Agnostic Online Learning

(Dávid Pál is eligible for Mark Fulk Award.)

Shai Ben-David and **Dávid Pál**
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, ON, Canada
{shai,dpal}@cs.uwaterloo.ca

Shai Shalev-Shwartz
Toyota Technological Institute
Chicago, IL, USA
shai@tti-c.org

Abstract

We study learnability of hypotheses classes in agnostic online prediction models. The analogous question in the PAC learning model [Valiant, 1984] was addressed by Haussler [1992] and others, who showed that the VC dimension characterization of the sample complexity of learnability extends to the agnostic (or "unrealizable") setting. In his influential work, Littlestone [1988] described a combinatorial characterization of hypothesis classes that are learnable in the online model. We extend Littlestone's results in two aspects. First, while Littlestone only dealt with the realizable case, namely, assuming there exists a hypothesis in the class that perfectly explains the entire data, we derive results for the non-realizable (agnostic) case as well. In particular, we describe several models of non-realizable data and derive upper and lower bounds on the achievable regret. Second, we extend the theory to include margin-based hypothesis classes, in which the prediction of each hypothesis is accompanied by a confidence value. We demonstrate how the newly developed theory seamlessly yields novel online regret bounds for the important class of large margin linear separators.

1 Introduction

The goal of this paper is to analyze the effects of the *structure* of hypothesis classes (or classes of experts) on their online learnability in agnostic (non-realizable) settings. In the online learning model, a learning algorithm observes instances in a sequential manner. On round t, after observing the tth instance, $\mathbf{x}_t \in \mathcal{X}$, the algorithm attempts to predict the label associated with the instance. For example, the instance can be a vector of barometric features and the learner should predict if it's going to rain tomorrow. Once the algorithm has made a prediction, it is told whether the prediction was correct (e.g. it was rainy today) and then uses this information to improve its prediction mechanism. The goal of the learning algorithm is simply to minimize the number of prediction mistakes it makes.

Littlestone [1988] studied online learnability in the realizable case. That is, it is assumed that the label associated with each instance in the sequence of examples is determined by a fixed, yet unknown, mapping rule, $h : \mathcal{X} \to \{0, 1\}$, taken from a predefined hypothesis class, denoted $\mathcal{H}$, which is known to the learner. For the realizable case, Littlestone provided a full combinatorial characterization of the learning complexity of a class. He defined a combinatorial measure, which we call Littlestone's dimension and denote $\mathrm{Ldim}(\mathcal{H})$, and showed that this quantity is the minimum (over all learning algorithms) of the maximum (over all sequences of examples) of the number of prediction mistakes. Furthermore, Littlestone's proof is constructive – there exists a generic optimal algorithm that, for every hypothesis class, $\mathcal{H}$, is guaranteed to make at most $\mathrm{Ldim}(\mathcal{H})$ prediction mistakes on any sequence of examples (even if the instances are chosen in adversarial manner).

Despite the elegance of Littlestone's theory, it received relatively little attention by online learning researchers. This might be partly attributed to the fact that the realizable assumption is rather strong. In recent years, much attention has been given to the unrealizable case, in which no hypothesis in $\mathcal{H}$ generates the labels, and we assume nothing about the "true" labeling mechanism. In the batch model of learning, the agnostic setting, formalized by Haussler [1992], has become the mainstream theoretical model for classification prediction and has been thoroughly analyzed. Rather than providing absolute error bounds, the bounds derived in that model are relative to the best error rates in some benchmark hypothesis class. The VC dimension of a class provides an almost-tight characterization of the worst case (relative) error rates of learning w.r.t. that class. Similarly, in the online learning model, in the unrealizable case there is no finite bound on the number of prediction mistakes. Instead, the analysis examines bounds which are relative to the performance of the benchmark hypothesis class. The learner's *regret* is the difference between the learner's number of mistakes and the number of mistakes of the optimal hypothesis in $\mathcal{H}$. The term 'regret' refers to how 'sorry' the learner is, in retrospect, for not to have followed the predictions of the optimal hypothesis. When analyzing the learner's performance using the notion of regret, the goal is to have a regret bound which grows sub-linearly with the number of examples. Put another way, the average per-round regret goes to zero as the number of examples tends to infinity.

A basic observation, noted by Cover [1965] in the context of universal prediction of individual sequences, is that even a trivial hypothesis class, containing just two functions,

cannot have a vanishing regret if one allows the labels to depend upon the learner's prediction. It is therefore common to impose a basic restriction on the way the labels are generated. Namely, to require that y_t will be determined by the (possibly adversarial) environment before the learner predicts $\hat{y}_t$ (or independently of that prediction)[1].

Using this assumption, and by allowing the learner to randomize his predictions, Littlestone and Warmuth [1994] presented a learning algorithm with vanishing expected regret for the case of finite hypothesis classes. The actual value of these regret bounds is determined by the size of the hypothesis class. However, the cardinality of $\mathcal{H}$ is a crude measure of its "learning complexity". For example, in the batch learning model, we know that the VC-dimension of a class yields tighter sample complexity bounds. Does there exist a similar combinatorial characterization for agnostic online learning? Are there infinite hypothesis classes that yield regret bounds that are sub-linear in the length of the instance sequence? And, given a class $\mathcal{H}$, what is the optimal online learning strategy? Our first contribution is to provide upper and lower bounds on the achievable regret for any hypothesis class $\mathcal{H}$ in terms of its Littlestone dimension - the same parameter that controls the number of mistakes in the realizable case.

The second contribution of this paper is the derivation of much tighter mistake bounds, for more restrictive label-generation scenarios. In particular, we derive absolute bounds on the number of mistakes (rather than regret bounds) assuming that the labels are generated by some $h \in \mathcal{H}$ but are contaminated by a stochastic bounded noise. This result is somewhat surprising as we can guarantee that the learner will make less mistakes than the number of noisy labels in the sequence of examples provided by the environment.

The third contribution of this paper is an extension of the online learnability to the case of margin-based hypotheses. That is, we now assume that each hypothesis is a mapping $h : \mathcal{X} \to \mathbb{R}$, where the actual prediction is based on $\mathrm{sign}(h(x))$ and the magnitude of $h(x)$ is a confidence level. This extension enables us to seamlessly derive online margin-based regret bounds for the class of linear separators. However, while most previous results bound the number of mistakes using a convex surrogate loss function (e.g. the hinge-loss) of a competing linear separator, we bound the number of mistakes using the number of margin mistakes of the competing linear separator.

1.1 Outline of Our Main Results

We denote by $(\mathbf{x}_t, y_t)$ the example observed on round t and by $\hat{y}_t$ the prediction of the learner on round t. The number of rounds is denoted by T. The bounds we present below depend on the Littlestone dimension of the hypothesis class $\mathcal{H}$, denoted $\mathrm{Ldim}(\mathcal{H})$. We formally define this combinatorial measure in Section 2. To simplify the presentation, we use the notation:

$$\tilde{\mathrm{L}}\mathrm{dim}(\mathcal{H}) \;=\; \min\left\{\ln(|\mathcal{H}|)\,,\, \mathrm{Ldim}(\mathcal{H})\ln(T)\right\}\,. \qquad (1)$$

[1]This dates back to early work in the context of game theory [Robbins, 1951, Blackwell, 1956, Hannan, 1957] and information theory [Cover and Shenhar, 1977, Feder et al., 1992].

Note that $\tilde{\mathrm{L}}\mathrm{dim}(\mathcal{H}) = \tilde{O}(\mathrm{Ldim}(\mathcal{H}))$, where the $\tilde{O}$ notation hides logarithmic factors. While for every class $\mathcal{H}$, $\mathrm{Ldim}(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$, there are classes for which $\mathrm{Ldim}(\mathcal{H})$ is much smaller than $\log_2(|\mathcal{H}|)$ (e.g., there are infinite size classes with a finite Littlestone dimension).
We first describe results for two main models of unrealizable data. At the end of this section, we discuss the margin-based extension.

Arbitrary labels In the first and less restrictive model, we make no assumptions regarding the origin of the sequence of examples.

The following theorem shows that if $\mathrm{Ldim}(\mathcal{H})$ is finite then there exist an online learning algorithm whose regret is bounded by $\tilde{O}(\sqrt{\mathrm{Ldim}(\mathcal{H})\,T})$. Our proof is constructive – in Section 3 we present a generic online algorithm which achieves this regret bound for every class, $\mathcal{H}$.

Theorem 1. *For any hypothesis class $\mathcal{H}$, there exists an online learning algorithm such that for any $h \in \mathcal{H}$ and any sequence of T examples we have*

$$\mathbb{E}\left[\sum_{t=1}^{T}|\hat{y}_t - y_t| - \sum_{t=1}^{T}|h(\mathbf{x}_t) - y_t|\right] \leq \sqrt{\tfrac{1}{2}\,\tilde{\mathrm{L}}\mathrm{dim}(\mathcal{H})\,T}\,,$$

where $\hat{y}_1, \ldots, \hat{y}_T$ are the learner's (randomized) predictions and expectation is w.r.t. the algorithm own randomization. Furthermore, no algorithm can achieve an expected regret bound smaller than $\Omega\left(\sqrt{\mathrm{Ldim}(\mathcal{H})\,T}\right)$.

As mentioned previously, in the realizable case, it is possible to derive the finite mistake bound $\mathrm{Ldim}(\mathcal{H})$. Our next theorem interpolates between the realizable case and the non-realizable case.

Theorem 2. *For any hypothesis class $\mathcal{H}$, and scalar $M^\star \geq 0$, there exists an online learning algorithm such that for any $h \in \mathcal{H}$ and any sequence of T examples that satisfies:*

$$\min_{h\in\mathcal{H}}\sum_{t=1}^{T}|h(\mathbf{x}_t) - y_t| \leq M^\star$$

we have

$$\mathbb{E}\left[\sum_{t=1}^{T}|\hat{y}_t - y_t|\right] \leq M^\star + \sqrt{2\,M^\star\,\tilde{\mathrm{L}}\mathrm{dim}(\mathcal{H})} + \tilde{\mathrm{L}}\mathrm{dim}(\mathcal{H})\,.$$

Note that in the realizable case $M^\star = 0$ and the above bound becomes $\tilde{O}(\mathrm{Ldim}(\mathcal{H}))$. Additionally, since for any sequence of T examples and any hypothesis h the cumulative number of mistakes of h is at most T, we can apply Theorem 2 with $M^\star = T$ and obtain a regret bound of the same order as of the bound in Theorem 1.

Bounded Stochastic Noise In the second model of unrealizable data, we do not allow the labels to be chosen arbitrarily. Instead, we assume that the labels are random variables, and we require that the Bayes optimal strategy for predicting the labels is in our hypothesis class $\mathcal{H}$. Formally, there exists $h \in \mathcal{H}$, such that $y_1, \ldots, y_T$ are random variables with $\Pr[y_t \neq h(\mathbf{x}_t)|\mathbf{x}_t] < 1/2$. Furthermore, we require that for some $\gamma \in [0, 1/2)$, $\Pr[y_t \neq h(\mathbf{x}_t)|\mathbf{x}_t] \leq \gamma$.

We provide a general online learning strategy for this case and analyze its performance. As in the first model, we allow the online learner to randomize its predictions. That is, now both $y_1, \ldots, y_T$ and $\hat{y}_1, \ldots, \hat{y}_T$ are sequences of random variables.

We can rewrite each label as $y_t = h(\mathbf{x}_t) + \nu_t$, where ν_t is a Bernoulli random variable with $\Pr[\nu_t = 1] \leq \gamma$ and the plus is modulus 2 (the xor operation). That is, y_t is a noisy version of the "true" label $h(\mathbf{x}_t)$. Based on this perspective, it makes sense to count the mistakes of the learner with respect to the "true" label $h(\mathbf{x}_t)$ rather than with respect to the noisy label y_t. The following theorem provides an absolute bound on the number of "true" mistakes of the learner.

Theorem 3. *For any hypothesis class $\mathcal{H}$, and noise level $\gamma \in [0, 1/2)$, there exists an online learning algorithm such that for any $h \in \mathcal{H}$ and a sequence of examples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_T, y_T)$, where each label is a random variable with $\Pr[y_t \neq h(\mathbf{x}_t)|\mathbf{x}_t] \leq \gamma$ we have*

$$\mathbb{E}\left[\sum_{t=1}^{T} |\hat{y}_t - h(\mathbf{x}_t)|\right] \;\leq\; \frac{\tilde{\mathrm{L}}\mathrm{dim}(\mathcal{H})}{1 - 2\sqrt{\gamma(1-\gamma)}}\,,$$

where expectation is with respect to the random labels $y_1, \ldots, y_T$ and the algorithm own randomization[2].

For this model, we do not have a tightly matching general lower bound. However, in Section 4.3 we prove lower bounds for this model and show that for some family of concept classes containing classes of arbitrary Littlestone dimension, for every $\mathcal{H}$ in that family, any learning algorithm for $\mathcal{H}$ makes $\Omega(\mathrm{Ldim}(\mathcal{H}) \ln(T))$ mistakes for some sequences of T instances.

It is interesting to compare the above mistake bound to the regret bound given in Theorem 1. Obviously, the conditions of Theorem 1 holds for the case of stochastic noise as well, and we can obtain the regret bound:

$$\mathbb{E}\left[\sum_{t=1}^{T} \left(|\hat{y}_t - y_t| - |y_t - h(\mathbf{x}_t)|\right)\right] \leq \sqrt{\tfrac{1}{2}\tilde{\mathrm{L}}\mathrm{dim}(\mathcal{H})\, T}\,.$$

To compare the above with Theorem 3, we use the triangle inequality to get that

$$|\hat{y}_t - y_t| = |\hat{y}_t - h(\mathbf{x}_t) + h(\mathbf{x}_t) - y_t| \leq |\hat{y}_t - h(\mathbf{x}_t)| + |y_t - h(\mathbf{x}_t)|$$

which gives

$$|\hat{y}_t - y_t| - |y_t - h(\mathbf{x}_t)| \leq |\hat{y}_t - h(\mathbf{x}_t)|\,.$$

Combining this with Theorem 3 we conclude that, for every $\mathcal{H}$ the algorithm promised by the theorem obtains

$$\mathbb{E}\left[\sum_{t=1}^{T} \left(|\hat{y}_t - y_t| - |y_t - h(\mathbf{x}_t)|\right)\right] \leq \frac{\tilde{\mathrm{L}}\mathrm{dim}(\mathcal{H})}{1 - 2\sqrt{\gamma(1-\gamma)}}\,. \tag{2}$$

That is, Theorem 3 also implies a regret bound, similar to the bound of Theorem 1. However, the dependence on T in the regret bound is exponentially better here: $\ln(T)$ instead of $\sqrt{\ln(T)\, T}$. This type of assumption, and the resulting fast rate, is somewhat similar to Massart noise condition in the agnostic PAC model (see e.g. Boucheron et al. [2005]).

Additionally, Theorem 3 gives us an absolute bound on the number of mistakes of the learner with respect to the "ground truth" labels, $h(\mathbf{x}_t)$. We usually think that the learner cannot hope to have an error rate which is smaller than the error rate of the teacher (environment). Quite surprisingly, in our case the number of mistakes the learner will make can be significantly smaller than the number of mistakes the teacher (environment) makes on the sequence. That is, the learner can overcome label noise very efficiently. For example, suppose that $\gamma = 0.25$ and $\mathcal{H}$ is a finite class. In this case, the expected error rate of the learner is upper bounded by $\frac{8\,\ln(|\mathcal{H}|)}{T}$, which goes to 0 with T, while the error rate of the environment is the constant 0.25.

Margin-based bounds So far, we assumed that each hypothesis is a mapping from $\mathcal{X}$ to $\{0, 1\}$. We now describe an extension for the case of margin-based hypotheses. We say that $\mathcal{H}$ is a *margin-based hypothesis class* if, each hypothesis $h \in \mathcal{H}$ is of the form $h : \mathcal{X} \to \mathbb{R}$, where the actual prediction is $\phi(h(\mathbf{x}))$ where

$$\phi(a) \;\stackrel{\text{def}}{=}\; \frac{\mathrm{sign}(a) + 1}{2} \in \{0, 1\}\,, \tag{3}$$

and $|h(\mathbf{x})|$ is a confidence in the prediction. We define a μ-margin mistake of a hypothesis $h : \mathcal{X} \to \mathbb{R}$ on an example $(\mathbf{x}, y)$ by $|h(\mathbf{x}) - y|_\mu$ where:

$$|a - y|_\mu \;\stackrel{\text{def}}{=}\; \begin{cases} 0 & \text{if } \phi(a) = y \;\wedge\; |a| \geq \mu \\ 1 & \text{otherwise} \end{cases} \tag{4}$$

That is, $|h(\mathbf{x}) - y|_\mu = 0$ if h classifies $\mathbf{x}$ correctly with a sufficient confidence.

We extend the notion of Littlestone's dimension to classes of margin-based hypotheses. The bounds we present below depend on this margin-based dimension of $\mathcal{H}$, denoted $\mathrm{Ldim}_\mu(\mathcal{H})$. Analogously to the definition of $\tilde{\mathrm{L}}\mathrm{dim}$, we use the notation

$$\tilde{\mathrm{L}}\mathrm{dim}_\mu(\mathcal{H}) \;=\; \min\left\{\ln(|\mathcal{H}|)\,,\, \mathrm{Ldim}_\mu(\mathcal{H}) \ln(T)\right\}\,. \tag{5}$$

The following two theorems are analogous to Theorem 1, Theorem 2, and Theorem 3.

Theorem 4. *For any margin-based hypothesis class $\mathcal{H}$, and any margin parameter $\mu > 0$, there exists an online learning algorithm such that for any $h \in \mathcal{H}$ and any sequence of T examples we have*

$$\mathbb{E}\left[\sum_{t=1}^{T} |\hat{y}_t - y_t| - \sum_{t=1}^{T} |h(\mathbf{x}_t) - y_t|_\mu\right] \leq \sqrt{\tfrac{1}{2}\,\tilde{\mathrm{L}}\mathrm{dim}_\mu(\mathcal{H})\, T}\,.$$

Additionally, for any $M^\star \geq 0$ there exists an online algorithm such that if $\min_h \sum_{t=1}^{T} |h(\mathbf{x}_t) - y_t|_\mu \leq M^\star$ we have

$$\mathbb{E}\left[\sum_{t=1}^{T} |\hat{y}_t - y_t|\right] \leq M^\star + \sqrt{2\, M^\star\, \tilde{\mathrm{L}}\mathrm{dim}(\mathcal{H})} + \tilde{\mathrm{L}}\mathrm{dim}(\mathcal{H})\,.$$

Theorem 5. *For any margin-based hypothesis class $\mathcal{H}$, margin parameter $\mu > 0$, and noise level $\gamma \in [0, 1/2)$, there exists an online learning algorithm such that for any*

[2]Recall that $\tilde{\mathrm{L}}\mathrm{dim}(\mathcal{H}) \;=\; \min\left\{\ln(|\mathcal{H}|)\,,\, \mathrm{Ldim}(\mathcal{H}) \ln(T)\right\}$.

$h \in \mathcal{H}$ and a sequence of examples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_T, y_T)$, such that $|h(\mathbf{x}_t)| \geq \mu$ for every $t \leq T$ and $\Pr[y_t \neq \phi(h(\mathbf{x}_t)) \,|\, \mathbf{x}_t] \leq \gamma$, we have

$$\mathbb{E}\left[\sum_{t=1}^{T} |\hat{y}_t - h(\mathbf{x}_t)|\right] \leq \frac{\tilde{\mathrm{Ldim}}_\mu(\mathcal{H})}{1 - 2\sqrt{\gamma(1-\gamma)}} .$$

An interesting margin-based class is the class of linear separators. In particular, let $\mathcal{X}$ be the unit L_2 ball of a Hilbert space and let $\mathcal{H} = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle \;:\; \|\mathbf{w}\|_2 \leq 1\}$. Then, we show that $\mathrm{Ldim}_\mu(\mathcal{H}) = \frac{1}{\mu^2}$. As a corollary, we obtain:

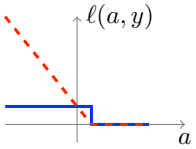
Corollary 6. *Let $\mu > 0$. There exists an online learning algorithm, such that for any $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_T, y_T)$, with $\|\mathbf{x}_t\|_2 \leq 1$ for all t, and for any $\mathbf{w} \in \mathbb{R}^d$ such that $\|\mathbf{w}\|_2 \leq 1$ we have*

$$\mathbb{E}\left[\sum_{t=1}^{T} |\hat{y}_t - y_t|\right] \leq \sum_{t=1}^{T} |\langle \mathbf{w}, \mathbf{x}_t \rangle - y_t|_\mu + \sqrt{\frac{T \ln(T)}{2\,\mu^2}} .$$

Additionally, for any $M^\star \geq 0$ there exists an online algorithm such that if $\min_{\mathbf{w}:\|\mathbf{w}\|_2 \leq 1} \sum_{t=1}^{T} |\langle \mathbf{w}, x_t \rangle - y_t|_\mu \leq M^\star$ we have

$$\mathbb{E}\left[\sum_{t=1}^{T} |\hat{y}_t - y_t|\right] \leq M^\star + \sqrt{2\,M^\star\,\ln(T)/\mu^2} + \ln(T)/\mu^2 .$$

We can also derive a bound for bounded stochastic noise using Theorem 5. It is important to notice that the function $|a - y_t|_\mu$ is non-convex. Bounds with this non-convex loss function are widely used in the batch setting. However, most if not all previous online learning bounds for linear separators are with the hinge-loss, $\ell(a, y_t) = \max\{0, 1 - (2y_t - 1)\frac{a}{\mu}\}$. This is mainly because the hinge-loss is a convex function, and previous learning algorithms assume (sometime implicitly) a convex loss function. The two loss functions, for the case $y = 1$, are illustrated on the left plots.



2 Background: Littlestone's Dimension and the Realizable Case

In this section we briefly overview Littlestone's results regarding online learnability in the realizable case and formally define Littlestone's dimension. See Littlestone [1988] for more details.

In the realizable case we assume that there exists $h \in \mathcal{H}$ such that for all t, $y_t = h(\mathbf{x}_t)$. The hypothesis h is called the *target*. We make no additional assumptions regarding the choice of the instances or the choice of the target hypothesis h. We look at worst case guarantees on the number of mistakes of an algorithm A. That is, we would like to upper bound

$$\mathcal{M}(A) = \max_{h \in \mathcal{H}} \max_{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T} \sum_{t=1}^{T} |\hat{y}_t - h(\mathbf{x}_t)| .$$

(where, $\hat{y}_t$ is the prediction of the online algorithm A on the instance x_t). We assume that the online algorithm is deterministic and so there is no expectation in the formula (see the discussion about randomized algorithms in the realizable case at the end of this section).

Given a hypothesis class $\mathcal{H}$, the natural question is how small $\mathcal{M}(A)$ can we make? In other words, what is the minimum of $\mathcal{M}(A)$ over all possible algorithms A. As mentioned above, Littlestone gave a nice combinatorial characterization of this quantity. He called the quantity *the optimal mistake bound*, and we refer to it as Littlestone's dimension, $\mathrm{Ldim}(\mathcal{H})$.

To define Ldim, we use the following notation. We consider trees whose internal nodes are labeled by instances. Therefore, any branch (i.e., a root-to-leaf path) can be described as a sequence of examples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_d, y_d)$, where $\mathbf{x}_i$ is the instances associated with the ith internal node in the path, and y_i is 1 if node $i+1$ in the path is the right child of the ith node, and otherwise $y_i = 0$. We require that there are no repetitions of an instance along any branch.

Definition 7 (Shattered tree)**.** *An instance-labeled tree is shattered by a class $\mathcal{H}$ if for any root-to-leaf path $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_d, y_d)$ there is some $h \in \mathcal{H}$ such that for all $i \leq d$, $h(\mathbf{x}_i) = y_i$ for all i.*

An illustration of an $\mathcal{H}$-shattered tree of depth 2 is given in Figure 1.

Definition 8 (Littlestone's dimension (Ldim))**.** *For a non-empty class, $\mathcal{H}$, $\mathrm{Ldim}(\mathcal{H})$ is the largest integer d such that there exist a full binary tree of depth d (i.e., any branch contains d-many non-leaf nodes) that is shattered by $\mathcal{H}$. (We define $\mathrm{Ldim}(\emptyset) = -1$.)*

The definition of Ldim immediately implies that for any class $\mathcal{H}$,

$$VC - dim(\mathcal{H}) \leq \mathrm{Ldim}(\mathcal{H}) \leq \log_2(\mathcal{H})$$

. The following two lemmas (due to Littlestone) establish the connection between$\mathrm{Ldim}(\mathcal{H})$ and online learnability of $\mathcal{H}$.:

Lemma 9. *The worst-case number of mistakes, $\mathcal{M}(A)$, of any deterministic learning algorithm A is at least $\mathrm{Ldim}(\mathcal{H})$.*

This can be seen by noting that the environment can choose an $\mathcal{H}$-shattered tree of depth $\mathrm{Ldim}(\mathcal{H})$ and, for any learning algorithm, present to the learner instances along a branch from the root to a leaf such that, for any instance, the label predicted by the algorithm turns out to be wrong.

Interestingly, there is a generic algorithm, A, such that for every class $\mathcal{H}$ with a finite Littlestone dimension, $\mathcal{M}(A) = \mathrm{Ldim}(\mathcal{H})$. Thus, $\mathrm{Ldim}(\mathcal{H})$ is the exact characterization of online learnability in the realizable case.

Algorithm 1 Standard Optimal Algorithm (SOA)

input: A hypothesis class $\mathcal{H}$
initialize: $V_0 = \mathcal{H}$
for $t = 1, 2, \ldots$
  receive $\mathbf{x}_t$
  for $r \in \{0, 1\}$ let $V_t^{(r)} = \{h \in V_{t-1} \;:\; h(\mathbf{x}_t) = r\}$
  predict $\hat{y}_t = \arg\max_r \mathrm{Ldim}(V_t^{(r)})$
  (in case of a tie predict $\hat{y}_t = 0$)
  receive true answer y_t
  update $V_t = V_t^{(y_t)}$

	h_1	h_2	h_3	h_4
$\mathbf{v}_1$	0	0	1	1
$\mathbf{v}_2$	0	1	?	?
$\mathbf{v}_3$	?	?	0	1

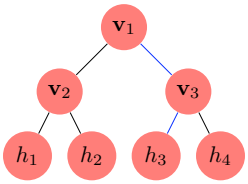


Figure 1: An illustration of an $\mathcal{H}$-shattered tree of depth 2. Internal nodes are labeled with instances $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ and the leaves are labeled with hypothesis h_1, h_2, h_3, h_4. The predictions of hypotheses on $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ is given in the table. A question mark means that $h_j(\mathbf{v}_i)$ can be either 1 or 0. For example, the path from the root to the leaf labeled h_3 corresponds to the sequence of examples $(\mathbf{v}_1, 1), (\mathbf{v}_3, 0)$, which can also be written as $(\mathbf{v}_1, h_3(\mathbf{v}_1)), (\mathbf{v}_3, h_3(\mathbf{v}_3))$.

The following lemma formally establishes the optimality of SOA.

Lemma 10. *For any class $\mathcal{H}$ with finite Littlestone dimension, the SOA algorithm makes at most $\mathrm{Ldim}(\mathcal{H})$ mistakes on any sequence of instances labeled by some $h \in \mathcal{H}$.*

The idea behind the proof is to note that, whenever SOA errs then the Ldim of the resulting version space (the space of all the hypotheses in $\mathcal{H}$ that are consistent with the labels presented by the environment, so far) goes down by at least 1.

Randomized Predictions To derive the lower bound, one can think of the environment as choosing y_t to be $\neg\hat{y}_t$. As we describe in the next section, in the unrealizable case, allowing the environment to base its label on the learner's prediction leads to non-vanishing regret. We circumvent this problem by assuming that the environment must decide on y_t before observing $\hat{y}_t$, and the learner is allowed to make randomized predictions (so the environment cannot predict these predictions).

This leads to the question, whether randomization helps in the realizable case. As it turns out, it does not help too much, since one can show that even if the learner is randomized, there exists a sequence of instances and a target hypothesis such the expected number of mistakes (with respect to the target) is at least $\mathrm{Ldim}(\mathcal{H})/2$. This easily follows by an averaging argument, since if the environment plays according to a root-to-leaf path chosen uniformly at random in a fixed mistake tree of depth $\mathrm{Ldim}(\mathcal{H})$, the expected number of mistakes is at least $\mathrm{Ldim}(\mathcal{H})/2$. Therefore, the distinction between randomized and deterministic learners in the realizable case is not significant.

3 Agnostic Online Learnability with Arbitrary Labels

In the previous section we have shown that Littlestone's dimension exactly characterizes the achievable mistake bounds in the realizable case. However, the realizable assumption is rather strong. The focus of this paper is the more realistic, unrealizable, case. In the unrealizable case, our goal is to minimize the regret with respect to a benchmark class of labeling functions, $\mathcal{H}$. That is, the difference between the learner's number of mistakes and the number of mistakes of the optimal hypothesis in $\mathcal{H}$.

As before, we are interested in a combinatorial measure that determines the optimal achievable regret bound for hypothesis classes. A natural candidate is the Littlestone's dimension. Recall that in the unrealizable case we assume that the environment must decide on y_t before observing $\hat{y}_t$, and the learner is allowed to make randomized predictions As a warm-up, we recall a well known expected regret bound under these assumptions for the case of finite hypotheses classes. That is in terms of the cardinality of $\mathcal{H}$. Then, we present the main result, constructing a generic online algorithm that has the expected regret bound $\sqrt{\tilde{\mathrm{Ldim}}(\mathcal{H})\, T}$ (regardless of the cardinality of $\mathcal{H}$). Finally, we provide a lower bound on the achievable regret.

3.1 An Expert Algorithm for Finite Classes

Let $\mathcal{H}$ be a finite hypothesis class. We can think on the hypotheses in $\mathcal{H}$ as "experts", and the goal of the online learning algorithm is to track the optimal expert. In the following we denote the set of experts by $\{f_1, f_2, \ldots, f_N\}$ (rather than by h_i's) since, the results apply to a more general case where "experts" do not necessarily have to be fixed functions.

One way to do this is by using the weighted majority algorithm [Littlestone and Warmuth, 1994]. The version of the algorithm we give here, as well as the regret bound, is based on [Cesa-Bianchi and Lugosi, 2006, Chapter 2].

Algorithm 2 Learning with Expert Advice

input: Number of experts N ; Learning rate $\eta > 0$
initialize: $\mathbf{w}^0 = (1, \ldots, 1) \in \mathbb{R}^N$; $Z_0 = N$
for $t = 1, 2, \ldots, n$
  receive expert advice $(f_1^t, f_2^t, \ldots, f_N^t) \in \{0,1\}^N$
  environment determine y_t without revealing it to learner
  define $\hat{p}_t = \frac{1}{Z_{t-1}} \sum_{i: f_i^t = 1} w_i^{t-1}$
  predict $\hat{y}_t = 1$ with probability $\hat{p}_t$
  receive label y_t
  update: $w_i^t = w_i^{t-1} \exp\left(-\eta |f_i^t - y_t|\right)$; $Z_t = \sum_{i=1}^N w_i^t$

The algorithm maintains a weight for each expert and makes a randomized prediction according to the relative mass of experts. Finally, the weights of experts that erred on the last example are diminished by a factor of $\exp(-\eta)$. The definition of $\hat{y}_t$ clearly implies that

$$\mathbb{E}[|\hat{y}_t - y_t|] = \frac{1}{Z_{t-1}} \sum_{i=1}^{N} w_i^{t-1} |f_i^t - y_t| \,. \tag{6}$$

The following theorem, whose proof can be easily derived from [Cesa-Bianchi and Lugosi, 2006, Chapter 2], analyzes the expected regret of the algorithm.

Theorem 11. *If we run Algorithm 2 with learning rate $\eta = \sqrt{8 \ln(N)/T}$ then the following expected regret bound holds:*

$$\sum_{t=1}^{T} \mathbb{E}[|\hat{y}_t - y_t|] - \min_{1 \le i \le N} \sum_{t=1}^{T} |f_i^t - y_t| \;\le\; \sqrt{\tfrac{1}{2} \ln(N)\, T} \,.$$

Similarly, if we run Algorithm 2 with learning rate $\eta = \ln(1+\sqrt{2\ln(d)/M^\star})$*, and if* $\min_i \sum_{t=1}^T |f_i^t - y_t| \le M^\star$ *then the following expected regret bound holds:*

$$\sum_{t=1}^{T} \mathbb{E}[|\hat{y}_t - y_t|] \le M^\star + \sqrt{2\,M^\star\,\ln(N)} + \ln(N)\,.$$

Note that since, for every finite class $\mathcal{H}$, $\tilde{\mathrm{Ldim}}(\mathcal{H}) \le \ln(|\mathcal{H}|)$ the bound of Theorem 11 is implied by Theorem 1 and Theorem 2.

3.2 Agnostic Online Learning and Littlestone's Dimension

The main result of this section replaces the parameter $\ln(|\mathcal{H}|)$ of Theorem 11 by the Littlestone dimension of $\mathcal{H}$. As mentioned above, this will allow us to improve the cardinality based bounds, as well as to extend them to infinite $\mathcal{H}$'s. We describe an algorithm for agnostic online learnability that achieves an expected regret bound of $\sqrt{\frac{1}{2}\,\mathrm{Ldim}(\mathcal{H})\,T\,\ln(T)}$.

As in the case of finite hypothesis classes, the main idea is to construct a set of experts and then to use the Learning-with-experts-advice algorithm. However, the expected regret of the Learning-with-experts-advice algorithm is $O(\sqrt{\ln(N)\,T})$, where N is the number of experts. Therefore, unless $|\mathcal{H}|$ is finite, we cannot use each $h \in \mathcal{H}$ as an expert. The challenge is therefore how to define a set of experts that on one hand is not excessively large while on the other hand contains experts that give accurate predictions.

We construct the set of experts so that for each hypothesis $h \in \mathcal{H}$ and every sequence of instances, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$, there exists at least one expert in the set which behaves exactly as h on these instances. For each $L \le \mathrm{Ldim}(\mathcal{H})$ and each sequence $1 \le i_1 < i_2 < \cdots < i_L \le T$ we define an expert. The expert simulates the game between SOA (Algorithm 1) and the environment on the sequence of instances $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$ assuming that SOA makes a mistake precisely in rounds $i_1, i_2, \ldots, i_L$. The expert is defined by the following algorithm.

Algorithm 3 $\mathrm{Expert}(i_1, i_2, \ldots, i_L)$

input A hypothesis class $\mathcal{H}$; Indices $i_1 < i_2 < \cdots < i_L$
initialize: $V_1 = \mathcal{H}$
for $t = 1, 2, \ldots, T$
  receive $\mathbf{x}_t$
  for $r \in \{0,1\}$ let $V_t^{(r)} = \{h \in V_t : h(\mathbf{x}_t) = r\}$
  define $\tilde{y}_t = \mathrm{argmax}_r\, \mathrm{Ldim}\left(V_t^{(r)}\right)$
    (in case of a tie set $\tilde{y}_t = 0$)
  if $t \in \{i_1, i_2, \ldots, i_L\}$
    predict $\hat{y}_t = \neg\tilde{y}_t$
  else
    predict $\hat{y}_t = \tilde{y}_t$
  update $V_{t+1} = V_t^{(\hat{y}_t)}$

The following key lemma shows that, on any sequence of instances, for each hypothesis $h \in \mathcal{H}$ there exists an expert with the same behavior.

Lemma 12. *Let* $\mathcal{H}$ *be any hypothesis class with* $\mathrm{Ldim}(\mathcal{H}) < \infty$*. Let* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$ *be any sequence of instances. For any* $h \in \mathcal{H}$*, there exists* $L \le \mathrm{Ldim}(\mathcal{H})$ *and indices* $1 \le i_1 < i_2 < \cdots < i_L \le T$ *such that when running* $\mathrm{Expert}(i_1, i_2, \ldots, i_L)$ *on the sequence* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$*, the expert predicts* $h(\mathbf{x}_t)$ *on each online round* $t = 1, 2, \ldots, T$*.*

Proof. Fix $h \in \mathcal{H}$ and the sequence $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$. We must construct L and the indices $i_1, i_2, \ldots, i_L$. Consider the Algorithm 1 (SOA) running on input $(\mathbf{x}_1, h(\mathbf{x}_1))$, $(\mathbf{x}_2, h(\mathbf{x}_2))$, $\ldots$, $(\mathbf{x}_T, h(\mathbf{x}_T))$. SOA makes at most $\mathrm{Ldim}(\mathcal{H})$ mistakes on such input. We define L to be the number of mistakes made by SOA and we define $\{i_1, i_2, \ldots, i_L\}$ to be the set of rounds in which SOA made the mistakes.

Now, consider the $\mathrm{Expert}(i_1, i_2, \ldots, i_L)$ running on the sequence $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$. By construction, the set V_t maintained by $\mathrm{Expert}(i_1, i_2, \ldots, i_L)$ equals to the set V_t maintained by SOA when running on the sequence $(\mathbf{x}_1, h(\mathbf{x}_1)), \ldots, (\mathbf{x}_T, h(\mathbf{x}_T))$. Since the predictions of SOA differ from the predictions of h if and only if the round is in $\{i_1, i_2, \ldots, i_L\}$, we conclude that the predictions of $\mathrm{Expert}(i_1, i_2, \ldots, i_L)$ are always the same as the predictions of h. ■

The above lemma holds in particular for the hypothesis in $\mathcal{H}$ that makes the least number of mistakes on the sequence of examples, and we therefore obtain the following:

Corollary 13. *Let* $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_T, y_T)$ *be a sequence of examples and let* $\mathcal{H}$ *be a hypothesis class with* $\mathrm{Ldim}(\mathcal{H}) < \infty$*. There exists* $L \le \mathrm{Ldim}(\mathcal{H})$ *and indices* $1 \le i_1 < i_2 < \cdots < i_L \le T$*, such that* $\mathrm{Expert}(i_1, i_2, \ldots, i_L)$ *makes at most as many mistakes as the best* $h \in \mathcal{H}$ *does. Namely,*

$$\min_{h \in \mathcal{H}} \sum_{t=1}^{T} |h(\mathbf{x}_t) - y_t|$$

mistakes on the sequence of examples.

Our agnostic online learning algorithm is now an application of the Learning-with-expert-advice algorithm.

Algorithm 4 Agnostic Online Learning Algorithm

input: A hypothesis class $\mathcal{H}$ with $\mathrm{Ldim}(\mathcal{H}) < \infty$;
  number of rounds T ; learning rate $\eta > 0$
for $L = 1, 2, \ldots, \mathrm{Ldim}(\mathcal{H})$
  foreach sub-sequence $1 \le i_1 < i_2 < \cdots < i_L \le T$
    construct an $\mathrm{Expert}(i_1, i_2, \ldots, i_L)$ as in Algorithm 3
run Algorithm 2 with the set of constructed experts and η

To analyze Algorithm 4 we combine Corollary 13 with the upper bound on the number of experts,

$$N = \sum_{L=0}^{\mathrm{Ldim}(\mathcal{H})} \binom{T}{L} \le T^{\mathrm{Ldim}(\mathcal{H})}\,, \tag{7}$$

and with Theorem 11. This proves the regret bounds given in Theorem 1 and Theorem 2.

3.3 A Matching Lower bound

In this section we prove the second part of Theorem 1 which states that no algorithm can achieve regret below $\Omega(\sqrt{\mathrm{Ldim}(\mathcal{H})T})$.

Lemma 14 (Lower Bound)**.** *Let $\mathcal{H}$ be any hypothesis class with a finite* $\mathrm{Ldim}(\mathcal{H})$*. For any (possibly randomized) algorithm, exists a sequence $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_T, y_T)$ such that*

$$\mathbb{E}\left[\sum_{t=1}^{T} |\hat{y}_t - y_t|\right] - \min_{h\in\mathcal{H}} \sum_{t=1}^{T} |h(\mathbf{x}_t) - y_t| \geq \sqrt{\frac{\mathrm{Ldim}(\mathcal{H})T}{8}} \,.$$

Proof. Let $d = \mathrm{Ldim}(\mathcal{H})$ and, for simplicity, assume that T is an integer multiple of d, say, $T = kd$ for some non-negative integer k. Consider a full binary $\mathcal{H}$-shattered tree of depth d. We construct the sequence $(\mathbf{x}_1, y_1)$, $(\mathbf{x}_2, y_2)$, $\ldots$, $(\mathbf{x}_T, y_T)$ by following a root-to-leaf path $(\mathbf{u}_1, z_1)$, $(\mathbf{u}_2, z_2)$, $\ldots$, $(\mathbf{u}_d, z_d)$ in the shattered tree. We pick the path in a top-down fashion starting at the root. The label $z_i \in \{0, 1\}$ determines whether the path moves to the left or to the right subtree of $\mathbf{u}_i$ and it thus determines $\mathbf{u}_{i+1}$.

Each node $\mathbf{u}_i$ on the path, $i = 1, 2, \ldots, d$, corresponds to a block $(\mathbf{x}_{(i-1)k+1}, y_{(i-1)k+1}), \ldots, (\mathbf{x}_{ik}, y_{ik})$ of k examples. We define $\mathbf{x}_{(i-1)k+1} = \mathbf{x}_{(i-1)k+2} = \cdots = \mathbf{x}_{ik} = \mathbf{u}_i$ and we choose $y_{(i-1)k+1}, \ldots, y_{ik}$ independently uniformly at random. For each block, let $T_i = \{(i-1)k + 1, \ldots, ik\}$ be the indices of the ith block. Denote $r = \sum_{t\in T_i} y_t$. We have

$$\min_{z_i\in\{0,1\}} \sum_{t\in T_i} |z_i - y_t| = \begin{cases} k - r & \text{if } r \geq k/2 \\ r & \text{if } r < k/2 \end{cases}$$

Therefore, $k/2 - \min_{z_i\in\{0,1\}} \sum_{t\in T_i} |z_i - y_t| = |r - k/2|$. Taking expectation over the y's and using Khinchine's inequality (see e.g. [Cesa-Bianchi and Lugosi, 2006, page 364]) we obtain

$$k/2 - \mathbb{E}\left[\min_{z_i\in\{0,1\}} \sum_{t\in T_i} |z_i - y_t|\right] = \mathbb{E}[|r - k/2|] \geq \sqrt{k/8} \,.$$

Next, we note that there exists $h \in \mathcal{H}$ such that for each block we have $h(\mathbf{u}_i) = z_i$. Thus, by summing over the blocks we get

$$\frac{dk}{2} - \mathbb{E}\left[\min_{h\in\mathcal{H}} \sum_{t=1}^{T} |h(\mathbf{x}_t) - y_t|\right] \geq d\sqrt{k/8} \,.$$

Finally, since $dk/2 = T/2 = \mathbb{E}[\sum_{t=1}^{T} |\hat{y}_t - y_t|]$, we conclude that the expected regret, w.r.t. the randomness of choosing the labels, is at least $d\sqrt{k/8} = \sqrt{dT/8}$. Therefore, there exists a particular sequence for which the regret is at least $\sqrt{dT/8}$, and this concludes our proof. ■

4 Online Learning with Bounded Stochastic Noise

We now turn to a second online learning model. While in this model the labels are still not required to be realizable by a hypothesis from $\mathcal{H}$, their generation is more restrictive than in the previously discussed "agnostic" model. The following setting models a scenario in which there exist some function $h \in \mathcal{H}$ that assigns "correct" labels to the instances. However, the information provided to the learner is a noisy version of these labels. In realistic situations, such a noise may stem from either communication issues or from an inherent weakness of the expert providing the labels. In a sense, this model is half way between the fully realizable model of Littlestone and the fully agnostic model we have discussed so far.

Formally, in the *Bounded Stochastic Noise* model we assume that there exists a hypothesis $h^* \in \mathcal{H}$, called the *target*, such that the labels provided to the learner, $y_1, y_2, \ldots, y_T$ are independent $\{0, 1\}$-valued random variables, such that for all t, $\Pr[h^*(\mathbf{x}_t) \neq y_t] \leq \gamma$, where $\gamma \in (0, 1/2)$ is a parameter of the model which we call the *noise rate* (or, more precisely, an upper bound on the noise rate).

Instead of regret, it is natural, in the stochastic model, to measure the performance of a learning algorithm A by the expected number of mistakes with respect to the target hypothesis,

$$\mathcal{M}(A) = \mathbb{E}\left[\sum_{t=1}^{T} |\hat{y}_t - h^\star(\mathbf{x}_t)|\right] \,,$$

where expectation is with respect to both the random choice of labels (the noise) and the internal randomization of the algorithm (and, as before, the $\hat{y}_t$'s are A's predictions for the instances, x_t). Anyway, the number $\mathcal{M}(A)$ is closely related to the regret,

$$\mathcal{R}(A) = \mathbb{E}\left[\sum_{t=1}^{T} |\hat{y}_t - y_t|\right] - \min_{h\in\mathcal{H}} \mathbb{E}\left[\sum_{t=1}^{T} |\hat{y}_t - h(\mathbf{x}_t)|\right] \,,$$

where the second expectation is taken with respect to the random choice of the labels y_t's and the first is taken with respect to both the random choice of y_t's and the internal randomization of the learning algorithm. Recall that, as shown in the introduction, it is always the case that $\mathcal{R}(A) \leq \mathcal{M}(A)$. On the other hand, note that $\mathcal{R}(A) \geq (1 - 2\gamma)\mathcal{M}(A)$, since any mistake with respect to true label $h(\mathbf{x}_t)$ adds to the regret at least $(1 - 2\Pr[h(\mathbf{x}_t) \neq y_t]) \geq 1 - 2\gamma$. In other words, $\mathcal{R}(A)$ and $\mathcal{M}(A)$ are within constant factor of each other.

To learn in the stochastic model, we use the learning algorithms which mimic those we used in the agnostic model (Section 3). The differences are in the learning rate η, and the resulting bounds on the number of mistakes. For a finite hypothesis class $\mathcal{H}$ and any fixed upper bound $\gamma < 1/2$ on the noise rate, we show that the expected number of mistakes is upper bounded by $O(\ln |\mathcal{H}|)$. When $\mathcal{H}$ is infinite, we use the Algorithm 4 which simulates a class of $O(T^{\mathrm{Ldim}(\mathcal{H})})$ experts, and apply the expert learning result for finite sets of experts to obtain an upper bound $O(\ln(T^{\mathrm{Ldim}(\mathcal{H})})) = O(\mathrm{Ldim}(\mathcal{H}) \ln(T))$ for any fixed (upper bound on the) noise rate $\gamma < 1/2$.

For this learning model, rather than providing a general lower bound on the expected number of mistakes, we provide a lower bound only for a specific family of hypothesis classes. Let $\mathcal{H}_k$ denote the hypothesis class which contains all the hypotheses that assign label 1 to k-many points of the domain. Note that, for every k, $\mathrm{Ldim}(\mathcal{H}_k) = k$. For

such classes over a finite domain (and hence finite hypothesis class), we show $\Omega(\ln(|\mathcal{H}_k|))$ lower bound on the number of mistakes for any $\gamma \in (0, 1/2)$. For infinite domains (and hence infinite classes $\mathcal{H}_k$), we prove a lower bound of $\Omega(k \ln(T))$. Note that for this particular family of classes (the $\mathcal{H}_k$'s) the two lower bounds match our upper bounds. It remains an open question to prove similar lower bounds for general hypothesis classes.

4.1 An Expert Algorithm for Finite Classes

This section is the counterpart of Section 3.1. We assume that $\mathcal{H} = \{h_1, h_2, \ldots, h_N\}$ is finite and think of the hypotheses as "experts". The goal of the learning algorithm, essentially, is to find the target expert that did not make any "true" mistake, despite that true labels are corrupted by random noise with rate up to γ. We use the Algorithm 2 for learning with expert advice.

More precisely, we think of the expert advice $f^1, f^2, \ldots$ as being deterministic and we assume that the feedback labels $y_1, y_2, \ldots$ are independent random variables. We assume that there exists $i \in \{1, 2, \ldots, N\}$ such that $\Pr[y_t \neq f_i^t] \leq \gamma$ for all t. The "true" mistakes are counted with respect to the predictions of the target expert. Formally, the true number of mistakes is $\sum_{t=1}^{T} |\hat{y}_t - f_i^t|$. The following theorem gives an upper bound on the expected number of true mistakes of the algorithm.

Theorem 15. *For $\gamma \in [0, 1/2)$, if we run Algorithm 2 with learning rate $\eta = \frac{1}{2}\ln(\frac{1-\gamma}{\gamma})$ with respect to a set of experts $\{f_1, \ldots, f_N\}$, then if for some $i \in \{1, \ldots, N\}$, the labels y_t (randomly generated by the environment) are such that $\Pr(y_t \neq f_i^t) \leq \gamma$ for all t then,*

$$\mathbb{E}\left[\sum_{t=1}^{T} |\hat{y}_t - f_i^t|\right] \leq \frac{1}{1 - 2\sqrt{\gamma(1-\gamma)}} \ln(N)$$

where the expectation is taken with respect to both y_t's and the internal randomization of the algorithm.

Proof. Fix γ, i, T. Suppose that at the end of round t the algorithm has computed a certain weight vector $\mathbf{w}_t$. We show that in the following rounds $s = t+1, t+2, \ldots, T$ the expected number of mistakes the algorithm makes is at most

$$\mathbb{E}\left[\sum_{s=t+1}^{T} |\hat{y}_s - f_i^s| \,\middle|\, \mathbf{w}^t\right] \leq C_\gamma \ln\left(\frac{Z_t}{w_i^t}\right) \tag{8}$$

where $Z_t = \sum_{i=1}^{d} w_i^t$ and $C_\gamma = \frac{1}{1-2\sqrt{\gamma(1-\gamma)}}$ is a positive constant that depends only on γ. If we substitute $t = 0$ in the inequality, the theorem follows, since $Z_0 = d$ and $w_i^0 = 1$.

We prove inequality (8) by backward inductions on t. The base case $t = T$ is trivial, since the left side is zero and the right side is positive because $Z_T / w_i^T > 1$. Suppose that the inequality holds true for t, we show it for $t-1$. Let

$$u = \sum_{\substack{1 \leq j \leq N \\ f_j^t = f_i^t}} w_j^{t-1} \qquad v = \sum_{\substack{1 \leq j \leq N \\ f_j^t \neq f_i^t}} w_j^{t-1}$$

be the total weights of the experts that in round t answer correctly and incorrectly respectively. Clearly, $v = Z_{t-1} - u$. The probability that in round t the algorithm makes a mistake is $\Pr[\hat{y}_t \neq f_i^t \mid \mathbf{w}_{t-1}] = v/Z_{t-1}$. Then, with probability $p = \Pr[y_t \neq f_i^t] \leq \gamma$ the algorithm receives incorrect feedback $y_t = \neg f_i^t$ and updates the total weight to $Z_t = e^{-\eta}u + v$ and target expert's weight $w_i^t = e^{-\eta} w_i^{t-1}$. With probability $1 - p$ the algorithm receives correct feedback $y_t = f_i^t$ and updates total weight to $Z_t = u + e^{-\eta}v$ and target expert's weight remains unchanged $w_i^t = w_i^{t-1}$. And so we have,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{s=t}^{T} |\hat{y}_s - f_i^s| \,\middle|\, \mathbf{w}^{t-1}\right] = v/Z_{t-1} + \mathbb{E}\left[\sum_{s=t+1}^{T} |\hat{y}_s - f_i^s| \,\middle|\, \mathbf{w}^t\right] \\
&\leq v/Z_{t-1} + \mathbb{E}\left[C_\gamma \ln\left(\frac{Z_t}{w_i^t}\right) \,\middle|\, \mathbf{w}^t\right] \\
&= v/Z_{t-1} + pC_\gamma \ln\left(\frac{e^{-\eta}u + v}{e^{-\eta} w_i^{t-1}}\right) \\
&\quad + (1-p)C_\gamma \ln\left(\frac{u + e^{-\eta}v}{w_i^{t-1}}\right)
\end{aligned}$$

where in the second step we have used the induction hypothesis. It remains to show that the last expression is bounded by $C_\gamma \ln(Z_{t-1}/w_i^t)$. Canceling w_i^{t-1} and recalling that $u + v = Z_{t-1}$, this is equivalent to showing that

$$\begin{aligned}
v/(u+v) + pC_\gamma \ln(u + e^\eta v) + (1-p)C_\gamma \ln(u + e^{-\eta}v) \\
\leq C_\gamma \ln(u+v)
\end{aligned}$$

Moreover, since for any $\alpha > 0$ the inequality holds for a pair $(\alpha u, \alpha v)$ if and only it holds for (u, v), we can without loss of generality assume that $u + v = 1$ (and $u, v \in (0, 1)$) and so we are left to show

$$v + pC_\gamma \ln(u + e^\eta v) + (1-p)C_\gamma \ln(u + e^{-\eta}v) \leq 0$$

Substituting $u = 1 - v$ we can define a real-valued function f of one real parameter v

$$f(v) := v + pC_\gamma \ln(1 - v + e^\eta v) + (1-p)C_\gamma \ln(1 - v + e^{-\eta}v)$$

and we thus must show that f is non-positive on the interval $(0, 1)$. We use the inequality $\ln(1 + x) \leq x$ valid for all $x > -1$ and obtain

$$\begin{aligned}
f(v) &\leq v + pC_\gamma(-v + e^\eta v) + (1-p)C_\gamma(-v + e^{-\eta}v) \\
&= v(1 + C_\gamma(e^\eta p - 1 + (1-p)e^{-\eta}))
\end{aligned}$$

Note that the last expression is a linear function in v. Thus, in order to show that last expression is non-positive for all $v \in (0, 1)$, it suffices to show that its slope is non-positive:

$$\begin{aligned}
&1 + C_\gamma(e^\eta p - 1 + (1-p)e^{-\eta}) \\
&= 1 + C_\gamma(p \underbrace{(e^\eta - e^{-\eta})}_{\geq 0} + e^{-\eta} - 1) \\
&\leq 1 + C_\gamma(\gamma(e^\eta - e^{-\eta}) + e^{-\eta} - 1) \\
&= 1 + C_\gamma\left(\gamma\left(\sqrt{\frac{1-\gamma}{\gamma}} - \sqrt{\frac{\gamma}{1-\gamma}}\right) + \sqrt{\frac{\gamma}{1-\gamma}} - 1\right) \\
&= 1 + C_\gamma\left(\gamma \frac{1-2\gamma}{\sqrt{(1-\gamma)\gamma}} + \sqrt{\frac{\gamma}{1-\gamma}} - 1\right) \\
&= 1 + C_\gamma\left(2\sqrt{(1-\gamma)\gamma} - 1\right) \\
&= 1 - 1 = 0\,.
\end{aligned}$$

This concludes the proof. ■

The bound we have just proved is a weakened version of Theorem 3, where $\ln(|\mathcal{H}|)$ replaces $\tilde{\mathrm{Ldim}}(\mathcal{H})$.

4.2 $\mathrm{Ldim}(H)$-Based Bound and Infinite Hypothesis Classes

For infinite hypothesis class $\mathcal{H}$, we take the same approach as in Section 3.2 and use Algorithm 4. The key ingredient is Lemma 12 which guarantees that regardless of the target hypothesis and the sequence of instances, there exists an $\mathrm{Expert}(i_1, i_2, \ldots, i_L)$ with exactly the same predictions as the target hypothesis. In other words, it does not matter whether we count learner's mistakes with respect to the target or with respect to $\mathrm{Expert}(i_1, i_2, \ldots, i_L)$. Combining this with the upper bound (7) on number of experts, Theorem 15 concludes the proof of Theorem 3.

4.3 Two Lower Bounds

In this section, we present two lower bounds on the expected number of mistakes in stochastic model. Both lower bounds are for the hypothesis classes of the form $\mathcal{H}_k$ consisting of all functions $h : \mathcal{X} \to \{0, 1\}$ that assign label 1 to exactly k domain points (formally $|h^{-1}(1)| = k$). It is easy to see that, for every value of k, $\mathrm{Ldim}(\mathcal{H}_k) = k$, provided the size of the domain is at least $2k$. The first, $\Omega(\ln(|\mathcal{H}_k|))$, lower bound is for the case when the domain (and hence $\mathcal{H}_k$) is finite. The second, $\Omega(k \ln(T))$, lower bound is for the case when the domain (and hence $\mathcal{H}_k$) is infinite. The main technical tool for both lower bounds is the following lemma, which takes care of the case when the domain is finite and $k = 1$.

Lemma 16. *Let $\gamma \in (0, 1/2)$, let A be any learning algorithm, and suppose $\mathcal{X}$ is finite. There exists a target $h^* \in \mathcal{H}_1$, time horizon $T = O(|\mathcal{X}| \ln(|\mathcal{X}|))$, a sequence $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$ of instances and a sequence of independent random variables $y_1, y_2, \ldots, y_T$ with $\Pr[y_t \neq h^*(\mathbf{x}_t)] = \gamma$ such the expected number of mistakes of A is at least*

$$\mathbb{E}\left[\sum_{t=1}^{T} |\hat{y}_t - h^*(\mathbf{x}_t)|\right] \geq \Omega(\ln(|\mathcal{X}|)) \ .$$

For a proof see the full version Ben-David et al. [2009].

Theorem 17. *Let $\gamma \in (0, 1/2)$, let A be any learning algorithm, and suppose $|\mathcal{X}| \geq 2k$ is finite. There exists a target $h^* \in \mathcal{H}_k$, time horizon $T = O(|\mathcal{X}| \log(|\mathcal{X}|/k))$, sequence of instances $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$ and sequence of independent random variables $y_1, y_2, \ldots, y_T \in \{0, 1\}$ with $\Pr[h^*(\mathbf{x}_t) \neq y_t] = \gamma$ for all t, such that the expected number of mistakes of A is at least*

$$\mathbb{E}\left[\sum_{t=1}^{T} |\hat{y}_t - h^*(\mathbf{x}_t)|\right] \geq \Omega(\log |\mathcal{H}_k|) \ .$$

Proof. Let $n = |\mathcal{X}|$. We split $\mathcal{X}$ into k disjoint subsets $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_k$ each of size $\Theta(n/k)$. The adversary picks the target $h^* \in \mathcal{H}_k$ such that in each $\mathcal{X}_i$ there is exactly one point $\mathbf{x}_i^*$ on which h^* attains value 1. According to Theorem 16, the adversary can choose the point in $\mathbf{x}_i^*$ and a sequence of $O((n/k) \log(n/k))$ instances in $\mathcal{X}_i$ such that learner makes in expectation at least $\Omega(\log(n/k))$ mistakes. Concatenating the sequences together the adversary obtains a sequence of $O(n \log(n/k))$ instances on which the learner makes $\Omega(k \log(n/k)) = \Omega(\log |\mathcal{H}_k|)$ mistakes. ■

Theorem 18. *Let $\gamma \in (0, 1/2)$, let A be any learning algorithm, and suppose $\mathcal{X}$ is infinite. For any time horizon $T \geq 2k$, there exists a target $h^* \in \mathcal{H}_k$ and a sequence of T instances such that A makes at least $\Omega(k \log T)$ mistakes in expectation.*

Proof. Consider subset $\mathcal{X}_0$ of the domain of size $n = \Theta(T/\log(T/k))$ points. By Theorem 17 there exists a target $h^* \in \mathcal{H}_k|_{\mathcal{X}_0}$ and sequence of instances in $\mathcal{X}_0$ of length

$$\begin{aligned}\Theta(n \log(n/k)) &= \Theta\left(\frac{T}{\log(T/k)} \log\left(\frac{T}{k \log(T/k)}\right)\right) \\ &= \Theta(T)\end{aligned}$$

such that the learning algorithm makes

$$\Omega\left(k \log(n/k)\right) = \Omega\left(k \log\left(\frac{T}{k \log(T/k)}\right)\right) = \Omega(k \log T)$$

mistakes. We choose the size of $\mathcal{X}_0$ so that constant hidden in the $\Theta(T/\log(T/k))$ notation guarantees that the length of the sequence of instances is exactly T. ■

5 Margin-based hypothesis classes

In this section we extend our results to the case of margin-based hypothesis classes. By doing so, we will obtain new regret bounds for the class of linear separators with large margin – which is maybe the most popular hypothesis class.

Recall that margin-based hypotheses are mappings $h : \mathcal{X} \to \mathbb{R}$, where the prediction is $\phi(h(\mathbf{x}))$, $\phi(a) = \frac{1}{2}(\mathrm{sign}(a)+1)$, and $|h(\mathbf{x})|$ is the confidence in the prediction. Recall also the definition of $|a - y|_\mu$ given in Eq. (4).

Our first step is to generalize the notion of a shattered tree to margin-based hypotheses. This is done by simply replacing the loss $|a - y|$ with the loss $|a - y|_\mu$.

Definition 19 (μ-shattered tree). *A d-depth full binary tree, with an instance associated with each node, is μ-shattered by a margin-based hypothesis class $\mathcal{H}$ if for any root-to-leaf path, $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_d, y_d)$, there exists $h \in \mathcal{H}$ such that $\sum_{t=1}^{d} |h(\mathbf{x}_t) - y_t|_\mu = 0$.*

Definition 20 (Margin-based Littlestone's dimension). *The margin-based Littlestone's dimension of a class $\mathcal{H}$, denoted $\mathrm{Ldim}_\mu(\mathcal{H})$, is the largest integer d such that there exists a d-depth tree that is μ-shattered by $\mathcal{H}$. (We define $\mathrm{Ldim}(\emptyset) = -\infty$.)*

It is easy to verify that if $\mathrm{Ldim}_\mu(\mathcal{H}) = L$ then for any learning algorithm, there exists a sequence $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_L, y_L)$ on which the algorithm makes L mistakes, while there exists $h \in \mathcal{H}$ such that $\sum_{t=1}^{d} |h(\mathbf{x}_t) - y_t|_\mu = 0$. The following theorem shows that we can also modify the SOA algorithm for margin based hypothesis classes.

Theorem 21. *Let $\mathcal{H}$ be a margin-based hypothesis class with $\mathrm{Ldim}_\mu(\mathcal{H}) < \infty$. Let $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_T, y_T)$ be a sequence of examples such that there exists $h \in \mathcal{H}$ for which $\sum_{t=1}^{T} |h(\mathbf{x}_t) - y_t|_\mu = 0$. Then, if we run SOA (Algorithm 1) on the sequence, while replacing Ldim with Ldim_μ, then the number of prediction mistakes is at most $\mathrm{Ldim}_\mu(\mathcal{H})$.*

The proof is analogous to the proof of Lemma 10 and is omitted due to lack of space.

The above theorem implies that as in previous sections, we can construct a not-to-large pool of experts, that mimics the predictions of all hypotheses in $\mathcal{H}$, as long as $\mathrm{Ldim}_\mu(\mathcal{H})$ is small. Formally, the following lemma is the analog of Lemma 12.

Lemma 22. *Let $\mathcal{H}$ be any margin-based hypothesis class with $\mathrm{Ldim}_\mu(\mathcal{H}) < \infty$. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$ be any sequence of instances. For any $h \in \mathcal{H}$, there exists $L \leq \mathrm{Ldim}_\mu(\mathcal{H})$ and indices $1 \leq i_1 < i_2 < \cdots < i_L \leq T$ such that when running $\mathrm{Expert}(i_1, i_2, \ldots, i_L)$ on the sequence $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$, while replacing Ldim with Ldim_μ, the expert predicts $\phi(h(\mathbf{x}_t))$ on each online round $t = 1, 2, \ldots, T$.*

The proof is again similar to the proof of Lemma 12 and is omitted due to lack of space. The proofs of Theorems 4 and 5 follow from the above lemma using the same technique as in previous sections.

Finally, we demonstrate the usefulness of our general theory by considering a specific margin based class. Let $\mathcal{X} = \{\mathbf{x} : \|\mathbf{x}\|_2 \leq 1\}$ be the unit ℓ_2 ball of some Hilbert space, and let the hypothesis class be linear separators, $\mathcal{H} = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle : \|\mathbf{w}\|_2 \leq 1\}$. The well known, Novikoff [1962], mistake bound for the Perceptron algorithm immediately implies that $\mathrm{Ldim}_\mu(\mathcal{H}) \leq 1/\mu^2$. Combining this with Theorem 4 yields Corollary 6.

6 Extensions and Discussion

We extended Littlestone's theory of online learnability in two ways. First, while Littlestone focused only on the realizable case, we give upper and lower regret bounds for the non-realizable case as well. Second, we also consider margin-based hypothesis classes. The elegance of the theory enables us to seamlessly derive novel online regret bounds for the important class of linear separators. There are several extensions to this work which we didn't discuss due to lack of space.

Automatically tuning parameters. All our algorithms receive the time horizon T (or the parameter $M^\star$) as part of their input, which might be not realistic. Using the standard doubling trick, this dependence on T or $M^\star$ can be easily removed. See for example [Cesa-Bianchi and Lugosi, 2006, Chapter 2] for details.

Adaptive vs. Oblivious Environments For simplicity of presentation, we implicitly assume an oblivious environment. However, our results can easily be adapted to adaptive environment (see the discussion of this point in [Cesa-Bianchi and Lugosi, 2006, Page 69].

Comparison to batch learning Many of the results we derived in this paper share similarity with results obtained for the batch learning model. For example, the per-round regret bound $\tilde{O}(\sqrt{\mathrm{Ldim}(\mathcal{H})/T})$ is similar to the generalization bound $O(\sqrt{\mathrm{VCdim}(\mathcal{H})/T})$ for batch learning. Also, margin based bounds, and faster rates under noise conditions also appear in the analysis of batch learning algorithms. It is therefore interesting to mirror additional results such as Tsybakov noise condition (see e.g. Boucheron et al. [2005]).

Computational complexity The focus of this paper was the existence of algorithms and the resulting regret bounds, rather than computational efficiency. The algorithms presented here are based on Littlestone's algorithm that needs to compute the Ldim of sub-classes of $\mathcal{H}$ repeatedly. It turns out that this is a computationally hard problem (at least as hard as computing the VC-dimension, see Frances and Litman [1998]). Ignoring this issue (we are computing the dimension only for sub-classes on a fixed $\mathcal{H}$, which may be much easier than the worst case general problem), we are running $T^{\mathrm{Ldim}(\mathcal{H})}$ many experts, each making a constant time computation for each label prediction.

Open Questions. The main open question is to close the $O(\sqrt{\log T})$ gap between $O(\sqrt{\mathrm{Ldim}(\mathcal{H}) \log(T) T})$ regret upper bound and $\Omega(\sqrt{\mathrm{Ldim}(\mathcal{H}) T})$ lower bound. It seems (and we believe it) that $\mathrm{Ldim}(\mathcal{H})$ is the key quantity characterizing the worst-case regret with respect to $\mathcal{H}$. Theoretically, however, the factor $O(\sqrt{\log T})$ can hide a surprise and defeat our belief.

The second open question is to find a lower bound on the expected mistakes in the stochastic model with noise rate γ for arbitrary class $\mathcal{H}$. It seems that for finite $\mathcal{H}$ the number of mistakes does not depend on $\mathrm{Ldim}(\mathcal{H})$ and rather it depends only on the cardinality of $\mathcal{H}$. The case when $\mathcal{H}$ is infinite is also interesting, however, the main obstacle seems to be the lack of examples of interesting infinite hypothesis classes with finite Littlestone's dimension.

Acknowledgements. We are grateful to Miroslava Sotáková for fruitful discussions about the stochastic model.

References

S. Ben-David, D. Pal, and S. Shalev-Shwartz. Agnostic online learning, 2009. Available at www.cs.uwaterloo.ca/∼shai/publications/agnostic-online.pdf.

D. Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6(1):1–8, Spring 1956.

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

T.M. Cover. Behavior of sequential predictors of binary sequences. *Trans. 4th Prague Conf. Information Theory Statistical Decision Functions, Random Processes*, 1965.

T.M. Cover and A. Shenhar. Compound Bayes predictors for sequences with apparent Markov structure. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-7 (6):421–424, June 1977.

M. Feder, N. Merhav, and M. Gutman. Universal prediction of individual sequences. *IEEE Transactions on Information Theory*, 38:1258–1270, 1992.

Moti Frances and Ami Litman. Optimal mistake bound learning is hard. *Inf. Comput.*, 144(1):66–82, 1998.

J. Hannan. Approximation to Bayes risk in repeated play. In M. Dresher, A. W. Tucker, and P. Wolfe, editors, *Contributions to the Theory of Games*, volume III, pages 97–139. Princeton University Press, 1957.

D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1988.

N. Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108:212–261, 1994.

A. B. J. Novikoff. On convergence proofs on perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*, volume XII, pages 615–622, 1962.

H. Robbins. Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the 2nd Berkeley symposium on mathematical statistics and probability*, pages 131–148, 1951.

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