Bridging Model Theory and Machine Learning

Ayushi XXX

Geboren am XX. XXXXXXX XXXX in XXXXXXX, XXXXXXXXX

17 September 2024

Bachelorarbeit Mathematik

Betreuer: Prof. Dr. Philipp Hieronymi

Zweitgutachter: Dr. Tingxiang Zou

Mathematisches Institut

Mathematisch-Naturwissenschaftliche Fakultät der

Rheinischen Friedrich-Wilhelms-Universität Bonn

Contents

Conventions

Throughout the text we make use of shorthand notation as common in model theory. We denote

- the set of integers $\{0, 1, \ldots, n-1\}$ as n,
- a finite number $n \in \mathbb{N}$ as $n < \omega$,
- the set of all subsets of X of size n as $\binom{X}{n}$,
- the set of all functions $f : X \to Y$ as ${}^{X}Y$.

For example, the set of all functions $f : X \to \{0, 1\}$ is denoted as ${}^{X}2$.

- log denotes the logarithm to the base 2,

For any set Y there exists only one function $f : \varnothing \to Y$, also known as the *empty function*.

Introduction

This thesis explores the fascinating connections between two seemingly disparate fields of mathematics and computer science: model theory and machine learning. At first glance, these areas may appear to have little in common — model theory is a branch of mathematical logic concerned with the formal study of mathematical structures, while machine learning focuses on algorithms that can learn and make predictions from data. However, there are deep and surprising links between fundamental concepts in these domains.

Our investigation centers on two key relationships:

- The connection between Probably Approximately Correct (PAC) learnability in computational learning theory and the model-theoretic notion of NIP (Non-Independence Property) formulas.
- The correspondence between online learnability in computational learning theory and stable formulas in model theory.

These connections allow us to bridge abstract logical properties of theories with concrete learnability guarantees for concept classes. By translating between the languages of logic and learning theory, we gain new insights into the theoretical foundations of machine learning and expand our understanding of the expressiveness of logical theories. This thesis is structured in two main parts:

In the first part, we prove the fundamental theorem of PAC learning, which establishes that a concept class is PAC learnable if and only if it has finite Vapnik-Chervonenkis (VC) dimension. We then introduce NIP theories and demonstrate the equivalence between finite VC dimension and the NIP property.

The second part explores online learning and the Littlestone dimension as a measure of concept class complexity. We present the Standard Optimal Algorithm for online learning and prove its optimality. On the model theory side, we establish the equivalence between finite Littlestone dimension and Shelah's 2-rank. This allows us to show that stable formulas correspond to concept classes with finite Littlestone dimension.

Finally, we illustrate these concepts with several examples of stable theories. These examples demonstrate how model theory provides a rich source of concrete, learnable concept classes.

While a deep understanding of these areas is not required, readers are expected to have some background in key areas. In model theory, familiarity with first-order logic, languages, structures, models and theories is beneficial. Knowledge of key theorems such as the Löwenheim-Skolem theorem and the compactness theorem will be particularly helpful. In probability theory, knowledge of basic concepts including probability spaces, measures, random variables and expectation will be helpful.

1 VC dimension and NIP theories

1.1 PAC learning framework

The Probably Approximately Correct (PAC) learning framework, introduced by Leslie Valiant in 1984, provides a formal foundation for analyzing machine learning problems. It offers a mathematical model to quantify when and how learning is possible, bridging the gap between computational learning theory and practical machine learning algorithms.

In the PAC framework, the learner receives a sample of labeled examples $\{(x_i, f(x_i)) : i \in n\}$, where each x_i is drawn independently from X according to the unknown distribution μ, and $f \in \mathcal{C}$ is the ground truth they aim to learn. The learner's goal is to output a hypothesis $h \in \mathcal{H}$ that, with high probability, closely approximates the target concept f on future examples drawn from the same distribution.

At its core, PAC learning addresses a fundamental question: Under what conditions can a learning algorithm reliably generalize from a finite set of examples to accurately predict outcomes on unseen data?

Definition 1.1 The framework formalizes this idea by introducing several key components:

- An *input space* X is a set of all possible instances,
- A *concept* f is a binary-valued function $X \to \{0, 1\}$,
- A *concept class* $\mathcal{C} \subseteq {}^{X}2$ is a class of concepts,
- A *target concept* $f \in \mathcal{C}$ is the true function to be learned,
- A *hypothesis* h is a function, representing the learner's prediction,
- A *hypothesis class* $\mathcal{H} \subseteq {}^{X}2$ is a set of hypotheses h,
- A *sample* $S = \{(x_1, f(x_1)), \dots, (x_n, f(y_n))\} \subseteq (X \times 2)^n$, $n < \omega$, corresponding to the restriction of target function f to $\{x_1, \dots, x_n\}$.
- A *hypothesis function* or a *learning function* $H : \mathcal{C}_{\text{fin}} \to \mathcal{H}$. It represents an algorithm or deterministic procedure, which given a sample S corresponding to the restriction $f|_S \in \mathcal{C}_{\text{fin}}$ outputs a prediction $H(f)$.
    - $\mathcal{C}_{\text{fin}} = \{\mathcal{C}|_Y : Y \subseteq X, Y \text{ finite}\}$ represents all possible labeled samples from X.

Remark 1.2 (Restrictions) In this thesis, we focus on PAC learning with two important properties:

- Consistency: A hypothesis h is *consistent* with a labeled sample, if it correctly classifies all instances in that sample. Formally, given a sample S, h satisfies $\forall x_i \in S : h(x_i) = f(x_i)$. Similarly, a hypothesis function H is *consistent*, if for all $f \in \mathcal{C}$ and all $S \subseteq X$ finite it holds $\forall x \in S : H(f|_S)(x) = f(x)$.

- Realizability: This assumes that there exists a hypothesis $h \in \mathcal{H}$ which perfectly classifies all instances, that is $\forall x \in S : h(x) = f(x)$.

Given our focus on the realizable case, we will assume that the hypothesis class $\mathcal{H}$ is equal to the concept class $\mathcal{C}$. Consequently, we can refine our notation and sometimes write $H : \mathcal{C}_{\text{fin}} \to \mathcal{C}$ or $H : \mathcal{C}_{\text{fin}} \to {}^{X}2$ in special cases.

Definition 1.3 (PAC learnability) Let $\mathcal{C}$ be a concept class on a set X. We say that $\mathcal{C}$ is *probably approximately correct (PAC) learnable* if there exists a hypothesis function $H : \mathcal{C}_{\text{fin}} \to {}^{X}2$ such that:

- For all $\varepsilon, \delta \in (0,1)$, there exists a natural number $N_{\varepsilon,\delta} < \omega$ satisfying the following condition:
- For all $n \geqslant N_{\varepsilon,\delta}$, all $f \in \mathcal{C}$, and all probability measures μ on X (with the correct sets being μ-measurable),

  $$\mu^n(\{\overline{a} \in X^n : \operatorname{err}_\mu(H, f, \overline{a}) > \varepsilon\}) \leqslant \delta$$

  where:

  - $\varepsilon \in (0,1)$ is the *accuracy parameter*, specifying the acceptable error rate,
  - $\delta \in (0,1)$ is the *confidence parameter* indicating the desired probability of successfull learning,
  - $N_{\varepsilon,\delta} : (0,1)^2 \to \mathbb{N}, (\varepsilon, \delta) \mapsto N_{\varepsilon,\delta}$ is the *sample complexity function*, which determines the minimum number of examples required to guarantee PAC learning.
  - $\operatorname{err}_\mu(H, f, \overline{a})$ is the error of the hypothesis function H predicting f given sample $\overline{a} = (a_1, \ldots, a_n)$, defined by

    $$\mu^n(\{x \in X : H(f|_{\overline{a}})(x) \neq f(x)\})$$

This definition ensures that with high probability $(1-\delta)$, the learning function H will produce a hypothesis that is approximately correct (up to an error of ε) when the sample size is at least $N_{\varepsilon,\delta}$.

1.2 The fundamental theorem of PAC Learning

The fundamental theorem of PAC learning, first proven by Blumer, Ehrenfeucht, Haussler and Warmuth in 1989, establishes that a concept class is PAC learnable if and only if it has finite VC dimension. This section presents a complete proof of this theorem.

1.2.1 VC dimension and shatter function

Definition 1.4 (VC dimension) Let X be an input space and let $\mathcal{C} \subseteq {}^{X}2$ be a concept class on X. For any $Y \subseteq X$:

- the *restriction* of $\mathcal{C}$ to Y is $\mathcal{C}|_Y := \{f|_Y : f \in \mathcal{C}\}$.
- $\mathcal{C}$ *cuts out* Y from X if $\exists f \in \mathcal{C} : f|_X = \mathbb{1}_Y$.
- $\mathcal{C}$ *shatters* Y if $\mathcal{C}|_Y = {}^{Y}2$ or, equivalently, if $\mathcal{C}$ cuts out every subset of Y.

The Vapnik-Chervonenkis (VC) dimension of $\mathcal{C}$, denoted $\text{VCdim}(\mathcal{C})$, is defined as

$$\sup\{|Y| : Y \subseteq X \text{ is finite and } \mathcal{C} \text{ shatters } Y\}.$$

If $\mathcal{C}$ shatters arbitrarily large finite sets, then $\text{VCdim}(\mathcal{C}) = \infty$. If $\mathcal{C}$ shatters no set, then $\text{VCdim}(\mathcal{C}) = -\infty$. A concept class $\mathcal{C}$ is called a *VC class* if it has finite VC dimension.

Remark 1.5 (Learning problems and set systems) There exists a natural correspondence between learning problems $(X, \mathcal{C})$ and set systems $(X, \mathcal{F})$. Each $f \in \mathcal{C}$ defines a unique set $A_f \subseteq X$ where $A_f = \{a \in X : f(a) = 1\}$. The set system perspective often simplifies proofs and combinatorial arguments, while the function-based view aligns more closely with the learning theory framework. We will sometimes use the notation $(X, \mathcal{F})$ interchangeably with $(X, \mathcal{C})$ in our proofs, as this leads to more concise and intuitive arguments.

Informally, the VC dimension measures the complexity or expressiveness of a concept class in relation to its domain X by looking at how many points it can label arbitrarily. A higher VC dimension indicates that the concept class can represent more complex decision boundaries.

Example 1.6 (Easy) Consider $X = \mathbb{R}$ and the class of threshold functions $\mathcal{C} = \{f_a(x) : f_a(x) = 1 \text{ if } x \geqslant a, f_a(x) = 0 \text{ if } x < a\}$. This class has VC dimension 2, since it can shatter any set of two points and cannot shatter any set of three points. Note that the definition of VC dimension requires only one set of maximum size to be shattered.

It is important to note that the VC dimension is not an equivalence.

$$\begin{aligned} &\text{VC dimension } n \not\Longrightarrow \text{any set of cardinality } n \text{ can be shattered} \\ &\text{VC dimension } n \Longleftarrow \text{any set of cardinality } n \text{ can be shattered} \end{aligned}$$

Example 1.7 (Intermediate) Consider $X = \mathbb{R}^2$ and the class of half-planes $\mathcal{C} = \{f_{a,b}(x) : f_{a,b}(x) = 1 \text{ if } \langle a, x\rangle \geqslant b, f_{a,b}(x) = 0 \text{ if } \langle a, x\rangle \leqslant b\}$. Using elementary geometry (specifically, Radon's theorem), one can prove that $\mathcal{C}$ has VC dimension 3.

Remark 1.8 Prominent mathematician Terry Tao writes in his blogpost:

> In the field of analysis, it is common to make a distinction between "hard", "quantitative", or "finitary" analysis on one hand, and "soft", "qualitative", or "infinitary" analysis on the other. "Hard analysis" is mostly concerned with finite quantities (e.g. the cardinality of finite sets, the measure of bounded sets, the value of convergent integrals, the norm of finite-dimensional vectors, etc.) and their quantitative properties (in particular, upper and lower bounds). "Soft analysis", on the other hand, tends to deal with more infinitary objects (e.g. sequences, measurable sets and functions, σ-algebras, Banach spaces, etc.) and their qualitative properties (convergence, boundedness, integrability, completeness, compactness, etc.). To put it more symbolically, hard analysis is the mathematics of ε, N, $O()$, and $\leqslant$; soft analysis is the mathematics of 0, ∞, $\in$, and $\to$.

This distinction also characterizes the approaches to VC dimension in computer science and model theory.

In computer science, researchers typically employ a quantitative approach to VC dimension. They often seek to determine or estimate the exact VC dimension of concept classes, as this provides explicit bounds on learning complexity. Their proofs typically follow a two-step approach: First, they show that any set of cardinality $n+1$ cannot be shattered and then they explicitly shatter a set of cardinality n.

In contrast, model theorists generally adopt a qualitative approach to VC dimension. Their primary concern is whether the VC dimension is finite or infinite, rather than its exact value. A model theorist might prove that a theory T has finite VC dimension (equivalently, is NIP) without necessarily computing the exact VC dimension of any particular formula in T.

The following example illustrates the "soft" approach in model theory:

Example 1.9 (Hard) The theory of real ordered field[1] with an exponential function is o-minimal, and thus NIP. This implies that any formula φ is NIP and the concept class uniformly defined by φ is a VC class. Such examples provide only qualitative information about the learnability, without specifying the exact VC dimension.

The next two theorems establish elementary properties of VC dimension. The first theorem demonstrates the monotonicity properties of VC dimension with respect to both the input space and the concept class. The second

[1] A reader without a background in model theory can safely skip this example until later sections, where we discuss it in detail.

theorem shows how VC dimension changes when concept classes are combined using different Boolean operations.

Theorem 1.1 (Basic properties I) *Let $\mathcal{C} \subseteq 2^X$ be a concept class on input space X with $\mathrm{VCdim}(\mathcal{C}) = d < \infty$.*

(1) Any subset of a shattered set A is shattered.

(2) For any set Y with $Y \subseteq X$: $\mathrm{VCdim}(\mathcal{C}|_Y) \leqslant \mathrm{VCdim}(\mathcal{C})$.

(3) For any concept classes $\mathcal{C}'$ with $\mathcal{C}' \subseteq \mathcal{C}$: $\mathrm{VCdim}(\mathcal{C}') \leqslant \mathrm{VCdim}(\mathcal{C})$.

Proof Let $(X, \mathcal{F})$ be the set system corresponding to $(X, \mathcal{C})$ by *Theorem* 1.5.

(1) Let $B \subseteq A$. To show B is shattered, we need to prove that $\forall S \subseteq B : \exists F \in \mathcal{F} : F \cap B = S$. Since B is a subset of A, S is also a subset of A. Since A is shattered by $\mathcal{F}$, there exists $F \in \mathcal{F} : F \cap A = S$. Now, $F \cap B = (F \cap A) \cap B = S \cap B = S$. Therefore, we have found $F \in \mathcal{F}$ such that F cuts out S. Since S was arbitrary, this holds for all subsets of B. Thus, B is shattered by $\mathcal{F}$.

(2) Let $A \subseteq Y$ be any set shattered by $\mathcal{F}|_Y$. We need to show that $|A| \leqslant d$. For every subset $S \subseteq A$, there exists a set $(F \cap Y) \in \mathcal{F}|_Y : (F \cap Y) \cap A = S$. Since $A \subseteq Y$, this implies $(F \cap Y) \cap A = F \cap A = S$. Therefore A is also shattered by $\mathcal{F}$. Since $\mathrm{VCdim}(\mathcal{C}) = d$, we must have $|A| \leqslant d$.

(3) Let $\mathcal{F}' \subseteq \mathcal{F}$ and let A be any set shattered by $\mathcal{F}'$. Since $\mathcal{F}' \subseteq \mathcal{F}$ this automatically implies that A is shattered by $\mathcal{F}$. Since $\mathrm{VCdim}(\mathcal{C}) = d$, we must have $|A| \leqslant d$.

Theorem 1.2 (Basic properties II) *Let X be an input space, $f \in {}^X2$ a function, and $\mathcal{C}_1$ and $\mathcal{C}_2$ concept classes of VC dimension $n_1 < \omega$ and $n_2 < \omega$. Then the following holds regarding concept classes and their VC dimensions:*

(1) intersection $\mathcal{C}_\cap = \{f_1 \cdot f_2 : f_1 \in \mathcal{C}_1, f_2 \in \mathcal{C}_2\}$, $\mathrm{VCdim}(\mathcal{C}_\cap) \leqslant \max\{n_1, n_2\}$,

(2) union $\mathcal{C}_\cup = \{f_1 + f_2 - (f_1 \cdot f_2) : f_1 \in \mathcal{C}_1, f_2 \in \mathcal{C}_2\}$, $\mathrm{VCdim}(\mathcal{C}_\cup) \geqslant \max\{n_1, n_2\}$

(3) negation $\mathcal{C}_\neg = \{1 - f_1 : f_1 \in \mathcal{C}_1\}$, $\mathrm{VCdim}(\mathcal{C}_\neg) = n_1$

(4) symmetric difference $\mathcal{C} \triangle f = \{|g - f| : g \in \mathcal{C}_1\}$, $\mathrm{VCdim}(\mathcal{C} \triangle f) = n_1$

Proof Let $(X, \mathcal{F}_1), (X, \mathcal{F}_2)$ be the set systems corresponding to $(X, \mathcal{C}_1), (X, \mathcal{C}_2)$. Let $(X, \mathcal{F}_1 \triangle B), B = \{x \in X : f(x) = 1\}$ be a set system corresponding to $(X, \mathcal{C} \triangle f)$. The statements to prove correspond to Boolean operations on $\mathcal{F}_1$ and $\mathcal{F}_2$.

(1) By Theorem 1.1(3), $\mathrm{VCdim}(\mathcal{C}_1 \cap \mathcal{C}_2) \leqslant \max\{n_1, n_2\}$.

(2) By Theorem 1.1(3), $\mathrm{VCdim}(\mathcal{C}_1 \cup \mathcal{C}_2) \geqslant \max\{n_1, n_2\}$.

(3) Let $A \subseteq X$ of cardinality n_1 be shattered by $\mathcal{C}_1$. For every subset $S \subseteq A$, $\exists F \in \mathcal{F}_1 : F \cap A = S$. This implies $\forall (A \backslash S) \subseteq A, \exists F \in \mathcal{F}_1 : A \backslash (F \cap A) = A \backslash F = A \backslash S$. Since each S is in one-to-one correspondence with $(A \backslash S)$, $\mathrm{VCdim}(\mathcal{C}_\neg) = n_1$.

(4) Let $A \subseteq X$ of cardinality n_1 be shattered by $\mathcal{C}_1$. The proof is a character-building exercise in basic set theory. We show that any two sets in $\mathcal{F}_1$ cut out the same subset from A if and only if they cut out the same subset in $(\mathcal{F}_1 \triangle B)$.

- For any $F_1, F_2 \in \mathcal{F}_1$, we have:

$$\begin{aligned} F_1 \cap A = F_2 \cap A \iff & F_1 \cap B \cap A = F_2 \cap B \cap A \text{ and} \\ & F_1 \cap (X \setminus B) \cap A = F_2 \cap (X \setminus B) \cap A \end{aligned}$$

  This equivalence holds because A can be partitioned into $A \cap B$ and $A \cap (X \setminus B)$.

$$\begin{aligned} \iff & (X \setminus F_1) \cap B \cap A = (X \setminus F_2) \cap B \cap A \text{ and} \\ & F_1 \cap (X \setminus B) \cap A = F_2 \cap (X \setminus B) \cap A \end{aligned}$$

  This equivalence holds because membership in F_1 completely determines membership in $X \setminus F_1$.

$$\begin{aligned} \iff & ((X \setminus F_1) \cap B) \cup (F_1 \cap (X \setminus B)) \cap A = \\ & ((X \setminus F_2) \cap B) \cup (F_2 \cap (X \setminus B)) \cap A \end{aligned}$$

  This step combines the two conditions using set union. This equivalence holds because the sets $(X \setminus F_1) \cap B$ and $F_1 \cap (X \setminus B)$ are disjoint (and similarly for F_2).

$$\iff (F_1 \triangle B) \cap A = (F_2 \triangle B) \cap A.$$

  The final step uses the definition of symmetric difference.
- Therefore, $\mathcal{C}_1|_A$ cuts out 2^{n_1} subsets from A if and only if $(\mathcal{C}_1 \triangle f)|_A$ cuts out 2^{n_1} subsets from A. This implies that they both shatter A and have the same VC dimension.

While VC dimension considers all possible subsets of X, it is often useful to analyze how the concept class behaves on subsets of specific sizes. This allows us to examine how the "shattering power" of the concept class grows as we increase the size of the subsets we consider.

Definition 1.10 (Shatter function) Define the *shatter function* $\pi_{\mathcal{C}}(m) : \mathbb{N} \to \mathbb{N}$ as

$$\pi_{\mathcal{C}}(m) := \max\left\{ |\mathcal{C}_Y| : Y \in \binom{X}{m} \right\}.$$

Lemma 1.11 (Sauer-Shelah, 1972) *Define* $\Phi_n(m) := \sum_{i=0}^{n} \binom{m}{i}$. *If* $\mathcal{C}$ *has VC dimension* n *and* $m > n$, *then* $\pi_{\mathcal{C}}(m) \leqslant \Phi_n(m)$.

The function $\Phi_n(m)$ counts the number of subsets of an m-element set that have size less than or equal to n. In other words, it represents the number of subsets of $\{1, 2, \ldots, m\}$ with cardinality at most n.

There exist numerous proofs and versions of this lemma, which has important applications in model theory, graph theory, computational geometry and other disciplines. We give a proof using the "shifting" technique commonly used in extremal set theory.

PROOF (LEMMA 1.5, [CHE16]) We proceed by contradiction. Fix some $m > n$ and suppose $\pi_{\mathcal{C}}(m) > \Phi_n(m)$.

- By Theorem 1.10, there exists $Y \subseteq X$ with $|Y| = m$ such that $|\mathcal{C}|_Y| = \pi_{\mathcal{C}}(m) > \Phi_n(m)$. Since $\pi_{\mathcal{C}}(m)$ depends only on the size of $|\mathcal{C}|_Y| \leqslant 2^Y$, we can without loss of generality assume that:
  - $\mathcal{C}$ is finite with $|\mathcal{C}| = \pi_{\mathcal{C}}(m)$,
  - $X = \{x_1, \ldots, x_m\}$ with $|X| = m$.

We construct a sequence of concept subclasses $\mathcal{C}_0, \ldots, \mathcal{C}_m$ of $\mathcal{C}$ using a "shifting" operation.

- Let $\mathcal{C}_0 := \mathcal{C}$.
- Given $\mathcal{C}_k$, construct $\mathcal{C}_{k+1}$ as follows:
  - For each $f \in \mathcal{C}_k$, if $f(x_{k+1}) = 1$ and exists $g \notin \mathcal{C}_k$ such that $g(x_{i \neq k+1}) = f(x_{i \neq k+1})$ and $g(x_{k+1}) = 0$ then replace f by g in $\mathcal{C}_{k+1}$. Otherwise, keep f in $\mathcal{C}_k$.
- This construction has three key properties:
  (1) For each l, $|\mathcal{C}_k| = |\mathcal{C}_{k+1}|$,
  (2) If A is shattered by $\mathcal{C}_{k+1}$, then A is shattered by $\mathcal{C}_k$,
  (3) If $f \in \mathcal{C}_m$, then $\mathrm{supp}(f)$ is shattered by $\mathcal{C}_m$.
- Proofs of properties:
  (1) Holds by construction: each replacement preserves cardinality.
  (2) Let A be shattered by $\mathcal{C}_{k+1}$. For any $B \subseteq A$:
    * If g cuts out B and $g \in \mathcal{C}_k$, then $\mathcal{C}_k$ cuts out B.
    * If g cuts out B but $g \notin \mathcal{C}_k$, then g must have been added to $\mathcal{C}_{k+1}$ to replace some $f \in \mathcal{C}_k$ during the shifting operation.
      · If $x_{k+1} \notin A$, then $f \in \mathcal{C}_k$ that g replaced cuts out B, since $g|_A = \mathbb{1}_B = f|_A$. Thus $\mathcal{C}_k$ cuts out B.
      · If $x_{k+1} \in A$, then $x_{k+1} \notin B$ (since $g(x_{k+1}) = 0$). Since $\mathcal{C}_{k+1}$ shatters A, there exists $h \in \mathcal{C}_{k+1}$ cutting out $B \cup \{x_{k+1}\}$. By construction, h must have been in $\mathcal{C}_k$ and h should have been replaced with $h' \notin \mathcal{C}_k$ cutting out B. Since it was not replaced, h' was already in $\mathcal{C}_k$, thus $\mathcal{C}_k$ cuts out B.

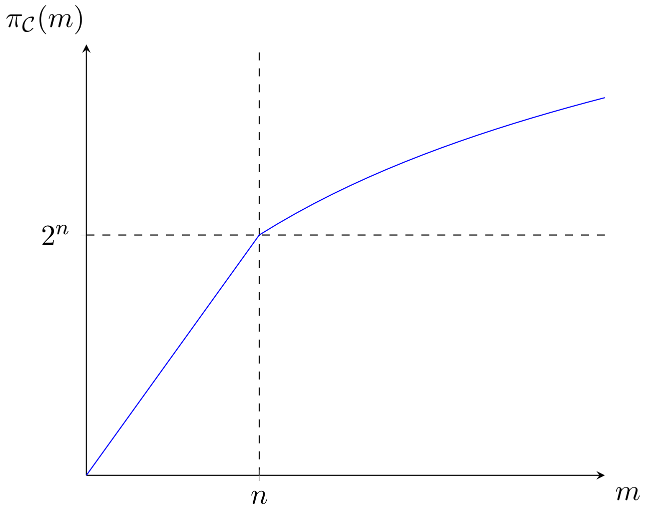


Figure 1: A schematic depiction of shatter function growth for a concept class $\mathcal{C}$ with $\mathrm{VCdim}(\mathcal{C}) = n$. The y-axis uses a logarithmic scale. For $m \leqslant n$, the function grows exponentially as 2^m. At $m = n$, the function transitions to slower growth, bounded by a polynomial function.

(3) Assume $\exists f \in \mathcal{C}_m$ with $\mathrm{supp}(f)$ not shattered by $\mathcal{C}_m$. Then $\exists x_{i+1} \in \mathrm{supp}(f)$ with no $g \in \mathcal{C}_m$ such that $g(x_{i+1}) = 0$. But f would have been replaced at step i by construction, removing x_{i+1} from $\mathrm{supp}\, f$. This contradicts the assumption $f \in \mathcal{C}_m$.

It follows from (2) that $\mathrm{VCdim}(\mathcal{C}_m) \leqslant n$. From (3), no $f \in \mathcal{C}_m$ has $|\,\mathrm{supp}(f)| > n$. Therefore: $\Phi_n(m) \geqslant |\mathcal{C}_m| \overset{(1)}{=} |\mathcal{C}| = |\pi_{\mathcal{C}}(m)|$.

Corollary 1.12 (Growth of the shatter function) *Let $\mathcal{C}$ be a concept class of VC dimension n. Then,*

$$\pi_{\mathcal{C}}(m) \begin{cases} = 2^m & m \leqslant n \\ \leqslant \Phi_n(m) & m > n \end{cases}$$

and, in particular, $\pi_{\mathcal{C}}(m) \in O(m^n)$.

Remark 1.13 In the special case where $\mathcal{C} = \binom{X}{n}$, this bound is tight for $m > n$.

Proof (Theorem 1.12)

- For $m \leqslant n$, the bound holds by Theorem 1.1(1).
- For $m > n$, the bound holds by Theorem 1.11.

- The growth estimate holds due to the following well-known binomial inequality:
$$\Phi_n(m) = \sum_{i=0}^{n} \binom{m}{i} \leqslant \sum_{i=0}^{n} \left(\frac{me}{i}\right)^i \leqslant m^n.$$
For each term in the sum above, we have
$$\binom{m}{i} = \frac{m!}{i!(m-i)!} = \frac{(i+1)(i+2)\dots(m)}{i!} \leqslant \frac{m^i}{i!} < \left(\frac{me}{i}\right)^i.$$

This result implies that for any given concept class $\mathcal{C}$, its shatter function can only grow either exponentially or polynomially in terms of m. As the cardinality of the set increases beyond n, the fraction of subsets of the set that can be shattered approaches 0.

1.2.2 ε-nets and the VC theorem

The next sections lie at the intersection of statistics, probability theory and computer science. Our primary objective is to examine and prove the Vapnik-Chervonenkis (VC) theorem, a fundamental result in statistical learning theory.

To provide additional context, we present a correspondence between some statistical and computational concepts, adapted from a widely-used textbook "All of Statistics: A Concise Course in Statistical Inference" by Wasserman:

Statistics	Computer Science	Meaning
estimation	learning	using data to estimate an unknown quantity
classification	supervised learning	predicting a discrete Y from X
large deviation bounds	PAC learning	uniform bounds on probability errors

The main goal of this section is the VC theorem, which is a variation on the theme of large deviation bounds. These bounds, which include the well-known Chernoff, Hoeffding, and Markov inequalities, provide estimates for the probability that an average of independent random variables deviates significantly from its expected value.

To build towards the VC theorem, we will introduce the concept of ϵ-nets. These are subsets of our sample space that, in a sense, approximate the entire space well.

Remark 1.14 In this section, we let $(X, \mathcal{A}, \mu)$ denote a probability space. Let $\mathcal{C}$ be a concept class on X so that each $f \in \mathcal{C}$ has μ-measurable support, i.e. for all $f \in \mathcal{C} : \text{supp}(f) \in \mathcal{A}$. For each $f \in \mathcal{C}$, let $\mu(f) = \mu(\text{supp}(f)) = \int_X f d\mu$.

In other words, $\mu(f)$ is the μ-probability that, given $a \in X, f(a) = 1$. For any $n < \omega$ we consider the product measure of μ on X^n, which we will denote by μ^n.

Lemma 1.15 (Lemma 2.3.4, [Gui13]) *Let $\mathcal{C}$ be a PAC learnable concept class on X and let H be a learning function for $f \in \mathcal{C}$ with sample complexity $N_{\varepsilon,\delta}$. Then, for all $\varepsilon \in (0,1), \delta \in (0,1)$, probability measures μ on X, and $n \geqslant N_{\varepsilon,\delta}$,*

$$\mathbb{E}(\overline{a} \mapsto \mathrm{err}_\mu(H, f, \overline{a})) \leqslant \delta + \varepsilon(1-\delta).$$

Proof Let $Y_0 = \{\overline{a} \in X^n : \mathrm{err}_\mu(H, f, \overline{a}) > \varepsilon\}$. Let $Y_1 = X \setminus Y_0$. By definition, since $\mathcal{C}$ is PAC learnable, the probability of sampling $\overline{a} \in Y_0$ with error $> \varepsilon$ is less than δ, so

$$\begin{aligned}\mathbb{E}(\overline{a} \mapsto \mathrm{err}_\mu(H, f, \overline{a})) &= \int_{X^n} \mathrm{err}_\mu(H, f, -) d\mu^n \\ &\leqslant \int_{Y_0} \mathrm{err}_\mu(H, f, -) d\mu^n + \int_{Y_1} \mathrm{err}_\mu(H, f, -) d\mu^n \\ &\leqslant 1 \cdot \mu^n(Y_0) + \varepsilon \cdot \mu^n(Y_1) \\ &\leqslant \delta + \varepsilon(1-\delta).\end{aligned}$$

Definition 1.16 (Definition 2.2.6, [Gui13]) For $\varepsilon \in (0,1)$, a subset $N \subseteq X$ is called an ε-net for $\mathcal{C}$ if for every $f \in \mathcal{C}$ with $\mu(f) = \mu(\mathrm{supp}(f)) \geqslant \varepsilon$, there exists $a \in N$ such that $f(a) = 1$.

Intuitively, an ε-net intersects every function $f \in \mathcal{C}$ whose support has μ-measure at least ε. This allows it to serve as an approximation of X with respect to $\mathcal{C}$ capturing all the “large” sets.

Fact 1.17 (Chebyshev's inequality) If $f : X \to \mathbb{R}$ is a random variable and $\varepsilon > 0$, then

$$\mu\left(\{a \in X : |f(a) - \mathbb{E}(f)| \geqslant \varepsilon\}\right) \leqslant \frac{\mathrm{Var}(f)}{\varepsilon^2}.$$

Lemma 1.18 (Lemma 2.2.5, [Gui13]) *Fix $p \in [0,1]$ and finite $n \geqslant \frac{8}{p}$. Let $(f_0, \dots, f_{n-1}) \in {}^X2$ such that $\mu(f_i) = p$ for all $i < n$. Then,*

$$\mu\left(\left\{(a_1, \dots, a_n) \in X^n : \sum_{i=1}^n f_i(a_i) \leqslant \frac{1}{2}np\right\}\right) \leqslant \frac{1}{2}.$$

Proof Define $f : X^n \to \mathbb{R}$ as $f(a_1, \dots, a_n) = \sum_{i=1}^n f_i(a_i)$. The expected value of f is equal to $\mathbb{E}(f) = \sum_{i=1}^n \mathbb{E}(f_i) = np$ and the variance of f is equal to $\operatorname{Var}(f) = \sum_{i=1}^{n} \operatorname{Var}(f) = np(1-p)$. Applying Theorem 1.17 with $\varepsilon = np/2$, we get:

$$\begin{aligned}
\mu\left(\left\{\overline{a} \in X^n : |f(\overline{a}) - np| \geqslant \frac{np}{2}\right\}\right) &\leqslant \frac{np(1-p)}{(np/2)^2} \\
&= \frac{4(1-p)}{np} \\
&\leqslant \frac{4}{np} \\
&\leqslant \frac{1}{2}.
\end{aligned}$$

Therefore, the μ-probability of $f(\overline{a}) \notin \left(\frac{np}{2}, \frac{3np}{2}\right)$ is at most $\frac{1}{2}$. Consequently, the probability of $f(\overline{a}) \notin (\frac{np}{2}, \infty)$ is also at most $\frac{1}{2}$, which proves the lemma.

The goal now is to show that with high probability, a randomly chosen set of points forms an ε-net for $\mathcal{C}$.

Theorem 1.3 (VC Theorem 2.2.7, [Gui13]) *Let $(X, \mathcal{B}, \mu)$ be a probability space, $\mathcal{C}$ be a concept class on X with each concept having μ-measurable support, $d, n < \omega$ and $\varepsilon \in (0,1)$. If the VC dimension of $\mathcal{C}$ is $\leqslant d$, then*

$$\mu(\{\overline{a} \in X^n : \{a_1, \ldots, a_n\} \text{ is not an } \varepsilon\text{-net for } \mathcal{C}\}) \leqslant 2(2n)^d 2^{-\frac{\varepsilon n}{2}}.$$

Proof The proof consists of two main parts: first, we define sets to characterize "bad" samples and derive an estimate, and second, we compute an upper bound on this estimate.

We start with the first part:

- Without loss of generality, we can assume $\mathcal{C} = \{f \in \mathcal{C} : \mu(f) \geqslant \varepsilon\}$, as functions with measure less than ε do not need to be witnessed by $\overline{a}$.
- Define $Y_0 = \{\overline{a} \in X^n : (\exists f \in \mathcal{C})(\forall i \leqslant n)(f(a_i) = 0)\}$. This set represents all samples $\overline{a}$ that fail to be an ε-net. Our goal is to show that $\mu(Y_0)$ is small.

We will set up a "coupling" argument to bound $\mu(Y_0)$ in terms of μ-measure of another set:

- For each $f \in \mathcal{C}$ define

  $$Y_f = \left\{\overline{a} \in X^{2n} : (\forall i \leqslant n)(f(a_i) = 0) \wedge \sum_{i=n+1}^{2n} f(a_i) \geqslant \frac{\varepsilon n}{2}\right\}.$$

  This set represents all $2n$-tuples where f assigns 0 to the first n elements and 1 to at least $\lceil \varepsilon n/2 \rceil$ of the remaining n elements.

- Let $Y_1 = \bigcup_{f \in \mathcal{C}} Y_f$. This set can be seen as a product $Y_0 \times Y_{\varepsilon n/2}$, where $Y_{\varepsilon n/2}$ denotes all a's where f has sufficiently big empirical measure.
- For any $\overline{a} \in X^n$ define the "section" of Y_1 corresponding to $\overline{a}$, that is $Y_1|_{\overline{a}} = \{\overline{b} \in X^n : (\overline{a}, \overline{b}) \in Y_1\}$.
  - If $\overline{a} \in Y_0$, then straightforward application of Theorem 1.18 with $p = \varepsilon$ yields:
    $$\mu\left(\left\{(a_1, \dots, a_n) \in X^n : \sum_{i=1}^{n} f_i(a_i) > \frac{1}{2}n\varepsilon\right\}\right) \geqslant \frac{1}{2}.$$
    The left-hand side of this inequality is precisely $\mu(Y_1|_{\overline{a}})$, therefore $\mu(Y_1|_{\overline{a}}) \geqslant 1/2$.
  - If $\overline{a} \notin Y_0$, then $\mu(Y_1|_{\overline{a}}) = 0$, as the first condition of Y_f fails.
  - This implies
    $$\mu(Y_1) = \int_{Y_0} \mu(Y_1|_{\overline{a}}) d\mu(\overline{a}) \geqslant \frac{1}{2}\int_{Y_0} d\mu(\overline{a}) = \frac{1}{2}\mu(Y_0).$$
    Therefore, $\mu(Y_0) \leqslant 2\mu(Y_1)$. So, to bound $\mu(Y_0)$, it suffices to compute an upper bound for $\mu(Y_1)$.

A measure-theory inclined reader might remember: If $X = A \times B$ is a measurable rectangle, then every A section is either B or $\varnothing$, according to whether $x \in A$ or not.

Now the second part:

- Consider the product space $X^{2n} \times \binom{2n}{n}$ with the product measure $\mu \otimes \nu$ where ν is the uniform probability measure on $\binom{2n}{n}$.
  - X^{2n} is the set of $2n$-sequences $(a_0, \dots, a_{2n-1}) \subseteq X^{2n}$,
  - $\binom{2n}{n}$ denotes the set of n-element subsets of $2n$.
  - For $I \subseteq \binom{2n}{n}$, let $\sigma|_I : 2n \to 2n$ be the permutation that maps I to $\{0, 1, \dots, n-1\}$ and its complement to $\{n, n+1, \dots, 2n-1\}$.
  - For $\overline{a} \in X^{2n}$ and $I \in \binom{2n}{n}$, define $\overline{a}_I = (a_{\sigma(0)}, \dots, a_{\sigma(2n-1)})$.
- Fix $\overline{a} \in X^{2n}$ and $f \in \mathcal{C}$. We compute the probability that $\overline{a}_I \in Y_f$ for some $I \in \binom{2n}{n}$.
  - If $\sum_{i=0}^{2n-1} f(a_i) < \frac{\varepsilon n}{2}$, then $\overline{a}_I \notin Y_f$ no matter how we choose I because it is impossible to satisfy the second condition of Y_f.
  - If $\sum_{i \leqslant 2n} f(a_i) \geqslant \frac{\varepsilon n}{2}$., then I must index elements where $f(a_i) = 0$ and avoid elements with $f(a_i) = 1$ because of the first condition of Y_f. By assumption, we have at least $k = \lceil \frac{\varepsilon n}{2} \rceil$ to avoid. Thus, the probability that $\overline{a}_I \in Y_f$ is at most
    $$\frac{\binom{2n-k}{n}}{\binom{2n}{n}} = \frac{\frac{(2n-k)!}{n!(n-k)!}}{\frac{(2n)!}{n!n!}} = \frac{(2n-k)!}{(2n)!} \cdot \frac{n!}{(n-k)!} = \frac{(n-k+1)\dots(n)}{(2n-k+1)\dots(2n)}.$$

Factoring out 2 from the denominator we can estimate each factor as less than $\frac{1}{2}$, thus

$$\frac{(n-k+1)\dots(n)}{(2n-k+1)\dots(2n)} \leqslant \left(\frac{1}{2}\right)^{k} \leqslant \left(\frac{1}{2}\right)^{\frac{\varepsilon n}{2}} = 2^{-\frac{\varepsilon n}{2}}.$$

Hence $2^{-\frac{\varepsilon n}{2}}$ is the upper bound for $\overline{a}_I \in Y_f$.

Now, we use the VC dimension to bound $\mu(Y_1)$:

- Since $\mathcal{C}$ has VC dimension $\leqslant d$, by Theorem 1.11, $\pi_{\mathcal{C}}(2n) \leqslant \Phi_d(2n)$. Therefore, for a fixed $\overline{a} \in X^{2n}$,

  $$|\{I \subseteq 2n : (\exists f \in \mathcal{C} : f \text{ cuts out } a_{i\in I})\}| \leqslant \Phi_d(2n) \leqslant (2n)^d.$$

  Therefore, for any $\overline{a} \in X^{2n}$, the probability $\mu(\overline{a}_I \in Y_1) \leqslant (2n)^d 2^{-\frac{\varepsilon n}{2}}$. This implies $\mu(Y_1) \leqslant (2n)^d 2^{-\frac{\varepsilon n}{2}}$.

Finally, we conclude $\mu(Y_0) < 2(2n)^d 2^{-\frac{\varepsilon n}{2}}$, which proves the theorem.

Remark 1.19 Exact measurability conditions are discussed in the appendix A1 and A2 of [Blu+89].

The VC Theorem provides a bound on the convergence of empirical measures to true measures uniformly over a class of sets (or functions), where the uniformity is controlled by the VC dimension.

1.2.3 Proof of the fundamental theorem of PAC learning

In this section, we prove the main theorem:

Theorem 1.4 (Theorem 2.1, [Blu+89]) *Let X be a set and $\mathcal{C}$ a concept class on X. Then, the following are equivalent:*

a) $\mathcal{C}$ is a VC class.

b) $\mathcal{C}$ is PAC learnable.

This theorem establishes that if the complexity measure of $\mathcal{C}$ (as measured by its VC dimension) is bounded, we can efficiently learn it by sampling data from X, independent of any other assumptions.

We prove this theorem by explicitly calculating a lower and an upper bound on the sample complexity $N_{\varepsilon,\delta}$ in Theorem 1.5 and Theorem 1.6.

Theorem 1.5 (Lower bound) *Let $d < \omega$ such that* $\mathrm{VCdim}(\mathcal{C}) \geqslant d$. *Then, any hypothesis function $H : \mathcal{C}_{fin} \to {}^{X}2$ that is a witness to PAC learnability of $\mathcal{C}$ has sample complexity*

$$N_{\varepsilon,\delta} \geqslant d(1 - 2(\varepsilon(1-\delta) + \delta)).$$

Theorem 1.6 (Upper bound) *Let $d < \omega$ such that $\mathrm{VCdim}(\mathcal{C}) \leqslant d$. Then, any consistent hypothesis function $H : \mathcal{C}_{\mathit{fin}} \to \mathcal{C}$ is a learning function for $\mathcal{C}$ with sample complexity*

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

Fact 1.20 Let $(X, \mathcal{A}, \mu)$ and $(Y, \mathcal{B}, \gamma)$ be two probability space and $f : (X \times Y) \to \mathbb{R}$ a random variable on the product space. Then, there exists $a \in X$ such that $f|_a : Y \to \mathbb{R}$, where $f|_a(b) = f(a, b)$, such that

$$\mathbb{E}(f|_a) \geqslant \mathbb{E}(f)$$

Proof (Theorem 1.5) Let $H : \mathcal{C}_{\text{fin}} \to {}^{X}2$ be any hypothesis function witnessing PAC learnability of $\mathcal{C}$. Our goal is to compute a lower bound on the expected value of hypothesis function error $\mathrm{err}_\mu(H, f, \overline{a})$ in terms of d and n, and then leverage Theorem 1.15 to extract a lower bound on $N_{\varepsilon,\delta}$.

- We begin by analyzing the hypothesis function error on $X \times \mathcal{C}$, viewing it as a function in two variables $(\overline{a}, f) \mapsto \mathrm{err}_\mu(H, f, \overline{a})$. To find a lower bound on the expected value of $\mathrm{err}_\mu(H, f, \overline{a})$, we consider the integral

$$\mathbb{E}(\mathrm{err}_\mu(H, f, \overline{a})) = \int_{X^n \times \mathcal{C}} \mathrm{err}_\mu(H, f, \overline{a}) d\mu((\overline{a}, f)).$$

  - Since $\mathcal{C}$ is PAC learnable, we can choose any probability measure μ on X and on $\mathcal{C}$. We opt for the uniform probability measure, which implies that for $a \in X : \mu(\{a\}) = 1/d$ and for $f \in \mathcal{C} : \mu(\{f\}) = 1/2^d$.
  - Since $\mathcal{C}$ has VC-dimension $\geqslant d$, there exists a subset of X with cardinality d that is shattered by $\mathcal{C}$. By restricting the measure to this shattered set, we may assume $|X| = d$ and $\mathcal{C} = {}^{X}2$.
  - Since H is consistent by Theorem 1.2, given sample $Y \subseteq X$, we can restrict the domain of the error function to the "unseen" subset $(X \setminus Y) \subseteq X^n$.
  - These reductions allow us to simplify the integral above to

$$\mathbb{E}(\mathrm{err}_\mu(H, f, \overline{a})) = \left(\frac{1}{d \cdot 2^d}\right)\left(\sum_{(\overline{a}, f) \in (X \setminus Y) \times \mathcal{C}} \mathrm{err}_\mu(H, f, \overline{a})\right).$$

- Now, we compute a lower bound of the expected value.

  - Fix $n \leqslant d$ and consider sequences $\overline{a} = (a_1, \dots, a_n) \in X^n$. Define $Y = \{a_1, \dots, a_k\}$, the set of distinct values of $\overline{a}$, noting that $|Y| = k \leqslant n$.
  - By Theorem 1.1(1), $\mathcal{C}$ shatters $Y \subseteq X$ so $\mathcal{C}|_Y = {}^Y2$. For any fixed concept $g \in {}^Y2$, it can be extended to a function in $\mathcal{C}$ in 2^{d-k} ways.

$1 \leqslant k \leqslant n < d \leqslant \mathrm{VCdim}(\mathcal{C})$

  - For a fixed $x \in (X \setminus Y)$, by symmetry, exactly half of $f \in \mathcal{C}$ which extend g will disagree with $H(g)$ on x. Summing over all possible elements $x \in X \setminus Y$, we have:

$$\sum_{x \in X \setminus Y} \left| \left\{ \left| H(g)(b) - f(b) \right| : b \in X, f \in {}^X2 \right\} \right| \geqslant \frac{1}{2} \left(2^{d-k} \right) (d-k)\,.$$

  - Since $|{}^Y2| = 2^k$, over all possible concepts $g \in {}^Y2$ we get:

$$\sum_{i=1}^{2^k} \left\{ \left| H(f|_Y)(b) - f(b) \right| : b \in X, f \in {}^X2 \right\} \geqslant \frac{1}{2} 2^d (d-k) \geqslant \frac{1}{2} 2^d (d-n)\,.$$

  - Thus, we arrive at the lower bound:

$$\mathbb{E}(\mathrm{err}_\mu(H, f, \overline{a})) \geqslant \frac{1}{d \cdot 2^d} \cdot \frac{1}{2} 2^d (d-n) = \frac{d-n}{2d}.$$

- We can now leverage this lower bound to obtain a lower bound on $N_{\varepsilon,\delta}$.

  - By Theorem 1.20, there exists $f \in \mathcal{C}$ with $\mathbb{E}(\overline{a} \mapsto \mathrm{err}_\mu(H, f, \overline{a})) \geqslant \frac{d-n}{2d}$.
  - By Theorem 1.15, $\mathbb{E}(a \mapsto \mathrm{err}_\mu(H, f, \overline{a})) \leqslant \varepsilon(1-\delta) + \delta$.
  - Since $N_{\varepsilon,\delta}$ must hold for any probability measure, it must hold specifically for μ. Therefore, we conclude that $N_{\varepsilon,\delta} \geqslant d(1 - 2(\varepsilon(1-\delta) + \delta))$.

Now we prove the other direction, i.e. Theorem 1.6 using ε-nets and the VC Theorem 1.3.

Proof (Theorem 1.6) The key idea of the proof is to use ε-nets to show that a consistent hypothesis function H can PAC learn a concept class $\mathcal{C}$ with finite VC dimension. This approach is somewhat consistent with the principle of Occam's Razor — a short explanation (i.e. a hypothesis function that is as simple as possible) tends to be more valid than a long explanation.

- Fix a target concept $f \in \mathcal{C}$ and a sample $\overline{a} \in X^n$.

  - For any prediction $h = H(f|_{\overline{a}}) \in \mathcal{C}$, the error function $|h - f|$ belongs to concept class $(\mathcal{C} \triangle f)$. Therefore we can describe the error in terms of the μ-measure of $|f - h|$, namely $\mathrm{err}_\mu(H, f, \overline{a}) = \mu\left(|f - h|\right)$.
  - By Theorem 1.2(4), $(\mathcal{C} \triangle f)$ has VC dimension d.

- By the VC Theorem 1.3, we can estimate the probability of $\overline{a}$ failing to be an ε-net, independent of μ:

$$\mu\left(\left\{\overline{a} \in X^n : \{a_1, \ldots, a_n\} \text{ not an } \varepsilon\text{-net for } (\mathcal{C} \triangle f)\right\}\right) \leqslant 2\left(2\frac{\mathrm{e}n}{d}\right)^d 2^{-\frac{\varepsilon n}{2}},$$

Here we use a sharper bound $(\frac{\mathrm{e}n}{d})^d$ on the shatter function's growth rate, as shown in Theorem 1.12.

- We need to choose n large enough so that:

$$\begin{aligned} & 2\left(2\frac{\mathrm{e}n}{d}\right)^d 2^{-\frac{\varepsilon n}{2}} \leqslant \delta \\ \iff \quad & \log(2) + d\log\left(2\frac{\mathrm{e}n}{d}\right) \leqslant \log(\delta) + \frac{\varepsilon n}{2} \\ \iff \quad & \frac{\varepsilon n}{2} \geqslant d\log\left(2\frac{\mathrm{e}n}{d}\right) + \log(\frac{2}{\delta}). \end{aligned}$$

We choose $n \geqslant \max\left\{\frac{4}{\varepsilon}\log\left(\frac{2}{\delta}\right), \frac{8d}{\varepsilon}\log\left(\frac{13}{\varepsilon}\right)\right\}$ and split the inequality in two parts:

$$\frac{\varepsilon n}{4} \geqslant \log(\frac{2}{\delta}) \quad \text{and} \quad \frac{\varepsilon n}{4} \geqslant d\log\left(2\frac{\mathrm{e}n}{d}\right).$$

  - The first inequality holds trivially by our choice of $n \geqslant \frac{4}{\varepsilon}\log(\frac{2}{\delta})$.
  - The second inequality holds for all $m > n$ if we can prove it for the lower bound $n \geqslant \frac{8d}{\varepsilon}\log\left(\frac{13}{\varepsilon}\right)$, implying $\frac{\varepsilon n}{4} \geqslant 2d\log\left(\frac{13}{\varepsilon}\right)$. so the inequality holds if we can prove

$$\begin{aligned} & 2\log\left(\frac{13}{\varepsilon}\right) \geqslant \log\left(16\left(\frac{\mathrm{e}}{\varepsilon}\right)\log\left(\frac{13}{\varepsilon}\right)\right) \\ \iff & \left(\frac{13}{\varepsilon}\right)^2 \geqslant \frac{16\mathrm{e}}{\varepsilon}\log\left(\frac{13}{\varepsilon}\right) \\ \iff & \frac{13^2}{16\mathrm{e}\varepsilon} \geqslant \log\left(\frac{13}{\varepsilon}\right). \end{aligned}$$

  This inequality holds for $\varepsilon = 1$ and all smaller values, completing the proof.

- Conclusion:
  - We have proven that for $n \geqslant \max\{\frac{4}{\varepsilon}\log\left(\frac{2}{\delta}\right), \frac{8d}{\varepsilon}\log\left(\frac{13}{\varepsilon}\right)\}$, $\overline{a} \in X^n$ is an ε-net for $(\mathcal{C} \triangle f)$ with probability greater than $1-\delta$.
  - If $\mathrm{err}_\mu(H, f, \overline{a}) \geqslant \varepsilon$, then by definition of an ε-net, $\overline{a}$ "catches" some $a_i \in X$ such that $f(a_i) \neq h(a_i)$.
  - This contradicts consistency of H, therefore

$$\mu\left(\left\{\overline{a} \in X^n : \mathrm{err}_\mu(H, f, \overline{a}) > \varepsilon\right\}\right) < 2(2n)^d 2^{-\frac{\varepsilon n}{2}} \leqslant \delta.$$

Proof (Theorem 1.4) We prove both directions of the equivalence:

$\Longrightarrow$: Suppose $\mathcal{C}$ is a VC class. By definition, there exists $d < \omega$ such that $\mathrm{VCdim}(\mathcal{C}) = d$. By Theorem 1.6, any consistent hypothesis function H is a PAC learning function for $\mathcal{C}$. Therefore, $\mathcal{C}$ is PAC learnable.

$\Longleftarrow$: Suppose $\mathcal{C}$ is not a VC class. Then $\mathcal{C}$ has infinite VC dimension. By Theorem 1.5, for any hypothesis function H and VC dimension $d \in \mathbb{N}$, the sample complexity satisfies $N_{\varepsilon,\delta} \geqslant d(1 - 2(\varepsilon(1-\delta) + \delta))$. Since the VC dimension of $\mathcal{C}$ is infinite, no finite sample size is sufficient for PAC learning $\mathcal{C}$. Therefore, $\mathcal{C}$ is not PAC learnable.

1.3 NIP theories

Now we move into the realm of model theory.

Remark 1.21 (Notation) We will work in a fixed signature L.

- $\mathcal{M}$ denotes an L-structure with universe M.
- x, y, z represent tuples of variables,
- $|x|$ denotes the length of tuple x.
- For $A \subseteq M$, A_x represents all tuples from A of length $|x|$.

In a partitioned L-formula $\varphi(x; y)$

- x represents object variables,
- y represents parameter variables,
- for $b \in M_y$, $A \subseteq M_x$ define $\varphi(A, b) = \{x \in A : \mathcal{M} \models \varphi(x, b)\}$,
- for $A \subseteq M_x$ and $B \subseteq M_y$ define $\varphi(A; B) = \{\varphi(a; B) : a \in A\}$.

Definition 1.22 (Uniformly definable family) Let $\mathcal{M}$ be an L-structure and $\varphi(x; y)$ any fixed L-formula. The formula $\varphi(x; y)$ generates a *uniformly definable family* on M_x as a collection of definable sets:

$$\mathcal{C}_\varphi = \{\varphi(M_x; b) : b \in M_y\}.$$

The VC dimension of $\varphi(x; y)$ is defined to be the VC dimension of the induced concept class $\mathcal{C}_\varphi$.

The term “uniformly definable” emphasizes that a single formula φ simultaneously defines all sets within the family, instead of using multiple formulas to define different sets. This uniformity allows us to study properties of the entire family by analyzing the single formula φ.

Example 1.23 Consider RCF, the theory of ordered real closed fields. Let $\varphi(x_1, x_2; y_1, y_2, y_3)$ be the formula:

$$(x_1 - y_1)^2 + (x_2 - y_2)^2 < y_3$$

This formula defines the interior of a circle in $\mathbb{R}^2$. Specifically, $\varphi(x_1, x_2; y_1, y_2, y_3)$ generates the family of sets

$$\mathcal{C}_\varphi = \{\{(x_1, x_2) \in \mathbb{R}^2 : (x_1 - a)^2 + (x_2 - b)^2 < r\} : a, b, r \in \mathbb{R}\}$$

Each set in this family is the interior of a circle with center (a, b) and radius $\sqrt{r}$. The corresponding concept class consists of indicator functions:

$$\mathcal{C} = \{f_{a,b,r} : \mathbb{R}^2 \to 0, 1 \mid a, b, r \in \mathbb{R}, r > 0\}$$

where

$$f_{a,b,r}(x_1, x_2) = \begin{cases} 1 & \text{if } (x_1 - a)^2 + (x_2 - b)^2 < r \\ 0 & \text{otherwise} \end{cases}$$

Thus, φ uniformly defines all open circles in $\mathbb{R}^2$, allowing us to study properties of this entire family through the single formula φ.

Definition 1.24 (Independence property)

- A formula $\varphi(x; y)$ has the *independence property* with respect to $\mathcal{M}$, if:

  - For every $n \in \mathbb{N}$, there exists a sequence $(b_0, \ldots, b_{n-1})$ of elements from M_y such that
  - For every $k \subseteq n$, there exists is an $a_k \in M_x$ such that

$i < k < n$

$$\mathcal{M} \models \varphi(a_k; b_i) \iff i \in k$$

- The *independence dimension* $I(\varphi)$ is defined as:
  - If φ does not have the independence property (*is NIP*), then $I(\varphi)$ is the greatest n for which the above condition holds.
  - If φ has the independence property *(is not NIP)*, then $I(\varphi) = \infty$.
- The *dual formula* $\psi(y; x)$ represents a dual formula; φ and ψ are identical as formulas but the roles of x and y are reversed.

Our main theorem is Proposition 1.3 from [Las92] establishing the equivalence between the independence property and VC dimension.

From now on, fix a structure $\mathcal{M}$ and a formula $\varphi(x; y)$. Let $\mathcal{C}_\varphi$ be the concept class associated with φ in $\mathcal{M}$.

Theorem 1.7 (Prop. 1.3, [Las92]) *If VC dimension of $\mathcal{C}_\varphi$ is d and the independence dimension of φ is n then $n \leqslant 2^d$ and $d \leqslant 2^n$, and the following are equivalent:*

a) $\mathcal{C}_\varphi$ is a VC class.

b) φ is NIP.

This theorem will immediately follow from two lemmas below.

Lemma 1.25 (Lemma 1.4, [Las92]) *Let $\psi(y; x)$ be the dual formula of $\varphi(x; y)$. Then* $\mathrm{VCdim}(\mathcal{C}_\varphi) \leqslant d \iff I(\psi) \leqslant d$

PROOF We will show that $\mathrm{VCdim}(\mathcal{C}_\varphi) > d \iff I(\psi) > d$, which is equivalent to the statement of the lemma.

- By definition, $\mathrm{VCdim}(\mathcal{C}_\varphi) > d$ if and only if there exists a set $A = \{a_0, \dots, a_d\} \subseteq M_x$ that is shattered by $\mathcal{C}_\varphi$. This holds if and only if for every $S \subseteq d$, there exists $b_S \in M_y$ such that:

$$\mathcal{M} \models \varphi(a_i; b_S) \iff i \in S.$$

- By the definition of the dual formula ψ, this condition is equivalent to:

$$\mathcal{M} \models \psi(b_S; a_i) \iff i \in S$$

This last statement is precisely the definition of $I(\psi) > d$. Therefore, $\mathrm{VCdim}(\mathcal{C}_\varphi) > d \iff I(\psi) > d$, which completes the proof.

Lemma 1.26 (Lemma 1.5, [Las92]) *Let $\psi(y; x)$ be the dual formula of $\varphi(x; y)$. If $I(\varphi) \leqslant n$ then $I(\psi) \leqslant 2^n$.*

The idea of the proof rests on the observation that a shattered set of size n corresponds to 2^n parameters that shatter it. Each element in this set implicitly defines a subset of these parameters — those corresponding to sets containing that element.

Proof We prove the contrapositive: $I(\psi) > 2^n \implies I(\varphi) > n$.

- Since $I(\psi) > 2^n$, there exict sequences $\{b_i : i \in 2^n\}$ and $\{a_k : k \subseteq 2^n\}$ such that for all $i \in 2^n$ and $k \subseteq 2^n$: $\mathcal{M} \models \psi(a_k, b_i) \iff i \in k$
- Dualizing ψ to φ, we obtain: $\mathcal{M} \models \varphi(b_i, a_k) \iff i \in k$. Note that $i \in 2^n$ is a number and $k \subseteq 2^n$ a subset.
- To obtain $I(\varphi) > n$, we need to reverse these roles:
  - For each $i \in n$, define $k_i = \{l \subseteq n : i \in l\}$, the set of all subsets of n containing i.
  - For each $i \in n$, define $c_i = a_{k_i}$.
- Now we can show that for all $i \in n$ and $k \subseteq n$:

$$\mathcal{M} \models \varphi(b_i; c_k) \iff \mathcal{M} \models \varphi(b_i; a_{k_i}) \iff k_i \in i$$

  - Now b_i encodes a subset of n and $k_i \in n$ a number.
  - The first equivalence holds by definition of c_k.
  - The second equivalence holds because $i \in k_i \iff i \in k$.
- Therefore, the sequence $(c_k : k \subseteq n)$ demonstrates that $I(\varphi) > n$, as it satisfies the independence property for φ with respect to the sequence $(b_i : i \in n)$.

By contraposition, we conclude that if $I(\varphi) \leqslant n$, then $I(\psi) \leqslant 2^n$, proving the lemma.

Proof (Theorem 1.7) We will prove both directions of the equivalence.

$a) \implies b)$ Assume $\mathcal{C}_\varphi$ is a VC class, so $\mathrm{VCdim}(\mathcal{C}_\varphi) = d < \infty$.

- By Theorem 1.25, $\mathrm{VCdim}(\mathcal{C}_\varphi) = d \iff I(\psi) \leqslant d$.
- Applying Theorem 1.26 to ψ, we get $I(\varphi) \leqslant 2^d < \infty$.
- Thus, $I(\varphi) \leqslant 2^d$, implying φ has finite independence dimension and is NIP.

$b) \impliedby a)$ Assume φ is NIP, so $I(\varphi) = n < \infty$.

- By Theorem 1.26, $I(\varphi) = n \implies I(\psi) \leqslant 2^n$.
- Applying Theorem 1.25, we get $\mathrm{VCdim}(\mathcal{C}_\varphi) \leqslant 2^n < \infty$.
- Thus, $\mathrm{VCdim}(C_\varphi) \leqslant 2^n$, implying $\mathcal{C}_\varphi$ has finite VC dimension and is a VC class.

Definition 1.27 (Independence property for theories) A theory *is NIP* if and only if every formula is NIP.

This implication establishes that any formula in a NIP theory, as well as finite boolean combinations of such formulas, has a finite VC dimension. As a direct consequence, these formulas (uniformly) define PAC learnable concept classes, significantly expanding our understanding of learnable concept classes. Model theory provides us with lots of interesting NIP theories, all of which are now known to be PAC learnable. This also recovers many standard computer science results, including halfspaces, threshold functions, circles, and convex n-gons, albeit without explicit bounds. We give more examples at the end of the Section 2.3.2.

2 Littlestone dimension and stable theories

2.1 Online learning framework

Unlike the previously discussed PAC learning model, which relies on a set of training examples to develop a hypothesis before applying it to new data, online learning operates in a dynamic, sequential manner. The online learning process can be thought of as a game between two players: the learner and the environment. This game proceeds over a series of rounds, each of which follows a particular pattern:

1) The environment selects an instance x_i.
2) The learner predicts a label $h(x_i) \in 2$.
3) The environment reveals the true label $f(x_i) \in 2$.

In this framework, the goal of the environment is to challenge the learner by selecting instances that lead to errors. To achieve this, the environment must select $y_i = 1 - h(x_i)$ for each round. However, the environment faces a constraint: it must select instances in a way that is consistent with the hypothesis class $\mathcal{C}$, ensuring that the target concept remains within that class.

As noted in the Theorem 1.2 about consistent and realizable learning, we will focus our attention on scenarios where the online learning process is both consistent and realizable. The complexity of the online learning task varies significantly depending on the hypothesis class $\mathcal{C}$:

- In the most challenging case, where $\mathcal{C} = {}^{X}2$, the learner has no chance of consistently predicting the correct label.
- In the simplest case, where $\mathcal{C} = \{f\}$, the environment has no flexibility in choosing instances that would lead the learner to make errors.

Between these extremes lies a spectrum of hypothesis classes with varying degrees of complexity. The study of online learning aims to understand how the structure of the hypothesis class affects the learner's ability to make accurate predictions and the environment's ability to present challenging instances.

The next several definitions formally describe the learning process and introduce the concepts of binary trees and their labelings. In these binary trees, which we will refer to as *mistake trees* in the context of online learning, each internal node is associated with a sample x_i from the input space, while the external nodes (leaves) correspond to functions f from the concept class. Each path through the tree represents a sequence of choices made by the environment. Branching left corresponds to choosing 1, while branching right corresponds to choosing 0. A labeling represents a valid assignment of $x_i \in$

X and $f \in \mathcal{C}$ to nodes in the tree. This tree structure provides a visual representation of the potential trajectories of the learning process.

The notation and terminology in this section follows [Bha21].

Definition 2.1 (Trees for learning problems) Given a learning problem $(X, \mathcal{C})$, we define the following terms:

- A *binary tree* is either a single leaf or a pair of subtrees.
- A *binary element tree*, denoted by T, is a rooted binary tree with nodes partitioned into leaves $L \subseteq T$ and non-leaves $N \subseteq T$. The leaves L are labeled with elements from $\mathcal{C}$, and the non-leaves N are labeled with elements from X.
- A *perfect binary element tree*, denoted by B_n, is a binary element tree T in which every non-leaf has exactly two children, and all leaves are located at the same level n.

Our definition of a binary tree allows for both finite trees and infinite trees of depth ω (and defines a coinductive datatype). In a traditional set-theoretic approach, a tree T would typically be defined as a nonempty prefix-closed subset of $2^{<\omega}$ such that for every $u \in 2^\omega, u0 \in T \iff u1 \in T$.

Despite our formal definition, we can still intuitively imagine a tree as a set of nodes, one of which is a root, some of which are leaves, and the set is equipped with a partial order that defines the ancestry relationship.

Definition 2.2 (Order relation on trees) We define a partial order relation on nodes of an ordered tree T. For any nodes $u, v \in T$, we say

- node v is *below* node u, denoted by $u < v$, if $u \neq v$ and v is contained in the subtree with root u,
- node v is *left below* u, denoted by $u <_L v$, if v is contained in the left subtree of u,
- node v is *right below* u, denoted by $u <_R v$ if v is contained in the right subtree of u.

At a casual glance, one might mistakenly interpret $u < v$ as u being below v, whereas in this context, v is actually lower than u. Rather, one should imagine $u <$ (literally) as a root of a subtree represented by the symbol $<$.

We analogously define the notion of *above*, *left above* and *right above*.
For an unordered tree T, where we cannot distinguish between *left* and *right* nodes, we instead define

- $v \sim_u w$ if v, w lie in the same subtree of u,
- $v \perp_u w$ if v, w lie in different subtrees of u.

Unordered trees will become relevant later in the context of model theory. In the contex of online learning we will focus on ordered trees.

Definition 2.3 (Tree embedding & dimension)

- an *embedding* of tree T_1 into the tree T_2 is an injection of nodes in T_1 into nodes in T_2 preserving the order structure given by $<_L$ and $<_R$ or by $<, \sim, \perp$ relations.

- the *dimension* d of a tree is the largest n such that B_n can be embedded into T or ∞ if there are arbitrarily large such trees.

In the definition 2.1 we split tree T in leaves and non-leaves labeled by elements from $\mathcal{C}$ and X, respectively. We now describe properties, constituting a valid labeling of T.

Definition 2.4 (Labelings of trees) For an ordered tree T, a *labeling* $\alpha : T \to X \cup \mathcal{C}$ is *valid* if:

- α consists of two disjoint injective maps $N \to X, L \to \mathcal{C}$ and
- for every leaf $v \in L$ the following conditions are satisfied:
  - if $u <_L v$, then $\alpha(v)(u) = 1$,
  - if $u <_R v$, then $\alpha(v)(u) = 0$.

For an unordered tree T, a *labeling* $\alpha : T \to X \cup \mathcal{C}$ is *valid* if:

- α consists of two disjoint maps $N \to X, L \to \mathcal{C}$ and
- for every non-leaf u and leaves $v, w \in L$ the following holds:
  - if $v \sim_u w$, then $\alpha(v)(u) = \alpha(w)(u)$,
  - if $v \perp_u w$, then $\alpha(v)(u) \neq \alpha(w)(u)$.

No requirements are imposed on other nodes where the order relation may be undefined. In some situations, if the labeling $N \to X$ is already given and we can extend it to a valid labeling, we say that T is *labeled* by $\mathcal{C}$ or $\mathcal{C}$ *shatters* T or T *admits* $\mathcal{C}$.

The concept of a tree admitting $\mathcal{C}$ or being shattered by $\mathcal{C}$ is particularly important. It means that the concept class is rich enough to realize all the classifications represented by the leaves of the tree, given the feature tests at the internal nodes. In other words, this means that the concept class has enough expressiveness to be consistent with the decision strategy represented by the tree structure.

2.2 Littlestone dimension and SOA

This section follows [Lit88].

Definition 2.5 (Littlestone dimension) The *Littlestone dimension* of $\mathcal{C}$, denoted as $\mathrm{Ldim}(\mathcal{C})$, is the largest n such that $\mathcal{C}$ shatters B_n. If $\mathcal{C}$ shatters arbitrarily large finite B_n, then $\mathrm{Ldim}(\mathcal{C}) = \infty$. If $\mathcal{C}$ shatters no B_n, then $\mathrm{Ldim}(\mathcal{C}) = -\infty$.

The Littlestone dimension of a concept class thus represents the maximum number of errors that the environment can force any learning algorithm to make in that scenario. This dimension provides a worst-case bound on the number of prediction errors that must be made when learning any concept from the class. A higher Littlestone dimension indicates a more complex concept class, which is inherently more difficult to learn in an online setting.

Remark 2.6 Littlestone dimension is a weaker notion of dimension than VC dimension. We offer three complementary perspectives:

- For a leaf v in a binary tree, define $P(v)$ as the set of nodes on the path from the root to v, where $P_L(v)$ and $P_R(v)$ represent left and right turns, respectively. A concept class $\mathcal{C}$ shatters a tree if, for each leaf v, there exists $f \in \mathcal{C}$ such that $f^{-1}(1) \cap P(v) = P_L(v)$. This definition is weaker than its VC analog because it requires f to cut out $P_L(v)$ from $P(v)$ and not from the entire input space X. In essence, we restrict the input space X to $P(v)$ for each leaf v, and make no requirements about how concepts should behave on nodes outside of these root-to-leaf paths.
- The Littlestone dimension becomes equivalent to the VC dimension if we require that internal nodes at the same level have $x \in X$. This shows that the VC dimension serves as a lower bound on the Littlestone dimension, a statement we will prove later in Theorem 2.10.
- As shown in the previous section, all concept classes with finite VC dimension correspond to NIP formulas. This section will establish that finite Littlestone dimension is equivalent to a formula being stable. This connection will not surprise readers familiar with stability theory, as stable theories are a subclass of NIP theories.

Definition 2.7 (Thicket shatter function) Define the *thicket shatter function* $\rho_{\mathcal{C}}(m) : \mathbb{N} \to \mathbb{N}$ as

We count the maximum number of leaves labeled correctly.

$$\rho_{\mathcal{C}}(m) := \max\{|L| : L \subseteq T, \mathcal{C} \text{ shatters } B_m\}.$$

The Sauer-Shelah lemma 1.11 also holds verbatim for the thicket shatter function $\rho_{\mathcal{C}}(m)$.

Lemma 2.8 (Sauer-Shelah lemma, 1972) *Let* $\Phi_n(m) := \sum_{i=0}^{n} \binom{m}{i}$. *If* $\mathcal{C}$ *has Littlestone dimension* n *and* $m > n$, *then* $\rho_{\mathcal{C}}(m) \leqslant \Phi_n(m)$ *and, in particular,* $\rho_{\mathcal{C}}(m) \in O(m^n)$.

Proof The proof is lengthy and inductive, offering limited insight, so we omit it and refer the reader to the proof of Theorem 4.2 in [Bha21].

Definition 2.9 (Mistake bound) Let $f \in \mathcal{C}$ be a target concept and H a hypothesis function. We define the *mistake bound* $\text{mis}(H, f, \overline{a})$, as the maximum number of mistakes H makes predicting f on the sequence $\overline{a} = \{a_0, \ldots, a_{n-1}\}$.

From now on fix a concept class $\mathcal{C}$ with Littlestone dimension d.

Theorem 2.1 (Lemma 1, [Lit88]) *The number of mistakes of any deterministic algorithm H is at least d.*

Proof By Theorem 2.5, $\mathcal{C}$ shatters a perfect mistake tree B_d. We will show that an adversarial environment can force any deterministic learning algorithm to make at least d mistakes:

- Present the instances x_i to the learner in the order they appear along a path from the root to a leaf in B_d.
- For each prediction $h(x_i)$ made by the learner, assign the opposite truth value $1 - h(x_i)$ to the instance.

This strategy ensures that the learner makes a mistake on every prediction and there always exists a hypothesis $f \in \mathcal{C}$ consistent with all the assigned labels, due to the tree being shattered. Since the height of the tree is d, the adversary can force the learner to make at least d mistakes before reaching a leaf. This holds true regardless of the specific algorithm used by the learner, as long as it is deterministic. Therefore, the number of mistakes of any deterministic algorithm on the concept class $\mathcal{C}$ is at least d, which is the Littlestone dimension of $\mathcal{C}$.

Theorem 2.2 (Algorithm 2 and Theorem 3, [Lit88])

There exists an algorithm H that makes at most d mistakes. The algorithm $\mathcal{H}$ is usually refered to as the Standard Optimal Algorithm (SOA).

Proof We provide a constructive proof by describing the Standard Optimal Algorithm (SOA) and demonstrating its optimality.

- **The algorithm:** The algorithm proceeds in rounds, maintaining a hypothesis class V_i in each round i.
  - Initialization: Set $V_0 = \mathcal{C}$.
  - For each round i:
    - Receive $x_i \in X$.
    - Partition V_i into two subclasses:

$$V_i^0 := \{f \in V_i : f(x_i) = 0\} \text{ and } V_i^1 := \{f \in V_i : f(x_i) = 1\},$$

$$V_i = V_i^0 \cup V_i^1$$

From a computational perspective, the computation of both VC dimension and Littlestone dimension is NP-hard. Furthermore, there is no polynomial-time algorithm that can approximate either of these dimensions to within a factor of $o(\log n)$, see arXiv:2211.01443.

  - ∗ Choose prediction $p_i = \arg\max_{j\in\{0,1\}} \operatorname{Ldim}(V_i^j)$.
  - ∗ Receive true label y_i.
  - ∗ Update $V_{i+1} = V_i^{y_i}$.

  It should be noted that if $\operatorname{Ldim}(V_i^0) \neq \operatorname{Ldim}(V_i^1)$, then $y_i = 1-p_i$ unless the environment is adversarial and maximizing mistakes, hence the use of "at most" in the theorem statement.

- **The optimality:** We prove that the Littlestone dimension of V_i strictly decreases in each round, implying that the algorithm makes at most d mistakes. Proof by contradiction:

  - Assume $\operatorname{Ldim}(V_{i+1}) \geqslant \operatorname{Ldim}(V_i)$ for some i. Then by construction, V_{i+1} is either V_i^0 or V_i^1.
  - Since we choose the subclass with greater Littlestone dimension, both $\operatorname{Ldim}(V_i^0)$ and $\operatorname{Ldim}(V_i^1)$ are greater than or equal to $\operatorname{Ldim}(V_i)$. However, V_i is the union of V_i^0 and V_i^1, which leads to the following inequality:

    $$\operatorname{Ldim}(V_i) \geqslant \min\{\operatorname{Ldim}(V_i^0), \operatorname{Ldim}(V_i^1)\} + 1 > \operatorname{Ldim}(V_i),$$

    This inequality holds because V_i is a mistake tree that is one level deeper than the smaller of the mistake trees V_i^0 and V_i^1. This contradiction proves our claim.

  Since the Littlestone dimension strictly decreases in each round and is initially d, the algorithm makes at most d mistakes.

Lemma 2.10 (Theorem 4, [Lit88]) *The Littlestone dimension of a concept class $\mathcal{C}$ is always greater than or equal to its VC dimension.*

The equivalent model-theoretic statement is: "If φ is stable, then φ is NIP."

PROOF Let d be the VC dimension of $\mathcal{C}$. By definition, there exists a set $A = \{a_0, \dots, a_{n-1}\}$ of size d shattered by $\mathcal{C}$. Construct a perfect binary tree of height d as follows: We use the elements $a_i \in A$ as labels for the internal nodes of the tree. All nodes at depth i are labeled with $a_i \in A$. The leaves of this tree, which are at depth d, are then labeled with functions from $\mathcal{C}|_A$. The resulting tree is shattered by $\mathcal{C}$ and has height d. This implies that the Littlestone dimension of $\mathcal{C}$ is at least d.

The lemma we proved demonstrates that the gap between the Littlestone dimension and the VC dimension of a concept class $\mathcal{C}$ is always non-negative. Moreover, it can be arbitrarily large, as the following example illustrates.

Example 2.11 Let $X = [0, 1]$ and $\mathcal{C}$ the class of threshold functions on X. The VC dimension of $\mathcal{C}$ is 2 and the Littlestone dimension is ∞.

To demonstrate that $\mathcal{C}$ has infinite Littlestone dimension, we can construct an infinitely deep binary tree as follows:

- Consider leaves $v_1 = \frac{1}{2}, v_2 = \frac{1}{4}, v_3 = \frac{1}{8}, \ldots, v_i = \frac{1}{2^i}$
- Label each leaf with
$$h_i(x) = \begin{cases} 1, & x \geqslant \frac{1}{2^i} \\ 0, & x < \frac{1}{2^i} \end{cases}$$

For the internal nodes, we can always find values in $[0, 1]$ that are consistent with the labeling of the leaves. This is possible due to the density of real numbers.

2.3 Stable theories

The next section is based on Chapter 5 from [Bha21].

2.3.1 Littlestone dimension and Shelah 2-rank

Remark 2.12 (Set systems and learning problems, revised) The fundamental objects in this section are set systems.

- A set system $(X, \mathcal{F}, \in)$ is a structure with universe $M = X \cup \mathcal{F}$, sorts M_X and $M_{\mathcal{F}}$ and equipped with a binary relation symbol $\in \subseteq M_X \times M_{\mathcal{F}}$. For brevity, we will write $(X, \mathcal{F})$, suppresing $\in$.
- Any such set system corresponds to a learning problem $(X, \mathcal{C})$, where X is the input space, $\mathcal{C} = \{\mathbb{1}_A(x) : A \in \mathcal{F}\}$ is the concept class. Each concept or hypothesis function is in correspondence with a set $A \in \mathcal{F}$.

Remark 2.13 (Model-theoretic setup) We discuss model-theoretic assumptions and preparations we need to make in order to prove the results. We use `typewriter` script to distinguish syntactic variables $\mathtt{x}, \mathtt{y}$ from values x, y.

- In this section, we fix a partitioned first-order formula $\varphi(\mathtt{x}; \mathtt{y})$ and the induced set system $(X, \mathcal{F})$ with $\mathcal{F} = \{\varphi(M_{\mathtt{x}}, b) : b \in M_{\mathtt{y}}\}$. The relation $\in$ in our set system is interpreted as $a \in F \iff \mathcal{M} \models \varphi(a; b)$ for the b that defines F.
- We fix a sufficiently saturated model $\mathcal{M}$ of the theory $\mathrm{Th}(X, \mathcal{F})$ to ensure that all types are realized, which may affect later rank calculations if, for example, we work in a model which doesn't realize many types. Whenever we assert that some sentence holds, we always mean relative to the model $\mathcal{M}$.

In our previous discussion (Theorem 1.2), we focused on consistent and realizable cases. To extend this concept into model-theoretic terms, we will now introduce the notion of partial φ-types.

Definition 2.14 Let φ be the formula $\mathtt{x} \in \mathtt{F}$. We define:

- A *φ-formula* is either $\varphi(x;\mathtt{F})$ or $\neg\varphi(x;\mathtt{F})$ for some $x \in M_x$. For brevity, we denote $\varphi(x;\mathtt{F})^1 = \varphi(x;\mathtt{F})$ and $\varphi(x;\mathtt{F})^0 = \neg\varphi(x;\mathtt{F})$.
- A *finite φ-type* p is a conjunction of φ-formulas, including the *empty conjunction* $\top$. We denote by $p(\mathcal{F})$ the subfamily of $\mathcal{F}$ satisfying the type p.
- Two finite φ-types p and q are *contradictory*, if there exists $x \in X$ and $t \in \{0,1\}$ such that:
  - $\varphi(x;\mathtt{F})^t \in p$,
  - $\varphi(x;\mathtt{F})^{1-t} \in q$

  In this case, we say p and q disagree on $\varphi(x;\mathtt{F})$.

How does the notion of a partial φ-type relate to online learning? As we receive observed data in the form of pairs (x_i, y_i), where $x_i \in X$ represents an instance and $y_i \in 0,1$ represents its label, we can enforce consistency in our structure $(X,\mathcal{F})$ by requiring either $\varphi(x_i;\mathtt{F})$ or $\neg\varphi(x_i;\mathtt{F})$ to hold in $(X,\mathcal{F})$, depending on the value of y_i. This process effectively restricts our concept class to only those concepts that agree with the observed data and the collection of these restrictions is what we now defined as a finite φ-type. The subfamily of $\mathcal{F}$ that satisfies these constraints is written as $p(\mathcal{F})$. This represents the set of concepts in our class that are consistent with the observed data so far.

Now, we formalize Theorem 2.4 describing conditions under which a tree T admits $(X,\mathcal{F})$ consistent with a finite φ-type p.

Definition 2.15 (Definition 5.1, [Bha21]) Let T be an unordered tree with non-leaves N and leaves L. We define

- a signature $\mathcal{L}_T = \{\in\} \cup \{a_u : u \in N\} \cup \{b_v : v \in L\}$, where
  - $\in$ is a binary relation symbol $\subseteq X \times \mathcal{F}$
  - $a_u : u \in N$ are constant symbols of sort X
  - $b_v : v \in L$ are constant symbols of sort $\mathcal{F}$
- a first-order $\mathcal{L}_T$-theory Adm_p^T with the following axioms:

  (1) $p(b_v)$ for any $v \in L$,

  (2) $\varphi(a_u;b_v) \not\leftrightarrow \varphi(a_u;b_w)$ if $v \perp_u w$,

  (3) $\varphi(a_u;b_v) \leftrightarrow \varphi(a_u;b_w)$ if $v \sim_u w$.

For simplicity, we will occasionally denote the constants a_u or b_v as the image $\alpha(u)$ or $\alpha(v)$ of the respective node under a valid labeling α.

If $(X, p(\mathcal{F}))$ admits T, then $\mathrm{Th}(X, p(\mathcal{F})) \cup \mathrm{Adm}_p^T$ is consistent. This means there exists a model which simultaneously satisfies both $\mathrm{Th}(X,\mathcal{F})$ and Adm_p^T. Since $\mathcal{M}$ is sufficiently saturated, the consistency of $\mathrm{Th}(X,\mathcal{F}) \cup \mathrm{Adm}_p^T$ is equivalent to the admissibility of T in $\mathcal{M} = (M_X, p(M_{\mathcal{F}}))$.

Example 2.16 Let $X = \{1,2,3,4\}$ and $\mathcal{F} = \{\{1,2\},\{3,4\},\{1,3\}\}$. We consider the structure $(X,\mathcal{F})$ with the signature $\mathcal{L} = \{\in\}$. The theory of $(X,\mathcal{F})$ is the set of all first-order L-sentences true in $(X,\mathcal{F})$. For example, the sentence "there are exactly 3 sets in $\mathcal{F}$" can be expressed as:

$$\exists x_1, x_2, x_3 : \left(\bigwedge_{i=1}^{3} x_i \in M_{\mathcal{F}}\right) \wedge \left(\bigwedge_{i=1}^{2} x_i \neq x_{i+1}\right) \wedge \left(\forall y : y \in M_{\mathcal{F}} \wedge \bigvee_{i=1}^{3} y = x_i\right).$$ [2]

Define a partial φ-type $p(\mathtt{F}) = \{\neg\varphi(4;\mathtt{F})\}$. This implies $p(\mathcal{F}) = \{\{1,2\},\{1,3\}\}$.

Now consider a tree T with one root u and two leaves v, w. We expand the signature to $\mathcal{L}_T = \{\in, a_u, b_v, b_w\}$, where a_u is a constant symbol of sort X and b_v, b_w are constant symbols of sort $\mathcal{F}$. The theory Adm_p^T consists of the following $\mathcal{L}_T$-sentences:

$$\{4 \notin b_v, 4 \notin b_w, a_u \in b_v \not\leftrightarrow a_u \in b_w\}.$$

We will now verify the equivalence between the consistency of $\mathrm{Th}(X,\mathcal{F}) \cup \mathrm{Adm}_p^T$ and T admitting $(X,\mathcal{F})$:

$\Longleftarrow$: If T admits $(X, p(\mathcal{F}))$, then there exists a valid labeling α. In our example, α is given by $\alpha(u) = 3, \alpha(v) = \{1,3\}, \alpha(w) = \{1,2\}$.

- The property (1) holds by construction, since $4 \notin \{1,3\}$ and $4 \notin \{1,2\}$ is both in $\mathrm{Th}(X,\mathcal{F})$ and Adm_p^T.
- The properties (2) and (3) hold since $3 \in \{1,3\}$ and $3 \notin \{1,2\}$, so $\mathrm{Th}(X,\mathcal{F})$ and Adm_p^T are consistent.

$\Longrightarrow$: If $\mathrm{Th}(X,\mathcal{F}) \cup \mathrm{Adm}_p^T$ is consistent, then there exists an interpretation in $\mathcal{M}$ of the constants a_u, b_v, b_w that satisfies property $(1), (2), (3)$. This interpretation provides a valid labeling α of T.

- This equivalence doesn't necessarily hold for infinite trees. To see why, suppose $\mathrm{Th}(X,\mathcal{F}) \cup \mathrm{Adm}_p^T$ is consistent. By the Löwenheim-Skolem theorem, it has a countable model $(X', \mathcal{F}')$. Since T is infinite, it has 2^ω leaves. For T to be admissible in $(X', \mathcal{F}')$, we would need an injective map from the leaves L to F'. This is equivalent to having an injection from 2^ω to ω, wihch is impossible.

The moral of the story is that, at least in the finite case, the consistency of $\mathrm{Th}(X,\mathcal{F}) \cup \mathrm{Adm}_p^T$ hinges on the existence of a valid labeling $\alpha : T \to X \cup p(\mathcal{F})$. This labeling exists only in a sufficiently saturated model where every leaf type is realized.

In essence, Adm_p^T is a formal way of saying that T can be properly labelled by elements of X and $\mathcal{F}$ that satisfy p, in a way that respects the branching structure of T. This is probably one of the reasons why [Hod97] describes Shelah 2-rank as the branching index.

[2] Here $x \in M_{\mathcal{F}}$ means "x is of sort $M_{\mathcal{F}}$".

Example 2.17 Imagine you are a network engineer working for a large university campus. The IT department wants to optimize WiFi coverage across the campus grounds. They need to understand the actual coverage area of each WiFi access point, which is theoretically circular and has uniform signal strength. At point $(x, y) = (1, 1)$ there is signal and at point $(x, y) = (3, 3)$ there is no signal. Then we can translate the concepts in online learning and model theory as in Table 1.

Online learning	**Model theory**
Input space: 2D real plane	$X = \mathbb{R}^2$
Concept: Interior of a circle with center (y_1, y_2) and radius y_3	$\varphi(x_1, x_2; y_1, y_2, y_3) := (x_1 - y_1)^2 + (x_2 - y_2)^2 < y_3^2$
Hypothesis class: Circles in $\mathbb{R}^2$	$\mathcal{F} = \{\mathbb{1}_{\varphi(M_x;b)} : b \in \mathbb{R}^3\}$
Labeled examples: $(1, 1)$ is positive, $(3, 3)$ is negative	$p(\mathtt{F}) = \{\varphi(1, 1; \mathtt{F}), \neg\varphi(3, 3; \mathtt{F})\}$
Hypothesis class consistent with observed data	$(X, p(\mathcal{F}))$
Mistake tree T	T
Conditions for a mistake tree T to admit $\mathcal{C}$	Adm_p^T
Littlestone dimension	Shelah 2-rank

Table 1: Dictionary of terms between online learning and model theory

Now we are ready to define a local version of Morley rank, the Shelah 2-rank.

Definition 2.18 (Shelah 2-rank) For any finite φ-type $p(\mathtt{F})$, define the *Shelah 2-rank* $R^\varphi(p)$

1) $R^\varphi(p) \geqslant 0$ if p is consistent, i.e. there exists some $b \in M_{\mathcal{F}}$ such that $p(b)$ holds.

2) For any finite k, $R^\varphi(p) \geqslant k + 1$ if there exists a pairwise contradictory family of types $\{p_i : i < \omega\}$ such that $R^\varphi(p \wedge p_i) \geqslant k$ for all $i < \omega$.

3) $R^\varphi(p) = \infty$ if $R^\varphi(p) \geqslant k$ for all $k < \omega$.

4) $R^\varphi(p) = -\infty$ if p is inconsistent.

Some definitions require only two contradicting types. The definitions are equivalent and we will implicitly prove this later.

The second condition in Theorem 2.18 allows us to infer the existence of an additional order structure on types $\{p_i : i < \omega\}$, which can be extracted as an infinite sequence.

Lemma 2.19 (Lemma 5.6, [Bha21]) *Suppose $\{p_i : i < \omega\}$ is a sequence of pairwise contradictory finite φ-types. Then there exists an infinite set $S \subseteq \omega$ such that for any $r \in S$, there exists a $a \in M_X$, such that for any $s > r$ in S, p_r and p_s disagree on $\varphi(a, \mathtt{F})$.*

Proof We will construct S inductively as follows:

- Let $S_0 = \omega$.
- For any $S_i \subseteq \omega$ let m_i be its least element.
  - Consider finite φ-type p_{m_i} consisting of finitely many formulas $\{\varphi(x_i; \mathtt{F})^t : i < \omega, t \in 2\}$.
  - Since any two types are pairwise contradictory, there exists set $S_{i+1} \subseteq S_i \setminus \{m_i\}$ such that $\varphi(x_i; \mathtt{F})^t$ occurs in p and $\varphi(x_i; \mathtt{F})^{1-t}$ occurs in infinitely many $(p_j)_{j \in S_{i+1}}$.
- This construction gives us
  - A descending chain of index sets $S_0 \supset S_1 \supset S_2 \supset \ldots$,
  - An increasing sequence of indices $m_0 < m_1 < m_2 < \ldots$,
  - A sequence of elements $x_0, x_1, x_2, \ldots$

  such that

  - For all $j > i$, p_{m_j} and p_{m_i} disagree (uniformly) on $\varphi(x_i, \mathtt{F})$,
  - For each $k < i$, p_{m_k} and p_{m_i} disagree (non-uniformly) on $\varphi(x_k, \mathtt{F})$.

Therefore $S = \{m_0, m_1, \ldots\}$ has precisely the qualities we seek.

The rank of $p(\mathcal{F})$ is determined by two key factors: its consistency and its ability to be split into disjoint sets. If $p(\mathcal{F})$ is non-empty, meaning there exists some $F \in \mathcal{F}$ that realizes p, its rank is 0. This corresponds to restricting the concept class $\mathcal{C}$ to functions that are consistent with the observed data. The rank is 1 if there exists some $a \in M_X$ such that both $p \wedge \varphi(a; \mathtt{F})^0$ and $\varphi(a; \mathtt{F})^1$ are consistent. In this case, we can split $p(\mathcal{F})$ into two disjoint sets that disagree on $\varphi(a; \mathtt{F})$. This splitting behavior is analogous to the mistake trees discussed in the previous section.

The ability to continue this branching process determines the number of distinct types that φ can exhibit. Each split potentially doubles the number of types, corresponding to the different ways of extending p consistently. In fact, this branching behavior is what ultimately determines the number of types φ exhibits. It is this limitation on infinite branching that characterizes stable theories as "tame", distinguishing them from theories where such branching can proceed without bound.

Definition 2.20 (k-branching trees) Define 0-branching tree as a single leaf. For $k \geqslant 1$, the k-branching tree T_k is the unordered binary tree with subtrees T_k and T_{k-1}. The dimension of T_k is k. Note, that any T_k contains a perfect binary tree B_k.

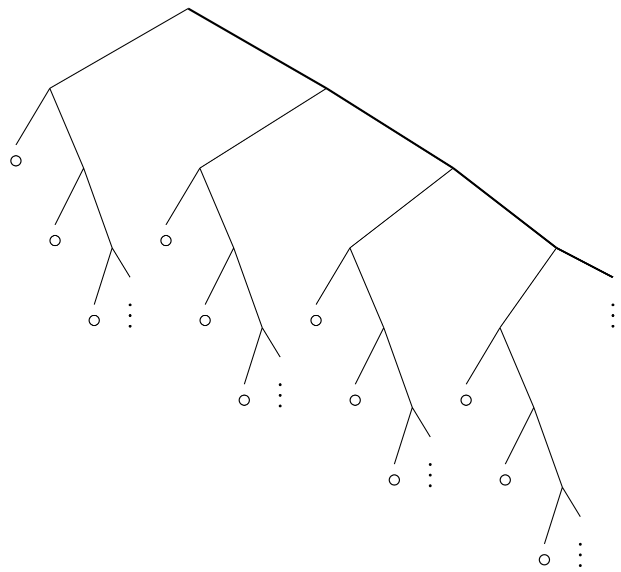


Figure 2: The infinite 2-branching tree. The *spine* is indicated by the thick edge. The vertices of any T_{k+1} can be partitioned into the vertices on the spine, plus countably many copies of T_k.

Theorem 2.3 (Theorem 5.9, [Bha21]) *The following conditions are equivalent:*

1) $R^{\varphi}(p) \geqslant k$*,*

2) $\mathrm{Th}(X, \mathcal{F}) \cup \mathrm{Adm}_p^{T_k}$ *is consistent.*

Proof We work in a sufficiently saturated model $\mathcal{M}$ of $\mathrm{Th}(X, \mathcal{F})$. Since it is sufficiently saturated, the admissibility of T_k in $\mathcal{M}$ is equivalent to the consistency of $\mathrm{Th}(X, \mathcal{F}) \cup \mathrm{Adm}_p^{T_k}$, as discussed in Theorem 2.16. We use induction:

- Base case: Show that the equivalence holds trivially for $k = 0$.
  - By Theorem 2.18, $R^{\varphi}(p) \geqslant 0$ iff there exists some $F \in M_{\mathcal{F}}$ which satisfies p, that is $\mathcal{M} \models p(F)$.
  - By Theorem 2.20, T_0 consists of a single leaf v and zero non-leaves, so the labeling $\alpha : v \mapsto F$ is valid and equivalent to the consistency of $\mathrm{Th}(X, \mathcal{F}) \cup \mathrm{Adm}_p^{T_0}$ by the discussion in Theorem 2.16.

Now comes the inductive step, where we assume the equivalence holds for k and prove it for $k + 1$.

- Assume $R^{\varphi}(p) \geqslant k + 1$:
  - By Theorem 2.18, there exists a family of pairwise contradictory types $\{p_i : i < \omega\}$ witnessing $R^{\varphi}(p) \geqslant k + 1$.

- By Theorem 2.19 there exists an infinite set $S \subseteq \omega$ such that $\{p_i : i \in S\}$ also witnesses $R^{\varphi}(p) \geqslant k+1$, with the following additional property:

  $$\forall r \in S : \exists a(r) \in M_X : \forall s > r : p_r \text{ and } p_s \text{ disagree on } a(r).$$

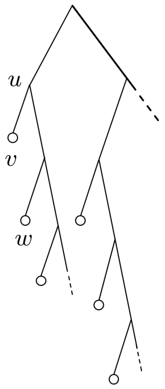
  We can assume without loss of generality that our original family already has this property. This is because we can relabel the types p_i using only indices from S and discard the unwanted types, resulting in a subfamily that still witnesses $R^{\varphi}(p) \geqslant k+1$ and has this additional structure.
- Consider the infinite $k+1$-branching tree T_{k+1}. It can be visualized as consisting of a single infinite spine with countably many copies of T_k branching off from it. We partition the nodes of T_{k+1} as follows:
  - The set of non-leaves N is partitioned as

    $$N = \left(\bigcup_{i<\omega} N_i \right) \cup N_s,$$

    where N_i is the set of non-leaves in the i-th copy of T_k and N_s is the set of vertices along the spine.
  - The set of leaves L is partitioned as

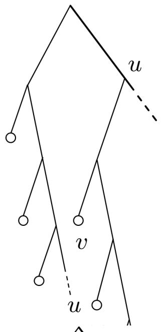
    $$L = \bigcup_{i<\omega} L_i,$$

    where L_i is the set of leaves in the i-th copy of T_k.
- By assumption, we have $R^{\varphi}(p \wedge p_i) \geqslant k$ for each $i < \omega$. Therefore by inductive hypothesis, $\mathrm{Th}(X, \mathcal{F}) \cup \mathrm{Adm}^{T_k}_{p \wedge p_i}$ is consistent for each i. This implies that each copy of T_k admits a valid labeling $\alpha_i : N_i \to M_X, L_i \to M_{\mathcal{F}}$.
- We now construct a labeling α of T_{k+1} by combining all valid labelings α_i. Define $\alpha : N \cup L \to M_X \cup M_{\mathcal{F}}$ as follows:
  - For each $i < \omega$ and $v \in N_i \cup L_i$ let $\alpha(v) = \alpha_i(v)$
  - For each $r < \omega$ and v the r-th node on the spine N_s let $\alpha(v) = a(r)$
- We will now verify that α is a valid labeling by checking axioms (1), (2), and (3) from Theorem 2.15:
  - For any leaf $v \in L$, $\alpha(v)$ satisfies $p \wedge p_i$ for some i, which implies $\alpha(v)$ satisfies p,
  - For any two leaves $v, w \in L$ and their common ancestor $u \in N$ there are four possible combinations:
    - If $v, w \in L_i$ and $u \in N_i$ for some i, the conditions are inherited from α_i.



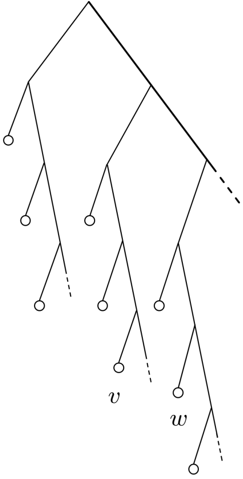
        - If $v, w \in L_i$ and $u \in N_s$, then v and w always will be in the same subtree relative to u and u must be the r-th node on the spine, $r \leqslant i$.
          Both $\alpha(v), \alpha(w)$ satisfy the type $p \wedge p_i$. By properties, established in Theorem 2.19, all realizations of p_i agree on all previous nodes $a(r)$ for all $r \leqslant i$. Therefore,

$$\varphi(a(r); \alpha(v)) \leftrightarrow \varphi(a(r); \alpha(w)).$$



        - If $v \in L_i$ and $w \in L_j$ for $i < j$ and v and w are in the same subtree relative to u.
          Then u must be the r-th node on the spine and $r < i$. By properties established in Theorem 2.19, $\alpha(v)$ and $\alpha(w)$ agree on all previous nodes $a(r)$ for all $r < i$. Therefore

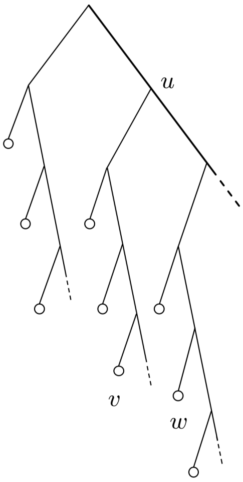
$$\varphi(a(r); \alpha(v)) \leftrightarrow \varphi(a(r); \alpha(w)).$$



        - If $v \in L_i$ and $w \in L_j$ for $i < j$ and v and w are in different subtrees relative to u.
          Then u is the r-th node on the spine and $r = i$. Yet again, by properties established in Theorem 2.19, $\alpha(v)$ and $\alpha(w)$ disagree on $a(i)$, therefore

$$\varphi(a(i); \alpha(v)) \not\leftrightarrow \varphi(a(i); \alpha(w)).$$

    - We have now proven that α is a valid labeling of T_{k+1} and since $\mathcal{M}$ is sufficiently saturated it is equivalent to $\mathrm{Th}(X, \mathcal{F}) \cup \mathrm{Adm}_p^{T_{k+1}}$ being consistent. This concludes the forward direction.



- Assume $\mathrm{Th}(X, \mathcal{F}) \cup \mathrm{Adm}_p^{T_{k+1}}$ is consistent in $\mathcal{M}$.
    - By discussion in Theorem 2.16, $(X, p(\mathcal{F}))$ admits T_{k+1}. This is equivalent to the existence of a valid labeling $\alpha : T_{k+1} \to X \cup p(\mathcal{F})$ satisfying axioms $(1), (2), (3)$ of Theorem 2.15.
    - Let N_s, N_i, L_i be as discussed above and let $a(i)$ be the label of the i-th node along the spine.
    - By axiom (2), for each $i < \omega$, all leaves v in i-th copy of T_k must agree on the truth value of $\varphi(a(i); \alpha(v))$, since for any two such leaves v, w, we have $v \sim_{a(i)} w$.
    - For $i < \omega$, define

$$p_i = \{\varphi^t(a(i), \mathcal{F})\} \cup \{\varphi^{1-t}(a(i'), \mathcal{F}) : i' < i\},$$

      where $t \in \{0, 1\}$ is chosen such that $\varphi^t(a(i), \mathtt{F})$ is satisfied by the leaves in the i-th copy of T_k.

- For each $i < \omega$, consider the restriction $\alpha|_{T_k^i}$ of α to the i-th copy of T_k. This restriction inherits properties (2) and (3) from the original labeling α. For property (1), observe that for each $v \in L_i$, $\alpha(v)$ satisfies $p \wedge p_i$ by construction of p_i.
- By the inductive hypothesis, this implies $R^\varphi(p \wedge p_i) \geqslant k$ for each $i < \omega$. Note that for each $i < j$, p_i and p_j disagree on $\varphi(a(i); \mathtt{F})$. Thus, $\{p_i : i < \omega\}$ forms a pairwise contradictory family of types. By Theorem 2.18, the existence of this family with $R^\varphi(p \wedge p_i) \geqslant k$ implies $R^\varphi(p) \geqslant k + 1$. This concludes the proof of the reverse direction.

Having established both directions, we have shown that $R^\varphi(p) \geqslant k+1$ if and only if $\mathrm{Th}(X, \mathcal{F}) \cup \mathrm{Adm}_p^{T_{k+1}}$ is consistent, completing the inductive step.

Our main result in this section establishes the equivalence between finite Littlestone dimension and finite Shelah 2-rank. The original paper by Bhaskar [Bha21] doesn't give a direct proof, so we will give it in full, bringing together the arguments.

To quantify the growth of the shatter function $\rho(m)$ (see Theorem 2.7), we introduce the intermediate concept of thicket density. This density is defined as the least exponent bounding the growth rate of $\rho(m)$, serving as a crucial link between finite and infinite structures.

Definition 2.21 (Definition 4.5, [Bha21]) Let $\rho(m)$ be the thicket shatter function of $(X, \mathcal{F})$, as defined in Theorem 2.7. The *thicket density* of $(X, \mathcal{F})$, denoted by $\mathrm{dens}(X, \mathcal{F})$, is defined as $\inf\{c \in \mathbb{R} : \rho(m) \in O(m^c)\}$.

If $\rho(m) \in O(m^c)$ for all $c \in \mathbb{R}$ then $\mathrm{dens}(X, \mathcal{F}) = -\infty$ and if $\rho(m) \notin O(m^c)$ for all $c \in \mathbb{R}$ then $\mathrm{dens}(X, \mathcal{F}) = \infty$.

The key challenge lies in transitioning from the inadmissibility of the infinite tree T_k to the inadmissibility of the finite tree B_k. While it's straightforward to show that admitting T_k implies admitting B_k, the reverse implication requires careful analysis of asymptotic behavior. Thicket density allows us to bridge this gap effectively.

Littlestone dimension	Thicket density	Shelah 2-rank
Admissibility of finite tree B_k	Rate of asymptotic growth	Admissibility of infinite tree T_k

The main challenge arises from the fact that T_k can be very deep and unbalanced, potentially having any finite number of leaves. Consequently, finding one admissible embedding of B_k in T_k doesn't guarantee that it will fully label T_k. Thicket density provides a way to overcome this obstacle by analyzing the asymptotic growth rate of admissible labelings, thereby connecting the finite structure of B_k to the infinite structure of T_k.

The original proof goes through for thicket density, briefly mentioning at the end that in case of Littlestone dimension, density and dimension are equivalent as $\mathbb{N} \cup \{\infty, -\infty\}$-valued quantities. We decided to change the arguments and instead give an "honest" proof, instead bounding Littlestone dimension by thicket density and thicket density by the Shelah 2-rank.

Theorem 2.4 (Theorem 4.4, [Bha21]) *If $R^{\varphi}(p) < k$, then* $\mathrm{dens}(\mathcal{C}_{\varphi}) < k$.

Proof Assume $R^{\varphi}(p) < k$:

- By Theorem 2.3 $\mathrm{Adm}_p^{T_k}$ is inconsistent with $\mathrm{Th}(X, \mathcal{F})$. By compactness theorem, there exists a finite subtree S of T_k such that Adm_p^S is inconsistent with $\mathrm{Th}(X, \mathcal{F})$. By discussion in Theorem 2.16, $(X, p(\mathcal{F}))$ forbids S. Since T_k has dimension k, S has dimension at most k. Therefore, $(X, p(\mathcal{F}))$ forbids some finite tree S of dimension k.

We prove that if $(X, p(\mathcal{F}))$ forbids a finite tree S of dimension k, then $\mathrm{dens}(X, p(\mathcal{F})) < k$. Proof by induction on construction of T:

- Base case: If S has dimension 0, it is the single leaf B_0. If $(X, p(\mathcal{F}))$ forbids S, then $p(\mathcal{F})$ must be empty, hence its Littlestone dimension is $-\infty < 0$.
- Inductive step: Suppose S has subtrees S_1 and S_2 with dimensions k_1 and k_2. By the inductive hypothesis, there exist functions $f_1(n) \in O(n^{k_1-1})$ and $f_2(n) \in O(n^{k_2-1})$ that serve as upper bounds on the thicket shatter functions $\rho_1(n)$ and $\rho_2(n)$ if they forbid S_1 or S_2 respectively.
  - For $x \in X$, let $\mathcal{F}_x = \{A \in \mathcal{F} : x \in A\}$ and $\mathcal{F}_{\overline{x}} = \{A \in \mathcal{F} : x \notin A\}$ with thicket shatter functions ρ_x and $\rho_{\overline{x}}$ respectively.
    - Let $P_i(x)$ express that $(X, \mathcal{F}_x)$ admits S_i.
    - Let $Q_i(x)$ express that $(X, \mathcal{F}_{\overline{x}})$ admits S_i.
  - Then $(X, p(\mathcal{F}))$ admits S if and only if:

$$\exists x \in X\, ((P_1(x) \wedge Q_2(x)) \vee (P_2(x) \wedge Q_1(x)))$$

  - Reasoning propositionally, $(X, p(\mathcal{F}))$ forbids S if and only if

$$\begin{aligned}\forall x \in X (\neg P_1(x) \wedge \neg P_2(x)) &\vee (\neg P_1(x) \wedge \neg Q_1(x)) \vee \\ (\neg Q_2(x) \wedge \neg P_2(x)) &\vee (\neg Q_2(x) \wedge \neg Q_1(x))\end{aligned}$$

  - By the inductive hypothesis, this implies:

$$\begin{aligned}\forall x \in X (\rho_x \leqslant f_1 \wedge \rho_x \leqslant f_2) &\vee (\rho_x \leqslant f_1 \wedge \rho_{\overline{x}} \leqslant f_1) \vee \\ (\rho_{\overline{x}} \leqslant f_2 \wedge \rho_x \leqslant f_2) &\vee (\rho_{\overline{x}} \leqslant f_1 \wedge \rho_{\overline{x}} \leqslant f_2),\end{aligned}$$

    where $\rho_x \leqslant f_1$ abbreviates $\forall n < \omega : \rho_x(n) \leqslant f_1 \in O(n^{k_1-1})$. Label the four cases (1)-(4) respectively.

S of dimension k
S_1, S_2 of dimension $k_1, k_2 < k$

$\rho_1 \leqslant f_1 \in O(n^{k_1-1})$
$\rho_2 \leqslant f_2 \in O(n^{k_2-1})$
$\rho \leqslant \rho_1 + \rho_2$
$\rho_x, \rho_{\overline{x}}$

- We consider two cases.
    - If (2) and (3) hold, then $\rho_x \leqslant f_1$ and $\rho_{\overline{x}} \leqslant f_1$, which implies $\rho \leqslant \rho_x + \rho_{\overline{x}} \leqslant 2f_1 \in O(n^{k-1})$.
    - If (1) and (4) hold, then reasoning propositionally:

$$
\begin{aligned}
&\forall x \in X\,((\rho_x \leqslant f_1 \wedge \rho_x \leqslant f_2) \vee (\rho_{\overline{x}} \leqslant f_1 \wedge \rho_{\overline{x}} \leqslant f_2)) \\
\Longleftrightarrow\ &\forall x \in X\,((\rho_x \leqslant f_1 \vee \rho_{\overline{x}} \leqslant f_1) \wedge (\rho_x \leqslant f_2 \vee \rho_{\overline{x}} \geqslant f_2)) \\
\Longleftrightarrow\ &\forall x \in X(\rho_x \leqslant f_1 \vee \rho_{\overline{x}} \leqslant f_1) \wedge \forall x \in X(\rho_x \leqslant f_2 \vee \rho_{\overline{x}} \leqslant f_2)
\end{aligned}
$$

- We claim that for any function g, $\forall x \in X(\rho_x \leqslant g \vee \rho_{\overline{x}} \leqslant g)$ implies $\rho \leqslant \int g$, where $\int g = 1 + \sum_{k<n} g(k)$. $g \in O(n^p) \Rightarrow \int g \in O(n^{p+1})$
    - If $g \in O(n^p)$ then $\int g \in O(n^{p+1})$, since $\sum_{k<n} k^p \in O(n^{p+1})$ by Faulhaber's formula.
    - If $k_1 = k_2$ then $d = k_1 + 1$ and ρ would be bounded by $\int f_1 \in O(n^{k_1})$.
    - If $k_1 \neq k_2$, assume without loss of generality that $k_1 < k_2$. Then ρ would be bounded above by both $\int f_1$ and $\int f_2$. Then ρ is bounded by $\int f_1 \in O(n^{k_1})$. $(\rho \leqslant \int f_1) \wedge (\rho \leqslant \int f_2)$
- To prove the above claim, we show $\rho(n) \leqslant (\int g)(n)$ by induction on n.
    - For $n = 0$, $\rho(0) \leqslant 1 \leqslant (\int g)(0)$.
    - For the inductive step, consider the labeled tree T witnessing $\rho(n)$ and let r be the label of the root. By hypothesis, either ρ_r or $\rho_{\overline{r}}$ is bounded above by g.
        - Therefore, the number of solutions in one of the subtrees must be bounded by $g(n-1)$.
        - The number of solutions in the remaining subtree is bounded by $\rho(n-1)$, which by induction is at most $(\int g)(n-1)$.
        - Therefore, the total number of solutions is at most $g(n-1) + (\int g)(n-1) = (\int g)(n)$.

Theorem 2.5 *The following conditions are equivalent for any finite type p:*

1) $\mathrm{Ldim}(\mathcal{C}_\varphi)$ *is finite.*

2) $R^\varphi(p)$ *is finite.*

Proof

1) $\Longrightarrow$ 2): Assume $R^\varphi(p) = \infty$. By Theorem 2.3, $\mathrm{Th}(X, \mathcal{F})$ is consistent with $\mathrm{Adm}_p^{T_k}$ for all $k < \omega$. Therefore there exists a valid labeling α_k of T_k for each $k < \omega$. Since each T_k has dimension k and there exists an embedding of a perfect binary tree B_k in T_k, $\mathcal{F}$ shatters each B_k by restriction of α_k. Therefore, the Littlestone dimension of $\mathcal{F}$ is ∞.

2) $\implies$ 1): By Theorem 2.4, $\text{dens}(\mathcal{C}_\varphi)$ is finite. Assume for contradiction that $\text{Ldim}(\mathcal{C}_\varphi) = \infty$, then $\mathcal{C}_\varphi$ shatters B_n for any $n \in \mathbb{N}$. By Theorem 2.8, this implies $\rho_{\mathcal{C}_\varphi}(n) = 2^n$. By definition, $\text{dens}(\mathcal{C}_\varphi) = \inf\{c \in \mathbb{R} : \rho(n) \in O(n^c)\}$. For any finite real number c, n^c grows more slowly then 2^n for sufficiently large n. Therefore, there is no finite c for which $\rho_{\mathcal{C}_\varphi} \in O(n^c)$. This means $\text{dens}(X, \mathcal{F}) = \infty$, contradiction. Therefore $\text{Ldim}(\mathcal{C}_\varphi) < \infty$.

The equivalence of Shelah's 2-rank and Littlestone dimension, illustrates a profound insight: the combinatorial structure of definable sets in a theory closely mirrors the learnability of concept classes in online learning. Both notions, at their core, measure the depth of nested binary choices that can be made before reaching an inherent limit — be it logical inconsistency or forced correct prediction.

2.3.2 Shelah 2-rank and stability

In the last section we have shown that formulas with finite Littlestone dimension have finite Shelah 2-rank. This notion of rank is precisely the dividing lane between stable and unstable formulas. To better undestand why stability theory is important, we provide the reader with a historic note on stability theory and its goals.

Remark 2.22 (Historic note on stability theory) In model theory, a fundamental question is how many models a (complete) theory with an infinite model can have. The Löwenheim-Skolem theorem tells us that every theory with an infinite model has models of arbitrary infinite cardinality.

Leopold Löwenheim, 1915
Albert Thoralf Skolem, 1920

The next question is: for a fixed infinite cardinal κ, how many non-isomorphic models of cardinality κ can a theory T have? We can consider the *spectrum function* $I(T, \kappa)$ which gives the number of non-isomorphic models of T of cardinality κ. Note that for any theory T and infinite cardinal κ larger than the cardinality of the language of T, we have $1 \leqslant I(T, \kappa) \leqslant 2^\kappa$.

A fundamental result of Morley states that if T is a countable theory and $I(T, \kappa) = 1$ for some uncountable κ, then $I(T, \kappa) = 1$ for all uncountable κ. He conjectured that $I(T, \kappa)$ is non-decreasing for uncountable cardinals. Saharon Shelah's deep and extensive work in the exploration and classification of all possible complete theories can be seen as motivated to a large extent by Morley's conjecture intended to generalize Morley's theorem to a computation of the possible "spectra" of complete first-order theories.

Michael D. Morley, 1965

In the process, he introduced several "dividing lines", separating theories that have maximum possible number of models from those whose models can be described by some "small" invariants (such as dimension in vector spaces). This classification project was closely related to understanding the behavior of types in these theories.

One of the main dividing lines is stability, introduced in his 1969 paper *Stable Theories.* His approach was motivated by the idea that "tame" theories should have "few types", since types describe the possible behaviors of elements in models or their elementary extensions. It emerged as a key property of theories with well-behaved spectrum functions.

Saharon Shelah, 1969

In "Stable Theories", Shelah proved a fundamental trichotomy for complete theories T (in a countable language):

- For any model $\mathcal{M}$, the number of types is bounded by $|A| + 2^{\aleph_0}$.
- For any model $\mathcal{M}$, the number of types is strictly greater than $|A| + 2^{\aleph_0}$ but bounded by $|A|^{\aleph_0}$.
- For any infinite λ, there exists a model $\mathcal{M}$ with $|A| = \lambda < |S_1(A)|$, where $S_1(A)$ is the set of types over A.

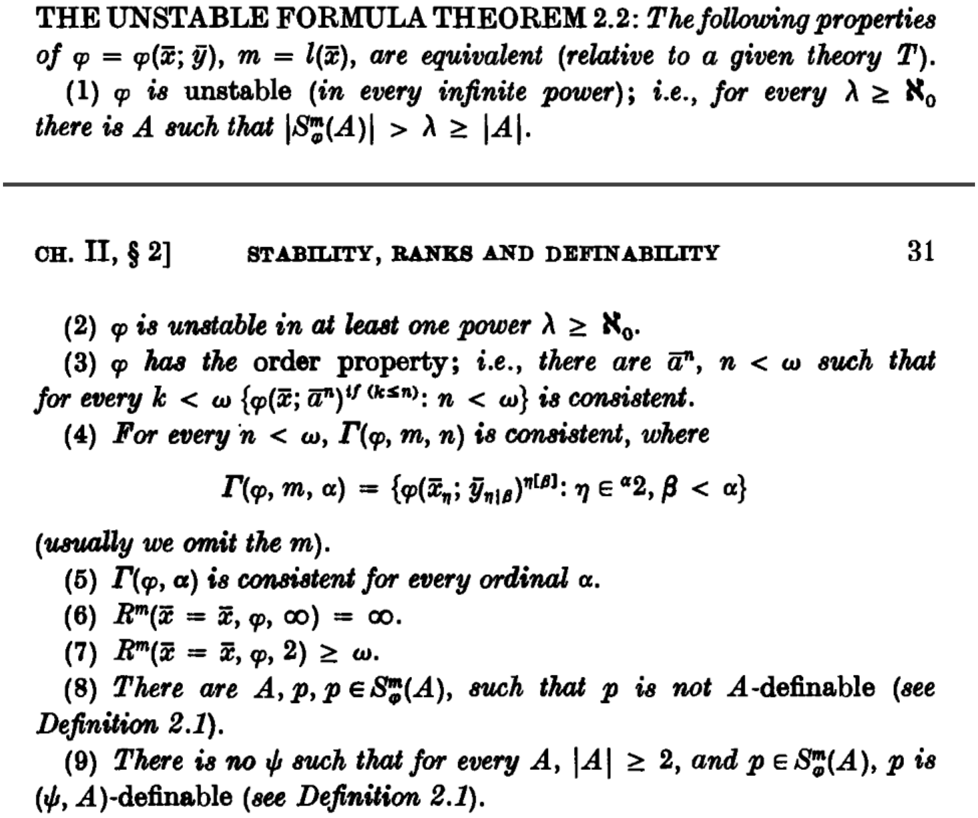
Shelah called the first two cases "stable", as the number of types remains controlled as a function of the cardinality of the parameter set. The third case represents the "wild" situation, where for any λ we can find a model that defines more than λ types. The connection between the number of types and the number of non-isomorphic models is fundamental to stability theory, since types characterize possible behaviors of elements in models. Many types suggest many different ways elements can behave across models of a theory. Moreover, models can be distinguished by the types they realize. A theory with many types thus potentially allows for many non-isomorphic models, each of which realizes a different combination of types.

This classification has laid the foundation for modern stability theory, providing a framework for understanding the complexity of first-order theories through the behavior of their types. This deep connection between types (local behavior) and models (global structure) is a key insight of stability theory, and forms the basis for much of the subsequent work in model theory.

A modern definition of stability is often given in terms of the k-order property, which originally was part of the Unstable Formula Theorem by Shelah, see Fig. 3.

Definition 2.23 (Stability)

- a formula φ has the *k-order-property* if there are k tuples $(a_i, b_i) \in M_x \times M_y$, such that $\models \varphi(a_i; b_j) \iff i < j$,
- a formula φ has the *order property* if it has the k-order-property for all finite k,
- a formula φ is *stable* if there exists some k such that φ does not have the k-order-property,
- a theory is *stable* if it implies that all formulas are stable.



THE UNSTABLE FORMULA THEOREM 2.2: *The following properties of $\varphi = \varphi(\bar{x}; \bar{y})$, $m = l(\bar{x})$, are equivalent (relative to a given theory T).*

(1) *φ is* unstable *(in every infinite power); i.e., for every $\lambda \geq \aleph_0$ there is A such that $|S^m_\varphi(A)| > \lambda \geq |A|$.*

CH. II, § 2] STABILITY, RANKS AND DEFINABILITY 31

(2) *φ is unstable in at least one power $\lambda \geq \aleph_0$.*

(3) *φ has the* order property; *i.e., there are $\bar{a}^n$, $n < \omega$ such that for every $k < \omega$ $\{\varphi(\bar{x}; \bar{a}^n)^{\mathrm{if}\,(k \leq n)}: n < \omega\}$ is consistent.*

(4) *For every $n < \omega$, $\Gamma(\varphi, m, n)$ is consistent, where*

$$\Gamma(\varphi, m, \alpha) = \{\varphi(\bar{x}_\eta; \bar{y}_{\eta \restriction \beta})^{\eta[\beta]}: \eta \in {}^{\alpha}2, \beta < \alpha\}$$

(usually we omit the m).

(5) *$\Gamma(\varphi, \alpha)$ is consistent for every ordinal α.*

(6) *$R^m(\bar{x} = \bar{x}, \varphi, \infty) = \infty$.*

(7) *$R^m(\bar{x} = \bar{x}, \varphi, 2) \geq \omega$.*

(8) *There are $A, p, p \in S^m_\varphi(A)$, such that p is not A-*definable *(see Definition 2.1).*

(9) *There is no ψ such that for every A, $|A| \geq 2$, and $p \in S^m_\varphi(A)$, p is (ψ, A)-*definable *(see Definition 2.1).*

Figure 3: The original statement of Shelah's unstable formula theorem as it appears in [She90].

Remark 2.24 Stability of a formula essentially means that its parameters cannot be used to define a linear order on arbitrarily large subsets of the domain. This notion has a natural interpretation in machine learning. The k-order property for a formula corresponds to the ability of a concept class to express k different threshold functions. In this context, the Littlestone dimension of a concept class can be understood as a measure of how many "threshold-like" concepts the class can contain.

The Shelah 2-rank, denoted as $R(\overline{x} = \overline{x}, \varphi, 2)$ in Shelah's original formulation, is related to the notion of stability by the Unstable Formula Theorem. This theorem, as shown in Figure 3, provides several equivalent characterizations of stability. Of particular interest to us are conditions (1) and (3) from Shelah's theorem.

Fact 2.25 (The Unstable Formula Theorem 2.2 in [She90]) For a formula φ, the following properties are equivalent:

- φ is stable,
- $R^{\varphi}(p) < \infty$ for any finite φ-type p.

The proof of this theorem is quite lengthy, and draws from both model theory and combinatorics, introducing Stone spaces of complete n-types and going back to Ramsey and Erdős-Makkai theorem (so we omit it). Nonetheless this theorem allows us to show that stable theories are a subset of NIP theories, discussed in Section 1.3.

Theorem 2.6 *Stable theories are NIP.*

Proof Let T be a stable theory. Consider an arbitrary formula φ in T. Since T is stable, φ is stable. We can establish that φ is NIP through the following chain of implications:

1) The Shelah 2-rank of φ is finite (Theorem 2.25).

2) Finite Shelah 2-rank implies that the concept class $\mathcal{C}_\varphi$ has finite Littlestone dimension (Theorem 2.5).

2) Finite Littlestone dimension of $\mathcal{C}_\varphi$ implies finite VC dimension of $\mathcal{C}_\varphi$ (Theorem 2.10).

3) Finite VC dimension of $\mathcal{C}_\varphi$ implies φ is NIP (Theorem 1.7).

Following this chain of implications, we conclude that φ is NIP. As φ was arbitrary, this holds for all formulas in T, therefore T is NIP.

Remark 2.26 Given a class $\mathcal{C}_\varphi$, there is a natural bipartite graph G_φ associated with any concept class. The node set of G_φ consists of two disjoint parts: the elements of the input set X and the concepts in $\mathcal{C}_\varphi$. We associate each concept $f \in \mathcal{C}_\varphi$ with the parameter set b_j defining it. The edge relation in G_φ is then defined by φ. Specifically, there is an edge between $a_i \in X$ and $b_j \in \mathcal{C}_\varphi$ if and only if $\varphi(a_i; b_j)$ is true (or $f(a_i) = 1$ if we view it as function in $\mathcal{C}_\varphi$). This corresponds to the k-order-property we defined in Theorem 2.23. The Littlestone dimension of $\mathcal{C}_\varphi$ is finite if and only if there exists an upper bound on the size of any half-graph that appears as an induced subgraph of G_φ.

We will now examine a number of examples of stable theories. They form a subclass of NIP theories and, as shown in Section 1.2, every formula φ in a stable theory has finite VC dimension. Consequently, it defines a PAC learnable concept class $\mathcal{C}_\varphi$ with finite Littlestone dimension.

The strength of our approach is that it draws on the extensive work of model theorists over many decades. By linking stability to learnability, we gain access to a rich collection of examples that are known to be stable and thus learnable. Without model theory, direct proof of these properties is very difficult and intractable.

Our exposition follows Section 5 in [CF19]. For readers interested in a deeper exploration of these connections, we recommend consulting standard textbooks on model theory for more detailed treatments.

Example 2.27 The theory of algebraically closed fields (ACF) is formulated in the ring signature $\mathcal{L}_r = \{+, -, \cdot, 0, 1\}$. It consists of the standard field axioms along with an infinite sequence of axioms, one for each positive integer n, stating that every polynomial of degree n has a root:

$$\varphi_n = \forall y_0 \ldots \forall y_{n-1} \exists x : x^n + \sum_{i=0}^{n-1} y_i x^i = 0.$$

- ACF can be further specialized to describe fields of specific characteristic:
  - For a prime p, ACF_p is the theory of algebraically closed fields of characteristic p. It includes the additional axiom $\psi_p = \forall x : px = 0$, where px denotes x added to itself p times.
  - For characteristic 0, ACF_0 is the theory of algebraically closed fields of characteristic 0. It includes the negation of ψ_p for all primes p, i.e., $\neg\psi_p = \exists x : px \neq 0$ for every prime p.

Both ACF_0 and ACF_p admit quantifier elimination. This allows any formula φ in these theories to be expressed as a quantifier-free formula, equivalent to a Boolean combination of polynomial equations and inequations.

In the context of learning theory, for any formula φ in ACF, the corresponding concept class $\mathcal{C}_\varphi$ is the solution set of this Boolean combination. In model-theoretic terms, this is a definable set (with parameters), while algebraic geometers refer to it as a constructible set. We can view this set as a subset of $\mathbb{A}^n$, where $\mathbb{A}$ is any algebraically closed field of the appropriate characteristic, and n is the number of free variables in φ.

A concrete example of a concept class in this theory is the family of elliptic curves over the complex numbers, defined by the formula $\varphi(x, y; a, b) = y^2 = x^3 + ax + b$, where a and b are parameters and x and y are variables. Given a set of points in $\mathcal{C}$, where each point is labeled based on whether it lies on a target elliptic curve, we can learn an approximation of that curve with high probability. The PAC learning guarantee ensures that with high probability, our learned curve will correctly classify most new points drawn from the same distribution as our training data.

Example 2.28 The second example is related to differential Galois theory. The theory of differentially closed fields in characteristic 0 (DCF_0) is formulated in the ring signature $\mathcal{L}_r$ together with a unary function symbol D dor the derivation. DCF_0 consists of the axioms for ACF_0, along with two axioms stating that D is an additive homomorphism satisfying the product rule and one additional axiom:

- $\forall x, y : yD(x + y) = D(x) + D(y)$
- $\forall x, y : yD(xy) = xD(y) + yD(x)$
- For any non-constant differential polynomials $p_1(x)$ and $p_2(x)$ of order $n_1 > n_2$ there is an x such that $f(x) = 0$ and $g(x) \neq 0$.

This last axiom is particularly important as it ensures that every consistent system of differential equations has a solution in the field. DCF_0 admits quantifier elimination, allowing any formula φ in this theory to be expressed as a quantifier-free formula, equivalent to a Boolean combination of differential polynomial equations and inequations.

A concrete example of a concept class in this theory is the family of solutions to linear differential equations, defined by the formula $\varphi(y; a, b) = D(y) = ay + b$, where a and b are parameters and y is a variable. Given a set of points in a differential field (e.g., the field of germs of meromorphic functions), where each point is labeled based on whether it satisfies a target linear differential equation, we can learn an approximation of that equation with high probability.

Example 2.29 The third example comes from group theory. The elementary theory of a non-abelian free group T_{fg}, is formulated in the group signature $\mathcal{L}_g = \{\cdot, {}^{-1}, 1\}$ and was shown to be stable by Zlil Sela in 2006. This theory cannot be axiomatized by first-order axioms, since the ultraproduct of free groups is not free. While this theory does not admit quantifier elimination, any formula in the language of groups is, modulo the theory of the free group, equivalent to a $\forall\exists$-formula.

A concrete example of a concept class in this theory could be the set of elements satisfying a certain word equation. For instance, we might consider $\varphi(x; a, b) = x = aba^{-1}b^{-1}$, where a and b are parameters and x is a variable. Given a set of elements in a free group, the stability of the theory guarantees that we can PAC learn the parameters a and b in this formula.

3 Further Research

In this thesis, we've proven the equivalence between finite VC dimension and NIP theories, finite Littlestone dimension and stable theories. This road opens up multiple ways to build upon these equivalences.

- The reliance of both the Littlestone dimension and the VC dimension on shatter functions, which are themselves based on different forms of the Sauer-Shelah lemma, raises a question: Is there a unifying principle that encompasses both concepts? This question is partially answered by Roland Walker in paper "Tree Dimension and the Sauer-Shelah Dichotomy" by defining a new tree dimension invariant called *leveled tree dimension* to measure the complexity of leaf sets in binary trees. arXiv:2203.12211v2, 2022
- The generalization to higher-arity trees has been explored by Hunter Chase and James Freitag in their paper "Model theory and combinatorics of banned sequences". In this work, they study 2^s-ary trees and apply it to the notion of model-theoretic op_s-rank introduced by Guingona and Hill. When $s = 1$, this generalization recovers the original Shelah's 2-rank. arXiv:1801.07640, 2018
- In the same paper, they address a question arising from the relationship between Littlestone dimension and VC dimension. Given that finite Littlestone dimension implies finite VC dimension, they investigate whether this stronger condition can lead to strengthening of the fundamental theorem of PAC learning. The authors provide a partial answer to this question by adapting the VC theorem to the context of finite Littlestone dimension. Their key contribution is showing that under these stronger assumptions, the VC theorem can be modified to allow for sampled elements to depend on the results of previous samples, in contrast to the independent sampling required in the standard VC theorem.
- A connection to o-minimality comes from the fact that o-minimal theories are NIP. This can be used to show that if the activation functions of a neural network are definable in an o-minimal expansion of the real numbers (such as $\mathbb{R}_{\text{an,exp}}$, which includes analytic and exponential functions), then the hypothesis class computed by the network has finite VC dimension. This theoretical result has practical implications: it ensures that, given enough training samples, such neural networks can learn to approximate any concept to the best possible representation within their hypothesis class. A comprehensive reference on this topic is the textbook "Neural Network Learning: Theoretical Foundations" by Anthony and Bartlett.
- A more recent application of this concept was done by D'Inverno et al. in their paper "VC dimension of Graph Neural Networks with Pfaffian

activation functions". Their work extends this analysis to common GNN architectures. The authors derive upper bounds on the VC dimension in terms of key architectural parameters like the number of layers, hidden feature size, and input dimension. They show that for GNNs with Pfaffian activations, the VC dimension grows as $O(p^4)$, where p is the number of parameters. Theoretical results are supported by experiments measuring the gap between training and test accuracy as network size increases. arXiv:2401.12362v2, 2024

- The VC dimension is closely related to the concept of compression schemes. A classic example is compressing rectangles in $\mathbb{R}^2$ to just four points – from an arbitrarily large set of labeled samples, one can select the four outermost positively labeled points and discard the rest. This compression retains all necessary information to recreate a labeling consistent with the full original sample, akin to the notion of sufficient statistics in probability and statistical theory. Just as sufficient statistics capture all relevant information about a parameter in a probabilistic model, these compression schemes encapsulate the essential information needed for learning. The now disproven Warmuth conjecture stated that every concept class of VC dimension at most d admits a compression scheme of size at most d. Moran and Yehudayoff in "Sample compression schemes for VC classes" proved that there exists a compression scheme whose size is exponential in the VC dimension, and Pálvölgyi and Tardos in their paper "Unlabeled Compression Schemes Exceeding the VC-dimension" disproved this conjecture by constructing an explicit counterexample. arXiv:1503.06960v2, 2015 arXiv:1811.12471v2, 2018

- The uniform definability of types over finite sets (UDTFS) conjecture is another important model-theoretic conjecture that was recently proven by Eshel and Kaplan in their paper "On uniform definability of types over finite sets for NIP formulas". This conjecture states that a formula φ has the non-independence property (NIP) in a theory T if and only if it has UDTFS in T. More specifically, UDTFS means that for any NIP formula φ, there exists another formula φ that can uniformly define all possible φ-types over any finite set of parameters. What's particularly interesting about the proof of this conjecture is that it combined two previously known results from machine learning theory. This is somewhat unusual, as typically model theory techniques are applied to machine learning problems rather than the reverse. arXiv:1904.10336v2, 2020

- Last but certainly not least, the model theorists strike again. In paper "The Unstable Formula Theorem revisited via algorithms" Malliaris and Moran focused on developing a complete algorithmic analog of the Unstable Formula Theorem. It summarizes all previous work on the subject and unifies it into one common framework. This provides a much clearer arXiv:2212.05050v2, 2023 picture of the connections between model theory and machine learning. Furthermore, they also propose a new approach to online learning called *probably eventually correct* learning (PEC) where the main difference between PEC and PAC learning is

- Error rate:
    - PAC learning: The output hypothesis has low (but potentially non-zero) error with high probability.
    - PEC learning: The output hypothesis eventually has zero error (up to measure zero) with probability 1.
- Sample complexity:
    - PAC learning: Requires a finite sample size to achieve the desired error/confidence.
    - PEC learning: Allows the sample size to be unbounded, only requiring the learner to eventually converge to the correct hypothesis.
- Stability:
    - PAC learning: No explicit stability requirement on the output hypotheses.
    - PEC learning: Requires *stable* learning in the sense that the algorithm changes its output hypothesis only a bounded number of times.
- Algorithm:
    - PAC learning: No specific canonical algorithm.
    - PEC learning: The Standard Optimal Algorithm (SOA) is shown to be a PEC learner for Littlestone classes.

References

[Bha21] Siddharth Bhaskar. "Thicket density". In: *J. Symb. Log.* 86.1 (2021), pp. 110–127.

[Blu+89] Anselm Blumer et al. "Learnability and the Vapnik-Chervonenkis dimension". In: *Journal of the ACM (JACM)* 36.4 (1989), pp. 929–965.

[CF19] Hunter Chase and James Freitag. "Model theory and machine learning". In: *Bulletin of Symbolic Logic* 25.3 (2019), pp. 319–332.

[Che16] Artem Chernikov. *Topics in combinatorics.* `https://www.math.ucla.edu/~chernikov/teaching/Combinatorics285N/CombinatoricsNotes.pdf`. 2016.

[Gui13] Vincent Guingona. *NIP Theories and Computational Learning Theory.* `https://tigerweb.towson.edu/vguingona/NIPTCLT.pdf`. 2013.

[Hod97] Wilfrid Hodges. *A shorter model theory.* Cambridge University Press, Cambridge, 1997.

[Las92] Michael C. Laskowski. "Vapnik-Chervonenkis classes of definable sets". In: *J. London Math. Soc. (2)* 45.2 (1992), pp. 377–384.

[Lit88] Nick Littlestone. "Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm". In: *Machine learning* 2 (1988), pp. 285–318.

[She90] Saharon Shelah. *Classification Theory and the number of non-isomorphic models.* North-Holland, 1978. Revised edition, 1990.

Deutsche Zusammenfassung / German summary

Die vorliegende Arbeit verfolgt das Ziel, die Zusammenhänge zwischen Modelltheorie und maschinellem Lernen an der Schnittstelle zwischen mathematischer Logik und Informatik zu beschreiben. Als Grundlage wurde der Artikel "Model Theory and Machine Learning" von Hunter und Chase verwendet, [CF19].

$$\begin{array}{ccc} \{\text{Mathematische Logik}\} & & \{\text{Informatik}\} \\ \bigcup & & \bigcup \\ \{\text{Modelltheorie}\} & \bigcap & \{\text{Maschinelles Lernen}\} \end{array}$$

Wir konzentrieren uns dabei auf zwei zentrale Themenbereiche: Zum einen untersuchen wir die Verbindung zwischen PAC-Lernbarkeit und der NIP-Eigenschaft, zum anderen betrachten wir den Zusammenhang zwischen Online-Lernbarkeit und Stabilität. Beide Themen haben ihre Wurzeln in der Kombinatorik, ein Thema, das uns in den Beweisen immer wieder begleiten wird. Die Arbeit besteht aus zwei Hauptteilen:

Im ersten Teil beweisen wir das fundamentale Theorem der PAC-Lernbarkeit. Dieses besagt, dass eine Konzeptklasse genau dann PAC-lernbar ist, wenn sie eine endliche VC-Dimension hat. Auf der modelltheoretischen Seite wird anschließend die NIP-Eigenschaft eingeführt und gezeigt, dass Formeln in NIP-Theorien den Konzeptklassen mit endlicher VC-Dimension entsprechen.

Der zweite Teil widmet sich der Online-Lernbarkeit und der Littlestone-Dimension als Maß für die Komplexität einer Konzeptklasse. Wir stellen den Standard-Optimal-Algorithm als optimale Strategie für das Online-Lernen vor. Auf der modelltheoretischen Seite wird die Äquivalenz von endlicher Littlestone-Dimension und Shelahs 2-Rang, einem wichtigen Konzept aus der Modelltheorie, bewiesen. In Folge demonstrieren wir, dass Formeln in stabilen Theorien den Konzeptklassen mit endlicher Littlestone-Dimension entsprechen.

$$\begin{array}{ccc} \text{endliche VC-Dimension} & \longleftrightarrow & \text{NIP Theorie} \\ \bigcup & & \bigcup \\ \text{endliche Littlestone-Dimension} & \longleftrightarrow & \text{Stabile Theorie} \end{array}$$

Abschließend befassen wir uns mit der Stabilitätstheorie und zeigen, dass stabile Theorien eine Untermenge von NIP-Theorien bilden. Zur Veranschaulichung präsentieren wir mehrere Beispiele für stabile Theorien. Dazu gehören die Theorie algebraisch geschlossener Felder, die Theorie differenziell geschlossener Felder und die elementare Theorie nicht-abelscher freier Gruppen. Diese Beispiele demonstrieren, wie die Modelltheorie eine Vielzahl konkreter, PAC-lernbarer Konzeptklassen zum Vorschein bringt.