THE BULLETIN OF SYMBOLIC LOGIC
Volume 25, Number 3, September 2019

MODEL THEORY AND MACHINE LEARNING

HUNTER CHASE AND JAMES FREITAG

Abstract. About 25 years ago, it came to light that a single combinatorial property determines both an important dividing line in model theory (NIP) and machine learning (PAC-learnability). The following years saw a fruitful exchange of ideas between PAC-learning and the model theory of NIP structures. In this article, we point out a new and similar connection between model theory and machine learning, this time developing a correspondence between *stability* and learnability in various settings of *online learning*. In particular, this gives many new examples of mathematically interesting classes which are learnable in the online setting.

§1. Introduction. The purpose of this note is to describe the connections between several notions of computational learning theory and model theory. The connection between *probably approximately correct* (PAC) learning and the nonindependence property (NIP) is well known and was originally noticed by Laskowski [8]. In the ensuing years, there have been numerous interactions between the combinatorics associated with PAC-learning and model theory in the NIP setting. Below, we provide a quick introduction to the PAC-learning setting as well as learning in general. Our main purpose, however, is to explain a *new connection* between the model theory and machine learning. Roughly speaking, our manuscript is similar to [8], but develops the connection between *stability* and *online learning*.

That the combinatorial quantity of VC-dimension plays an essential role in isolating the main dividing line in both PAC-learning and perhaps the second most prominent dividing line in model-theoretic classification theory (NIP/IP) is a remarkable fact. This connection has been the subject of numerous works in recent years [5–7, 11]. In the setting of *online learning* (described below), another combinatorial notion, the *Littlestone dimension*, isolates the dividing line between learnability and nonlearnability of a concept class. Given how well studied the connection between model theory and the combinatorics associated with machine learning is, it is surprising that it hasn't been noticed until now that the same combinatorial quantity isolates what is perhaps the most prominent dividing line in classification theory (stable/unstable).

Received February 6, 2018.
2010 *Mathematics Subject Classification.* 03C98, 97P20.
Key words and phrases. model theory, machine learning.

1079-8986/19/2503-0002
DOI:10.1017/bsl.2018.71

Now, we roughly describe the PAC setting, in part to contrast the setting with that of online learning. Given an infinite set X with a probability measure μ on X and a collection of measurable subsets of X, denoted by $\mathcal{C}$, one attempts to "learn" a fixed but unknown $A \in \mathcal{C}$ by sampling from X. For some large n, n elements of X are randomly sampled, and the goal is to estimate the probability $\mu(A)$ by the proportion of elements of the sample which lie in A. For some $\varepsilon > 0$ fixed ahead of time, we say that the sample estimates the set A *ε-well* if the proportion of elements of the sample which lie in A is within ε of $\mu(A)$. The class $\mathcal{C}$ is *learnable* if for any δ there is a large enough n such that the measure of the samples of size n (computed using the product measure μ^n) which estimate the sample ε-well is greater than $1 - \delta$. Roughly, *for large enough sample size, we can get arbitrarily high likelihood that a sample estimates the true probability arbitrarily well.* That is, for a large enough sample size, predictions are *probably approximately correct,* hence the name PAC. It turns out that there is a purely combinatorial characterization of $\mathcal{C}$ being PAC-learnable (which remarkably does not depend on the distribution μ); the collection $\mathcal{C}$ is PAC-learnable if and only if $\mathcal{C}$ has finite *VC-dimension*.

The connection to model theory is as follows: when X is taken to be $\mathcal{M}$, a model of a first order theory T and $\phi(x, y)$ is a formula in the language of T, we let $\mathcal{C} = \{\phi(\mathcal{M}, a) \mid a \in \mathcal{M}\}$. Then, the VC-dimension of $\mathcal{C}$ is finite if and only if $\phi(x, y)$ is NIP.

In the most straightforward (and restrictive) setup of online learning, we are given an infinite set X (with no distribution) along with a collection $\mathcal{C}$ of subsets of X. The collection $\mathcal{C}$ is known to the learner. Fix some $A \in \mathcal{C}$ which is not known to the learner. Fixing some large n, there will be n rounds. In round i, an element x_i is selected, and the learner must predict the value of $1_A(x_t)$, that is, whether or not x_i is in the unknown set A. We call the value of the learner's prediction $\hat{y}_i$. The goal of online learning is to minimize the number of mistakes made during these predictions

$$\sum_{i=1}^{n} |\hat{y}_i - 1_A(x_i)|.$$

In this setting, there is no assumption about how the elements $\bar{x} = (x_1, \ldots, x_n)$ are chosen, and the choice of x_{i+1} is allowed to depend on the predictions made by the learner in the previous rounds. One seeks to minimize the number of mistakes over all possible sequences of samples. This setting of computational learning often arises when the data become available in sequential order or the data are chosen by a process which is assumed to be adversarial to the learner (a process or opponent seeking to make the number of mistakes large). Variations on how the samples are chosen are possible as well; for instance, a certain limited amount of randomness is often injected into how the elements x_i are chosen without moving the sampling back into the PAC context.

It turns out that the number of mistakes that the best deterministic algorithm makes (over all possible samples) can be bounded in terms of a

https://doi.org/10.1017/bsl.2018.71 Published online by Cambridge University Press

combinatorial quantity associated with the collection $\mathcal{C}$, the Littlestone dimension. When X is taken to be $\mathcal{M}$, a model of a first-order theory T, $\phi(x, y)$ is a formula in the language of T, and $\mathcal{C} = \{\phi(\mathcal{M}, a) \mid a \in \mathcal{M}\}$, the Littlestone dimension (also called thicket dimension) is precisely the Shelah 2-rank of $\phi(x, y)$, which is finite if and only if $\phi(x, y)$ is stable. A number of variants of this basic setup have much less restrictive assumptions (sometimes with a certain amount of randomness similar to the PAC setting) while also having the property that learnability is characterized by stability. In Section 4, we will give an exposition of the various settings in which stability characterizes learnability.

It seems surprising to the authors that the connection pointed out in the previous paragraph has not been previously noticed, but the following quote of [15] offers something of an explanation:

> A reflection on the past two decades of research in learning theory reveals (in our somewhat biased view) an interesting difference between Statistical Learning Theory and Online Learning. In the former, the focus has been primarily on understanding complexity measures rather than algorithms... In contrast, Online Learning has been mainly centered around algorithms.

The dividing lines in model-theoretic classification theory are more naturally associated with combinatorial properties and the various complexity measures associated with PAC-learning than with algorithms, and in the less restrictive online setups, the role of Littlestone dimension is perhaps somewhat more hidden than the role of VC-dimension in the PAC setup.

The correspondence between online learnability and stability is similar to the correspondence between PAC-learnability and NIP, but it should be mentioned that the fields (online learning and stability theory) are in rather different positions than in PAC-learning correspondence with NIP. At this point, stability theory has been extensively developed, while at the time of [8], the study of theories without the independence property was in its infancy, while PAC-learning was much more developed. Various notions from PAC-learning eventually played a big role in the development of structural results for NIP structures. In the case of the correspondence between stability and online learning, there seems to be more potential for the application of model theoretic ideas in online learning. For instance, in the final sentence of [2], the authors mention that one of the main open questions in the theory is to close the gap between the lower bounds and upper bounds for the expected number of mistakes a learner makes in various online contexts, and that this question seems to have as a main obstacle a lack of interesting infinite concept classes with finite Littlestone dimension. Model theory offers a remedy for this obstacle; a great many mathematically interesting theories have been proven to be stable over the last forty plus years of classification theory, often with highly nontrivial proofs. So, following our discussion of online learning, we give some prominent examples of stable theories, giving various new examples of classes of finite Littlestone dimension.

https://doi.org/10.1017/bsl.2018.71 Published online by Cambridge University Press

Now, we describe the organization of this manuscript. In Section 2, we describe the setting of computational learning in very general terms. In Section 3 we specialize to the PAC-setting. In Section 4, we specialize to the setting of online learning before describing several variants. In the final section, we survey some stable theories, and use the connection pointed out earlier in the paper to give many new examples of classes with finite Littlestone dimension.

§2. Machine learning generalities. In this section, we describe the generalities of machine learning, in quite a general setup, while mentioning the cases of particular interest to us. Let Y be a set, which we will call the set of *labels*. Let Y' be another set, which we will refer to as the *predictions.* Fix a function

$$L : Y \times Y' \to \mathbb{R}_{\geq 0}$$

which we call the *loss function*.

REMARK 2.1. The most common setup occurs when $Y = Y' = \{0, 1\}$ and $L(y, y') = |y - y'|$.

Another common example occurs when $Y = Y' = I \subseteq \mathbb{R}$, with I a bounded interval. In this case, a common loss function is given by $L(y, y') = (y - y')^2$. Settings in which $Y, Y' \subset \mathbb{R}$ are sometimes called margin-based. These settings are less natural to connect directly to model theory, though it might make sense to study margin-based machine learning in the context of continuous model theory [20].

Let X be another set, which we call the set of *examples* (also sometimes called *inputs* or *instances*). A *concept* is a map $A : X \to Y$. In the example given above with $Y = \{0, 1\}$, a concept is simply a subset of X. A concept class $\mathcal{C}$ is a collection of concepts.

Fix some concept A. The learner will make a series of predictions about a sample of inputs from X by selecting a prediction $\hat{y}_i$ for the label of each element x_i from the sample. The learner incurs a loss for each element x_i of the sample, by evaluating $L(A(x_i), \hat{y}_i)$. If the elements of the sample are indexed by the set I, then the total loss incurred is given by

$$\sum_{i \in I} L(A(x_i), \hat{y}_i).$$

The goal of the learner is always the same—minimize the total loss coming from making predictions about a series of elements of X. Besides the objects described above, the differences in various settings of learning theory are derived from the assumptions about what data the learner have available and how the elements of the sample are chosen.

§3. PAC-learning and NIP. In this section, we will quickly explain the connection between PAC-learning and NIP. Our presentation essentially follows [6]. Fix a concept class $\mathcal{C}$ on a set X with $Y = Y' = \{0, 1\}$. Let $\mathcal{C}_{fin} = \{A|_Z \mid Z \subset X, Z \text{ finite}, A \in \mathcal{C}\}$. Let μ be a probability measure on X such that each element of $\mathcal{C}$ is measurable. We will think of the learner

https://doi.org/10.1017/bsl.2018.71 Published online by Cambridge University Press

as having complete knowledge of the elements of $\mathcal{C}$, and the elements for a sample being drawn randomly with respect to the distribution given by μ.

Let $G : \mathcal{C}_{fin} \to 2^X$ be a function. Let $\bar{a} = (a_1, \ldots, a_n)$. Define

$$err_\mu(G, A, \bar{a}) := \mu(\{c \in X \mid f(c) \neq G(A|_{\bar{a}})(c)\}).$$

Here, one should think that G is a function being used to generate predictions, while the error is the probability that the next prediction is incorrect.

We say that $\mathcal{C}$ is *probably approximately correct learnable* (PAC-learnable) if there is a $G : \mathcal{C}_{fin} \to 2^X$ such that for all $\varepsilon > 0$ and all $\delta > 0$, there is $N_{\varepsilon,\delta} \in \mathbb{N}$ such that for all $A \in \mathcal{C}$, and all μ on X such that all elements of $\mathcal{C}$ measurable,

$$\mu^{N_{\varepsilon,\delta}}\left(\{\bar{a} \in X^{N_{\varepsilon,\delta}} \mid err_\mu(G, A, \bar{a}) > \varepsilon\}\right) < \delta,$$

where $\mu^{N_{\varepsilon,\delta}}$ is the product measure. That is, the probability that the error is high (bigger than ε) is small (less than δ). Supposing that the class $\mathcal{C}$ is PAC-learnable, there is a minimal $N_{\varepsilon,\delta}$ for which the inequality holds, which is called the *sample complexity*.

The following theorem establishes the connection between VC-dimension and PAC-learnability:

THEOREM 3.1. *Let $\mathcal{C}$ be a concept class on X. Then the following are equivalent*:

(1) *$\mathcal{C}$ has finite VC-dimension.*
(2) *$\mathcal{C}$ is PAC-learnable, and*

$$N_{\varepsilon,\delta} \leq max\left\{\frac{4}{\varepsilon}\log_2\left(\frac{2}{\delta}\right), \frac{8d}{\varepsilon}\log_2\left(\frac{13}{\varepsilon}\right)\right\}.$$

In fact, even more is true—if $\mathcal{C}$ is PAC-learnable with sample complexity $N_{\varepsilon,\delta}$, then one can show that the expected value of the function $\bar{a} \mapsto err_\mu(G, A, \bar{a})$ is bounded by $\delta + \varepsilon(1 - \delta)$.

In the years since Laskowski's paper [8], connections between the VC theory and NIP have developed extensively with important notions from VC-theory adapted to the model-theoretic setting and vice versa [5–7, 11].

§4. Online learning and stability. The initial setting of online learning which we describe is due to Littlestone [9]; the particular setting received relatively little attention, perhaps due to the very strong assumptions ([9] is in fact famous for several other contributions). Littlestone's work was generalized in various ways in the ensuing years, with the assumptions being significantly weakened. We will begin with the original setup of [9], and eventually describe two settings laid out in [2]. First, we set up some of the combinatorial notions pertinent in each of the settings we consider.

The next several definitions follow the notation and terminology of Bhaskar [3].

DEFINITION 4.1. A *binary element tree of height h*, denoted by $\mathcal{T}_h$, is a rooted complete binary tree of height h whose nonleaf vertices are labeled

https://doi.org/10.1017/bsl.2018.71 Published online by Cambridge University Press

by elements of the set X and whose leaves are labeled by elements of $\mathcal{C}$ (see Figure 1).

For the following definitions, fix a binary element tree of height h.

DEFINITION 4.2. A vertex v_1 is *below* a vertex v_2 if v_2 lies on the (unique) path from v_1 to the root of the tree. We say that v_1 is *left-below* v_2 if v_1 is below v_2 and the first edge along the path from v_2 to v_1 goes down and to the left. The notion of *right-below* is defined analogously. When a vertex labeled by b is left-below a vertex labeled by a, we write $a <_L b$. Similarly, when a vertex labeled by b is right-below a vertex labeled by a, we write $a <_R b$.

DEFINITION 4.3. A leaf, labeled by $A \in \mathcal{C}$ is said to be *well labeled* if for each vertex above Y, say labeled by a,

$$a \in A \text{ if and only if } a <_L A.$$

DEFINITION 4.4. The *thicket shatter function* $\rho_{\mathcal{F}} : \mathbb{Z}^{\geq 0} \to \mathbb{Z}^{\geq 0}$ is defined by letting $\rho_{\mathcal{F}}(n)$ be the maximum number of well-labeled leaves on a binary element tree of height n, $\mathcal{T}_n$, whose leaves are labeled with elements of $\mathcal{F}$. The *thicket dimension* $Ldim(\mathcal{F})$ is the maximum integer n such that $\rho_{\mathcal{F}}(n) = 2^n$, or else $Ldim(\mathcal{F}) = \infty$.

Thicket dimension has appeared in at least several other contexts under different names; in fact Bhaskar [3] was aware of the terminology and definitions of [18], which we reproduce next:

DEFINITION 4.5. Let $\mathcal{M}$ be a monster model of a complete $\mathcal{L}$-theory. Fix a consistent partial type $\pi(x)$ and a partitioned formula $\phi(x; y)$. Then the ordinal $R(\pi, \phi, 2)$, called the Shelah 2-rank, is defined as follows:

- $R(\pi, \phi, 2) \geq 0$.
- For any limit ordinal λ, $R(\pi, \phi, 2) \geq \lambda$ if $R(\pi, \phi, 2) \geq \alpha$ for all $\alpha < \lambda$.
- For any ordinal α, $R(\pi, \phi, 2) \geq \alpha + 1$ if there is some $\phi(x, a)$ such that $R(\pi \cup \{\phi(x, a)\}, \phi, 2) \geq \alpha$ and $R(\pi \cup \{\neg\phi(x, a)\}, \phi, 2) \geq \alpha$.

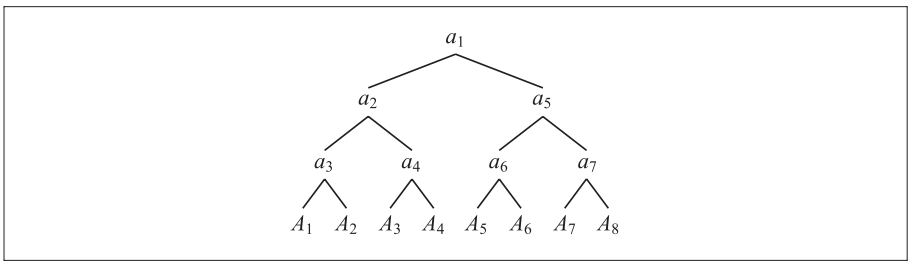

FIGURE 1. A binary element tree of height three. Here, $a_i \in X$ and $A_i \in \mathcal{C}$. The leaf labeled with A_4 is well-labeled if and only if $a_1 \in A_4$ and $a_2, a_4 \notin A_4$. For all other a_i, there is no requirement about membership in A_4.

https://doi.org/10.1017/bsl.2018.71 Published online by Cambridge University Press

In general, $R(\pi, \Delta, 2)$ can also be defined for a finite collection of formulas Δ, but this case can be shown to reduce to the case of a single formula. The formula $\phi(x, y)$ is *stable* if and only if $R(\{x = x\}, \phi, 2)$ is finite [18]; a theory is stable if every formula is stable. It is reasonably clear that the $R(\pi, \phi, 2)$ is the thicket dimension of the set system on $\mathcal{M}^{|y|}$ given by the collection of sets $\{\phi(b, \mathcal{M}) \mid b \in \pi(\mathcal{M})\}$; for more details, see [3].

The thicket dimension also appears for the first time in the context of learning theory in [9]; the quantity came to be called the *Littlestone dimension* [2].

4.1. The realizable case. Fix a set system $\mathcal{C}$ on a set X. Assume that $Y = Y' = \{0, 1\}$ and the loss function for a prediction $\hat{y}$ and concept (that is, a set) A on input x is given by $|\hat{y} - 1_A(x)|$. Over all possible algorithms, we seek to minimize our loss, that is, the number of mistakes we make over n rounds of predictions. In the *realizable* case, we assume that $A \in \mathcal{C}$, so that the true concept is among the set of concepts $\mathcal{C}$ accessible to the learner. There are *no* assumptions on the choices of the instances x_t. The goal is to minimize the worst case number of mistakes made by our predictions over all possible samples of the instances and choice of the concept. So, we seek to bound

$$M = \max_{A \in \mathcal{C}} \max_{\bar{x} = (x_1, \ldots, x_n)} \sum_{t=1}^{n} |\hat{y}_t - 1_A(x_t)|,$$

where $\hat{y}_t$ is chosen by some deterministic algorithm.

For applications and purposes of discussing the bounds, one often views the entity selecting the instances $\bar{x}$ as antagonistic to the learner—and in our current simplified setting, bounding the worst case number of mistakes bounds the actual number of mistakes made when the antagonistic sampling entity has perfect information about the prediction process.

THEOREM 4.6. [9] *The worst case number of mistakes of any deterministic algorithm in the online learning setting with concept class $\mathcal{C}$ is at least the Littlestone dimension of $\mathcal{C}$, and there is an algorithm that makes at most this many mistakes.*

REMARK 4.7. The algorithm which minimizes the number of worst-case mistakes in the above setting is referred to as the Standard Optimal Algorithm (SOA), and we describe it briefly here. Begin with $V_0 = \mathcal{C}$. At each stage, the learner inductively defines V_i. At stage t, the learner receives x_t, and sets, for $r = 0, 1$,

$$V_t^{(r)} := \{A \in V_{t-1} \mid 1_A(x_t) = r\}.$$

The learner predicts $\hat{y}_t = r$ which maximizes the Littlestone dimension of $V_t^{(r)}$ (ties are predicted in some fixed manner, say $\hat{y}_t = 0$ in the case of a tie). Then the learner gets the value of $1_A(x_t)$ and realizes whether a mistake has been made. At this point, set $V_t = V_t^{1_A(x_t)}$.

The essential point here is that if a mistake is made, it must be the case that the Littlestone dimension of V_t is strictly less than the Littlestone dimension of V_{t-1} (proving this is an easy exercise). Of course, this bounds the total

https://doi.org/10.1017/bsl.2018.71 Published online by Cambridge University Press

number of mistakes which the algorithm can ever make under any choice of $\bar{x}$ by the Littlestone dimension.

Where $\mathcal{C}$ is generated by a stable formula $\phi(z,x)$, say $\mathcal{C} = \{\phi(b,\mathcal{M}) \mid b \in \mathcal{M}\}$, the algorithm equivalently functions as follows. Begin with the partial type $\pi_0(z) = \{z = z\}$, and inductively define $\pi_i(z)$. When the learner receives x_t, the learner predicts $\hat{y}_t = r$, where r maximizes $R(\pi_{t-1} \cup \{\phi(z,x_t)^r\}, \phi, 2)$, where $\phi(z,x)^1 = \phi(z,x)$ and $\phi(z,x)^0 = \neg\phi(z,x)$. Upon receiving $1_A(x_t)$, set $\pi_t(z) = \pi_{t-1}(z) \cup \{\phi(z,x_t)^{1_A(x_t)}\}$. Again, a mistake on x_t will mean $R(\pi_t,\phi,2) < R(\pi_{t-1},\phi,2)$.

4.2. Learning from experts. The case in which we assume that the learner has access to true concept $A \in \mathcal{C}$ is often referred to as the *realizable* case of online learning. For various applications, this assumption is too strong (as are other assumptions from the previous subsection which we will deal with in later sections). In this section, we will explain a context of online learning which removes the realizability assumption.

The goal again is to minimize mistakes, but here, the minimization will be relative to a particular class of $\{0,1\}$-valued functions, which we will call $\mathcal{H}$. That is, we wish to minimize, for any sampling of instances, $\bar{x} = (x_1,\ldots,x_T)$, the difference between the number of mistakes made by the learner and the minimal number of mistakes made by any of the functions in $\mathcal{H}$. So, in this case, the loss function is taken to be

$$\sum |\hat{y}_t - y_t| - \min_{h\in\mathcal{H}} \sum |h(x_t) - y_t|.$$

Here, one often thinks intuitively that the functions in $\mathcal{H}$ are experts making predictions, and the learner's job is to choose which expert's prediction to believe.

Littlestone and Warmuth [10] consider this problem in the case that $\mathcal{H}$ is finite via a probabilistic weighted majority algorithm. We will now describe their algorithm. At the outset, each of the N many experts $\{f_i\}_{i=1}^N = \mathcal{H}$ is assigned weight 1, and the weight of expert i at stage t will be denoted by w_i^t. We fix the learning rate $\eta > 0$, which dictates how much we discount the weight of an expert for providing incorrect advice. At each stage, the learner receives the expert advice, $(f_1(x_t),\ldots,f_N(x_t))$, a tuple in $\{0,1\}^N$. The learner predicts 1 with probability

$$p_t = \frac{1}{\sum_{i=1}^N w_i^{t-1}} \sum_{i=1}^N w_i^{t-1} f_i(x_t).$$

Then, once the actual value y_t is revealed, the weights are updated via: $w_i^t = w_i^{t-1} e^{-\eta\cdot|f_i(x_t)-y_t|}$. That is, those experts who were wrong see their weight drop by a factor of $e^{-\eta}$.

The expected value of the loss function of their algorithm with a sample of size T is

$$\sum_{t=1}^T E(|\hat{y}_t - y_t|) - \min_{h\in\mathcal{H}} \sum_{t=1}^T |h(x_t) - y_t| \leq \sqrt{\frac{1}{2}\ln(N)T}.$$

https://doi.org/10.1017/bsl.2018.71 Published online by Cambridge University Press

Here, the assumption that $\mathcal{H}$ is finite is often too strong for applications, however, [2] generalize the setup to the case in which $\mathcal{H}$ is infinite, but of finite Littlestone dimension, proving:

THEOREM 4.8. *There is an algorithm such that for all $h \in \mathcal{H}$ and any sequence of instances $\bar{x} = (x_1, \ldots, x_T)$,*

$$\sum_{t=1}^{T} E(|\hat{y}_t - y_t|) - \min_{h \in \mathcal{H}} \sum_{t=1}^{T} |h(x_t) - y_t| \leq \sqrt{\frac{1}{2} Ldim(\mathcal{H}) \cdot T \ln(T)}.$$

In [2] it is also shown that no algorithm (even allowing randomization) can achieve an expected bound better than $\sqrt{\frac{1}{8} Ldim(\mathcal{H}) T}$. Closing the gap between the lower and upper bounds for the loss function (sometimes called *regret* in this context) is one of the main open problems mentioned in [2], where the authors remark that there are few known interesting examples of infinite classes with finite Littlestone dimension. Certainly, the model theory provides a large array of mathematically interesting examples of such classes which may be useful in providing examples which improve various bounds discussed above.

4.3. Bounded stochastic noise. Suppose that we work in the general setup from the previous section (again, not assuming realizability), but with a difference in the way we generate labels and measure mistakes. Suppose that there is a function $h \in \mathcal{H}$ such that the labels $y_1, \ldots, y_T$ are independent $\{0, 1\}$-valued random variables with the property that for all t, $Pr(h(x_t) \neq y_t) \leq \gamma$ with $\gamma \in (0, \frac{1}{2})$. This value γ will be called the noise rate.

In this setting, one seeks to minimize the difference between the predictions and the output of the noisy function on the samples:

$$E\left(\sum_{t=1}^{T} |\hat{y}_t - y_t|\right).$$

Note here that there are two sources of randomness—the choices of the algorithm may be randomized and the labels y_t are random variables. The expectation is taken with respect to both of these.

THEOREM 4.9. *For any concept class $\mathcal{H}$, and any $\gamma \in [0, \frac{1}{2})$, there is an algorithm (possibly randomized) so that for any $h \in \mathcal{H}$, and a sequence of examples $(x_1, y_1), \ldots, (x_T, y_T)$ with each y_t a random variable as described above,*

$$E\left(\sum_{t=1}^{T} |\hat{y}_t - h(x_t)|\right) \leq \frac{Ldim(\mathcal{H}) \cdot \ln(T)}{1 - 2\sqrt{\gamma(1-\gamma)}}.$$

That is, the expected number of mistakes grows only logarithmically in the sample size. In [2], the authors give an example of a class $\mathcal{H}$ which shows that the left hand side of the inequality in the theorem is bounded below by $\Omega(Ldim(\mathcal{H}) \cdot \ln(T))$.

https://doi.org/10.1017/bsl.2018.71 Published online by Cambridge University Press

§5. Stability theory. In this section, we use stability theory to point out various mathematically interesting examples of classes which have finite Littlestone dimension. We will assume some basic familiarity with first order logic, but we provide some reminders for the nonmodel theorist for whom this section is written.

Fix some complete theory T in a language $\mathcal{L}$ and let $\mathcal{M}$ be a monster model of T. The nonmodel theorist can simply loosely assume that $\mathcal{M}$ is a very large structure in which over a small subset A (say of cardinality at most κ) for any tuple c in any model of T containing A, there is some $b \in \mathcal{M}$ such that $\text{tp}(c/A) = \text{tp}(b/A)$. Here, $\text{tp}(c/A)$ denotes the collection of all first order formulas in the language $\mathcal{L}$ with parameters from A which are satisfied by c.

For $n \in \mathbb{N}$, the space of types of n-tuples of $\mathcal{M}$ over some subset $A \subset \mathcal{M}$ is denoted by $S_n(A)$. It comes naturally equipped with a topology in which the basic open sets correspond to first order formulas with parameters in A. Rather than considering all formulas, sometimes it is natural to restrict to the ϕ-type of a tuple, denoted $\text{tp}_\phi(c/A)$, the collection of instances of ϕ with parameters in A which hold of c. When $\phi(x; y)$ is a formula, the space of ϕ-types over A (treating the variables y as parameters) is denoted by $S_\phi(A)$.

The theory T is called κ-stable if for every set $A \subseteq \mathcal{M}$ with $|A| \leq \kappa$, we have $|S_n(A)| \leq \kappa$ for all $n \in \mathbb{N}$. The theory is *stable* if it is κ-stable for some $\kappa \geq |T|$. Part of the utility of the notion is that it can be characterized in several disparate ways (this is far from being an exhaustive list):

FACT 5.1. [18] *The following conditions are equivalent*:

(1) *T is κ-stable for some κ.*
(2) *For any countable set $A \subset \mathcal{M}$, $S_\phi(A)$ is countable.*
(3) *Every formula $\phi(x; y)$ has finite Shelah 2-rank—that is, $R(\{x = x\}, \phi, 2)$ is a finite ordinal (recall that Shelah 2-rank is equal to Littlestone dimension).*
(4) *No formula $\phi(x; y)$ has the* order property. *A formula $\phi(x; y)$ has the order property if there are tuples $(a_1, b_1), (a_2, b_2), \ldots$ from $\mathcal{M}$ so that $\mathcal{M} \models \phi(a_i; b_j)$ if and only if $i \leq j$.*

When κ in the first condition of the above definition is be taken to be $\aleph_0$, the theory is (somewhat enigmatically) called ω-stable. Not every stable theory is ω-stable, even when making strong assumptions about various aspects of the language or structure. For instance, the theory of the integers where the language consists of the additive group operation as a binary function is stable, but not ω-stable.

Stability is one of the dividing lines (probably the most prominent one) which in certain contexts, model-theorists view as the border between "tame" and "wild" structures; stability allows for the development of various structural results, which are (often provably) impossible in the case of unstable theories. Stability has various nonobvious interactions with algebraic structure, and understanding these interactions has been the subject of a huge amount of model theoretic work over the past fifty years (for instance, there is a deep structure theory of stable groups [14]).

https://doi.org/10.1017/bsl.2018.71 Published online by Cambridge University Press

Consider the concept class $\mathcal{C}_\phi$ on $\mathcal{M}^{|y|}$ given by the collection of sets $\{\phi(b, \mathcal{M}) \mid b \in \mathcal{M}\}$. The theory T is stable precisely if each concept class of this form has finite Littlestone dimension (see Section 4 for an explanation).

We will elaborate on condition (4). Given a class $\mathcal{C}_\phi$, there is a natural bipartite graph G_ϕ associated with any concept class. The sets of vertices consist of 1) the elements of the underlying set and 2) concepts, with an edge between an element and a concept if and only if the element is in the concept. Finite Littlestone dimension of the concept class $\mathcal{C}_\phi$ is equivalent to there being an upper bound on the size of any half-graph which appears as an induced subgraph of G_ϕ.

5.1. Examples of notable stable theories. We now make a list (very far from comprehensive) of some notable stable theories and offer some explanation of the set systems (families of definable sets) which arise in the various settings. From our list, many mathematically interesting classes $\mathcal{C}_\phi$ with finite Littlestone dimension can be obtained.

(1) ACF, the theory of algebraically closed fields. By quantifier elimination for algebraically closed fields, the concept classes which appear as $\mathcal{C}_\phi$ in the theory of algebraically closed fields are precisely the uniform families of affine constructible sets. That is, when $f : V \to W$ is a rational map (everything defined over some fixed algebraically closed field), the corresponding family of constructible sets is the collection of fibers of the function f. More concretely, one can think of such a family as being given by solutions sets of families of polynomial equations and inequations:

$$f_1(x, a) = f_2(x, a), \ldots, f_n(x, a) = 0, f(x, a) \neq 0$$

where x is a tuple of indeterminates and a is a tuple which varies over some constructible subset of $\mathbb{A}^{|a|}$.

(2) DCF_0, the theory of differentially closed fields of characteristic zero, was first investigated by Robinson [16] and Blum [4] gave an elegant axiomatization from which it was straightforward to notice that the theory is stable. See [12] for a more comprehensive discussion of DCF_0, as we will be brief here. Differentially closed fields are universal domains for algebraic differential equations; that is, if a system of equations has a solution in some field of functions, it already has a solution in the differential closure of the field generated by the coefficients of the equations. By quantifier elimination for differentially closed fields, the concept classes which appear as $\mathcal{C}_\phi$ in the theory of differentially closed fields are precisely the uniform families of constructible sets in the Kolchin topology (boolean combinations of the zero sets of algebraic differential equations). That is, when $f : V \to W$ is a differential rational map between affine constructible sets V, W in the Kolchin topology (everything defined over some fixed differentially closed field), the corresponding family of constructible sets is the collection of fibers of the function f. Such a family is alternatively given by a collection of differential equations

https://doi.org/10.1017/bsl.2018.71 Published online by Cambridge University Press

and inequations

$$f_1(x,a) = f_2(x,a), \ldots, f_n(x,a) = 0, f(x,a) \neq 0$$

where x is a tuple of indeterminates from $\mathcal{M} \models DCF_0$ and $a \in \mathcal{M}$ is a tuple which varies over some Kolchin-constructible subset of $\mathbb{A}^{|a|}$.

(3) The theory of separably closed fields with characteristic $p \neq 0$ and fixed degree of imperfection $e \in \mathbb{N}$ (which we will describe here) is complete and was shown to be stable by Wood [19]. When a field F of characteristic p is closed under separable extensions, we say F is separably closed. A set $B \subseteq F$ is a *p-basis* of F if the collection of products of powers of elements of B of degree at most $p-1$ forms a basis for F as an F^p-vector space. The cardinality of such a set B is called the degree of imperfection of F (which we assume to be finite). Now, let $\{a_1, \ldots, a_e\}$ be a p-basis of F, and let $\{m_1, \ldots, m_{p^e}\}$ be the collection of monomials in $\{a_1, \ldots, a_e\}$ of degree at most $p-1$ in each element. Every element of F can be written uniquely in the form

$$x = \sum_{i=1}^{p^e} x_{(i)}^p m_i$$

where $x_i \in F$. For each element x_i in the above sum, we can repeat the process, writing

$$x_{(i)} = \sum_{j=1}^{p^e} x_{(i,j)}^p m_i.$$

Naturally, one can continue to iterate this process, defining x_σ for any σ a finite tuple of elements from $\{1, \ldots, p^e\}$. Let λ_σ be the unary function $x \mapsto x_\sigma$.

Let $\mathcal{L}_{p,e}$ be the language $\{+, -, \cdot, ^{-1}, 0, 1\} \cup \{a_1, \ldots, a_e\} \cup \{\lambda_\sigma : \sigma \in (p^e)^{<\omega}\}$. The theory of separably closed fields of characteristic p with degree of imperfection e eliminates quantifiers in the language $\mathcal{L}_{p,e}$. So, in one variable, definable sets correspond to boolean combinations of the zero sets of ideals in $F[x, \lambda_\sigma(x)]_{\sigma \in (p^e) \leq n}$, for some n.

(4) Let X be a compact complex manifold. Consider the structure $\mathcal{A}(X)$ where the basic relations are the complex analytic subsets of X^n for any $n \in \mathbb{N}$; we call a subset $A \subseteq X^n$ complex analytic if it is, for any point $p \in X^n$ there is a neighborhood U of p such that $A \cap U$ is given by the zero set of some fixed finite number of holomorphic functions on U. The model theory of compact complex manifolds began with Zilber's observation [21] that if one adds as a relation all complex analytic subsets of X^n for all n, then, the induced structure is stable. For an overview of the model theory of compact complex manifolds, see [13].

(5) Let R be a ring and $\mathcal{L}_R$ be the language of right R-modules, consisting of a symbol for addition and a unary function f_r for each $r \in R$, which is interpreted as scalar multiplication by r. Let T be any complete theory of right R-modules in the language $\mathcal{L}_R$. By a result of Baur [1],

https://doi.org/10.1017/bsl.2018.71 Published online by Cambridge University Press

every formula $\phi(x)$ is equivalent to a boolean combination of positive primitive formulas, that is, formulas of the form $\exists y \psi(x, y)$, where ψ is a conjunction of atomic formulas. In particular, every definable subset of an R-module M is a boolean combination of cosets of positive primitive definable subgroups of M. An abelian group can be viewed as a $\mathbb{Z}$-module, and from this characterization of definable sets, it is not hard to show that every abelian group has a stable theory in the language of groups.

(6) The theory of the nonabelian free group T_{fg} in the language of groups was shown to be stable by Sela [17] (Sela shows the same for any torsion-free hyperbolic group). Every formula in the language of groups is, modulo the theory of the free group, equivalent to a $\forall\exists$-formula. The strategy of the proof is complicated and is developed by Sela over a series of seven previous papers; see [17] for complete references.

REMARK 5.2. Recall that Shelah 2-rank is a local property—that is, it is a property of a formula, rather than an entire theory. In particular, mathematically interesting stable formulas can be found in unstable theories, and these generate uniformly definable families with finite Littlestone dimension just as well. For example, in RCF, the theory of real closed fields, one can examine the family of solution sets of $f(x, y) = 0$, or $f(x, y) \neq 0$, for a polynomial f. These families have finite Littlestone dimension, even though RCF is unstable. Of course, one must take care not to modify a stable formula so as to make it unstable. Whereas $y = x + z^2$ is a stable formula, $\exists z(y = x + z^2)$ is not.

§6. **Acknowledgments.** The authors would like to thank Siddharth Bhaskar, Alex Kruckman, Dimitrios Diochnos, Dave Marker, Lev Reyzin, Dhruv Mubayi, Maryanthe Malliaris, and Gyorgy Turan for useful suggestions and conversations during the preparation of this article. James Freitag was supported by NSF grant no. 1700095.

REFERENCES

[1] W. BAUR, *Elimination of quantifiers for modules.* ***Israel Journal of Mathematics***, vol. 25 (1976), no. 1, pp. 64–70.

[2] S. BEN-DAVID, D. PÁL, and S. SHALEV-SHWARTZ, *Agnostic online learning,* ***Proceedings of the 22nd Annual Conference on Learning Theory COLT, 2009.***

[3] S. BHASKAR, *Thicket density*, arXiv preprint, 2017, arXiv:1702.03956.

[4] L. BLUM, ***Generalized algebraic structures: A model theoretical approach***, Ph.D. thesis, MIT, 1968.

[5] A. CHERNIKOV and P. SIMON, *Externally definable sets and dependent pairs.* ***Israel Journal of Mathematics***, vol. 194 (2013), no. 1, pp. 409–425.

[6] V. GUINGONA, *Nip theories and computational learning theory*, `https://tigerweb.towson.edu/vguingona/NIPTCLT.pdf`.

[7] H. R. JOHNSON and M. C. LASKOWSKI, *Compression schemes, stable definable families, and o-minimal structures.* ***Discrete & Computational Geometry***, vol. 43 (2010), no. 4, pp. 914–926.

[8] M. C. Laskowski, *Vapnik-Chervonenkis classes of definable sets.* ***Journal of the London Mathematical Society***, vol. 2 (1992), no. 2, pp. 377–384.

[9] N. Littlestone, *Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm.* ***Machine learning***, vol. 2 (1988), no. 4, pp. 285–318.

[10] N. Littlestone and M. K. Warmuth, *The weighted majority algorithm.* ***Information and computation***, vol. 108 (1994), no. 2, pp. 212–261.

[11] R. Livni and P. Simon, *Honest compressions and their application to compression schemes*, ***Conference on Learning Theory***, 2013, pp. 77–92.

[12] D. Marker, M. Messmer, and A. Pillay, ***Model Theory of Fields***, A. K. Peters/CRC Press, Natick, MA, 2005.

[13] R. Moosa, *Model theory and complex geometry.* ***Notices of the AMS***, vol. 57 (2010), no. 2, pp. 230–235.

[14] B. Poizat, ***Stable Groups***, Mathematical Surveys and monographs, vol. 87, American Mathematical Society, Providence, RI, 1987.

[15] A. Rakhlin, K. Sridharan, and A. Tewari, *Online learning: Beyond regret*, ***Proceedings of the 24th Annual Conference on Learning Theory***, 2011, pp. 559–594.

[16] A. Robinson, *On the concept of a differentially closed field.* ***Bulletin of the Research Council of Israel Section F***, vol. 8F (1959), pp. 113–128.

[17] Z. Sela, *Diophantine geometry over groups viii: Stability*, arXiv preprint, 2006, arXiv:math/0609096.

[18] S. Shelah, ***Classification Theory and the Number of Nonisomorphic Models***, Studies in Logic and the Foundations of Mathematics, vol. 92, North-Holland, New York, 1978.

[19] C. Wood, *Notes on the stability of separably closed fields.* ***The Journal of Symbolic Logic***, vol. 44 (1979), no. 3, pp. 412–416.

[20] I. B. Yaacov, A. Berenstein, C. W. Henson, and A. Usvyatsov, *Model theory for metric structures*. Retrieved from `https://faculty.math.illinois.edu/~henson/cfo/mtfms.pdf`, 2006.

[21] B. Zilber, *Model theory and algebraic geometry*, ***Proceedings of the 10th Easter Conference on Model Theory, Humboldt Universitat***, 1993, pp. 93–117.

DEPARTMENT OF MATHEMATICS
UIC, CHICAGO IL, USA
E-mail: hchase2@uic.edu
E-mail: freitagj@gmail.com

https://doi.org/10.1017/bsl.2018.71 Published online by Cambridge University Press