',n-1} \le \frac{3\alpha_\lambda}{\alpha_{\lambda'} \cdot C_\Lambda} \cdot W_{\lambda',n-1}. \tag{23}$$

Using that $\lfloor 1/\alpha_{\lambda'} \rfloor + 2 \le 2/\alpha_{\lambda'}$ for each $\lambda' \in [\Lambda]$, it follows that

$$\begin{aligned} W_{\lambda',n} \le & A \cdot W_{\lambda',n-1} + 2 \sum_{\lambda=1}^{\Lambda} \sum_{j=1}^{m_\lambda} \left(\sum_{v_{\lambda'} \in \mathcal{S}_{\lambda,0}} w_{\lambda',\lambda}^{v_{\lambda'},j} + \frac{2}{\alpha_{\lambda'}} \sum_{v_{\lambda'} \in \mathcal{S}_{\lambda,1}} w_{\lambda',\lambda}^{v_{\lambda'},j} \right) \\ \le & W_{\lambda',n-1} \cdot \left(A + \sum_{\lambda=1}^{\Lambda} \frac{3\alpha_\lambda m_\lambda}{\alpha_{\lambda'} \cdot C_\Lambda} \right) \\ \le & W_{\lambda',n-1} \cdot \left(A + \frac{3\mu_1 \Lambda}{\alpha_{\lambda'} \cdot C_\Lambda} \right), \end{aligned}$$

as desired. The second claim (22) of the lemma follows in an identical manner, except with the leading term $W_{\lambda',n-1} \cdot A$ above deleted. ☐

Lemma 5.11 forms a crucial part of the cumulative loss bound proof for `Multi-scale Proper Learner`. For each step $n+1$ of `Multi-scale Proper Learner` for which the condition in step 2a holds, Lemma 5.11 upper bounds the expected error $\mathbb{E}_{f \sim \bar{f}_{t(n+1)}} \left[|\bar{f}_{t(n+1)}(x) - y| \right]$ of $\bar{f}_{t(n+1)}$ on any point (x, y) as the sum of 3 terms: the third term is the truncated error of $\mathcal{T}$ on (x, y) at some later step (namely, step $n + n_0 - 1$ for some $n_0 \ge 1$), and the first two terms depend on the behavior of the algorithm between steps n and $n + n_0 - 1$ (in particular, the first two terms are 0 if $n_0 = 0$, i.e., $n = n + n_0 - 1$).

Lemma 5.11. *Fix integers $n \ge 0$ and $n_0 \ge 1$. Let $\mathcal{S}_0 \subset \{n+1, n+2, \dots, n+n_0-1\}$ be the subset of rounds n in which the condition on step 2a of `Multi-scale Proper Learner` (Algorithm 2) holds, and $\mathcal{S}_1 = \{n+1, n+2, \dots, n+n_0-1\} \backslash \mathcal{S}_0$. Suppose further the condition on step 2a holds at step $n+1$. For any point $(x, y) \in \mathcal{X} \times \mathcal{Y}$, it holds that*

$$\mathbb{E}_{f \sim \bar{f}_{t(n+1)}} \left[|\bar{f}(x) - y| \right] \le \sum_{n' :\ n+n' \in \mathcal{S}_0} 80 \cdot C_\Lambda \sum_{\lambda=1}^{\Lambda} \alpha_\lambda \cdot \ln \left(\frac{W_{\lambda,n+n'-1}}{W_{\lambda,n+n'}} \right) + \frac{|\mathcal{S}_1|}{A} \cdot 150 C_\Lambda \mu_1 \Lambda^2 + 5 \cdot C_\Lambda \cdot \mathrm{TErr}(\mathcal{T}^{n+n_0-1}, x, y). \tag{24}$$

In particular, for any round $n+1$ on which the condition in step 2a holds, $\mathcal{S}_0 = \mathcal{S}_1 = \emptyset$ and so

$$\mathbb{E}_{f \sim \bar{f}_{t(n+1)}} \left[|\bar{f}(x) - y| \right] \le 2 \cdot C_\Lambda \cdot \mathrm{TErr}(\mathcal{T}^n, x, y).$$

Proof. Fix any $1 \le n' \le n_0 - 1$. We consider two cases:

- The condition in 2a holds at round $n+n'$, i.e., $n+n' \in \mathcal{S}_0$. For $\lambda \in [\Lambda]$, write

$$\zeta_\lambda := \frac{\sum_{(\mathcal{G}_\lambda^{v_\lambda}, w_\lambda^{v_\lambda}) \in \mathcal{T}_\lambda^{n+n'-1}} w_\lambda^{v_\lambda} \cdot \mathbb{1}[|\operatorname{SOA}(\mathcal{G}_\lambda^{v_\lambda}, \alpha_\lambda)(x_{t(n+n')}) - y_{t(n+n')}| > \alpha_\lambda]}{\sum_{(\mathcal{G}_\lambda^{v_\lambda}, w_\lambda^{v_\lambda}) \in \mathcal{T}_\lambda^{n+n'-1}} w_\lambda^{v_\lambda}}.$$

  Then by the update in step 2(a)iiA,

$$\frac{W_{\lambda,n+n'}}{W_{\lambda,n+n'-1}} \le 1 - \zeta_\lambda \cdot (1-\gamma) \le 1 - \zeta_\lambda/2 \le \exp(-\zeta_\lambda/2).$$

  Therefore,

$$\Delta(\mathcal{T}_\lambda^{n+n'}, \mathcal{T}_\lambda^{n+n'-1}) \le \zeta_\lambda \le 2 \ln\left(\frac{W_{\lambda,n+n'-1}}{W_{\lambda,n+n'}}\right).$$

  Therefore, by Lemma 5.7, it follows that for any $x \in \mathcal{X}$,

$$\left|\operatorname{Vote}_{\mathcal{T}^{n+n'}}(x) - \operatorname{Vote}_{\mathcal{T}^{n+n'-1}}(x)\right| \le 16 \sum_{\lambda=1}^{\Lambda} \alpha_\lambda \cdot \ln\left(\frac{W_{\lambda,n+n'-1}}{W_{\lambda,n+n'}}\right), \tag{25}$$

  and for any $(x,y) \in \mathcal{X} \times [0,1]$,

$$|\operatorname{TErr}(\mathcal{T}^{n+n'}, x, y) - \operatorname{TErr}(\mathcal{T}^{n+n'-1}, x, y)| \le 16 \sum_{\lambda=1}^{\Lambda} \alpha_\lambda \cdot \ln\left(\frac{W_{\lambda,n+n'-1}}{W_{\lambda,n+n'}}\right). \tag{26}$$

- The condition in step 2a does not hold at round $n+n'$, i.e., $n+n' \in \mathcal{S}_1$. By (22) of Lemma 5.10, for any $\lambda, \lambda' \in [\Lambda]$,

$$\sum_{(\lambda,j,b,v_{\lambda'}): \mathcal{G}_{\lambda',\lambda}^{v_{\lambda'},j,b} \neq \emptyset} w_{\lambda',\lambda}^{v_{\lambda'},j} \le \frac{3\mu_1\Lambda}{\alpha_{\lambda'} \cdot C_\Lambda} \cdot W_{\lambda',n+n'-1} \le \frac{3\mu_1\Lambda}{\alpha_{\lambda'} \cdot C_\Lambda} \cdot \frac{W_{\lambda',n+n'}}{A},$$

  where the summation on the left-hand side is over $\lambda \in [\Lambda], j \in [m_\lambda], 0 \le b \le \lfloor 1/\alpha_{\lambda'} \rfloor + 1$, and $v_{\lambda'} \in [|\mathcal{T}_{\lambda'}^{n-1}|]$. Then for each $\lambda' \in [\Lambda]$,

$$\begin{aligned}\Delta(\mathcal{T}_{\lambda'}^{n+n'}, \mathcal{T}_{\lambda'}^{n+n'-1}) &\le \frac{\sum_{(\lambda,j,b,v_{\lambda'}): \mathcal{G}_{\lambda',\lambda}^{v_{\lambda'},j,b} \neq \emptyset} w_{\lambda',\lambda}^{v_{\lambda'},j}}{W_{\lambda',n+n'}} \\ &\le \frac{1}{A} \cdot \frac{3\mu_1\Lambda}{\alpha_{\lambda'} \cdot C_\Lambda}.\end{aligned}$$

  Therefore, by Lemma 5.7, it follows that for any $x \in \mathcal{X}$,

$$|\operatorname{Vote}_{\mathcal{T}^{n+n'}}(x) - \operatorname{Vote}_{\mathcal{T}^{n+n'-1}}(x)| \le \frac{8}{A} \sum_{\lambda'=1}^{\Lambda} \alpha_{\lambda'} \cdot \frac{3\mu_1\Lambda}{\alpha_{\lambda'} \cdot C_\Lambda} \le \frac{1}{A} \cdot \frac{30\mu_1\Lambda^2}{C_\Lambda} \tag{27}$$

  and for any $(x,y) \in \mathcal{X} \times [0,1]$,

$$|\operatorname{TErr}(\mathcal{T}^{n+n'}, x, y) - \operatorname{TErr}(\mathcal{T}^{n+n'-1}, x, y)| \le \frac{1}{A} \cdot \frac{30\mu_1\Lambda^2}{C_\Lambda}. \tag{28}$$

Fix any pair $(x, y) \in \mathcal{X} \times [0, 1]$. By Lemma 5.1, we have

$$\mathrm{TErr}(\mathcal{T}^{n+n_0-1}, x, \mathrm{Vote}_{\mathcal{T}^{n+n_0-1}}(x)) \leq 2\,\mathrm{TErr}(\mathcal{T}^{n+n_0-1}, x, y), \tag{29}$$

and so, using the triangle inequality,

$$\begin{aligned}
&\mathrm{TErr}(\mathcal{T}^n, x, \mathrm{Vote}_{\mathcal{T}^n}(x)) \\
\leq&|\mathrm{Vote}_{\mathcal{T}^{n+n_0-1}}(x) - \mathrm{Vote}_{\mathcal{T}^n}(x)| + \mathrm{TErr}(\mathcal{T}^n, x, \mathrm{Vote}_{\mathcal{T}^{n+n_0-1}}(x)) \\
\leq&|\mathrm{Vote}_{\mathcal{T}^{n+n_0-1}}(x) - \mathrm{Vote}_{\mathcal{T}^n}(x)| + \left|\mathrm{TErr}(\mathcal{T}^{n+n_0-1}, x, \mathrm{Vote}_{\mathcal{T}^{n+n_0-1}}(x)) - \mathrm{TErr}(\mathcal{T}^n, x, \mathrm{Vote}_{\mathcal{T}^{n+n_0-1}}(x))\right| \\
&+ \mathrm{TErr}(\mathcal{T}^{n+n_0-1}, x, \mathrm{Vote}_{\mathcal{T}^{n+n_0-1}}(x)) \\
\leq&\sum_{n':\ n+n' \in \mathcal{S}_0} 32 \sum_{\lambda=1}^{\Lambda} \alpha_\lambda \cdot \ln\left(\frac{W_{\lambda,n+n'-1}}{W_{\lambda,n+n'}}\right) + \frac{|\mathcal{S}_1|}{A} \cdot \frac{60\mu_1\Lambda^2}{C_\Lambda} + 2\,\mathrm{TErr}(\mathcal{T}^{n+n_0-1}, x, y) =: \delta,
\end{aligned}$$

where the final inequality uses (25), (26), (27), (28), and (29). Thus, $(x, \mathrm{Vote}_{\mathcal{T}^n}(x)) \in \mathrm{HighVote}(\mathcal{T}^n, \delta)$. Since the condition on step 2a holds in round $n+1$ (by assumption), it follows that $\bar{f}_{t(n+1)}$ satisfies

$$\mathbb{E}_{f\sim\bar{f}_{t(n+1)}}\left[|f(x) - \mathrm{Vote}_{\mathcal{T}^n}(x)|\right] \leq \max\left\{64 \cdot \alpha_\Lambda, 2C_\Lambda \cdot \delta\right\} = 2C_\Lambda \cdot \delta, \tag{30}$$

where the equality above follows since $C_\Lambda \cdot \delta \geq C_\Lambda \cdot \mathrm{TErr}(\mathcal{T}^{n+n_0-1}, x, \mathrm{Vote}_{\mathcal{T}^{n+n_0-1}}(x)) \geq 32 \cdot \alpha_\Lambda$, since $C_\Lambda \geq 32$. It then follows that, writing $\mathcal{T}^n = [\mathcal{G}, w]$,

$$\begin{aligned}
&\mathbb{E}_{f\sim\bar{f}_{t(n+1)}}\left[|f(x) - y|\right] \\
\leq&\mathbb{E}_{f\sim\bar{f}_{t(n+1)}}\left[|f(x) - \mathrm{Vote}_{\mathcal{T}^n}(x)|\right] + |y - \mathrm{Vote}_{\mathcal{T}^n}(x)| \\
\leq&\mathbb{E}_{f\sim\bar{f}_{t(n+1)}}\left[|f(x) - \mathrm{Vote}_{\mathcal{T}^n}(x)|\right] + \sum_{v\in\mathcal{N}_\Lambda(\mathcal{T}^n)} P_w(v) \cdot |y - \mathrm{SOA}(\mathcal{G}^v_{1:\Lambda})(x)| \\
\leq&\mathbb{E}_{f\sim\bar{f}_{t(n+1)}}\left[|f(x) - \mathrm{Vote}_{\mathcal{T}^n}(x)|\right] + \sum_{v\in\mathcal{N}_\Lambda(\mathcal{T}^n)} P_w(v) \cdot \mathrm{TErr}(\mathcal{G}^v_{1:\Lambda}, x, y) \\
=&\mathbb{E}_{f\sim\bar{f}_{t(n+1)}}\left[|f(x) - \mathrm{Vote}_{\mathcal{T}^n}(x)|\right] + \mathrm{TErr}(\mathcal{T}^n, x, y) \\
\leq&\mathbb{E}_{f\sim\bar{f}_{t(n+1)}}\left[|f(x) - \mathrm{Vote}_{\mathcal{T}^n}(x)|\right] + \mathrm{TErr}(\mathcal{T}^{n+n_0-1}, x, y) \\
&+ \sum_{n':\ n+n' \in \mathcal{S}_0} 16 \sum_{\lambda=1}^{\Lambda} \alpha_\lambda \cdot \ln\left(\frac{W_{\lambda,n+n'-1}}{W_{\lambda,n+n'}}\right) + \frac{|\mathcal{S}_1|}{A} \cdot \frac{30\mu_1\Lambda^2}{C_\Lambda}.
\end{aligned} \tag{31}$$

The claim of the lemma then follows from (30) and (31).

□

Recall that for each $\lambda \in [\Lambda]$, we use the parameter $W_{\lambda,n} = \sum_{(\mathcal{G}^{v_\lambda}_\lambda, w^{v_\lambda}_\lambda) \in \mathcal{T}^n_\lambda} w^{v_\lambda}_\lambda$ as a potential function; in particular, Lemmas 5.9 and 5.10 show an upper bound for the values of $W_{\lambda,n}$ in each round n. In order to bound the total number of rounds n (thus showing that the algorithm converges), as well as the total error, it is necessary to have a lower bound for the weights $w^{v_\lambda}_\lambda$ as well; such a lower bound is provided by Lemma 5.12 below.

Lemma 5.12. *For all rounds n, and all $\lambda \in [\Lambda]$, it holds that for each pair $(\mathcal{G}^{v_\lambda}_\lambda, w^{v_\lambda}_\lambda) \in \mathcal{T}_\lambda$, $w^{v_\lambda}_\lambda \geq \gamma^{\mathrm{sfat}_{\alpha_\lambda}(\mathcal{F})}$.*

Proof. We will prove the stronger statement that at any step of the algorithm, for all pairs $(\mathcal{G}_\lambda^{v_\lambda}, w_\lambda^{v_\lambda}) \in \mathcal{T}_\lambda$, it holds that $w_\lambda^{v_\lambda} \geq \gamma^{\mathrm{sfat}_{\alpha_\lambda}(\mathcal{F}) - \mathrm{sfat}_{\alpha_\lambda}(\mathcal{G}_\lambda^{v_\lambda})}$. Note that the only two points in the algorithm where any pair $(\mathcal{G}_\lambda^{v_\lambda}, w_\lambda^{v_\lambda}) \in \mathcal{T}_\lambda$ is changed, for any $\lambda \in [\Lambda]$, are steps 2(a)iiA and 2(b)ii. Moreover, note that when any weight $w_\lambda^{v_\lambda}$ is changed to a new value $w_\lambda^{v_\lambda\prime}$, we always have $w_\lambda^{v_\lambda\prime}/w_\lambda^{v_\lambda} \geq \gamma$. Thus, it suffices to show that whenever any weight $w_\lambda^{v_\lambda}$ is changed, the sequential fat-shattering dimension of $\mathcal{G}_\lambda^{v_\lambda}$ (at the scale α_λ) always decreases by at least 1. To do so, we consider each of the two possibilities in turn:

- If we decrease w by a factor of γ in step 2(a)iiA, then we also replace $\mathcal{G}$ with $\mathcal{G}|^{\alpha_\lambda}_{(x_t,y_t)}$ (and it must be the case that $|\mathrm{SOA}(\mathcal{G},\alpha_\lambda)(x_t) - y_t| > \alpha_\lambda$). By Lemma A.2, it holds that $\mathrm{sfat}_{\alpha_\lambda}(\mathcal{G}|^{\alpha_\lambda}_{(x_t,y_t)}) < \mathrm{sfat}_{\alpha_\lambda}(\mathcal{G})$, i.e., the α_λ-sequential fat shattering dimension of $\mathcal{G}$ must strictly decrease.

- Now suppose we are at step 2(b)iA, where we will add $(\mathcal{G}_{\lambda',\lambda}^{v_{\lambda'},j,b}, w_{\lambda',\lambda}^{v_{\lambda'},j,b})$ to $\mathcal{T}_{\lambda'}$ (in step 2(b)ii), for some choices of $\lambda, \lambda' \in [\Lambda], v_{\lambda'} \in [|\mathcal{T}_{\lambda'}|], j \in [m_\lambda], b \in \{0,1,\ldots,\lfloor 1/\alpha_{\lambda'}\rfloor + 1\}$. Then for the pair $(\mathcal{G}_{\lambda'}^{v_{\lambda'}}, w_{\lambda'}^{v_{\lambda'}})$ which was previously in $\mathcal{T}_{\lambda'}$, we have $w_{\lambda',\lambda}^{v_{\lambda'},j,b} = \gamma \cdot w_{\lambda'}^{v_{\lambda'}}$, and $\mathcal{G}_{\lambda',\lambda}^{v_{\lambda'},j,b} = \mathcal{G}_{\lambda'}^{v_{\lambda'}}|^{\alpha_{\lambda'}}_{(\tilde{x}_j^\lambda, b\alpha_{\lambda'})}$, where $|\tilde{y}_j^\lambda - b\alpha_{\lambda'}| > 6\alpha_{\lambda'}$. We have $|\mathrm{SOA}(\mathcal{G},\alpha_{\lambda'})(\tilde{x}_j^\lambda) - \tilde{y}_j^\lambda| \leq 5\alpha_{\lambda'}$, and therefore, $|\mathrm{SOA}(\mathcal{G},\alpha_{\lambda'})(\tilde{x}_j^\lambda) - b\alpha_{\lambda'}| > \alpha_{\lambda'}$, so by Lemma A.2, we have $\mathrm{sfat}_{\alpha_{\lambda'}}(\mathcal{G}_{\lambda',\lambda}^{v_{\lambda'},j,b}) < \mathrm{sfat}_{\alpha_{\lambda'}}(\mathcal{G})$.

□

Lemma 5.12 above shows a lower bound on the size of the individual weights $w_\lambda^{v_\lambda}$ in the weighted subclass collections $\mathcal{T}^n$; in order to lower bound the weights $W_{\lambda,n}$, we need a lower bound on the number of pairs $(\mathcal{G}_\lambda^{v_\lambda}, w_\lambda^{v_\lambda})$ remaining in the multisets $\mathcal{T}_\lambda^n$; Lemma 5.13 below aids in obtaining such a lower bound.

Lemma 5.13. *Consider any round n for which the condition in step 2a of* `Multi-scale Proper Learner` *(Algorithm 2) fails. Suppose that before this round, for each $\lambda \in [\Lambda]$, there are $q_\lambda \in \mathbb{N}$ tuples $(\mathcal{G}_\lambda^{v_\lambda}, w_\lambda^{v_\lambda}) \in \mathcal{T}_\lambda$ with $f^\star \in \mathcal{G}_\lambda^{v_\lambda}$. Then after round n, for some $\lambda'' \in [\Lambda]$, there are at least $q_{\lambda''} \cdot \left(A + \frac{\mu_0}{128\Lambda\alpha_{\lambda''}}\right)$ tuples $((\mathcal{G}_{\lambda''}^{v_{\lambda''}})', (w_{\lambda''}^{v_{\lambda''}})') \in \mathcal{T}_{\lambda''}$ with $f^\star \in (\mathcal{G}_{\lambda''}^{v_{\lambda''}})'$.*

Proof. Set $\lambda \in [\Lambda], \lambda' \in [\Lambda - 3]$ to be so that $\frac{1}{m_\lambda}\sum_{j=1}^{m_\lambda} \mathbb{1}[|f^\star(\tilde{x}_j^\lambda) - \tilde{y}_j^\lambda| > \alpha_{\lambda'}] > \frac{\alpha_\lambda}{16\Lambda\alpha_{\lambda'}}$. (This is possible by the property in step 2b of `Multi-scale Proper Learner`.) Set $\lambda'' = \lambda' + 3 \in [\Lambda]$. Then $\frac{1}{m_\lambda}\sum_{j=1}^{m_\lambda} \mathbb{1}[|f^\star(\tilde{x}_j^\lambda) - \tilde{y}_j^\lambda| > 7\alpha_{\lambda''}] > \frac{1}{8} \cdot \frac{\alpha_\lambda}{16\Lambda\alpha_{\lambda''}}$. Note that if $|f^\star(\tilde{x}_j^\lambda) - \tilde{y}_j^\lambda| > 7\alpha_{\lambda''}$, then $f^\star(\tilde{x}_j^\lambda) \in [b\alpha_{\lambda''}, (b+1)\alpha_{\lambda''})$ for some b satisfying $|b\alpha_{\lambda''} - \tilde{y}_j^\lambda| > 6\alpha_{\lambda''}$.

Therefore, for any $v_{\lambda''}$ such that $f^\star \in \mathcal{G}_{\lambda''}^{v_{\lambda''}}$, there are at least $\frac{\alpha_\lambda m_\lambda}{128\Lambda\alpha_{\lambda''}}$ tuples $(j,b) \in [m_\lambda] \times \{0,1,\ldots,\lfloor\alpha_{\lambda''}\rfloor + 1\}$ so that $f^\star \in \mathcal{G}_{\lambda'',\lambda}^{v_{\lambda''},j,b}$: in particular, these correspond to the (at least) $\frac{\alpha_\lambda m_\lambda}{128\Lambda\alpha_{\lambda''}}$ values of j for which $|f^\star(\tilde{x}_j^\lambda) - \tilde{y}_j^\lambda| > 7\alpha_{\lambda''}$, each of which is handled as follows:

- If $|\mathrm{SOA}(\mathcal{G}_{\lambda''}^{v_{\lambda''}}, \alpha_{\lambda''})(\tilde{x}_j^{\lambda''}) - \tilde{y}_j^{\lambda''}| \leq 5\alpha_{\lambda''}$, then as we have remarked above, there is some b satisfying $|b\alpha_{\lambda''} - \tilde{y}_j^\lambda| > 6\alpha_{\lambda''}$ so that $f^\star \in \mathcal{G}_{\lambda'',\lambda}^{v_{\lambda''},j,b} = \mathcal{G}_{\lambda''}^{v_{\lambda''}}|^{\alpha_{\lambda''}}_{(\tilde{x}_j^\lambda, b\alpha_{\lambda''})}$ (this corresponds to step 2(b)iA of `Multi-scale Proper Learner`).

- Otherwise, we have $f^\star \in \mathcal{G}_{\lambda'',\lambda}^{v_{\lambda''},j,0} = \mathcal{G}_{\lambda''}^{v_{\lambda''}}$ (this corresponds to step 2(b)iB of `Multi-scale Proper Learner`).

Since $\mathcal{T}_{\lambda''}$ after step n contains A copies of the collection $\mathcal{T}_{\lambda''}$ before step n in addition to the tuples $(\mathcal{G}_{\lambda'',\lambda}^{v_{\lambda''},j,b}, w_{\lambda'',\lambda}^{v_{\lambda''},j})$ for each $\lambda \in [\Lambda]$, $j \in [m_\lambda], b \in \{0,1,\ldots,\lfloor 1/\alpha_{\lambda''} \rfloor + 1\}, v_{\lambda''} \in [|\mathcal{T}_{\lambda''}|]$ (so that $\mathcal{G}_{\lambda'',\lambda}^{v_{\lambda''},j,b} \neq \emptyset$), it follows that the number of tuples $((\mathcal{G}_{\lambda''}^{v_{\lambda''}})', (w_{\lambda''}^{v_{\lambda''}})') \in \mathcal{T}_{\lambda''}$ so that $f^\star \in (\mathcal{G}_{\lambda''}^{v_{\lambda''}})'$ is at least

$$q_{\lambda''} \cdot \left(A + \frac{\alpha_\lambda m_\lambda}{128\Lambda\alpha_{\lambda''}}\right) \geq q_{\lambda''} \cdot \left(A + \frac{\mu_0}{128\Lambda\alpha_{\lambda''}}\right).$$

□

For $n \in \mathbb{N}$, let N_n be the number of rounds, up to and including round n, for which the condition in step 2a fails. Also let T_n be the value of t immediately before executing round n of the algorithm. Finally, let $\mathcal{S}_n$ denote the set of rounds $n' \leq n$ for which the condition in step 2a holds. Note that $|\mathcal{S}_n| = T_n$.

The following lemma uses the previous lemmas of this section to bound various parameters that show up in our error bounds. The first two bounds (on δ_t and $\log(W_{\lambda,n'-1}/W_{\lambda,n'})$) will be used together with Lemma 5.11 to bound the error of the predictors $\bar{f}_t$, and the final bound (on N_n) shows that `Multi-scale Proper Learner` terminates after a finite number of steps.

Lemma 5.14. *For each $n \in \mathbb{N}$, it holds that*

$$\sum_{t=1}^{T_n} \delta_t \leq 32\alpha_\Lambda T_n + 64\Lambda \log(1/\gamma) \cdot \sum_{\lambda\in[\Lambda]} \alpha_\lambda \cdot \mathrm{sfat}_{\alpha_\lambda}(\mathcal{F})$$

$$\sum_{n'\in\mathcal{S}_n} \sum_{\lambda\in[\Lambda]} \alpha_\lambda \cdot \log\left(\frac{W_{\lambda,n'-1}}{W_{\lambda,n'}}\right) \leq \log(1/\gamma) \cdot \sum_{\lambda\in[\Lambda]} \alpha_\lambda \cdot \mathrm{sfat}_{\alpha_\lambda}(\mathcal{F})$$

$$N_n \leq 1024\Lambda \cdot \frac{A}{\mu_1} \cdot \log(1/\gamma) \sum_{\lambda\in[\Lambda]} \alpha_\lambda \cdot \mathrm{sfat}_{\alpha_\lambda}(\mathcal{F}).$$

Proof. The realizability assumption gives us that for each t, $y_t = f^\star(x_t)$. Thus, in each round n in which the condition in step 2a holds, for each $\lambda \in [\Lambda]$ and each $(\mathcal{G}_\lambda^{v_\lambda}, w_\lambda^{v_\lambda}) \in \mathcal{T}_\lambda$ for which $f^\star \in \mathcal{G}_\lambda^{v_\lambda}$ at the beginning of round n, after restricting $\mathcal{G}_\lambda^{v_\lambda} \leftarrow \mathcal{G}_\lambda^{v_\lambda}|_{(x_t,y_t)}^{\alpha_\lambda}$, it still holds that $f^\star \in \mathcal{G}$. For each round n in which the condition in step 2a fails, for each $\lambda \in [\Lambda]$, if $\mathcal{T}_\lambda$ has q_λ tuples $(\mathcal{G}_\lambda^{v_\lambda}, w_\lambda^{v_\lambda})$ so that $f^\star \in \mathcal{G}_\lambda^{v_\lambda}$ at the beginning of round n, then step 2(b)ii ensures that after round n, $\mathcal{T}_\lambda$ has $A \cdot q_\lambda$ tuples $(\mathcal{G}_\lambda^{v_\lambda}, w_\lambda^{v_\lambda})$ so that $f^\star \in \mathcal{G}_\lambda^{v_\lambda}$. Moreover, we have the following two facts:

- By Lemma 5.13, for each round n in which the condition in step 2a fails, there is some value of $\lambda' = \lambda'(n) \in [\Lambda]$ so that after round n, $\mathcal{T}_{\lambda'}$ has at least $q_{\lambda'} \cdot \left(A + \frac{\mu_0}{128\Lambda\alpha_{\lambda'}}\right)$ tuples $(\mathcal{G}_\lambda^{v_\lambda}, w_\lambda^{v_\lambda})$ with $f^\star \in \mathcal{G}_\lambda^{v_\lambda}$. (If the condition in step 2a holds in round n, $\lambda'(n)$ is not defined; we write $\lambda'(n) = \perp$.) For each $\lambda \in [\Lambda]$, and $n \geq 1$, let $N_{\lambda,n}$ denote the number of rounds $n' \leq n$, for which $\lambda'(n') = \lambda$. Note that $\sum_{\lambda\in[\Lambda]} N_{\lambda,n} = N_n$.
- By Lemma 5.9, for each round n in which the condition in step 2a holds, if we let t be the value of $t = t(n)$ at step 2(a)ii, then the following holds: if $\delta_t > 32\alpha_\Lambda$, then for some $\lambda'' = \lambda''(n) \in [\Lambda]$, $W_{\lambda'',n} \leq W_{\lambda'',n-1} \cdot \left(1 - \frac{\delta_t}{64\Lambda\alpha_{\lambda''}}\right) \leq W_{\lambda'',n-1} \cdot \exp\left(\frac{-\delta_t}{64\Lambda\alpha_{\lambda''}}\right)$. (If $\delta_t \leq 32\alpha_\Lambda$ or the condition in step 2a fails, then $\lambda''(n)$ is not defined; we write $\lambda''(n) = \perp$ for such n.) For each $\lambda \in [\Lambda]$ and $n \geq 1$, let $\mathcal{S}_{\lambda,n}$ denote the set of rounds $n' \leq n$ for which $\lambda''(n') = \lambda$.

By definition of $N_{\lambda,n}$, $\mathcal{T}_\lambda$ has at least $A^{N_n} \cdot \left(1 + \frac{\mu_0}{128\Lambda\alpha_\lambda A}\right)^{N_{\lambda,n}}$ tuples $(\mathcal{G}_\lambda^{v_\lambda}, w_\lambda^{v_\lambda})$ with $f^\star \in \mathcal{G}$. Combining this fact with Lemma 5.12, we get that the total weight of tuples in $\mathcal{T}_\lambda$ is lower bounded as follows:

$$W_{\lambda,n} \geq A^{N_n} \cdot \left(1 + \frac{\mu_0}{128\Lambda\alpha_\lambda A}\right)^{N_{\lambda,n}} \cdot \gamma^{\mathrm{sfat}_{\alpha_\lambda}(\mathcal{F})}. \tag{32}$$

We next proceed to compute an upper bound on $W_{\lambda,n}$. By Lemma 5.9, for all rounds n' in which the condition in step 2a holds, we have $W_{\lambda,n} \leq W_{\lambda,n-1}$. Further, by Lemma 5.10, for all rounds n' in which the condition in step 2a fails, we have $W_{\lambda,n} \leq W_{\lambda,n-1} \cdot \left(A + \frac{3\mu_1\Lambda}{\alpha_\lambda C_\Lambda}\right)$ for all $\lambda \in [\Lambda]$. Combining these facts with the definition of $\mathcal{S}_{\lambda,n}$ above, we get that

$$W_{\lambda,n} \leq A^{N_n} \cdot \left(1 + \frac{3\mu_1\Lambda}{\alpha_\lambda C_\Lambda A}\right)^{N_n} \cdot \prod_{n' \in \mathcal{S}_n} \frac{W_{\lambda,n'}}{W_{\lambda,n'-1}} \leq A^{N_n} \cdot \left(1 + \frac{3\mu_1\Lambda}{\alpha_\lambda C_\Lambda A}\right)^{N_n} \cdot \prod_{n' \in \mathcal{S}_{\lambda,n}} \exp\left(\frac{-\delta_{t(n')}}{64\Lambda\alpha_\lambda}\right) \tag{33}$$

Combining (32) and (33), we obtain that, for each $\lambda \in [\Lambda]$,

$$\begin{aligned}
\mathrm{sfat}_{\alpha_\lambda}(\mathcal{F}) \cdot \ln(1/\gamma) &\geq \sum_{n' \in \mathcal{S}_n} \ln\left(\frac{W_{\lambda,n'-1}}{W_{\lambda,n'}}\right) + N_{\lambda,n} \cdot \ln\left(1 + \frac{\mu_0}{128\Lambda\alpha_\lambda A}\right) - N_n \cdot \ln\left(1 + \frac{3\mu_1\Lambda}{\alpha_\lambda C_\Lambda A}\right) \\
&\geq \sum_{n' \in \mathcal{S}_n} \ln\left(\frac{W_{\lambda,n'-1}}{W_{\lambda,n'}}\right) + N_{\lambda,n} \cdot \frac{\mu_1}{512\Lambda\alpha_\lambda A} - N_n \cdot \frac{3\mu_1\Lambda}{\alpha_\lambda C_\Lambda A} \qquad (34) \\
&\geq \sum_{n' \in \mathcal{S}_{\lambda,n}} \frac{\delta_{t(n')}}{64\Lambda\alpha_\lambda} + N_{\lambda,n} \cdot \frac{\mu_1}{512\Lambda\alpha_\lambda A} - N_n \cdot \frac{3\mu_1\Lambda}{\alpha_\lambda C_\Lambda A}, \qquad (35)
\end{aligned}$$

where the inequality (34) uses that $A \geq \frac{\mu_0}{\alpha_\lambda}$ for all $\lambda \in [\Lambda]$ and $\mu_0 \geq \mu_1/2$. Note that our choice of $C_\Lambda = 3072\Lambda^3$ gives

$$\sum_{\lambda \in [\Lambda]} \left(N_{\lambda,n} \cdot \frac{\mu_1}{512\Lambda A_\lambda} - N_n \cdot \frac{3\mu_1\Lambda}{C_\Lambda A_\lambda}\right) = N_n \cdot \left(\frac{\mu_1}{512\Lambda A} - \frac{3\mu_1\Lambda^2}{C_\Lambda A}\right) \geq N_n \cdot \frac{\mu_1}{1024\Lambda A} > 0. \tag{36}$$

Thus, multiplying (35) by α_λ and summing it over $\lambda \in [\Lambda]$ gives that the following hold:

$$\begin{aligned}
\sum_{t=1}^{T_n} \delta_t &\leq 32\alpha_\Lambda T_n + \sum_{\lambda \in [\Lambda]} \sum_{n' \in \mathcal{S}_{\lambda,n}} \delta_{t(n')} \leq 32\alpha_\Lambda T_n + 64\Lambda \ln(1/\gamma) \cdot \sum_{\lambda \in [\Lambda]} \alpha_\lambda \cdot \mathrm{sfat}_{\alpha_\lambda}(\mathcal{F}) \\
\sum_{n' \in \mathcal{S}_n} \sum_{\lambda \in [\Lambda]} \alpha_\lambda \cdot \ln\left(\frac{W_{\lambda,n'-1}}{W_{\lambda,n'}}\right) &\leq \ln(1/\gamma) \cdot \sum_{\lambda \in [\Lambda]} \alpha_\lambda \cdot \mathrm{sfat}_{\alpha_\lambda}(\mathcal{F}) \\
N_n &\leq 1024\Lambda \cdot \frac{A}{\mu_1} \cdot \ln(1/\gamma) \sum_{\lambda \in [\Lambda]} \alpha_\lambda \cdot \mathrm{sfat}_{\alpha_\lambda}(\mathcal{F}).
\end{aligned}$$

□

Fix any smoothing parameter $\eta > 0$ so that $1/\eta \in \mathbb{N}$, and write $\xi := 1/\eta$. Consider the distributions $\bar{f}_1, \ldots, \bar{f}_T$ output by `Multi-scale Proper Learner`. Fix arbitrary distributions $\bar{f}_0, \ldots, \bar{f}_{1-\xi} \in \Delta^\circ(\mathcal{F})$. We define the *$\eta$-stabilized distributions* $\bar{h}_1, \ldots, \bar{h}_T$ as follows: for $t \in [T]$, we define

$$\bar{h}_t = \eta \sum_{s=0}^{\xi-1} \bar{f}_{t-s}. \tag{37}$$

Clearly it holds that $\left\|\bar{h}_t - \bar{h}_{t+1}\right\|_1 \le 2\eta$ for all $T \in [T-1]$, and $\bar{h}_t \in \Delta^\circ(\mathcal{F})$ for each $t \in [T]$. The below lemma gives a bound on the cumulative loss for the sequence $\bar{h}_t$:

Theorem 5.15 (Near-optimal stable proper learning)**.** *Fix any $\eta > 0$. In the realizable setting, the η-stabilized distributions $\bar{h}_t$ defined in (37) from the output of `Multi-scale Proper Learner` (Algorithm 2) satisfy:*

$$\sum_{t=1}^{T} \mathbb{E}_{h\sim\bar{h}_t}\left[|h(x_t) - y_t|\right] \le \frac{1}{\eta} \cdot O\left(\log^6(T) \cdot \min_{\alpha\in[0,1]}\left\{\alpha T + \int_\alpha^1 \mathrm{sfat}_\eta(\mathcal{F}) d\eta\right\}\right). \tag{38}$$

Further, $\left\|\bar{h}_t - \bar{h}_{t+1}\right\|_1 \le 2\eta$ for all $t \in [T-1]$ and $\bar{h}_t \in \Delta^\circ(\mathcal{F})$ for all $T \in [T]$.

Proof. As in the proof of Proposition 4.1, choose $\alpha \in [1/T, 1]$ minimizing the expression on the right-hand side of (38), and set $\Lambda = \lfloor 1/(2\alpha) \rfloor \le \log T$. Since $C_\Lambda = 3072\Lambda^3 = O(\log^3 T)$ and $\log(1/\gamma) = O(\log T)$, it suffices to show

$$\sum_{t=1}^{T} \mathbb{E}_{h\sim\bar{h}_t}\left[|h(x_t) - y_t|\right] \le O\left(\frac{C_\Lambda \cdot \Lambda^2}{\eta}\right) \cdot \left(\alpha_\Lambda T + \log(1/\gamma) \sum_{\lambda=1}^{\Lambda} \alpha_\lambda \,\mathrm{sfat}_{\alpha_\lambda}(\mathcal{F})\right),$$

when `Multi-scale Proper Learner` is run with the chosen scale parameter Λ.

For $t \in [T]$ and $0 \le s < \xi$, let $\mathcal{M}_{0,t,s}$ be the set of rounds n starting at the round when x_t is observed and up to (but not including) the round where x_{t+s} is observed, so that the condition in step 2a holds in round n. Let $\mathcal{M}_{1,t,s}$ be the set of such rounds n (i.e., starting at x_t, and up to but not including x_{t+s}) for which the condition in step 2a fails in round n. Let $\mathcal{M}_0$ be the set of all rounds n (up to, and including, the round that x_T is observed) for which the condition in step 2a holds in round n, and let $\mathcal{M}_1$ be the set of all rounds n for which the condition in step 2a fails in round n. Furthermore, recall the definition of δ_t in step 2(a)ii of Algorithm 2. By the definition of

$\bar{h}_t$, we have

$$
\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}_{h\sim\bar{h}_t}\left[|h(x_t)-y_t|\right] \\
\leq &\eta \sum_{t=1}^{T}\sum_{s=0}^{\xi-1} \mathbb{E}_{f\sim\bar{f}_{t-s}}\left[|f(x_t)-y_t|\right] \\
\leq &\xi + \eta \sum_{t=1}^{T-\xi+1}\sum_{s=0}^{\xi-1} \mathbb{E}_{f\sim\bar{f}_t}\left[|f(x_{t+s})-y_{t+s}|\right] \\
\leq &\xi + \eta \sum_{t=1}^{T-\xi+1}\sum_{s=0}^{\xi-1} \left[80\cdot C_\Lambda \sum_{n'\in\mathcal{M}_{0,t,s}} \sum_{\lambda\in[\Lambda]} \alpha_\lambda \cdot \ln\left(\frac{W_{\lambda,n'-1}}{W_{\lambda,n'}}\right) + \frac{|\mathcal{M}_{1,t,s}|}{A}\cdot 150 C_\Lambda \mu_1 \Lambda^2 + 5C_\Lambda\cdot\delta_{t+s}\right] \qquad (39)\\
\leq &\xi + \frac{80C_\Lambda}{\eta}\cdot \sum_{n'\in\mathcal{M}_0}\sum_{\lambda\in[\Lambda]} \alpha_\lambda \ln\left(\frac{W_{\lambda,n'-1}}{W_{\lambda,n'}}\right) + \frac{|\mathcal{M}_1|\cdot 150 C_\Lambda\mu_1\Lambda^2}{A\eta} + 5C_\Lambda\sum_{t=1}^{T}\delta_t \qquad (40)\\
\leq &\xi + \frac{80C_\Lambda}{\eta}\cdot\log(1/\gamma)\sum_{\lambda=1}^{\Lambda}\alpha_\lambda \operatorname{sfat}_{\alpha_\lambda}(\mathcal{F}) + 1024\Lambda\log(1/\gamma)\cdot\frac{150C_\Lambda\Lambda^2}{\eta}\sum_{\lambda=1}^{\Lambda}\alpha_\lambda\operatorname{sfat}_{\alpha_\lambda}(\mathcal{F}) \qquad (41)\\
&+ 5C_\Lambda\cdot\left(32\alpha_\Lambda T + 64\Lambda\log(1/\gamma)\sum_{\lambda=1}^{\Lambda}\alpha_\lambda\operatorname{sfat}_{\alpha_\lambda}(\mathcal{F})\right)\\
\leq &O\left(\frac{C_\Lambda\cdot\Lambda^2}{\eta}\right)\cdot\left(\alpha_\Lambda T + \log(1/\gamma)\sum_{\lambda=1}^{\Lambda}\alpha_\lambda\operatorname{sfat}_{\alpha_\lambda}(\mathcal{F})\right),
\end{aligned}
$$

where:

- (39) follows from Lemma 5.11;
- (40) follows from exchanging the order of summation and noting that for each $n'\in\mathcal{M}_0$, there are at most ξ^2 values of (t,s) so that $n'\in\mathcal{M}_{0,t,s}$ and for each $n'\in\mathcal{M}_1$, there are at most ξ^2 values of (t,s) so that $n'\in\mathcal{M}_{1,t,s}$;
- (41) follows from Lemma 5.14; notice that we have used here that $|\mathcal{M}_1| = N_n$, where n is total number of iterations of the outer while loop of `Multi-scale Proper Learner`.

□

6 Path-length regret bound for a stable proper learner

In this section we prove Theorem 6.1, obtaining a proper agnostic learner that gets a path-length regret bound. As we do throughout the paper, we assume that the given function class $\mathcal{F}$ has finite sequential fat-shattering dimension at all scales.

Theorem 6.1 (Path-length regret bound for a stable online learner)**.** *Suppose that α is chosen so that $1\leq\alpha T\leq\operatorname{sfat}_\alpha(\mathcal{F})$ and $\alpha\leq\kappa$. Moreover suppose that for all $t<T$, the examples (x_t,y_t)*

satisfy $\|x_t - x_{t+1}\|_{\infty,\mathcal{F}} \leq \kappa$ *and* $|y_t - y_{t+1}| \leq \kappa$. *Then, for any* $\Gamma \geq 1$ *`Optimistic SOA-Experts` with step size* $\eta_{OH} = \eta_{PSR} = \kappa/\Gamma$ *obtains a regret of*

$$\sum_{t=1}^{T} \mathbb{E}_{f_t \sim \bar{f}_t}[|f_t(x_t) - y_t|] - \sum_{t=1}^{T} |f_\star(x_t) - y_t| \leq O\left(\frac{\Gamma \cdot \mathrm{sfat}_\alpha(\mathcal{F}) \cdot \log^6 T}{\kappa} + \frac{\kappa^3 \cdot T}{\Gamma}\right) \tag{42}$$

Further, for any choices of $\eta_{OH}, \eta_{PSR} > 0$ *(and without restriction on the* (x_t, y_t)*), the iterates* $\bar{f}_t$ *of `Optimistic SOA-Experts` belong to* $\Delta^\circ(\mathcal{F})$ *and are stable in the following sense: for all* $t < T$,

$$\left\|\bar{f}_t - \bar{f}_{t+1}\right\|_1 \leq 5\eta_{OH} + 3\eta_{PSR}.$$

In particular, if we have $\Gamma\eta_{OH} = \Gamma\eta_{PSR} = \kappa = (\mathrm{sfat}_\alpha(\mathcal{F})/T)^{1/4} \cdot \log^{3/4}(T)$, *then the regret is bounded above by* $\tilde{O}\left(\Gamma \cdot \mathrm{sfat}_\alpha(\mathcal{F})^{3/4} \cdot T^{1/4}\right)$.

The main ingredient in the proof is Theorem 5.15 from the previous section which gives an (optimal) stable and proper learner for the setting of realizable online regression. Given this result, the proof of Theorem 5.15 is mostly standard, using results of [BDPSS09, RST15a] and [SALS15].

6.1 Defining the experts

As discussed in Section 3, the general idea of the proof is to use the SOA-experts framework of [BDPSS09, RST15a].[13] We begin by defining the experts in this setting in Definition 6.1 below. Let $\mathcal{X}^\star$ be the set of all finite sequences of elements of $\mathcal{X}$. Each expert is a function $E : \mathcal{X}^\star \to [0,1]$; $E(x_1, \ldots, x_t)$ should be interpreted as the label that the expert E predicts for x_t given that it has already seen $x_1, \ldots, x_{t-1}$.

Definition 6.1 ([RST15a])**.** Fix $T \in \mathbb{N}$, any $\mathcal{F} \subset [0,1]^{\mathcal{X}}$ and $\alpha \in (0,1)$, and set $d_\alpha := \mathrm{sfat}_\alpha(\mathcal{F})$. For each tuple (I, σ), where I is a subset $I \subset [T]$ of size $|I| \leq d_\alpha$, and $\sigma \in \{0, 1, \ldots, \lceil 1/\alpha \rceil - 1\}^{|I|}$, define the expert $E_{(I,\sigma)} : \mathcal{X}^\star \to [0,1]$ by

$$E_{(I,\sigma)}(x_1, \ldots, x_t) = \mathrm{SOA}(\mathcal{F}(t), \alpha)(x_t),$$

where $\mathcal{F}(t)$ is defined inductively via $\mathcal{F}(1) = \mathcal{F}$ and

$$\mathcal{F}(t+1) = \begin{cases} \mathcal{F}(t) & : t \notin I \\ \mathcal{F}(t)|^{\alpha}_{(x_t, \sigma_{i_t} \cdot \alpha)} & : t \in I, \end{cases}$$

where for $t \in I$, $i_t \in \{1, \ldots, |I|\}$ is defined so that t is the i_tth smallest element of I. We denote the set of experts $E_{(I,\sigma)}$ given T, α by $\mathscr{E}_{T,\alpha}$ (the class $\mathcal{F}$ is implicit in our notation).

The set of all experts $E_{(I,\sigma)}$ of Definition 6.1 can be seen as an algorithmic version of a sequential cover [RST15b, Definition 4]. Lemma 6.2 bounds the number of experts in $\mathscr{E}_{T,\alpha}$.

Lemma 6.2. *Given* $T \in \mathbb{N}$, $\mathcal{F} \subset [0,1]^{\mathcal{X}}$ *and* $\alpha \in (0,1)$, *the number of experts in the set* $\mathscr{E}_{T,\alpha}$ *of Definition 6.1 is at most* $\left(\frac{2eT}{\alpha}\right)^{\mathrm{sfat}_\alpha(\mathcal{F})}$.

[13] For references in this section to [RST15a], see in particular the version at `https://arxiv.org/pdf/1006.1138v1.pdf`.

Proof. The number of experts (I, σ) is at most

$$\sum_{s=1}^{\mathrm{sfat}_\alpha(\mathcal{F})} \binom{T}{s} \cdot \left\lceil \frac{1}{\alpha} \right\rceil^s \leq \left(\frac{2eT}{\alpha} \right)^{\mathrm{sfat}_\alpha(\mathcal{F})}.$$

□

Lemma 6.3 shows that the set of experts covers the class $\mathcal{F}$ in an online sense.

Lemma 6.3 (Lemma 15, [RST15a])**.** *Given $T \in \mathbb{N}$, $\mathcal{F} \subset [0,1]^{\mathcal{X}}$, and $\alpha \in (0,1)$, for each $f \in \mathcal{F}$ and any sequence $(x_1, \ldots, x_T)$, there exists some expert $E \in \mathscr{E}_{T,\alpha}$ so that for all $t \in [T]$,*

$$|f(x_t) - E(x_1, \ldots, x_t)| \leq \alpha.$$

Since Lemma 6.3 only proimses that some expert has error α with respect to any given hypothesis f, yet Lemma Theorem 5.15 (for proper learning) requires that its input sequence be *exactly* realizable, we need to work with the α-augmented class for a given class $\mathcal{F}$, defined below.

Definition 6.2 (α-augmented class)**.** For a real-valued class $\mathcal{F} \subset [0,1]^{\mathcal{X}}$ and $\alpha > 0$, define the *α-augmented class* $\mathcal{F}^\alpha$ by

$$\mathcal{F}^\alpha := \left\{ f' \in [0,1]^{\mathcal{X}} : \exists f \in \mathcal{F} \text{ such that } \left\| f' - f \right\|_{\infty,\mathcal{X}} \leq \alpha \right\}.$$

For each element $f' \in \mathcal{F}^\alpha$, fix some element $\mathrm{can}_\alpha(f') \in \mathcal{F}$ (a "canonical element") so that $\|f' - f\|_{\infty,\mathcal{X}} \leq \alpha$. We may extend this definition to elements of $\Delta^\circ(\mathcal{F}^\alpha)$ as follows: for $\bar{f}' = \sum_{i=1}^K w_i \cdot \delta_{f_i'}$, $f_i' \in \mathcal{F}^\alpha$, set $\mathrm{can}_\alpha(\bar{f}') := \sum_{i=1}^K w_i \cdot \delta_{\mathrm{can}_\alpha(f_i')}$. This definition will be used to ensure stability of the learner `Optimistic SOA-Experts`.

Lemma 6.4 bounds the α-sequential fat-shattering dimension of an augmented class in terms of that of the original class.

Lemma 6.4. *For any class $\mathcal{F} \subset [0,1]^{\mathcal{X}}$, and any $\alpha > 0$, it holds that, for all $\alpha' \geq 4\alpha$, $\mathrm{sfat}_{\alpha'}(\mathcal{F}^\alpha) \leq \mathrm{sfat}_{\alpha'/2}(\mathcal{F})$.*

Proof. Set $d_0 := \mathrm{sfat}_{\alpha'/2}(\mathcal{F})$. Suppose for the purpose of contradiction that there were some trees $\mathbf{x}, \mathbf{s}$ of depth $d > d_0$ so that for all $k_{1:d} \in \{-1,1\}^d$, there is some $f \in \mathcal{F}^\alpha$ so that $k_t \cdot (f(\mathbf{x}_t(k_{1:t-1})) - \mathbf{s}_t(k_{1:t-1})) \geq \alpha'/2$ for all $t \in [d]$. Then there is some $f' \in \mathcal{F}$ so that $k_t \cdot (f'(\mathbf{x}_t(k_{1:t-1})) - \mathbf{s}_t(k_{1:t-1})) \geq \alpha'/2 - \alpha \geq \alpha'/4$ for all $t \in [d]$, i.e., the trees $\mathbf{x}, \mathbf{s}$ witness an $\alpha'/2$-shattering of $\mathcal{F}$, which is a contradiction to $d_0 < d$. □

Lemma 6.5. *For any $t < T$, the following hold:*

- *For all $E \in \mathscr{E}_{T,\alpha}$, $\|\bar{g}_t(E) - \bar{g}_{t+1}(E)\|_1 \leq \eta_{\mathtt{PSR}}$;*
- *Suppose that $\|x_t - x_{t+1}\|_{\infty,\mathcal{F}} \leq \kappa$ and $|y_t - y_{t+1}| \leq \kappa$, and that `PSR-Learner` is run with step size $\eta_{\mathtt{PSR}}$. Then, for all $E \in \mathscr{E}_{T,\alpha}$, $|\ell_t(E) - \ell_{t+1}(E)| \leq 2\kappa + \eta_{\mathtt{PSR}}$.*

Algorithm 3: `Optimistic SOA-Experts`

Input: Function class $\mathcal{F} \subset [0,1]^{\mathcal{X}}$, time horizon $T \in \mathbb{N}$, scale $\alpha > 0$, step size $\eta > 0$.

1. Set $\mathcal{F}^{\alpha} := \{f \in [0,1]^{\mathcal{X}} : \exists f' \in \mathcal{F} \text{ such that } \|f' - f\|_{\infty,\mathcal{X}} \le \alpha\}$.
2. Initialize $\omega_{E,1} = 1/|\mathscr{E}_{T,\alpha}|$ for all $E \in \mathscr{E}_{T,\alpha}$.
3. For $1 \le t \le T$:
   (a) For each $E \in \mathscr{E}_{T,\alpha}$, define $\bar{h}_t(E) \in \Delta^{\circ}(\mathcal{F}^{\alpha})$ to be the $\eta_{\texttt{PSR}}/2$-smoothed hypotheses (as defined in (37)) of the output of `Multi-scale Proper Learner` (Algorithm 2) given the class $\mathcal{F}^{\alpha}$, the parameter $\Lambda = \lfloor \log 1/(4\alpha) \rfloor$, and the input sequence $(x_1, E(x_1)), (x_2, E(x_{1:2})), \ldots, (x_{t-1}, E(x_{1:t-1}))$.
   (b) For each $E \in \mathscr{E}_{T,\alpha}$, set $\bar{g}_t(E) := \mathrm{can}_{\alpha}(\bar{h}_t(E)) \in \Delta^{\circ}(\mathcal{F})$ to be canonical randomized hypothesis for $\bar{h}_t(E)$.
   (c) Predict the hypothesis $\bar{f}_t := \sum_{E \in \mathscr{E}_{T,\alpha}} \omega_{E,t} \cdot \bar{g}_t(E) \in \Delta^{\circ}(\mathcal{F})$.
   (d) Receive (x_t, y_t), draw $f_t \sim \bar{f}_t$ and suffer loss $|f_t(x_t) - y_t|$.
   (e) For each expert $E \in \mathscr{E}_{T,\alpha}$, compute the loss $\ell_t(E) := \mathbb{E}_{g \sim \bar{g}_t(E)}[|g(x_t) - y_t|]$.
   (f) Update the weights $\{\omega_{E,t}\}_{E \in \mathscr{E}_{T,\alpha}}$ using Optimistic Exponential Weights, i.e.,

$$\omega_{E,t+1} := \frac{\omega_{E,t} \cdot \exp\left(-\eta \cdot (2\ell_t(E) - \ell_{t-1}(E))\right)}{\sum_{E' \in \mathscr{E}_{T,\alpha}} \omega_{E',t} \cdot \exp\left(-\eta \cdot (2\ell_t(E') - \ell_{t-1}(E'))\right)}.$$

Proof. By Theorem 5.15, we have that $\left\|\bar{h}_t(E) - \bar{h}_{t+1}(E)\right\|_1 \le \eta_{\texttt{PSR}}$ for all experts E. Thus, by the definition of $\mathrm{can}_\alpha(\cdot)$ and the data processing inequality, $\|\bar{g}_t(E) - \bar{g}_{t+1}(E)\|_1 \le \eta_{\texttt{PSR}}$ for all experts E.

We may now compute

$$\begin{aligned}
|\ell_t(E) - \ell_{t+1}(E)| =& \left| \mathbb{E}_{g\sim\bar{g}_t(E)}[|g(x_t) - y_t|] - \mathbb{E}_{g\sim\bar{g}_{t+1}(E)}[|g(x_{t+1}) - y_{t+1}|] \right| \\
\le& \|\bar{g}_t(E) - \bar{g}_{t+1}(E)\|_1 + \left| \mathbb{E}_{g\sim\bar{g}_{t+1}(E)} \left[|g(x_t) - y_t| - |g(x_{t+1}) - y_{t+1}|\right] \right| \\
\le& \|\bar{g}_t(E) - \bar{g}_{t+1}(E)\|_1 + |y_t - y_{t+1}| + \mathbb{E}_{g\sim\bar{g}_{t+1}(E)} \left[|g(x_t) - g(x_{t+1})|\right] \\
\le& \|\bar{g}_t(E) - \bar{g}_{t+1}(E)\|_1 + |y_t - y_{t+1}| + \|x_t - x_{t+1}\|_{\infty,\mathcal{F}} \\
\le& 2\kappa + \eta_{\texttt{PSR}},
\end{aligned}$$

where the final inequality uses that $\|\bar{g}_t(E) - \bar{g}_{t+1}(E)\|_1 \le \eta_{\texttt{PSR}}$, $|y_t - y_{t+1}| \le \kappa$, and $\|x_t - x_{t+1}\|_{\infty,\mathcal{F}} \le \kappa$. □

Finally, we are ready to prove Theorem 6.1.

Proof of Theorem 6.1. Without loss of generality we may assume that $\kappa \ge (\mathrm{sfat}_\alpha(\mathcal{F})/T)^{1/4} \cdot \log^{3/4} T$ (since the expression on the right-hand side of (42) is minimized at $\kappa = (\mathrm{sfat}_\alpha(\mathcal{F})/T)^{1/4} \cdot \log^{3/4} T$, meaning that we can make κ larger if it is less than $(\mathrm{sfat}_\alpha(\mathcal{F})/T)^{1/4} \cdot \log^{3/4} T$).

By Lemma 6.5 and the fact that $\max\{\alpha, \eta_{\texttt{OH}}, \eta_{\texttt{PSR}}\} \le \kappa$, we have that for each $t < T$ and each expert $E \in \mathscr{E}_{T,\alpha}$, $|\ell_t(E) - \ell_{t+1}(E)| \le 5\kappa$. Set

$$f_\star = \arg\min_{f\in\mathcal{F}} \sum_{t=1}^{T} |f(x_t) - y_t|.$$

By Lemma 6.3, there is some expert $E_\star \in \mathscr{E}_{T,\alpha}$ so that for all $t \in [T]$, $|f_\star(x_t) - E_\star(x_1, \dots, x_t)| \le \alpha$. Thus, there is some $f^\alpha_\star \in \mathcal{F}^\alpha$ so that for all $t \in [T]$, $f^\alpha_\star(x_t) = E_\star(x_1, \dots, x_t)$. By Theorem 5.15 with $\eta = \eta_{\texttt{PSR}}/2$, it follows that

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}_{g\sim\bar{g}_t(E_\star)} \left[|g(x_t) - E_\star(x_1, \dots, x_t)|\right] \le& \alpha T + \sum_{t=1}^{T} \mathbb{E}_{h\sim\bar{h}_t(E_\star)} \left[|h(x_t) - E_\star(x_1, \dots, x_t)|\right] \\
\le& O\left(\frac{\log^6 T}{\eta_{\texttt{PSR}}} \cdot (\alpha T + \mathrm{sfat}_{4\alpha}(\mathcal{F}^\alpha))\right) \\
\le& O\left(\frac{\log^6 T}{\eta_{\texttt{PSR}}} \cdot (\alpha T + \mathrm{sfat}_\alpha(\mathcal{F}))\right), \qquad (43)
\end{aligned}$$

where the final inequality above follows from Lemma 6.4.

By [SALS15, Theorem 11], we have that

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}_{f_t \sim \bar{f}_t}\left[|f_t(x_t) - y_t|\right] &= \sum_{t=1}^{T} \sum_{E \in \mathscr{E}_{T,\alpha}} \omega_{E,t} \cdot \mathbb{E}_{g \sim \bar{g}_t(E)}[|g(x_t) - y_t|] \\
&\leq \min_{E \in \mathscr{E}_{T,\alpha}} \left\{ \sum_{t=1}^{T} \mathbb{E}_{g \sim \bar{g}_t(E)}[|g(x_t) - y_t|] \right\} + \frac{\log |\mathscr{E}_{T,\alpha}|}{\eta_{\texttt{OH}}} + \eta_{\texttt{OH}} \cdot (5\kappa)^2 \cdot T \\
&\leq \sum_{t=1}^{T} \mathbb{E}_{g \sim \bar{g}_t(E_\star)}[|g(x_t) - y_t|] + \frac{\log |\mathscr{E}_{T,\alpha}|}{\eta_{\texttt{OH}}} + \eta_{\texttt{OH}} \cdot (5\kappa)^2 \cdot T \\
&\leq \sum_{t=1}^{T} \mathbb{E}_{g \sim \bar{g}_t(E_\star)}[|g(x_t) - E_\star(x_1, \ldots, x_t)|] + \sum_{t=1}^{T} |E_\star(x_1, \ldots, x_t) - f_\star(x_t)| \qquad (44) \\
&\quad + \sum_{t=1}^{T} |y_t - f_\star(x_t)| + \frac{\log |\mathscr{E}_{T,\alpha}|}{\eta_{\texttt{OH}}} + \eta_{\texttt{OH}} \cdot (5\kappa)^2 \cdot T \\
&\leq \sum_{t=1}^{T} |y_t - f_\star(x_t)| + O\left(\frac{\log^6 T}{\eta_{\texttt{PSR}}} \cdot (\alpha T + \mathrm{sfat}_\alpha(\mathcal{F})) \right) \qquad (45) \\
&\quad + O\left(\frac{\mathrm{sfat}_\alpha(\mathcal{F}) \cdot \log(T/\alpha)}{\eta_{\texttt{OH}}} \right) + \eta_{\texttt{OH}} \cdot (5\kappa)^2 \cdot T,
\end{aligned}$$

where (44) uses the triangle inequality and (45) uses (43) and Lemma 6.2 (which bounds $|\mathscr{E}_{T,\alpha}|$).

By choosing $\eta_{\texttt{OH}} = \eta_{\texttt{PSR}} = \kappa/\Gamma \leq \kappa$ and using that $\alpha \geq 1/T$ and $\alpha T \leq \mathrm{sfat}_\alpha(\mathcal{F})$, we obtain

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}_{f_t \sim \bar{f}_t}[|f_t(x_t) - y_t|] - \sum_{t=1}^{T} |f_\star(x_t) - y_t| &\leq O\left(\frac{\Gamma \cdot \log^6 T}{\kappa} \cdot (\alpha T + \mathrm{sfat}_\alpha(\mathcal{F})) + \frac{\kappa^3 \cdot T}{\Gamma} \right) \\
&\leq O\left(\frac{\Gamma \cdot \mathrm{sfat}_\alpha(\mathcal{F}) \cdot \log^6 T}{\kappa} + \frac{\kappa^3 \cdot T}{\Gamma} \right).
\end{aligned}$$

Finally, when $\kappa = \frac{1}{\Gamma} \cdot (\mathrm{sfat}_\alpha(\mathcal{F})/T)^{1/4} \cdot \log^{3/4} T$, we obtain

$$\sum_{t=1}^{T} \mathbb{E}_{f_t \sim \bar{f}_t}[|f_t(x_t) - y_t|] - \sum_{t=1}^{T} |f_\star(x_t) - y_t| \leq \tilde{O}\left(\Gamma \cdot \mathrm{sfat}_\alpha(\mathcal{F})^{3/4} \cdot T^{1/4} \right).$$

Since each $\bar{f}_t$ is a finite convex combination of the collection of $\bar{g}_t(E)$, each of which is an element of $\Delta^\circ(\mathcal{F})$, it holds that $\bar{f}_t \in \Delta^\circ(\mathcal{F})$ as well. Finally we bound the stability of the iterates $\bar{f}_t$: for

any $t < T$,

$$\begin{aligned}
\left\|\bar{f}_t - \bar{f}_{t+1}\right\|_1 &= \left\| \sum_{E \in \mathscr{E}_{T,\alpha}} \left(\omega_{E,t} \cdot \bar{g}_t(E) - \omega_{E,t+1} \cdot \bar{g}_{t+1}(E)\right) \right\|_1 \\
&\leq \left\| \sum_{E \in \mathscr{E}_{T,\alpha}} \left(\omega_{E,t} \cdot \bar{g}_t(E) - \omega_{E,t+1} \cdot \bar{g}_t(E)\right) \right\|_1 + \left\| \sum_{E \in \mathscr{E}_{T,\alpha}} \left(\omega_{E,t+1} \cdot \bar{g}_t(E) - \omega_{E,t+1} \cdot \bar{g}_{t+1}(E)\right) \right\|_1 \\
&\leq \sum_{E \in \mathscr{E}_{T,\alpha}} |\omega_{E,t} - \omega_{E,t+1}| + \sum_{E \in \mathscr{E}_{T,\alpha}} \omega_{E,t+1} \cdot \|\bar{g}_t(E) - \bar{g}_{t+1}(E)\|_1 \\
&\leq (\exp(2\eta_{\texttt{OH}}) - 1) \cdot \frac{1}{\exp(-2\eta_{\texttt{OH}})} + \max_{E \in \mathscr{E}_{T,\alpha}} \|\bar{g}_t(E) - \bar{g}_{t+1}(E)\|_1 \qquad (46) \\
&\leq 5\eta_{\texttt{OH}} + 3\eta_{\texttt{PSR}}, \qquad (47)
\end{aligned}$$

where (46) follows from the Optimistic Exponential Weights updates in step 3f of Algorithm 3, and (47) follows from the fact that $\exp(4\eta) - \exp(2\eta) \leq 5\eta$ for $0 < \eta \leq 1/4$, and Lemma 6.5.

□

7 Fast rates for learning in games

In this section we present a key application of the stable proper learner `Optimistic SOA-Experts` in Section 6: we show that when multiple agents in a game each run the algorithm `Optimistic SOA-Experts`, then they can converge to equilibrium at faster rates than the typical $1/\sqrt{T}$ ones.

7.1 Problem setting: Littlestone games

We begin by defining the notion of games we consider, which generalizes finite-action normal form games to the case of extremely large or infinite action spaces. Further, we focus on the case of games for which the payoff for each player is in $\{0, 1\}$ under any pure strategy profile; our setup generalizes that of [HLM21], which considered the special case of 2-player 0-sum Littlestone games.

Definition 7.1 (General-sum Littlestone games)**.** Consider any integer $K \in \mathbb{N}$, denoting the number of players, and sets $\mathcal{F}_1, \ldots, \mathcal{F}_K$. Write $\mathcal{F}_{-k} := \prod_{j \in [K] \setminus \{k\}} \mathcal{F}_j$. A function $\ell : \mathcal{F}_1 \times \cdots \mathcal{F}_K \to \{0, 1\}$ is said to define a *Littlestone payoff function* if the following holds: for each $k \in [K]$, the class

$$\mathcal{F}_k^{\ell} := \{f_{-k} \mapsto \ell(f_k, f_{-k}) \ : \ f_k \in \mathcal{F}_k\} \subset \{0, 1\}^{\mathcal{F}_{-k}} \qquad (48)$$

has finite Littlestone dimension. A *Littlestone (general-sum)* game is a K-tuple of Littlestone payoff functions, namely a tuple $\ell = (\ell_1, \ldots, \ell_K)$. We say that *Littlestone dimension of the game* is $\max_{k \in [K]} \{\mathrm{Ldim}(\mathcal{F}_k^{\ell_k})\}$.

For $k \in [K]$, the payoff function ℓ_k in Definition 7.1 denotes the payoff function for player k in the Littlestone game. It is immediate that all finite-action normal form games are Littlestone games.

Before proceeding, we make the following definition: for a class $\mathcal{F} \subset \{0,1\}^{\mathcal{X}}$, define a class $\mathrm{mix}(\mathcal{F}) \subset [0,1]^{\Delta^\circ(\mathcal{X})}$, in bijection with $\mathcal{F}$, as follows: for each $f \in \mathcal{F}$, the corresponding $f \in \mathrm{mix}(\mathcal{F})$ is defined by, for $P \in \Delta^\circ(\mathcal{X})$, $f(P) := \mathbb{E}_{x\sim P}[f(x)]$.

To help describe how we apply `Optimistic SOA-Experts` in the context of learning in Littlestone games, we need to make an additional definition: for a Littlestone game $\ell = (\ell_1, \dots, \ell_K)$ with action sets $\mathcal{F}_1, \dots, \mathcal{F}_K$, for each $k \in [K]$, define the *loss set* $\mathcal{L}_k^{\ell_k}$ of player k as follows:

$$\mathcal{L}_k^{\ell_k} := \{f_k \mapsto \ell(f_k, f_{-k}) \ : \ f_{-k} \in \mathcal{F}_{-k}\} \subset \{0,1\}^{\mathcal{F}_k}.$$

In words, $\mathcal{L}_k^{\ell_k}$ is the set of mappings from f_k to $\{0,1\}$ which may be realized as the loss of player k given some valid actions of all other players. To avoid confusion, we denote elements of $\mathcal{L}_k^{\ell_k}$ with a capital L. It is evident that $\mathcal{L}_k^{\ell_k}$ is the dual class of $\mathcal{F}_k^{\ell_k}$; thus we may view $\mathcal{F}_k^{\ell_k}$ as a set of mappings from $\mathcal{L}_k^{\ell_k}$ to $\{0,1\}$, i.e., for $f_k \in \mathcal{F}_k^{\ell_k}$ and $L_k \in \mathcal{L}_k^{\ell_k}$, we have $f_k(L_k) := L_k(f_k)$.

We consider the following *independent learning setting* in Littlestone games, which directly generalizes the setting of independent learning in normal-form games. Consider a Littlestone game $\ell = (\ell_1, \dots, \ell_K)$, with action sets $\mathcal{F}_1, \dots, \mathcal{F}_K$:

- For each time step $1 \le t \le T$:
  1. Each player k plays a distribution over actions $\bar{f}_k^t \in \Delta^\circ(\mathcal{F}_k)$.
  2. Each player k observes its loss function $L_k^t \in \Delta^\circ(\mathcal{L}_k^{\ell_k})$ at time step t, namely the mapping $L_k^t(f_k) := \mathbb{E}_{f_{-k}\sim \bar{f}_{-k}^t}[\ell_k(f_k, f_{-k})]$.
  3. Each player k suffers loss $\ell_k(\bar{f}_k^t, \bar{f}_{-k}^t)$; notice that this loss value may also be written as $\bar{f}_k^t(L_k^t)$, by viewing $\bar{f}_k^t$ as an element of $\Delta^\circ(\mathrm{mix}(\mathcal{F}_k^{\ell_k}))$.

7.2 Independent learning algorithm for fast rates in games

Lemma 7.1. *Given a class $\mathcal{F} \subset \{0,1\}^{\mathcal{X}}$, it holds that* $\mathrm{sfat}_\alpha(\mathrm{mix}(\mathcal{F})) \le O\left(\mathrm{Ldim}(\mathcal{F}) \cdot \log(\mathrm{Ldim}(\mathcal{F})/\alpha)\right)$.

Proof. Denote $L = \mathrm{Ldim}(\mathcal{F}), V = \mathrm{VCdim}(\mathcal{F})$. Suppose $\mathbf{p}$ is an α-shattered $\Delta^\circ(\mathcal{X})$-valued tree of depth d for $\mathrm{mix}(\mathcal{F})$, witnessed by $\mathbf{s}$. For some constant $C > 1$, let us form a new $\mathcal{X}$-valued tree, $\mathbf{p}'$, by replacing each node v of $\mathbf{p}$, labeled by P_v, with a new tree $\mathbf{t}_v$ of depth $m := \lceil C \cdot V/\alpha^2 \rceil$. For $i \in [m]$, each node on the ith level of $\mathbf{t}_v$ is labeled by x_v^i, where the points $x_v^1, \dots, x_v^m \in \mathcal{X}$ satisfy the following: for all $f \in \mathcal{F}$,

$$\left| \mathbb{E}_{x\sim P_v}[f(x)] - \frac{1}{m}\sum_{i=1}^m f(x_v^i) \right| \le \frac{\alpha}{4}. \tag{49}$$

By classic uniform convergence bounds [Tal94, vdVW96] such points $x_v^1, \dots, x_v^m \in \mathcal{X}$ exist as long as C is sufficiently large (this holds even in the absence of additional measurability assumptions on $\mathcal{X}$ since P_v is finite-support). Let $s_v \in [0,1]$ be the label of the node of $\mathbf{s}$ corresponding to node v of $\mathbf{p}$. For each of the 2^m leaves of the tree $\mathbf{t}_v$, indexed by $(\delta_1, \dots, \delta_m) \in \{-1,1\}^m$, we will assign to each such leaf the subbtree rooted by either the left (-1) or right $(+1)$ child of v, as follows: if $\frac{1}{m} \cdot \sum_{i=1}^m \left(\frac{1+\delta_i}{2}\right) \ge s_v$, then use the subtree rooted by the right child of v, and otherwise use the subtree rooted by the left child of v. Formally, we have the following: for any sequence

$\delta_1, \dots, \delta_{dm} \in \{-1,1\}^{dm}$, and any $i \in [dm]$, writing $i = mt + j$ for $1 \le j \le m$, then $\mathbf{p}'_i(\delta_{1:i-1}) = x_v^j$, where v is the node of $\mathbf{p}$ corresponding to the sequence $(\epsilon_1, \dots, \epsilon_t)$, where

$$\epsilon_\ell = \operatorname{sign}\left(\frac{1}{m} \cdot \sum_{i=1}^{m} \left(\frac{1+\delta_{(\ell-1)m+i}}{2}\right) - s_{(\epsilon_1,\dots,\epsilon_{\ell-1})}\right) \qquad \forall \ell \in [t]. \tag{50}$$

The depth of the new tree $\mathbf{p}'$ we have constructed is dm. By (49) and (50), $\mathbf{p}'$ satisfies the following property: for any $t \in [d]$, and $\epsilon \in \{-1,1\}^d$, consider any function $f \in \mathcal{F}$ so that for $i < t$, $\epsilon_i \cdot (f(\mathbf{p}_i(\epsilon_{1:i-1})) - \mathbf{s}_i(\epsilon_{1:i-1})) > \alpha/2$. Let v be the node of $\mathbf{p}$ corresponding to the sequence $\epsilon_1, \dots, \epsilon_{t-1}$, $P_v = \mathbf{p}_t(\epsilon_{1:t-1})$ (as above), and consider the sequence $x_v^1, \dots, x_v^m$. Now define the sequence $\delta \in \{-1,1\}^{dm}$ inductively via $\delta_i = 2 \cdot f(\mathbf{p}'_i(\delta_{1:i-1})) - 1$ for $i \ge 1$. Then the sequence $\mathbf{p}'_{tm+1}(\delta_{1:tm}), \dots, \mathbf{p}'_{tm+m}(\delta_{1:tm+m-1})$ is exactly the sequence $x_v^1, \dots, x_v^m$; we will say that *f, f' encounter the sequence $x_v^1, \dots, x_v^m$ in the tree $\mathbf{p}'$.*

Now consider $f, f' \in \operatorname{mix}(\mathcal{F})$ which lead to different leaves of the tree $\mathbf{p}$, in the sense that there are $\epsilon \ne \epsilon' \in \{-1,1\}^d$ so that, for each $t \in [d]$, $\epsilon_t \cdot (f(\mathbf{p}_t(\epsilon_{1:t-1})) - \mathbf{s}_t(\epsilon_{1:t-1})) > \alpha/2$ and $\epsilon'_t \cdot (f'(\mathbf{p}_t(\epsilon'_{1:t-1})) - \mathbf{s}_t(\epsilon'_{1:t-1})) > \alpha/2$. Let $t_0 \in [d]$ be as small as possible so that $\epsilon_{t_0} \ne \epsilon'_{t_0}$, and let v be the node of $\mathbf{p}$ corresponding to the sequence $\epsilon_1, \dots, \epsilon_{t_0-1}$; let $P_v = \mathbf{p}_{t_0}(\epsilon_{1:t_0-1}) \in \Delta(\mathcal{X})$ be the label of v and $s_v = \mathbf{s}_{t_0}(\epsilon_{1:t_0-1}) \in [0,1]$ be the label of the corresponding node of $\mathbf{s}$, and $\mathbf{t}_v$ be the tree constructed in place of v (as above). By the choice of v, it holds that $f(P_v) > s_v + \alpha/2$ and $f'(P_v) < s_v - \alpha/2$. Thus, letting $x_v^1, \dots, x_v^m$ be the sequence constructed as above for the node v, we have $\sum_{i=1}^m f(x_v^i) > s > \sum_{i=1}^m f'(x_v^i)$.

Thus f, f' lead to different leaves of the tree $\mathbf{t}_v$, and hence (since f, f' both encounter the sequence $x_v^1, \dots, x_v^m$ in the tree $\mathbf{p}'$) also to different leaves of the tree $\mathbf{p}'$. Thus the sequential 0-covering number of the tree $\mathbf{p}'$ (see [RS14b, Definition 13.2]) is at least 2^d.[14] On the other hand, by the Sauer-Shelah lemma for trees [RS14b, Theorem 13.7], the sequential 0-covering number of the (depth-dm) tree $\mathbf{p}'$ is at most $(edm)^L$.

Summarizing, we have that $2^d \le (edm)^L$, i.e., $d \le L \log(edm)$, meaning that $d \le O(L \log(Lm)) \le O(L \log(LV/\alpha^2)) \le O(L \log(L/\alpha))$. ☐

Theorem 7.2. *Fix a Littlestone game with K players and a time horizon T. If the players play according to Algorithm 4 with each player using the algorithm* `Optimistic SOA-Experts` *(Algorithm 3) with step sizes $\eta_{\texttt{PSR}}, \eta_{\texttt{OH}}$ as in (51) below and scale $\alpha = 1/T$, then each player $k \in [K]$ suffers regret $\tilde{O}(\operatorname{Ldim}(\mathcal{F}_k^{\ell_k})^{3/4} \cdot \sqrt{K} \cdot T^{1/4})$, where the $\tilde{O}(\cdot)$ hides logarithmic factors in T and $\operatorname{Ldim}(\mathcal{F}_k^{\ell_k})$.*

Proof. Set

$$\eta = \eta_{\texttt{PSR}} = \eta_{\texttt{OH}} = \frac{\operatorname{Ldim}(\mathcal{F}_k^{\ell_k}) \cdot \log(\operatorname{Ldim}(\mathcal{F}_k^{\ell_k}) \cdot T)}{K^{1/2} \cdot T^{1/4}} \tag{51}$$

and $\alpha = 1/T$. Also write $D_k = \operatorname{Ldim}(\mathcal{F}_k^{\ell_k})$. In Algorithm 4, each player k applies `Optimistic SOA-Experts` with function class $\mathcal{F} = \operatorname{mix}(\mathcal{F}_k^{\ell_k})$ with feature space $\mathcal{X} = \Delta^\circ(\mathcal{L}_k^{\ell_k})$; by Lemma 7.1, it holds that $\operatorname{sfat}_\alpha(\operatorname{mix}(\mathcal{F}_k^{\ell_k})) \le O(D_k \cdot \log(D_k T))$.

By Theorem 6.1, the hypotheses $\bar{f}_k^t \in \Delta^\circ(\mathcal{F}_k^{\ell_k})$ output by each player k satisfy $\left\|\bar{f}_k^t - \bar{f}_k^{t+1}\right\|_1 \le 8\eta$.

[14] In more detail, what we have directly shown is that the *thicket shatter function* of the tree $\mathbf{p}'$ is at least 2^d; then [GGKM21, Lemma 2.7] implies that the sequential 0-covering number of the tree $\mathbf{p}'$ is at least 2^d.

Algorithm 4: `Independent Learning in a Game`

Input: Littlestone game $\ell = (\ell_1, \dots, \ell_K)$ with action sets $\mathcal{F}_1, \dots, \mathcal{F}_K$, time horizon $T \in \mathbb{N}$.
Input to each player: Each player $k \in [K]$ only knows its action set $\mathcal{F}_k$ and its loss class $\mathcal{L}_k^{\ell_k}$, as well as the horizon T.

1. Each player $k \in [K]$ initializes some online proper learning algorithm $\mathscr{A}_k$ (e.g., `Optimistic SOA-Experts`, Algorithm 3) with function class $\mathcal{F} = \mathrm{mix}(\mathcal{F}_k^{\ell_k})$ and feature space $\mathcal{X} = \Delta^\circ(\mathcal{L}_k^{\ell_k})$.
2. For $1 \le t \le T$:
   (a) Each player $k \in [K]$ plays a distribution $\bar{f}_k^t \in \Delta^\circ(\mathcal{F}_k)$ according to their respective algorithm $\mathscr{A}_k$.
   (b) Each player $k \in [K]$ observes the loss function $L_k^t \in \Delta^\circ(\mathcal{L}_k^{\ell_k}) = \mathcal{X}$ (defined as $L_k^t(f_k) = \mathbb{E}_{f_{-k} \sim \bar{f}_{-k}^t}[\ell_k(f_k, f_{-k})]$), and feeds the example $(L_k^t, 0)$ to its algorithm $\mathscr{A}_k$.
   (c) Each player k suffers loss $\ell_k(\bar{f}^t) = \mathbb{E}_{f \sim \bar{f}_k^t}[f(L_k^t)]$.

Now let us consider any player $k \in [K]$; by symmetry we may assume $k = 1$; then for any $t < T$ and any $f_1 \in \mathcal{F}_1$, abbreviating $f = (f_1, \dots, f_K)$, we have

$$
\begin{aligned}
&|L_1^{t+1}(f_1) - L_1^t(f_1)| \\
\le &\left| \mathbb{E}_{f_2 \sim \bar{f}_2^{t+1}, \dots, f_K \sim \bar{f}_K^{t+1}} [\ell_1(f)] - \mathbb{E}_{f_2 \sim \bar{f}_2^t, \dots, f_K \sim \bar{f}_K^t} [\ell_1(f)] \right| \\
\le &\sum_{j=2}^{K} \left| \mathbb{E}_{f_2 \sim \bar{f}_2^{t+1}, \dots, f_j \sim \bar{f}_j^{t+1}, f_{j+1} \sim \bar{f}_{j+1}^t, \dots, f_K \sim \bar{f}_K^t} [\ell_1(f)] - \mathbb{E}_{f_2 \sim \bar{f}_2^{t+1}, \dots, f_{j-1} \sim \bar{f}_{j-1}^{t+1}, f_j \sim \bar{f}_j^t, \dots, f_K \sim \bar{f}_K^t} [\ell_1(f)] \right| \\
\le &\sum_{j=2}^{K} \left| \mathbb{E}_{f_2 \sim \bar{f}_2^{t+1}, \dots, f_{j-1} \sim \bar{f}_j^{t+1}, f_{j+1} \sim \bar{f}_{j+1}^t, \dots, f_K \sim \bar{f}_K^t} \left[(\mathbb{E}_{f_j \sim \bar{f}_j^{t+1}} - \mathbb{E}_{f_j \sim \bar{f}_j^t})[\ell_1(f)] \right] \right| \\
\le &\sum_{j=2}^{K} \left\| \bar{f}_j^{t+1} - \bar{f}_j^t \right\|_1 \\
\le &8\eta K.
\end{aligned}
$$

It follows that $\left\| L_1^{t+1} - L_1^t \right\|_{\infty, \mathcal{F}_1^{\ell_1}} \le 8\eta K$ for all $t < T$. Since the choice of player $k = 1$ here is arbitrary, we have in a similar manner that for all $k \in [K]$, $\left\| L_k^{t+1} - L_k^t \right\|_{\infty, \mathcal{F}_k^{\ell_k}} \le 8\eta K$. Since, by assumption, each player runs `Optimistic SOA-Experts` with step size $\eta_{\texttt{OH}} = \eta_{\texttt{PSR}} = \eta$, we may apply Theorem 6.1 with $\kappa = 8\eta K$ and $\Gamma = 8K$ to obtain that each player's regret is bounded above by

$$
O\left(\frac{K \cdot \mathrm{sfat}_\alpha(\mathcal{F}_k^{\ell_k}) \cdot \log^3 T}{\eta K} + \eta^3 K^2 T \right) \le O\left(\sqrt{K} \cdot T^{1/4} \cdot D_k^{3/4} \cdot \log^3(D_k T) \right).
$$

□

8 On real-valued games satisfying the minimax theorem

In this section we show that all online learnable (real-valued) classes satisfy the minimax theorem, in the absense of any topological assumptions on the class $\mathcal{F}$ or the space $\mathcal{X}$, thus generalizing a corresponding result from [HLM21] which treated the binary setting.

8.1 Additional preliminaries

We first introduce some additional preliminaries. We begin by describing a way to discretize a hypothesis class $\mathcal{F} \subset [0,1]^{\mathcal{X}}$ at some scale $\eta > 0$. Roughly speaking, this is done by subdividing the interval $[0,1]$ into $\lceil 1/\eta \rceil$ intervals each of length $1/\lceil 1/\eta \rceil \leq \eta$, and rounding the output of each hypothesis to its interval. Formally, we make the following definitions: For a real number $y \in [0,1]$, define the discretiztion of y at scale η, denoted $\lfloor y \rfloor_\eta \in \{1/\lceil 1/\eta \rceil, 2/\lceil 1/\eta \rceil, \dots, 1\}$, as follows: $\lfloor y \rfloor_\eta := \frac{1}{\lceil 1/\eta \rceil} \cdot (1 + \lfloor y \cdot \lceil 1/\eta \rceil \rfloor)$ for $0 \leq y < 1$ and $\lfloor y \rfloor_\eta = 1$ for $y = 1$. It is straightforward from this definition that for all $y \in [0,1]$,

$$|y - \lfloor y \rfloor_\eta| \leq 1/\lceil 1/\eta \rceil \leq \eta.$$

Definition 8.1 (Thresholds with margin; similar to [JKT20], Definition 7)**.** Consider a hypothesis class $\mathcal{F} \subset [0,1]^{\mathcal{X}}$, $\alpha > 2\beta > 0$, and $d \in \mathbb{N}$. $\mathcal{F}$ is said to contain d *thresholds with margin* α *and tightness* β (respectively, *infinitely many thresholds with margin* α and tightness β) if there are $x_1, \dots, x_d \in \mathcal{X}$ and $f_1, \dots, f_d \in \mathcal{F}$ (respectively, $x_1, x_2, \dots \in \mathcal{X}$ and $f_1, f_2, \dots \in \mathcal{F}$) as well as $u, u' \in [0,1]$ so that:

- $|u - u'| \geq \alpha$;
- $|f_i(x_j) - u| \leq \beta$ for $i \leq j$ and $|f_i(x_j) - u'| \leq \beta$ for $i > j$.

We further say that $\mathcal{F}$ contains d (or infinitely) many *ordered* thresholds with margin α and tightness β if the above conditions hold and furthermore $u' > u$.

The following lemma, which gives a lower bound on the sequential fat-shattering dimension for a class with many thresholds, is standard, but we include a proof for completeness.

Lemma 8.1. *Suppose that $\mathcal{F}$ contains d thresholds with margin α and tightness β. Then* $\mathrm{sfat}_{\alpha - 2\beta}(\mathcal{F}) \geq \lfloor \log d \rfloor$.

Proof. The proof closely follows the analogous result for Littlestone dimension (see [She78, HH97, ALMM19]). Set $m = \lfloor \log d \rfloor$, and suppose that $x_1, \dots, x_{2^m}$ and $f_1, \dots, f_{2^m}$ are a collection of 2^m thresholds with margin α and tightness β, together with the values u, u' as in Definition 8.1; we may assume $u' > u$ without loss of generality (otherwise we can reverse the order of the thresholds). We construct a tree $\mathbf{x}$ of depth m that is shattered (together with the witness tree $\mathbf{s}$) as follows: the labels of the tree $\mathbf{x}$ correspond to the binary search process on $[2^m]$, so that $\mathbf{x}_t(\epsilon_1, \dots, \epsilon_{t-1}) = x_{2^{m-1} + \epsilon_1 \cdot 2^{m-2} + \dots + \epsilon_{t-1} \cdot 2^{m-t}}$. All nodes of the tree $\mathbf{s}$ are labeled by $(u+u')/2$. It is straightforward to see that the function f_i leads to the leaf which is i spots from the left (viewing -1 as the left child and 1 as the right child for each node).

The lower bound on the sequential fat-shattering dimension then follows from the fact that for $i > j$, we have $f_i(x_j) \geq (u+u')/2 + \alpha/2 - \beta = (u+u')/2 + (\alpha - 2\beta)/2$ and for $i \leq j$, we have $f_i(x_j) \leq (u+u')/2 - (\alpha - 2\beta)/2$. ∎

Lemma 8.2 below provides a sort of converse to Lemma 8.1, giving a lower bound on the number of thresholds in a real-valued class of large sequential fat-shattering dimension.

Lemma 8.2. *For some constant $c > 0$ the following holds. Suppose that $\alpha \geq 4\eta > 0, d \in \mathbb{N}$ are so that $\mathrm{sfat}_\alpha(\mathcal{F}) \geq d$. Then $\mathcal{F}$ contains $c \cdot \frac{\eta \log(\eta \log d)}{\log 1/\eta}$ thresholds with margin $\alpha/4$ and tightness η.*

A similar result to Lemma 8.2 was claimed in [JKT20, Theorem 8], though with a stronger quantitative bound (namely, the lower bound on the number of thresholds was $\Omega_\eta(\log d)$, not $\Omega_\eta(\log \log d)$, as we show). Unfortunately, there appears to be a gap in the proof [JKT20, Theorem 8]: in particular, the proof of Proposition 5 in [JKT20] (which is used to prove Theorem 8) begins with the following claim: "Since $\mathrm{sfat}_\eta(\mathcal{F}) \geq d$, in the online learning setting an adversary can force any deterministic learner to suffer $\eta/2$ loss for d rounds." This sentence is incorrect, even if the adversary only reveals the discretized labels to the learner: in particular, fix $\eta > 0$, $X := \log(\lfloor 1/\eta \rfloor/2)$, and set $\mathcal{X} = \{1, 2, \dots, X\}$. Consider the following class $\mathcal{F}$ which consists of 2^X hypotheses: for each $(\epsilon_1, \dots, \epsilon_X) \in \{-1, 1\}^X$, let $n(\epsilon) \in \{1, 2, \dots, \lfloor 1/\eta \rfloor/2\}$ be the integer corresponding to ϵ in base 2. Then there is a hypothesis $f_\epsilon \in \mathcal{F}$ so that $f_\epsilon(i) = \frac{1}{2} + \epsilon_i \cdot \eta \cdot n(\epsilon)$ for each $i \in \mathcal{X}$. It is evident that $\mathrm{sfat}_\eta(\mathcal{F}) \geq \Omega(\log 1/\eta)$, yet no matter which point $x_1 \in \mathcal{X}$ which the adversary first reveals to the learner, the value $\lfloor f^\star(x_1) \rfloor_\eta$ reveals the identity of $f^\star$, meaning that the learner will always make at most 1 mistake. This gap is filled in our proof of Lemma 8.2, at the cost of a weaker quantitative bound; the question of whether $\mathcal{F}$ contains $\Omega_\eta(\log d)$ thresholds (in the context of Lemma 8.2) is left open for future work.

Proof of Lemma 8.2. We will first construct a weaker notion of a collection of thresholds: in particular, we will first prove the following claim:

Claim 8.3. *For some constant $C_1 > 0$, the following holds. Fix $\alpha \geq 4\eta > 0$ and $m \in \mathbb{N}$. If $\mathrm{sfat}_\alpha(\mathcal{F}) \geq (C_1/\eta)^m$, then there is a collection of hypotheses $f_1, \dots, f_m \in \mathcal{F}$ and points $x_1, \dots, x_m \in \mathcal{X}$ so that, for each $i \in [m]$, the following holds: there are some $\nu_i, \mu_i \in \mathcal{D}_\eta$ satisfying $|\nu_i - \mu_i| \geq \alpha/4$, so that one of the below options holds:*

- *$\nu_i > \mu_i$; moreover, for $j \geq i$, we have $\lfloor f_i(x_j) \rfloor_\eta = \nu_i$ and for $j > i$ we have $\lfloor f_j(x_i) \rfloor_\eta \leq \mu_i$; or*
- *$\nu_i < \mu_i$; moreover, for $j \geq i$, we have $\lfloor f_i(x_j) \rfloor_\eta = \nu_i$, and for $j > i$, we have $\lfloor f_j(x_i) \rfloor_\eta \geq \mu_i$.*

Proof of Claim 8.3. We use induction on m. In the case $m = 1$, then since $\mathrm{sfat}_\alpha(\mathcal{F}) \geq 1$, $\mathcal{F}, \mathcal{X}$ are nonempty so the proof is completed by choosing any $x \in \mathcal{X}$ and $f \in \mathcal{F}$.

Now suppose $m > 1$ and the claim statement holds for all values $m' < m$. Set $d = (C_1/\eta)^m$, and write $\mathcal{D}_\eta := \{1/\lceil 1/\eta \rceil, 2/\lceil 1/\eta \rceil, \dots, 1\}$ to denote the set of discretized points with discretization of η, and $k = \lceil 1/\eta \rceil = |\mathcal{D}_\eta|$. Before continuing, we need to introduce the notion of *subtree*: given a tree $\mathbf{x}$, a subtree of $\mathbf{x}$ of depth t is defined inductively as follows. Any node of $\mathbf{x}$ is a subtree of depth 0. A subtree of depth t is obtained by taking any internal node v of $\mathbf{x}$ together with a subtree of the trees rooted at the left and right children of v. Note that if the tree $\mathbf{x}$ is α-shattered by a hypothesis class $\mathcal{F}$, then so is any subtree of $\mathbf{x}$.

Let $\mathbf{x}$ be an $\mathcal{X}$-valued tree of depth d shattered by $\mathcal{F}$, witnessed by a $[0, 1]$-valued tree $\mathbf{s}$. Let f be an arbitrary hypothesis in $\mathcal{F}$, and define a k-coloring of the nodes of $\mathbf{x}$ as follows: color a node corresponding to the sequence $\epsilon_{1:t-1}$ by the element $\lfloor f(\mathbf{x}_t(\epsilon_{1:t-1})) \rfloor_\eta \in \mathcal{D}_\eta$. By [JKT20, Lemma 16], there is a subtree $\mathbf{x}'$ of $\mathbf{x}$ of depth $d' := \lceil (d+1)/k \rceil \geq d/k$ so that all nodes are colored by some color $\nu^\star \in \mathcal{D}_\eta$. Denote the corresponding subtree of $\mathbf{s}$ by $\mathbf{s}'$. Set $\mathcal{X}' := \{x \in \mathcal{X} : \lfloor f(x) \rfloor_\eta = \nu^\star\}$, so that $\mathbf{x}'$ is $\mathcal{X}'$-valued and is shattered by $\mathcal{F}$, as witnessed by $\mathbf{s}'$.

Define the following subclasses of $\mathcal{F}$, viewed as classes of hypotheses on the restricted set $\mathcal{X}'$:

$$\mathcal{F}_+ := \left\{f \in \mathcal{F} \ : \ f(\mathbf{x}'_1) \geq \mathbf{s}'_1 + \alpha/2\right\} \subset [0,1]^{\mathcal{X}'}$$
$$\mathcal{F}_- := \left\{f \in \mathcal{F} \ : \ f(\mathbf{x}'_1) \leq \mathbf{s}'_1 - \alpha/2\right\} \subset [0,1]^{\mathcal{X}'}.$$

It is immediate that $\mathrm{sfat}_\alpha(\mathcal{F}_+) \geq d' - 1$ and $\mathrm{sfat}_\alpha(\mathcal{F}_-) \geq d' - 1$ (in particular, $\mathcal{X}'$-valued trees shattering $\mathcal{F}_+, \mathcal{F}_-$ are obtained by taking the subtrees of $\mathbf{x}'$ rooted at the right and left children, respectively, of its root). Set $\nu_+ = \lfloor \mathbf{s}'_1 + \alpha/2 \rfloor_\eta \geq \mathbf{s}'_1 + \alpha/2 - \eta$ and $\nu_- = \lfloor \mathbf{s}'_1 - \alpha/2 \rfloor_\eta \leq \mathbf{s}'_1 - \alpha/2 + \eta$.

We must have either $\nu^\star \geq \mathbf{s}'_1$ or $\nu^\star \leq \mathbf{s}'_1$. We consider each of the cases in turn:

- If $\nu^\star \geq \mathbf{s}'_1$, then we apply the inductive hypothesis on the class $\mathcal{F}_-$ and the data (feature) space $\mathcal{X}'$. We have that $\mathrm{sfat}_\alpha(\mathcal{F}_-) \geq d' - 1 \geq \frac{d}{2k} \geq (C_1/\eta)^{m-1}$ (as long as C_1 is chosen sufficiently large), meaning that, by the inductive hypothesis with the value $m-1$ (and the same values of α, η), we can find $f_2, \ldots, f_m \in \mathcal{F}_-$, $x_2, \ldots, x_m \in \mathcal{X}'$ so that the constraints of the claim statement are staisfied. Now we add $f_1 = f$, $x_1 = \mathbf{x}'_1$ to this collection. Note that, for $i \geq 1$, we have $\lfloor f_1(x_i) \rfloor_\eta = \lfloor f(x_i) \rfloor_\eta = \nu^\star$ since all x_i (including $\mathbf{x}'_1$) lie in $\mathcal{X}'$. Further, for $i > 1$, we have $\lfloor f_i(x_1) \rfloor_\eta = \lfloor f_i(\mathbf{x}'_1) \rfloor_\eta \leq \nu_-$ by definition of $\mathcal{F}_-$. Since $|\nu^\star - \nu_-| \geq \alpha/2 - \eta \geq \alpha/4$, we have verified the inductive step in this case; in particular, we may set $\nu_1 = \nu^\star$ and $\mu_1 = \nu_-$ (f_1, x_1 correspond to the first case in the claim statement).

- If $\nu^\star \leq \mathbf{s}'_1$, then we apply exactly the same argument except with $\mathcal{F}_+$ replacing $\mathcal{F}_-$. Again setting $f_1 = f, x_1 = \mathbf{x}'_1$, we have, for $i \geq 1$, $\lfloor f_1(x_i) \rfloor_\eta = \lfloor f(x_i) \rfloor_\eta = \nu^\star$, while for $i > 1$, we have $\lfloor f_i(x_i) \rfloor_\eta = \lfloor f_i(\mathbf{x}'_1) \rfloor_\eta \geq \nu_+$. Since $|\nu^\star - \nu_+| \geq \alpha/2 - \eta \geq \alpha/4$, we have verified the inductive step; in particular, we may set $\nu_1 = \nu^\star$ and $\mu_1 = \nu_+$ (f_1, x_1 now correspond to the second case in the claim statement).

□

Given Claim 8.3, we may now complete the proof of Lemma 8.2, as follows. Given that $\mathrm{sfat}_\alpha(\mathcal{F}) \geq d$, set $m = \left\lfloor \frac{\log d}{\log C_1/\eta} \right\rfloor$, wheree C_1 is the constant of Claim 8.3. Then we may consider a collection $f_1, \ldots, f_m \in \mathcal{F}$ and $x_1, \ldots, x_m \in \mathcal{X}$ satisfying the guarantee of Claim 8.3. Since there are $\lceil 1/\eta \rceil$ possibilities for the value ν_i, for $\ell := m/\lceil 1/\eta \rceil$, we can extract a subset $g_1 := f_{i_1}, \ldots, g_\ell := f_{i_\ell}, w_1 := x_{i_1}, \ldots, w_\ell := x_{i_\ell}$ so that, for some fixed $\nu \in \mathcal{D}_\eta$, $\nu_{i_j} = \nu$ for all $j \in [\ell]$; in particular, for $1 \leq i \leq j \leq m$, it holds that $\lfloor g_i(w_j) \rfloor_\eta = \nu$.

Now we color each tuple (i,j) with $1 \leq i < j \leq \ell$ with the value $\lfloor g_j(w_i) \rfloor_\eta \in \mathcal{D}_\eta$; note that for all such i, j, by our choice of ν, we must have that $|\lfloor g_j(w_i) \rfloor_\eta - \nu| \geq \alpha/4$. By Ramsey's theorem,[15] there is some $\mu \in \mathcal{D}_\eta$ and a sub-collection $h_1 := g_{i_1}, \ldots, h_p := g_{i_p}, v_1 = w_{i_1}, \ldots, v_p := w_{i_p}$ for some $p \geq \frac{\log \ell}{\lceil 1/\eta \rceil \log \lceil 1/\eta \rceil}$, so that for all $1 \leq i < j \leq p$, $\lfloor h_j(v_i) \rfloor_\eta = \mu$. Further, it must be the case that $|\nu - \mu| \geq \alpha/4$.

Summarizing, we have found a collection of thresholds (namely, $h_1, \ldots, h_p, v_1, \ldots, v_p$) with margin $\alpha/4$, tightness η, and of size

$$p \geq \Omega\left(\frac{\eta \log \ell}{\log 1/\eta}\right) \geq \Omega\left(\frac{\eta \log(\eta m)}{\log 1/\eta}\right) \geq \Omega\left(\frac{\eta \log(\eta \log(d))}{\log 1/\eta}\right).$$

□

[15] In particular, we use the following estimate on the multi-color Ramsey numbers [GG55]: for $N \geq c^{rc}$, if the edges of the complete graph on N vertices are colored with c colors, there is a monochromatic clique of size r.

Finally, we may combine Lemmas 8.1 and 8.2 to show that the sequential fat-shattering dimension of a class is finite if and only if the sequential fat-shattering dimension of the dual class is finite.

Lemma 8.4. *Suppose $\mathcal{F} \subset [0,1]^{\mathcal{X}}$. Then for any $\alpha > 0$, the dual class $\mathcal{F}^\star$ satisfies $\mathrm{sfat}_{\alpha/8}(\mathcal{F}^\star) \geq \Omega(\log(\alpha \cdot \log\log(\mathrm{sfat}_\alpha(\mathcal{F}))))$.*

For binary valued classes, it is known (see [Bha21]) that $\mathrm{Ldim}(\mathcal{F}^\star) \geq \Omega(\log\log(\mathrm{Ldim}(\mathcal{F})))$. The additional logarithm in Lemma 8.4 is due to the double logarithm in the lower bound of Lemma 8.2; we leave the question of improving the quantitative bound in Lemma 8.4 to future work.

Proof of Lemma 8.4. Set $\eta = \alpha/16$. Write $d = \mathrm{sfat}_\alpha(\mathcal{F})$. By Lemma 8.2, for some constant $c > 0$, $\mathcal{F}$ contains $m := c \cdot \frac{\alpha \log(\alpha \log d)}{\log 1/\alpha}$ thresholds with margin $\alpha/4$ and tightness η, which we denote $f_1, \ldots, f_m \in \mathcal{F}$, $x_1, \ldots, x_m \in \mathcal{X}$. Thus, the functions in $\mathcal{F}^\star$ corresponding to $x_m, \ldots, x_1$ furnish m thresholds in the dual class on the points $f_m, \ldots, f_m$, with margin $\alpha/4$ and tightness η. Then by Lemma 8.1, we have that

$$\mathrm{sfat}_{\alpha/8}(\mathcal{F}^\star) \geq \lfloor \log m \rfloor \geq \left\lfloor \log\left(c \cdot \frac{\alpha \log(\alpha \log d)}{\log 1/\alpha} \right) \right\rfloor \geq \Omega\left(\log(\alpha \cdot \log\log d)\right).$$

□

8.2 A minimax theorem for online learnable games

In this section we will consider infinite two-player zero-sum games: in particular, fix sets $\mathcal{X}, \mathcal{F}$ and a loss function $\ell : \mathcal{X} \times \mathcal{F} \to [0,1]$. The loss ℓ defines a function class in bijection with $\mathcal{F}$, namely the class $\mathcal{F}^\ell := \{x \mapsto \ell(x,f) \ : \ f \in \mathcal{F}\}$, as well as its dual class $\mathcal{X}^\ell$, namely the class $\{f \mapsto \ell(x,f) \ : \ x \in \mathcal{X}\}$. We say that the game $(\mathcal{X}, \mathcal{F}, \ell)$ is a *GC game* if $\mathrm{fat}_\alpha(\mathcal{F}^\ell) < \infty$ for all $\alpha > 0$ (here "GC" stands for "Glivenko-Cantelli", refelcting the fact that the hypothesis class $\mathcal{F}^\ell$ is a Glivenko-Cantelli class). It is folklore (see [KS21, Corollary 3.8]) that for any real-valued hypothesis class $\mathcal{G} \subset [0,1]^{\mathcal{X}}$, its dual class $\mathcal{G}^\star$ satisfies $\mathrm{fat}_{\alpha/2}(\mathcal{G}^\star) \geq \Omega(\log(\alpha \cdot \mathrm{fat}_\alpha(\mathcal{G})))$. Thus, for a GC game $(\mathcal{X}, \mathcal{F}, \ell)$, we have $\mathrm{fat}_\alpha(\mathcal{X}^\ell) < \infty$ for all α. For $\alpha > 0$, we say that $(\mathcal{X}, \mathcal{F}, \ell)$ is an *α-GC game* if $\max\left\{\mathrm{fat}_\alpha(\mathcal{F}^\ell), \mathrm{fat}_\alpha(\mathcal{X}^\ell)\right\} < \infty$.

We further define sequential analogues of the above notions: $(\mathcal{X}, \mathcal{F}, \ell)$ is defined to be an *SGC game* ("Sequential Glivenko-Cantelli") if $\mathrm{sfat}_\alpha(\mathcal{F}^\ell) < \infty$ for all $\alpha > 0$; by Lemma 8.4 this implies that $\mathrm{sfat}_\alpha(\mathcal{X}^\ell) < \infty$ for all $\alpha > 0$. Further, the game $(\mathcal{X}, \mathcal{F}, \ell)$ is said to be an *α-SGC game* if $\max\left\{\mathrm{sfat}_\alpha(\mathcal{F}^\ell), \mathrm{sfat}_\alpha(\mathcal{X}^\ell)\right\} < \infty$.

Lemma 8.5. *There is a constant $C_0 > 2$ so that the following holds. Fix any $\alpha > 0$, and suppose that $(\mathcal{X}, \mathcal{F}, \ell_0)$ is a $[0,1]$-valued (α/C_0)-GC game that does not contain infinitely many ordered thresholds with margin α and tightness α/C_0. Then*

$$\inf_{P_X \in \Delta(\mathcal{X})} \sup_{P_F \in \Delta(\mathcal{F})} \mathbb{E}_{(x,f) \sim P_X \times P_F}[\ell_0(x,f)] \leq \sup_{P_F \in \Delta(\mathcal{F})} \inf_{P_X \in \Delta(\mathcal{X})} \mathbb{E}_{(x,f) \sim P_X \times P_F}[\ell_0(x,f)] + 4\alpha.$$

Furthermore, the same statement holds if P_X, P_F are restricted to $\Delta^\circ(\mathcal{X}), \Delta^\circ(\mathcal{F})$, respectively.

Proof. The proof closely follows the technique of [HLM21, Proposition 9]. Fix a $[0,1]$-valued GC game $(\mathcal{X},\mathcal{F},\ell_0)$, set $\eta := \alpha/C_0$, and define the discretization $(\mathcal{X},\mathcal{F},\ell)$ as follows: for $(x,f)\in\mathcal{X}\times\mathcal{F}$,

$$\ell(x,f) := \lfloor \ell_0(x,f) \rfloor_\eta .$$

Since $\|\ell-\ell_0\|_{\infty,\mathcal{X}\times\mathcal{F}} \le \eta$, it follows that

$$\left| \inf_{P_X\in\Delta(\mathcal{X})} \sup_{P_F\in\Delta(\mathcal{F})} \mathbb{E}_{(x,f)\sim P_X\times P_F}[\ell_0(x,f)] - \inf_{P_X\in\Delta(\mathcal{X})} \sup_{P_F\in\Delta(\mathcal{F})} \mathbb{E}_{(x,f)\sim P_X\times P_F}[\ell(x,f)] \right| \le \eta \tag{52}$$

$$\left| \sup_{P_F\in\Delta(\mathcal{F})} \inf_{P_X\in\Delta(\mathcal{X})} \mathbb{E}_{(x,f)\sim P_X\times P_F}[\ell_0(x,f)] - \sup_{P_F\in\Delta(\mathcal{F})} \inf_{P_X\in\Delta(\mathcal{X})} \mathbb{E}_{(x,f)\sim P_X\times P_F}[\ell(x,f)] \right| \le \eta. \tag{53}$$

It is furthermore straightforward to see that $(\mathcal{X},\mathcal{F},\ell)$ is a 3η-GC game; one may see this by noting, for instance, that $\mathrm{fat}_{3\eta}(\{x\mapsto \ell(x,f) \ : \ f\in\mathcal{F}\}) \le \mathrm{fat}_\eta(\{x\mapsto \ell_0(x,f) \ : \ f\in\mathcal{F}\})$, and similarly for the dual class. Thus, by uniform convergence (i.e., Theorem A.3), there is a universal constant C so that, for all $\eta,\mathcal{X},\mathcal{F},\ell$, there is an integer V[16], so that for all finite-support distributions $P_X\in\Delta^\circ(\mathcal{X}), P_F\in\Delta^\circ(\mathcal{F})$, there are elements $x_1,\dots,x_V\in\mathcal{X}, f_1,\dots,f_V\in\mathcal{F}$, so that

$$\sup_{f\in\mathcal{F}} \left| \mathbb{E}_{x\sim P_X}[\ell(x,f)] - \frac{1}{V}\sum_{i=1}^V \ell(x_i,f) \right| \le C\eta \tag{54}$$

$$\sup_{x\in\mathcal{X}} \left| \mathbb{E}_{f\sim P_F}[\ell(x,f)] - \frac{1}{V}\sum_{i=1}^V \ell(x,f_i) \right| \le C\eta. \tag{55}$$

Now set

$$\theta = \sup_{P_F\in\Delta(\mathcal{F})} \inf_{P_X\in\Delta(\mathcal{X})} \mathbb{E}_{(x,f)\sim P_X\times P_F}[\ell(x,f)]$$
$$\omega = \inf_{P_X\in\Delta(\mathcal{X})} \sup_{P_F\in\Delta(\mathcal{F})} \mathbb{E}_{(x,f)\sim P_X\times P_F}[\ell(x,f)].$$

Suppose for the purpose of contradiction that $\omega > \theta + 4\alpha - 2\eta$ (if this is not the case, then by (52) and (53) the proof of the lemma is complete). We next construct two sequences of finite-support distributions P_X^t, P_F^t, $t\in\mathbb{N}$, where each of P_X^t, P_F^t is a uniform distribution over exactly V elements of $\mathcal{X},\mathcal{F}$, respectively (possibly with some elements being duplicates), and so that the following inequalities hold:

$$\sup_{P_F\in\Delta(\bigcup_{i<t}\mathrm{supp}(P_F^i))} \mathbb{E}_{(x,f)\sim P_X^t\times P_F}[\ell(x,f)] \le \theta + \frac{4\alpha-2\eta}{3} \qquad \forall t>1 \tag{56}$$

$$\inf_{P_X\in\Delta(\bigcup_{i\le t}\mathrm{supp}(P_X^i))} \mathbb{E}_{(x,f)\sim P_X\times P_F^t}[\ell(x,f)] \ge \theta + 2\cdot\frac{4\alpha-2\eta}{3} \qquad \forall t\ge 1. \tag{57}$$

We construct P_X^t, P_F^t inductively as follows:

[16] In particular, we may take $V = O\left(\frac{\max\{\mathrm{fat}_\eta(\{x\mapsto\ell_0(x,f):f\in\mathcal{F}\}),\mathrm{fat}_\eta(\{f\mapsto\ell_0(x,f):x\in\mathcal{X}\})\}\cdot\log 1/\eta}{\eta^2}\right)$, which is finite, by assumption.

- Suppose we have constructed $P_X^1, \dots, P_X^{t-1}, P_F^1, \dots, P_F^{t-1}$ satisfying (56) and (57) up to step $t-1$. To construct P_X^t so as to satisfy (56) at step t, we argue as follows: if $t = 1$, set $P_X^1 = \delta_x$ for any $x \in \mathcal{X}$ (i.e., the point mass at x), which suffices because (56) is vacuous for $t = 1$. Otherwise, let $F_{<t} = \bigcup_{i<t} \operatorname{supp}(P_F^i)$, which is a finite set. Since $\ell(x, f)$ takes values in a finite set, there are a finite number of distinct sequences $\{\ell(x, f)\}_{f \in F_{<t}}$ given by elements $x \in \mathcal{X}$. Thus, there is a finite subset $\mathcal{X}' \subset \mathcal{X}$ so that for each $x \in \mathcal{X}$, there is some $x' \in \mathcal{X}'$ so that $\ell(x, f) = \ell(x', f)$ for all $f \in F_{<t}$. Thus, by the von Neumann minimax theorem applied to the finite game $(\mathcal{X}', F_{<t}, \ell)$[17], there is a distribution $P_X^\star \in \Delta(\mathcal{X}')$ so that

$$\begin{aligned}
&\sup_{P_F \in \Delta(F_{<t})} \mathbb{E}_{(x,f)\sim P_X^\star \times P_F}[\ell(x, f)] = \sup_{P_F \in \Delta(F_{<t})} \inf_{P_X \in \Delta(\mathcal{X}')} \mathbb{E}_{(x,f)\sim P_X \times P_F}[\ell(x, f)] \\
&= \sup_{P_F \in \Delta(F_{<t})} \inf_{P_X \in \Delta(\mathcal{X})} \mathbb{E}_{(x,f)\sim P_X \times P_F}[\ell(x, f)] \le \sup_{P_F \in \Delta(\mathcal{F})} \inf_{P_X \in \Delta(\mathcal{X})} \mathbb{E}_{(x,f)\sim P_X \times P_F}[\ell(x, f)] = \theta,
\end{aligned} \tag{58}$$

  where the first equality follows from the choice of $P_X^\star$ as a minimax strategy for the X-player in the finite game, the second equality follows from the defining property of $\mathcal{X}'$, and the inequality follows from the fact that points on $\mathcal{F}$ are measurable, meaning that every measure $P_F \in \Delta(F_{<t})$ may be realized as the corresponding distribution on $\mathcal{F}$ restricted to $F_{<t}$.

  By (54) and using the fact that $P_X^\star$ is a finite-support measure, there is a sequence $x_1, \dots, x_V \in \mathcal{X}$ so that, setting P_X^t to be the empirical measure $P_X^t(S) := \frac{1}{V}\sum_{i=1}^V \mathbb{1}[x_i \in S]$, every $f \in F_{<t} \subset \mathcal{F}$ satisfies

$$\mathbb{E}_{x\sim P_X^t}[\ell(x, f)] - \mathbb{E}_{x\sim P_X^\star}[\ell(x, f)] \le C\eta \le \frac{4\alpha - 2\eta}{3},$$

  where the final inequality may be ensured by choosing C_0 so that $C_0 \ge 3C + 2$ (which implies that $C\eta \le \frac{\alpha - 2\eta}{3}$). Thus

$$\sup_{P_F \in \Delta(F_{<t})} \mathbb{E}_{(x,f)\sim P_X^t \times P_F}[\ell(x, f)] \le \sup_{P_F \in \Delta(F_{<t})} \mathbb{E}_{(x,f)\sim P_X^\star \times P_F}[\ell(x, f)] + \frac{4\alpha - 2\eta}{3} \le \theta + \frac{4\alpha - 2\eta}{3},$$

  showing that (56) holds at step t.

- Next suppose we have constructed $P_X^1, \dots, P_X^t, P_F^1, \dots, P_F^{t-1}$ satisfying (56) up to step t and satisfying (57) up to step $t-1$. We then construct P_F^{t+1} so as to satisfy (57) in a very similar manner as to the previous case: setting $X_{\le t} := \bigcup_{i \le t} \operatorname{supp}(P_X^i)$, we get that there is a finite set $\mathcal{F}' \subset \mathcal{F}$ and a distribution $P_F^\star \in \Delta(\mathcal{F}')$ so that

$$\begin{aligned}
&\inf_{P_X \in \Delta(X_{\le t})} \mathbb{E}_{(x,f)\sim P_X \times P_F^\star}[\ell(x, f)] = \inf_{P_X \in \Delta(X_{\le t})} \sup_{P_F \in \Delta(\mathcal{F}')} \mathbb{E}_{(x,f)\sim P_X \times P_F}[\ell(x, f)] \\
&= \inf_{P_X \in \Delta(X_{\le t})} \sup_{P_F \in \Delta(\mathcal{F})} \mathbb{E}_{(x,f)\sim P_X \times P_F}[\ell(x, f)] \ge \inf_{P_X \in \Delta(\mathcal{X})} \sup_{P_F \in \Delta(\mathcal{F})} \mathbb{E}_{(x,f)\sim P_X \times P_F}[\ell(x, f)] = \omega.
\end{aligned} \tag{59}$$

[17] With a slight abuse of notation, the loss function ℓ is restricted to $\mathcal{X}' \times F_{<t}$.

By (55) and the fact that $P_F^\star$ is a finite support measure, there is a sequence $f_1, \dots, f_V \in \mathcal{F}$ so that, setting P_F^{t+1} to be the empirical measure $P_F^{t+1}(S) := \frac{1}{V}\sum_{i=1}^V \mathbb{1}[f_i \in S]$, every $x \in X_{\le t} \subset \mathcal{X}$ satisfies $\mathbb{E}_{f\sim P_F^{t+1}}[\ell(x,f)] \ge \mathbb{E}_{f\sim P_F^\star}[\ell(x,f)] - \frac{4\alpha-2\eta}{3}$. Thus

$$\inf_{P_X\in\Delta(X_{\le t})} \mathbb{E}_{(x,f)\sim P_X\times P_F^{t+1}} \ge \omega - \frac{4\alpha-2\eta}{3} \ge \theta + 2\cdot\frac{4\alpha-2\eta}{3},$$

thus verifying (57) since we have assume $\omega - \theta \ge 4\alpha - 2\eta$.

As we have constructed each of P_X^i, P_F^i to be a uniform distribution over V elements of $\mathcal{X}, \mathcal{F}$, respectively, we may denote these elements as $x_{i,1}, \dots, x_{i,V}$ and $f_{i,1}, \dots, f_{i,V}$, for each $i \in \mathbb{N}$. For each $i, j \in \mathbb{N}$ with $i < j$, we define matrices $A^{ij}, B^{ij} \in \{1/\lceil 1/\eta\rceil, 2/\lceil 1/\eta\rceil, \dots, 1\}^{V\times V}$, as follows: for $k, m \in [V]$, we set

$$A^{ij}_{km} := \ell(x_{i,k}, f_{j,m}), \qquad B^{ij}_{km} := \ell(x_{j,k}, b_{i,m}).$$

The number of different possible matrix pairs (A^{ij}, B^{ij}) is $\lceil 1/\eta\rceil^{2V^2}$, which is finite. The infinite Ramsey theorem implies that there exists an infinite increasing sequence $i_1, i_2, \dots, \in \mathbb{N}$ and a pair of matrices $A^\star, B^\star \in \{1/\lceil 1/\eta\rceil, 2/\lceil 1/\eta\rceil, \dots, 1\}^{V\times V}$ so that for all $s, t \in \mathbb{N}$ with $s < t$, we have $(A^{i_s i_t}, B^{i_s i_t}) = (A^\star, B^\star)$.

We next claim that there are $k^\star, m^\star \in [V]$ so that $A^\star_{k^\star m^\star} - B^\star_{k^\star m^\star} \ge \frac{4\alpha-2\eta}{3}$. To see this, note that, for any $s, t \in \mathbb{N}$ with $s < t$, we have

$$\begin{aligned} &\inf_{P_X\in\Delta(\mathrm{supp}(P_X^{i_s}))} \sup_{P_F\in\Delta(\mathrm{supp}(P_F^{i_t}))} \mathbb{E}_{(x,f)\sim P_X\times P_F}[\ell(x,f)] \qquad (60)\\ &\ge \inf_{P_X\in\Delta(X_{\le i_t})} \mathbb{E}_{(x,f)\sim P_X\times P_F^{i_t}}[\ell(x,f)] \ge \theta + 2\cdot\frac{4\alpha-2\eta}{3}, \end{aligned}$$

and

$$\begin{aligned} &\inf_{P_X\in\Delta(\mathrm{supp}(P_X^{i_t}))} \sup_{P_F\in\Delta(\mathrm{supp}(P_F^{i_s}))} \mathbb{E}_{(x,f)\sim P_X\times P_F}[\ell(x,f)] \qquad (61)\\ &\le \sup_{P_F\in\Delta(F_{<i_t})} \mathbb{E}_{(x,f)\sim P_X^{i_t}\times P_F}[\ell(x,f)] \le \theta + \frac{4\alpha-2\eta}{3}. \end{aligned}$$

Notice that the quantity in (60) is the value of the game represented by the matrix $\{\ell(x_{i_s,k}, f_{i_t,m})\}_{k,m\in[V]}$, and this matrix is $A^{i_s i_t} = A^\star$. Similarly, the quantity in (61) is the value of the game represented by the matrix $\{\ell(x_{i_t,k}, f_{i_s,m})\}_{k,m\in[V]}$, and this matrix is $B^{i_s i_t} = B^\star$. Hence the value of the game $A^\star$ is at least $\frac{4\alpha-2\eta}{3}$ greater than the value of the game $B^\star$. Thus some entry of $A^\star$ must be at least $\frac{4\alpha-2\eta}{3}$ greater than the corresponding entry of $B^\star$. Indeed, if this were not the case, then we would have that

$$\min_{p\in\Delta([V])}\max_{q\in\Delta([V])} p^\top A^\star q < \frac{4\alpha-2\eta}{3} + \min_{p\in\Delta([V])}\max_{q\in\Delta([V])} p^\top B^\star q,$$

a contradiction to the previous sentence. This shows the existence of the $k^\star, m^\star$ as desired.

Finally we may construct a collection of infinitely many thresholds with margin α and tightness β. For $t \ge 1$, define $x_t^\star := x_{i_{2t},k^\star}$ and $f_t^\star = f_{i_{2t-1},m^\star}$. For $s, t \in \mathbb{N}$ with $s < t$, we have $i_{2s} < i_{2t-1}$,

and so $\ell(x_s^\star, f_t^\star) = \ell(x_{i_{2s},k^\star}, f_{2t-1,m^\star}) = A^\star_{k^\star m^\star}$. For $s,t \in \mathbb{N}$ with $s \geq t$, we have $i_{2s} > i_{2t-1}$, and so $\ell(x_s^\star, f_t^\star) = \ell(x_{i_{2s},k^\star}, f_{2t-1,m^\star}) = B^\star_{k^\star m^\star}$.

Recalling that $|\ell(x,f) - \ell_0(x,f)| \leq \eta$ for all x, f, it follows that $|\ell_0(x_s^\star, f_t^\star) - A^\star_{k^\star m^\star}| \leq \eta$ for $s \leq t$ and $|\ell_0(x_s^\star, f_t^\star) - B^\star_{k^\star m^\star}| \leq \eta$ for $s > t$. Further, since $2\eta \leq \alpha$ (as $C_0 > 2$), we have $A^\star_{k^\star m^\star} - B^\star_{k^\star m^\star} \geq \alpha$, as desired (in particular, in the context of Definition 8.1, we may take $u' = A^\star_{k^\star m^\star},\ u = B^\star_{k^\star m^\star}$).

Finally, to establish the statement of the lemma about finite support measures P_X, P_F, we note that exactly the same proof presented above works; the only difference is that in (58) and (59), we replace $\Delta(\mathcal{F})$ with $\Delta^\circ(\mathcal{F})$ and $\Delta(\mathcal{X})$ with $\Delta^\circ(\mathcal{F})$; it is evident that the claimed inequalities in (58) and (59) hold even with these substitutions. □

We next show a converse to Lemma 8.5, thus obtaining a necessary and sufficient condition for the minimax theorem to hold in all subgames of a GC game.

Lemma 8.6 (Converse to Lemma 8.5)**.** *For any $\alpha \in (0,1)$, any $[0,1]$-valued game $(\mathcal{X}, \mathcal{F}, \ell_0)$ which contains infinitely many ordered thresholds with margin α and tightness β satisfies, for some $\mathcal{X}' \subset \mathcal{X},\ \mathcal{F}' \subset \mathcal{F}$,*

$$\inf_{P_X \in \Delta(\mathcal{X}')} \sup_{P_F \in \Delta(\mathcal{F}')} \mathbb{E}_{(x,f)\sim P_X \times P_F}[\ell_0(x,f)] > \sup_{P_F \in \Delta(\mathcal{F}')} \inf_{P_X \in \Delta(\mathcal{X}')} \mathbb{E}_{(x,f)\sim P_X \times P_F}[\ell_0(x,f)] + \alpha - 2\beta.$$

Proof. Let $x_1, x_2, \ldots \in \mathcal{X}$ and $f_1, f_2, \ldots, \in \mathcal{F}$ denotes a collection of infinitely many ordered thresholds with margin α and tightness β. Write $\mathcal{X}' = \{x_1, x_2, \ldots\}$ and $\mathcal{F}' = \{f_1, f_2, \ldots\}$. By definition there are $u, u' \in [0,1]$ so that $u' - u \geq \alpha$, $|f_i(x_j) - u| \leq \beta$ for $i \leq j$ and $|f_i(x_j) - u'| \leq \beta$ for $i > j$. Then for any $P_X \in \Delta(\mathcal{X}')$, we have

$$\begin{aligned}
&\sup_{P_F \in \Delta(\mathcal{F}')} \mathbb{E}_{(x,f)\sim P_X \times P_F}[\ell_0(x,f)] \geq \sup_{i \geq 1} \mathbb{E}_{x_j \sim P_X}[\ell_0(x_j, f_i)] \geq \liminf_{i\to\infty} \mathbb{E}_{x_j \sim P_X}[\ell_0(x_j, f_i)] \\
&\geq \mathbb{E}_{x_j \sim P_X}[\liminf_{i\to\infty} \ell_0(x_j, f_i)] \geq u' - \beta,
\end{aligned}$$

where the second-to-last inequality follows from Fatou's lemma.

On the other hand, for any $P_F \in \Delta(\mathcal{F}')$, we have

$$\begin{aligned}
&\inf_{P_X \in \Delta(\mathcal{X}')} \mathbb{E}_{(x,f)\sim P_X \times P_F}[\ell_0(x,f)] \leq \inf_{j \geq 1} \mathbb{E}_{f_i \sim P_F}[\ell_0(x_j, f_i)] \leq \limsup_{j\to\infty} \mathbb{E}_{f_i \sim P_F}[\ell_0(x_j, f_i)] \\
&\leq \mathbb{E}_{f_i \sim P_F}[\limsup_{j\to\infty} \ell_0(x_j, f_i)] \leq u + \beta,
\end{aligned}$$

where again the second-to-last inequality follows from Fatou's lemma. The two displays above complete the proof. □

By combining Lemmas 8.5 and 8.6, we are able to show the following necessary and sufficient condition for all subgames of an infinite GC game to satisfy the minimax theorem:

Theorem 8.7. *Let $C_0 > 2$ be the constant of Lemma 8.5. A $[0,1]$-valued GC game $(\mathcal{X}, \mathcal{F}, \ell_0)$ satisfies*

$$\inf_{P_X \in \Delta(\mathcal{X}')} \sup_{P_F \in \Delta(\mathcal{F}')} \mathbb{E}_{(x,f)\sim P_X \times P_F}[\ell_0(x,f)] = \sup_{P_F \in \Delta(\mathcal{F}')} \inf_{P_X \in \Delta(\mathcal{X}')} \mathbb{E}_{(x,f)\sim P_X \times P_F}[\ell_0(x,f)] \tag{62}$$

for all $\mathcal{X}' \subset \mathcal{X},\ \mathcal{F}' \subset \mathcal{F}$ if and only if it does not contain infinitely many ordered thresholds with margin α and tightness α/C_0, for all $\alpha > 0$.

Proof. First suppose that (62) holds for all $\mathcal{X}', \mathcal{F}'$. If the game $(\mathcal{X}, \mathcal{F}, \ell_0)$ contained infinitely many ordered thresholds with margin α and tightness α/C_0, then by Lemma 8.6, there would be some $\mathcal{X}', \mathcal{F}'$ so that the left-hand side of (62) is at least the sum of $\alpha - 2\alpha/C_0 > 0$ and the right-hand isde of (62). This is a contradiction.

Conversely, suppose that the game $(\mathcal{X}, \mathcal{F}, \ell_0)$ does not contain infinitely many ordered thresholds with margin α and tightness α/C_0 for all $\alpha > 0$. Since, for each $\alpha > 0$ and $\mathcal{X}' \subset \mathcal{X}$, $\mathcal{F}' \subset \mathcal{F}$, $(\mathcal{X}', \mathcal{F}', \ell_0)$ is a (α/C_0)-GC game, Lemma 8.5 gives that for each $\alpha > 0$,

$$\inf_{P_X \in \Delta(\mathcal{X}')} \sup_{P_F \in \Delta(\mathcal{F}')} \mathbb{E}_{(x,f)\sim P_X \times P_F}[\ell_0(x,f)] \leq \sup_{P_F \in \Delta(\mathcal{F}')} \inf_{P_X \in \Delta(\mathcal{X}')} \mathbb{E}_{(x,f)\sim P_X \times P_F}[\ell_0(x,f)] + 4\alpha.$$

Then (62) follows by taking $\alpha \downarrow 0$. ☐

It is also immediate to show the minimax theorem for online learnable games, as follows:

Theorem 8.8 (Minimax theorem for online learnable games)**.** *Any SGC game $(\mathcal{X}, \mathcal{F}, \ell)$ satisfies the min-max theorem, i.e.,*

$$\inf_{P_X \in \Delta(\mathcal{X})} \sup_{P_F \in \Delta(\mathcal{F})} \mathbb{E}_{(x,f)\sim P_X \times P_F}[\ell(x,f)] = \sup_{P_F \in \Delta(\mathcal{F})} \inf_{P_X \in \Delta(\mathcal{X})} \mathbb{E}_{(x,f)\sim P_X \times P_F}[\ell(x,f)].$$

Further, the above equality remains true even if P_X, P_F are restricted to lie in $\Delta^\circ(\mathcal{X}), \Delta^\circ(\mathcal{F})$, respectively.

Proof. By Lemma 8.4 and since the given game is an SGC game, we have that the sequential fat-shattering dimension of the classes $\mathcal{F}^\ell := \{x \mapsto \ell(x,f) : f \in \mathcal{F}\}$ and $\mathcal{X}^\ell := \{f \mapsto \ell(x,f) : x \in \mathcal{X}\}$ is finite at all scales. Thus $(\mathcal{X}, \mathcal{F}, \ell)$ is a GC game. The inequality

$$\inf_{P_X \in \Delta(\mathcal{X})} \sup_{P_F \in \Delta(\mathcal{F})} \mathbb{E}_{(x,f)\sim P_X \times P_F}[\ell(x,f)] \geq \sup_{P_F \in \Delta(\mathcal{F})} \inf_{P_X \in \Delta(\mathcal{X})} \mathbb{E}_{(x,f)\sim P_X \times P_F}[\ell(x,f)]$$

is immediate. To see the opposite direction, fix any $\alpha > 0$, and let C_0 be the constant of Lemma 8.5. Certainly $(\mathcal{X}, \mathcal{F}, \ell)$ is a (α/C_0)-GC game. Further, by Lemma 8.1, the maximum number of thresholds in the game $(\mathcal{X}, \mathcal{F}, \ell)$ with margin α and tightness α/C_0 is $2^{O(\mathrm{sfat}_{\alpha\cdot(1-2/C_0)}(\mathcal{F}^\ell))} < \infty$. It follows from Lemma 8.5 that

$$\inf_{P_X \in \Delta(\mathcal{X})} \sup_{P_F \in \Delta(\mathcal{F})} \mathbb{E}_{(x,f)\sim P_X \times P_F}[\ell(x,f)] \leq \sup_{P_F \in \Delta(\mathcal{F})} \inf_{P_X \in \Delta(\mathcal{X})} \mathbb{E}_{(x,f)\sim P_X \times P_F}[\ell(x,f)] + 4\alpha,$$

and even if P_X, P_F are restricted to $\Delta^\circ(\mathcal{X}), \Delta^\circ(\mathcal{F})$, respectively. The statement of the theorem follows since $\alpha > 0$ may be taken arbitrarily small. ☐

A Miscellaneous lemmas

In this section we state some miscellaneous lemmas on the fat-shattering dimension of real-valued hypothesis classes. Many of these lemmas are well-known (see for instance [Gol21]), but we state the proofs for completeness.

Lemma A.1. *Fix a class $\mathcal{F} \subset [0,1]^{\mathcal{X}}$ and $\alpha \in (0,1)$. There are at most 2 integers j, $0 \leq j < \lfloor 1/\alpha \rfloor + 1$ so that*

$$\mathrm{sfat}_\alpha(\mathcal{F}) = \mathrm{sfat}_\alpha\left(\{f \in \mathcal{F} : f(x) \in [j\alpha, (j+1)\alpha)\}\right).$$

Moreover, if there are 2 such integers j, they differ by 1.

Proof. Suppose for the purpose of contradiction that for some j_1, j_2 with $|j_2 - j_1| \geq 2$, we have

$$\mathrm{sfat}_\alpha(\mathcal{F}) = \mathrm{sfat}_\alpha\left(\{f \in \mathcal{F} : f(x) \in [j_1\alpha, (j_1+1)\alpha)\}\right) = \mathrm{sfat}_\alpha\left(\{f \in \mathcal{F} : f(x) \in [j_2\alpha, (j_2+1)\alpha)\}\right).$$

Set $j' = (j_1 + j_2)/2$. We therefore have that

$$\mathrm{sfat}_\alpha(\mathcal{F}) = \mathrm{sfat}_\alpha(\{f \in \mathcal{F} : f(x) \geq j'\alpha + \alpha/2\}) = \mathrm{sfat}_\alpha(\{f \in \mathcal{F} : f(x) \leq j'\alpha - \alpha/2\}),$$

which is a contradiction. □

Lemma A.2. *For $\mathcal{F} \subset [0,1]^{\mathcal{X}}$ and $\alpha \in (0,1)$, and $(x,y) \in \mathcal{X} \times [0,1]$, if $|y - \mathrm{SOA}(\mathcal{F},\alpha)(x)| > \alpha$, then $\mathrm{sfat}_\alpha(\mathcal{F}|^\alpha_{(x,y)}) < \mathrm{sfat}_\alpha(\mathcal{F})$.*

Proof. Suppose for the purpose of contradiction that $\mathrm{sfat}_\alpha(\mathcal{F}|^\alpha_{(x,y)}) = \mathrm{sfat}_\alpha(\mathcal{F})$. Let $j = \lfloor y/\alpha \rfloor$. Then by definition of $\mathcal{F}|^\alpha_{(x,y)}$, we have that

$$\mathrm{sfat}_\alpha(\mathcal{F}) = \mathrm{sfat}_\alpha(\{f \in \mathcal{F} : f(x) \in [j\alpha, (j+1)\alpha)\}).$$

By definition of $\mathrm{SOA}(\mathcal{F},\alpha)$, we have that $\mathrm{SOA}(\mathcal{F},\alpha)(x) = j^\star\alpha$ for some $0 \leq j^\star < \lfloor 1/\alpha \rfloor + 1$. It must hold that $\mathrm{sfat}_\alpha(\{f \in \mathcal{F} : f(x) \in [j^\star\alpha, (j^\star+1)\alpha)\}) = \mathrm{sfat}_\alpha(\mathcal{F})$. Since $|y - j^\star\alpha| > \alpha$, we have that $y \notin [(j^\star-1)\alpha, (j^\star+1)\alpha)$, meaning that $j \notin \{j^\star, j^\star - 1\}$. By Lemma A.1 we must have $j = j^\star + 1$. But the definition of $\mathrm{SOA}(\mathcal{F},\alpha)$ requires that in this case that $\mathrm{SOA}(\mathcal{F},\alpha)(x) = (j^\star+1)\alpha$, which is a contradiction. □

Uniform convergence. Next we state a uniform convergence bound for real-valued hypothesis classes. The below bound is not optimal (as it only considers the fat-shattering dimension at a single scale), but as it does not quantitatively affect our statistical rates for online learning, it will suffice for our purposes.

For uniform convergence (which implies learnability under i.i.d. data), finiteness of the *fat-shattering dimension* [ABDCBH97], which is smaller than the sequential fat-shattering dimension, is sufficient (and necessary). The fat-shattering dimension of a hypothesis class $\mathcal{F} \subset [0,1]^{\mathcal{X}}$ at scale $\alpha > 0$, denoted $\mathrm{fat}_\alpha(\mathcal{F})$, is defined as follows. It is the largest positive integer d so that there are $x_1, \ldots, x_d \in \mathcal{X}$ and $s_1, \ldots, s_d \in [0,1]$ so that for each choice of $\epsilon_1, \ldots, \epsilon_d \in \{-1, 1\}$ it holds that there is some $f \in \mathcal{F}$ so that, for each $i \in [d]$, $\epsilon_i \cdot (f(x_i) - s_i) \geq \alpha/2$.

Theorem A.3 (Uniform convergence; [MV02][18]). *There are constants $C_0 \geq 1$ and $0 < c_0 \leq 1$ so that the following holds. For any $\mathcal{F} \subset [0,1]^{\mathcal{X}}$, and finite-support distribution P[19] on $\mathcal{X}$, and any $\gamma \in (0,1/2), \eta \in (0,1/2)$, it holds that for any*

$$n \geq C_0 \cdot \frac{\mathrm{fat}_{c_0\eta}(\mathcal{F}) \log(1/\eta) + \log(1/\gamma)}{\eta^2},$$

we have

$$\mathbb{P}_{x_1,\ldots,x_n \sim P}\left[\sup_{f \in \mathcal{F}} \left| \mathbb{E}_{x \sim P}[f(x)] - \frac{1}{n}\sum_{i=1}^n f(x_i) \right| > \eta \right] \leq \gamma.$$

[18] For an explanation of how the theorem follows from [MV02], see [Gol21, Corollary 20].

[19] The finite-suportedness assumption can be dropped if $\mathcal{F}$ is countable.

Closure bound for the sequential fat-shattering dimension. Next we establish a closure bound for the sequential fat-shattering dimension; the result is the real-valued analogue of [GGKM21, Proposition 2.3], and is also similar to [RST15a, Lemma 4], which proves an analogue for the sequential Rademacher complexity. To begin, we establish some additional preliminaries, following [RST15a]: for some set $\mathcal{Z}$ and a function class $\mathcal{F} \subset [0,1]^{\mathcal{Z}}$, fix a $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth d, and consider a set $\mathcal{V}$ of $\mathbb{R}$-valued trees of depth d. For $\alpha > 0$, the set $\mathcal{V}$ is defined to be a *sequential α-cover* of $\mathcal{F}$ on the tree $\mathbf{z}$ if for all $f \in \mathcal{F}$, and all $\epsilon \in \{-1,1\}^d$, there is some $\mathbf{v} \in \mathcal{V}$ so that

$$\max_{t \in [d]} |\mathbf{v}_t(\epsilon_{1:t-1}) - f(\mathbf{z}_t(\epsilon_{1:t-1}))| \leq \alpha. \tag{63}$$

Given a class $\mathcal{F}$, the *sequential α-covering number* (with respect to ℓ_∞) for the tree $\mathbf{z}$ is defined as follows:

$$\mathcal{N}_\infty(\mathcal{F}, \mathbf{z}, \alpha) := \min \{|\mathcal{V}| \ : \ \mathcal{V} \text{ is a sequential } \alpha\text{-cover of } \mathcal{F} \text{ on the tree } \mathbf{z}\}.$$

Next we need a few basic lemmas that related the sequential covering numbers of classes and their sequential fat-shattering dimension.

Lemma A.4 (Theorem 14.5, [RS14b])**.** *Consider a class $\mathcal{F} \subset [0,1]^{\mathcal{Z}}$. Then for any $\alpha > 0$, and $d \in \mathbb{N}$, and any $\mathcal{Z}$-valued tree $\mathbf{z}$ of depth $d \geq \mathrm{sfat}_\alpha(\mathcal{F})$,*

$$\mathcal{N}_\infty(\mathcal{F}, \mathbf{z}, \alpha) \leq \left(\frac{2ed}{\alpha \cdot \mathrm{sfat}_\alpha(\mathcal{F})} \right)^{\mathrm{sfat}_\alpha(\mathcal{F})}.$$

We remark that in the statement of [RS14b, Theorem 14.5] the term $\mathrm{sfat}_\alpha(\mathcal{F})$ does not appear in the denominator in the upper bound on $\mathcal{N}_\infty(\mathcal{F}, \mathbf{z}, \alpha)$. However, a close inspection of their proof shows that they establish $\mathcal{N}_\infty(\mathcal{F}, \mathbf{z}, \alpha) \leq \left(\frac{2ed}{\alpha m}\right)^m$ for some $m \leq \mathrm{sfat}_\alpha(\mathcal{F})$ (namely, m is the pararmeter called $\mathrm{fat}_2(\mathcal{G})$ therein). The statement of Lemma A.4 then follows by noting that the function $m \mapsto \left(\frac{2ed}{\alpha m}\right)^m$ is non-decreasing for $m \leq d$.

Lemma A.5. *Suppose a tree $\mathbf{z}$ of depth d is α-shattered by a class $\mathcal{F}$. Then $\mathcal{N}_\infty(\mathcal{F}, \mathbf{z}, \beta) \geq 2^d$ for any $\beta < \alpha/2$.*

Proof. Let $\mathbf{s}$ be an $\mathbb{R}$-valued tree that witnesses the shattering of $\mathbf{z}$. Let $\mathcal{V}$ be a β-cover of $\mathcal{F}$ on the tree $\mathbf{z}$. Consider any two leaves $\epsilon, \epsilon' \in \{-1,1\}^d$ of the tree $\mathbf{z}$, and let corresponding functions in $\mathcal{F}$ be denoted f, f'. (For a leaf $\epsilon \in \{-1,1\}^d$, a corresponding function $f \in \mathcal{F}$ is any function so that $\epsilon_t \cdot (f(\mathbf{z}_t(\epsilon_{1:t-1})) - \mathbf{s}_t(\epsilon_{1:t-1})) \geq \alpha/2$ for each $t \in [d]$.) Let $\mathbf{v}, \mathbf{v}' \in \mathcal{V}$ be the elements of the cover $\mathcal{V}$ as guaranteed by (63) for the leaves ϵ, ϵ'. We claim that $\mathbf{v} \neq \mathbf{v}'$, which would immediately complete the proof; so suppose to the contrary that $\mathbf{v} = \mathbf{v}'$.

Choose t as small as possible so that $\epsilon_t \neq \epsilon'_t$. Then (perhaps after interchanging the roles of f, f'), it holds that

$$f(\mathbf{z}_t(\epsilon_{1:t-1})) \geq \mathbf{s}_t(\epsilon_{1:t-1}) + \alpha/2, \qquad f'(\mathbf{z}_t(\epsilon_{1:t-1})) \leq \mathbf{s}_t(\epsilon_{1:t-1}) - \alpha/2. \tag{64}$$

On the other hand, since $\mathcal{V}$ is a β-cover of $\mathcal{F}$, we have (since $\mathbf{v} = \mathbf{v}'$) that

$$|\mathbf{v}_t(\epsilon_{1:t-1}) - f(\mathbf{z}_t(\epsilon_{1:t-1}))| \leq \beta, \qquad |\mathbf{v}_t(\epsilon_{1:t-1}) - f'(\mathbf{z}_t(\epsilon_{1:t-1}))| \leq \beta. \tag{65}$$

Using that $\beta < \alpha/2$, we get that (64) and (65) lead to a contradiction, thus completing the proof of the lemma. ∎

Given some $k \in \mathbb{N}$, some function $\phi : \mathbb{R}^k \times \mathcal{Z} \to \mathbb{R}$, and function classes $\mathcal{F}_1, \dots, \mathcal{F}_k \subset [0,1]^{\mathcal{X}}$, define the *$\phi$-composition* of $\mathcal{F}_1, \dots, \mathcal{F}_k$ as follows:

$$\phi(\mathcal{F}_1, \dots, \mathcal{F}_k) := \{z \mapsto \phi(f_1(z), \dots, f_k(z), z) \ : \ f_1 \in \mathcal{F}_1, \dots, f_k \in \mathcal{F}_k\}.$$

Lemma A.6. *Consider classes $\mathcal{F}_1, \dots, \mathcal{F}_k \subset [0,1]^{\mathcal{Z}}$, and consider a function $\phi : \mathbb{R}^k \times \mathcal{X} \to \mathbb{R}$ so that $\phi(\cdot, z)$ is L-Lipschitz for each $z \in \mathcal{Z}$. Fix any $\alpha > 0$, and suppose that $\mathrm{sfat}_{\alpha/(4L)}(\mathcal{F}_i) \le d$ for each $i \in [k]$ and some $d \in \mathbb{N}$. Then*

$$\mathrm{sfat}_\alpha(\phi(\mathcal{F}_1, \dots, \mathcal{F}_k)) \le O\left(dk \log\left(\frac{Lk}{\alpha}\right)\right).$$

Proof. Write $\mathcal{G} := \phi(\mathcal{F}_1, \dots, \mathcal{F}_k)$ to denote the composed class. Fix $\alpha > 0$, and write $N := \mathrm{sfat}_\alpha(\phi(\mathcal{F}_1, \dots, \mathcal{F}_k))$. Let $\mathbf{z}$ be a $\mathcal{Z}$-valued binary tree of depth N that is α-shattered by $\mathcal{G}$. By Lemma A.4, for each $i \in [k]$, we have that, for $\beta = \alpha/(4L)$,

$$\mathcal{N}_\infty(\mathcal{F}_i, \mathbf{z}, \beta) \le \left(\frac{2eN}{\beta \cdot \mathrm{sfat}_\beta(\mathcal{F}_i)}\right)^{\mathrm{sfat}_\beta(\mathcal{F}_i)}.$$

For each $i \in [k]$, let $\mathcal{V}_i$ be a minimal β-cover for the class $\mathcal{F}_i$ on the tree $\mathbf{z}$. Now consider the set

$$\mathcal{V} := \left\{\mathbf{v} = \phi(\mathbf{v}^1, \dots, \mathbf{v}^k) \ : \ \mathbf{v}^1 \in \mathcal{V}_1, \dots, \mathbf{v}^k \in \mathcal{V}_k\right\},$$

where $\phi(\mathbf{v}^1, \dots, \mathbf{v}^k)$ denotes the $\mathbb{R}$-valued tree defined by

$$\phi(\mathbf{v}^1, \dots, \mathbf{v}^k)_t(\epsilon_{1:t-1}) := \phi(\mathbf{v}^1_t(\epsilon_{1:t-1}), \dots, \mathbf{v}^k_t(\epsilon_{1:t-1}), \mathbf{z}_t(\epsilon_{1:t-1})).$$

Now fix any $g \in \mathcal{G}$; it can be written as $g(z) = \phi(f_1(z), \dots, f_k(z), z)$ for some $f_1 \in \mathcal{F}_1, \dots, f_k \in \mathcal{F}_k$. For each $i \in [k]$, let $\mathbf{v}^i \in \mathcal{F}_i$ denote a representative for f_i in the sense that for each $i \in [k]$,

$$\max_{t \in [d]} \left|\mathbf{v}^i_t(\epsilon_{1:t-1}) - f_i(\mathbf{z}_t(\epsilon_{1:t-1}))\right| \le \beta. \tag{66}$$

Then

$$\begin{aligned}
&\max_{t \in [d]} \left|\phi(\mathbf{v}^1, \dots, \mathbf{v}^k)_t(\epsilon_{1:t-1}) - g(\mathbf{z}_t(\epsilon_{1:t-1}))\right| \\
=&\max_{t \in [d]} \left|\phi(\mathbf{v}^1_t(\epsilon_{1:t-1}), \dots, \mathbf{v}^k_t(\epsilon_{1:t-1}), \mathbf{z}_t(\epsilon_{1:t-1})) - \phi(f_1(\mathbf{z}_t(\epsilon_{1:t-1})), \dots, f_k(\mathbf{z}_t(\epsilon_{1:t-1})), \mathbf{z}_t(\epsilon_{1:t-1}))\right| \\
\le& L\beta = \alpha/4. \qquad \text{(Using (66) and } L\text{-Lipschitzness of } \phi)
\end{aligned}$$

By Lemma A.5, since $\mathbf{z}$ is α-shattered by $\mathcal{G}$, we have that $\mathcal{N}_\infty(\mathcal{G}, \mathbf{z}, \alpha/4) \ge 2^N$. On the other hand, as we have shown above, the set $\mathcal{V}$ is a sequential $\alpha/4$-cover for the class $\mathcal{G}$ on the tree $\mathcal{V}$. Thus,

$$2^N \le \mathcal{N}_\infty(\mathcal{G}, \mathbf{z}, \alpha/4) \le |\mathcal{V}| \le \prod_{i=1}^k \left(\frac{4LeN}{\alpha \cdot \mathrm{sfat}_{\alpha/(4L)}(\mathcal{F}_i)}\right)^{\mathrm{sfat}_{\alpha/(4L)}(\mathcal{F}_i)},$$

which implies that

$$N \le \sum_{i=1}^k \mathrm{sfat}_{\alpha/(4L)}(\mathcal{F}_i) \cdot \log\left(\frac{4LeN}{\alpha \cdot \mathrm{sfat}_{\alpha/(4L)}(\mathcal{F}_i)}\right).$$

Recalling the assumption that $\mathrm{sfat}_{\alpha/(4L)}(\mathcal{F}_i) \le d$ for each i and using that $m \mapsto m \cdot \log\left(\frac{4LeN}{\alpha m}\right)$ is a non-decreasing function for $m \le N$, we obtain that $N \le O\left(dk \log\left(\frac{Lk}{\alpha}\right)\right)$.

□

Corollary A.7. *There is a constant $C \ge 1$ so that the following holds. Suppose $\mathcal{F}_1, \dots, \mathcal{F}_k \subset [0,1]^{\mathcal{X}}$ satisfy $\mathrm{sfat}_{\alpha/4}(\mathcal{F}_i) \le d$ for $i \in [k]$. Then for any $\alpha > 0$,*

$$\mathrm{sfat}_{\alpha}(\mathcal{F}_1 \cup \cdots \cup \mathcal{F}_k) \le C \cdot dk \cdot \log(k/\alpha). \tag{67}$$

Proof. For $i \in [k]$, define $\mathcal{F}_i' := \mathcal{F}_i \cup \{\mathbf{0}\}$, where $\mathbf{0}$ denotes the function that is identically 0 on $\mathcal{Z}$. Then clearly $\mathrm{sfat}_{\alpha/4}(\mathcal{F}_i') \le \mathrm{sfat}_{\alpha/4}(\mathcal{F}_i) + 1 \le d+1$. For $a_1, \dots, a_k \in \mathbb{R}$ and $z \in \mathcal{Z}$, define $\phi(a_1, \dots, a_k, z) := a_1 + \cdots + a_k$, which is clearly 1-Lipschitz with respect to $\|\cdot\|_\infty$. Further,

$$\mathcal{F}_1 \cup \cdots \cup \mathcal{F}_k \subset \phi(\mathcal{F}_1', \dots, \mathcal{F}_k'),$$

since for each $i \in [k]$ and $f_i \in \mathcal{F}_i$, $\phi(\mathcal{F}_1', \dots, \mathcal{F}_k')$ contains the function $z \mapsto \phi(\mathbf{0}, \dots, f_i(z), \dots, \mathbf{0}, z) = f_i(z)$. The result now follows from Lemma A.6 with $L = 1$. □

B Proof of Proposition 1.2

In this section we prove Proposition 1.2, thus showing that the bound of Theorem 5.15 is optimal up to a poly $\log T$ factor

Proof of Proposition 1.2. By compactness of $[1/T, 1]$, the infimum $M := \inf_{\alpha \in [1/T,1]} \left\{\alpha T + \int_\alpha^1 s(\eta) d\eta\right\}$ is obtained at some $\alpha_0 \in [1/T, 1]$.

Note also that the mapping $\alpha \mapsto \alpha T + \int_\alpha^1 s(\eta)d\eta$ is convex (its derivative is $T - s(\alpha)$, which is non-decreasing), meaning that for all $\alpha > \alpha_0$, $T - s(\alpha) \ge 0$, and for $1/T \le \alpha < \alpha_0$, $T - s(\alpha) \le 0$. Thus, by increasing α_0 by a factor of $3/2$, we can ensure that $T - s(\alpha_0) \ge 0$ and for all $1/T \le \alpha \le \alpha_0/2$, $T - s(\alpha) \le 0$ (further, doing so can only increase $\alpha T + \int_\alpha^1 s(\eta)d\eta$ by definition of α_0).

For any $\alpha \in [0,1]$, note that

$$\alpha T + \int_\alpha^1 s(\eta)d\eta \le \alpha T + \sum_{i \ge 0:\ 2^i\alpha \le 1} 2^i \alpha \cdot s(2^i \alpha).$$

Thus setting $\alpha = \alpha_0$ in the above display, one of the following possibilities holds:

1. $\alpha_0 T \ge M/2$. In this case, set $\alpha_0' = \alpha_0/2$, and set $d := T \le s(\alpha_0')$.
2. There is some $i \le \lceil \log 1/\alpha_0 \rceil \le \lceil \log T \rceil$ so that $2^i \alpha_0 \cdot s(2^i \alpha_0) \ge M/(2\lceil \log T \rceil)$. In this case, set $d := s(2^i\alpha_0) \le T$ (by definition of α_0, and using that $s(\alpha_0) \le T$). Now set $\alpha_0' := 2^i \cdot \alpha_0$.

Set $\mathcal{X} = \{1, 2, \dots, d\}$, and let $\mathcal{F}$ be the class of all functions on $\mathcal{X}$ so that for each $x \in \mathcal{S}$, $f(x) \in \{(1-\alpha_0')/2, (1+\alpha_0')/2\}$. Clearly, $\mathrm{sfat}_\alpha(\mathcal{F}) = d$ for all $\alpha \le \alpha_0'$, and $\mathrm{sfat}_\alpha(\mathcal{F}) = 0$ for all $\alpha > \alpha_0'$. Thus $\mathrm{sfat}_\alpha(\mathcal{F}) \le s(\alpha)$ for all $\alpha \in [0,1]$.

Further, the adversary can clearly force a cumulative loss of at least $\frac{\alpha_0'}{2} \cdot d$: simply feed each of the examples $x_1, \dots, x_d$ (using that $d \le T$), and set y_t to be whichever of $(1-\alpha_0'/2), (1+\alpha_0')/2$ is further from the algorithm's prediction at time t. In case 1 above, this cumulative loss becomes $\alpha_0 T/2 \ge \Omega(M)$, and in case 2 above, this cumulative loss becomes $2^i\alpha_0 \cdot s(2^i\alpha_0) \ge \Omega(M/\log T)$. Thus, in both cases, we get a cumulative loss of $\Omega(M/\log T)$, as desired. □

Acknowledgements

We are grateful to Sasha Rakhlin for helpful suggestions and to Steve Hanneke for a useful conversation.

References

[ABDCBH97] Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, 44(4):615–631, July 1997.

[ABED+21] Noga Alon, Omri Ben-Eliezer, Yuval Dagan, Shay Moran, Moni Naor, and Eylon Yogev. Adversarial Laws of Large Numbers and Optimal Regret in Online Classification. *arXiv:2101.09054 [cs, math, stat]*, January 2021. arXiv: 2101.09054.

[ACH+19] Scott Aaronson, Xinyi Chen, Elad Hazan, Satyen Kale, and Ashwin Nayak. Online Learning of Quantum States. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124019, December 2019. arXiv: 1802.09025.

[ALMM19] Noga Alon, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private PAC learning implies finite Littlestone dimension. *arXiv:1806.00949 [cs, math, stat]*, March 2019. arXiv: 1806.00949.

[Ang88] Dana Angluin. Queries and Concept Learning. *Machine Learning*, 2(4):319–342, April 1988.

[BBM05] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, August 2005. arXiv: math/0508275.

[BDPSS09] Shai Ben-David, David Pal, and Shai Shalev-Shwartz. Agnostic Online Learning. In *Proceedings of the 2009 Conference on Learning Theory*, 2009.

[BDR21] Adam Block, Yuval Dagan, and Sasha Rakhlin. Majorizing Measures, Sequential Complexities, and Online Learning. *arXiv:2102.01729 [cs, stat]*, February 2021. arXiv: 2102.01729.

[BEHW89] Anselm Blumer, A. Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, October 1989.

[BFR20] Blair Bilodeau, Dylan J. Foster, and Daniel M. Roy. Tight Bounds on Minimax Regret under Logarithmic Loss via Self-Concordance. *arXiv:2007.01160 [cs, stat]*, August 2020. arXiv: 2007.01160.

[Bha21] Siddharth Bhaskar. Thicket Density. *The Journal of Symbolic Logic*, 86(1):110–127, March 2021. arXiv: 1702.03956.

[BKP04] Olivier Bousquet, Vladimir Koltchinskii, and Dmitry Panchenko. Some Local Measures of Complexity of Convex Hulls and Generalization Bounds. *arXiv:math/0405340*, May 2004. arXiv: math/0405340.

[BLLW19] Sebastien Bubeck, Yuanzhi Li, Haipeng Luo, and Chen-Yu Wei. Improved Path-length Regret Bounds for Bandits. In *Proceedings of Machine Learning Research*, page 21, 2019.

[BLM20] Mark Bun, Roi Livni, and Shay Moran. An Equivalence Between Private Classification and Online Prediction. *arXiv:2003.00563 [cs, stat]*, March 2020. arXiv: 2003.00563.

[BT03] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, May 2003.

[CBL06] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, Cambridge; New York, 2006. OCLC: 70056026.

[CP20] Xi Chen and Binghui Peng. Hedging in games: Faster convergence of external and swap regrets. *arXiv:2006.04953 [cs]*, October 2020. arXiv: 2006.04953.

[DDK11] Constantinos Daskalakis, Alan Deckelbaum, and Anthony Kim. Near-Optimal No-Regret Algorithms for Zero-Sum Games. In *Proceedings of the 2011 Symposium on Discrete Algorithms*, page 20, 2011.

[DFG21] Constantinos Daskalakis, Maxwell Fishelson, and Noah Golowich. Near-Optimal No-Regret Learning in General Games. *arXiv:2108.06924 [cs]*, August 2021. arXiv: 2108.06924.

[DSS14] Amit Daniely and Shai Shalev-Shwartz. Optimal Learners for Multiclass Problems. *arXiv:1405.2420 [cs]*, May 2014. arXiv: 1405.2420.

[Dud78] R. M. Dudley. Central Limit Theorems for Empirical Measures. *The Annals of Probability*, 6(6):899–929, December 1978. Publisher: Institute of Mathematical Statistics.

[DYM21] Kefan Dong, Jiaqi Yang, and Tengyu Ma. Provable Model-based Nonlinear Bandit and Reinforcement Learning: Shelve Optimism, Embrace Virtual Curvature. *arXiv:2102.04168 [cs]*, May 2021. arXiv: 2102.04168.

[FK18] Dylan J. Foster and Akshay Krishnamurthy. Contextual bandits with surrogate losses: Margin bounds and efficient algorithms. *arXiv:1806.10745 [cs, stat]*, November 2018. arXiv: 1806.10745.

[FKL+18] Dylan J. Foster, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Logistic Regression: The Importance of Being Improper. *arXiv:1803.09349 [cs, stat]*, December 2018. arXiv: 1803.09349.

[FKS19] Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Optimistic Regret Minimization for Extensive-Form Games via Dilated Distance-Generating Functions. *arXiv:1910.10906 [cs, math]*, October 2019. arXiv: 1910.10906.

[FR20] Dylan J. Foster and Alexander Rakhlin. Beyond UCB: Optimal and Efficient Contextual Bandits with Regression Oracles. *arXiv:2002.04926 [cs, math, stat]*, February 2020. arXiv: 2002.04926.

[GG55] R. E. Greenwood and A. M. Gleason. Combinatorial Relations and Chromatic Graphs. *Canadian Journal of Mathematics*, 7:1–7, 1955. Publisher: Cambridge University Press.

[GGKM21] Badih Ghazi, Noah Golowich, Ravi Kumar, and Pasin Manurangsi. Near-tight closure bounds for the Littlestone and threshold dimensions. In *Proceedings of Machine Learning Research*, pages 1–11, 2021.

[Gol21] Noah Golowich. Differentially Private Nonparametric Regression Under a Growth Condition. In *Proceedings of Thirty Fourth Conference on Learning Theory*, pages 2149–2192. PMLR, July 2021. ISSN: 2640-3498.

[HAM21] Yu-Guan Hsieh, Kimon Antonakopoulos, and Panayotis Mertikopoulos. Adaptive Learning in Continuous Games: Optimal Regret Bounds and Convergence to Nash Equilibrium. *arXiv:2104.12761 [cs, math]*, April 2021. arXiv: 2104.12761.

[Han16] Steve Hanneke. The Optimal Sample Complexity of PAC Learning. *arXiv:1507.00473 [cs, stat]*, February 2016. arXiv: 1507.00473.

[HH97] Wilfrid Hodges and School of Mathematical Sciences Wilfrid Hodges. *A Shorter Model Theory*. Cambridge University Press, April 1997. Google-Books-ID: S6QYeuo4p1EC.

[HK16] Elad Hazan and Tomer Koren. The Computational Power of Optimization in Online Learning. *arXiv:1504.02089 [cs]*, January 2016. arXiv: 1504.02089.

[HKW98] D. Haussler, J. Kivinen, and M.K. Warmuth. Sequential prediction of individual sequences under general loss functions. *IEEE Transactions on Information Theory*, 44(5):1906–1925, September 1998. Conference Name: IEEE Transactions on Information Theory.

[HLM15] Elad Hazan, Roi Livni, and Yishay Mansour. Classification with Low Rank and Missing Data. *arXiv:1501.03273 [cs]*, January 2015. arXiv: 1501.03273.

[HLM21] Steve Hanneke, Roi Livni, and Shay Moran. Online Learning with Simple Predictors and a Combinatorial Characterization of Minimax in 0/1 Games. *arXiv:2102.01646 [cs, stat]*, February 2021. arXiv: 2102.01646.

[HLW94] D. Haussler, N. Littlestone, and M. K. Warmuth. Predicting {0, 1}-Functions on Randomly Drawn Points. *Information and Computation*, 115(2):248–292, December 1994.

[HM16] Elad Hazan and Tengyu Ma. A Non-generative Framework and Convex Relaxations for Unsupervised Learning. *arXiv:1610.01132 [cs, stat]*, December 2016. arXiv: 1610.01132.

[HRS15] Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. *arXiv:1509.01240 [cs, math, stat]*, September 2015. arXiv: 1509.01240.

[JKT20] Young Hun Jung, Baekjin Kim, and Ambuj Tewari. On the Equivalence between Online and Private Learnability beyond Binary Classification. *arXiv:2006.01980 [cs, stat]*, October 2020. arXiv: 2006.01980.

[Kol11] Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[KP04] Vladimir Koltchinskii and Dmitry Panchenko. Rademacher processes and bounding the risk of function learning. *arXiv:math/0405338*, May 2004. arXiv: math/0405338.

[KS21] Pieter Kleer and Hans Simon. Primal and Dual Combinatorial Dimensions. *arXiv:2108.10037 [cs, math]*, August 2021. arXiv: 2108.10037.

[KV05] Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, October 2005.

[Lit88] Nick Littlestone. Learning Quickly When Irrelevant Attributes Abound: A New Linear-threshold Algorithm. *Machine Learning*, 2:285–318, 1988.

[LRS15] Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with Square Loss: Localization through Offset Rademacher Complexity. *arXiv:1502.06134 [cs, math, stat]*, June 2015. arXiv: 1502.06134.

[Men02] S. Mendelson. Improving the sample complexity using global data. *IEEE Transactions on Information Theory*, 48(7):1977–1991, July 2002.

[Men14] Shahar Mendelson. Learning without concentration. In *Proceedings of The 27th Conference on Learning Theory*, pages 25–39. PMLR, May 2014. ISSN: 1938-7228.

[MHS19] Omar Montasser, Steve Hanneke, and Nathan Srebro. VC Classes are Adversarially Robustly Learnable, but Only Improperly. *arXiv:1902.04217 [cs, stat]*, July 2019. arXiv: 1902.04217.

[MV02] S. Mendelson and R. Vershynin. Entropy and the Combinatorial Dimension. *arXiv:math/0203275*, September 2002. arXiv: math/0203275.

[RS13] Alexander Rakhlin and Karthik Sridharan. Optimization, Learning, and Games with Predictable Sequences. *arXiv:1311.1869 [cs]*, November 2013. arXiv: 1311.1869.

[RS14a] Alexander Rakhlin and Karthik Sridharan. Online Nonparametric Regression. *arXiv:1402.2594 [cs, math, stat]*, February 2014. arXiv: 1402.2594.

[RS14b] Alexander Rakhlin and Karthik Sridharan. *Statistical Learning and Sequential Prediction*. 2014.

[RSS12] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Relax and Localize: From Value to Algorithms. *arXiv:1204.0870 [cs, stat]*, April 2012. arXiv: 1204.0870.

[RST11] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online Learning: Stochastic and Constrained Adversaries. *arXiv:1104.5070 [cs, stat]*, April 2011. arXiv: 1104.5070.

[RST15a] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online Learning via Sequential Complexities. *Journal of Machine Learning Research*, 16(1):155–186, 2015.

[RST15b] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Sequential complexities and uniform martingale laws of large numbers. *Probability Theory and Related Fields*, 161(1-2):111–153, February 2015.

[RST17] Alexander Rakhlin, Karthik Sridharan, and Alexandre B. Tsybakov. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23(2), May 2017. arXiv: 1308.1147.

[SALS15] Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E. Schapire. Fast Convergence of Regularized Learning in Games. *arXiv:1507.00407 [cs]*, December 2015. arXiv: 1507.00407.

[She78] Saharon Shelah. *Classification Theory and the Number of Non-isomorphic Models*. North-Holland Publishing Company, 1978. Google-Books-ID: rq6EAAAAIAAJ.

[Sim15] Hans U. Simon. An Almost Optimal PAC Algorithm. In *Proceedings of The 28th Conference on Learning Theory*, pages 1552–1563. PMLR, June 2015. ISSN: 1938-7228.

[SS11] Shai Shalev-Shwartz. *Online Learning and Online Convex Optimization*, volume 4. 2011.

[SST12] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Optimistic Rates for Learning with a Smooth Loss. *arXiv:1009.3896 [cs]*, November 2012. arXiv: 1009.3896.

[Tal94] Michel Talagrand. Sharper Bounds for Gaussian and Empirical Processes. *Annals of Probability*, 22(1):28–76, 1994.

[vdVW96] Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer New York, New York, NY, 1996.

[vEGM+15] Tim van Erven, Peter D Grunwald, Nishant A Mehta, Mark D Reid, and Robert C Williamson. Fast Rates in Statistical and Online Learning. *Journal of Machine Learning Research*, 16:1793–1861, 2015.

[VK06] Vladimir Naumovich Vapnik and Samuel Kotz. *Estimation of dependences based on empirical data*. Information science and statistics. Springer, New York, 2nd ed edition, 2006.

[Vov95] V. G. Vovk. A game of prediction with expert advice. In *Proceedings of the eighth annual conference on Computational learning theory*, COLT '95, pages 51–60, New York, NY, USA, July 1995. Association for Computing Machinery.

[Vov01] Volodya Vovk. Competitive On-Line Statistics. *International Statistical Review / Revue Internationale de Statistique*, 69(2):213–248, 2001. Publisher: [Wiley, International Statistical Institute (ISI)].

[WL18] Chen-Yu Wei and Haipeng Luo. More Adaptive Algorithms for Adversarial Bandits. In *Proceedings of Machine Learning Research*, page 29, 2018.

[WLA20] Chen-Yu Wei, Haipeng Luo, and Alekh Agarwal. Taking a hint: How to leverage loss predictors in contextual bandits? In *Proceedings of Machine Learning Research*, page 52, 2020.