6.962 Week 7 Summary: More Learning Theory: Characterizing Uniform Glivenko-Cantelli Classes using Scale-Sensitive Dimension

Michael Rosenblum

March 22, 2001

1. Abstract

I present a recent paper on learning theory by Alon, Ben-David, Cesa-Bianchi and Haussler [1997]. The main result is an exact characterization of the classes of real-valued loss functions $\mathcal{F}$ which are learnable in the probabilistic concept model. This characterization is purely combinatorial, "easy to check," and is a direct extension of VC-dimension to real-valued functions.

2. Introduction

A main problem in statistical learning theory is to estimate an unknown distribution from a finite number of i.i.d. samples. In the situation we will study, we are given:
• i.i.d. samples $x_1, ..., x_n$ drawn from an unknown distribution P,
• a class of hypothesis functions $\mathcal{G}$ taking values in $[0, 1]$,
• a class of loss functions $\mathcal{F}$ derived from $\mathcal{G}$ taking values in $[0, 1]$.
A learning algorithm is defined to be a mapping from observed finite sequences $x_1, ..., x_n$ to a hypothesis $g_n \in \mathcal{G}$. Let f_n be the loss function associated with g_n.

We'd like to know if g_n will generalize well. In other words, will the expected risk of f_n (denoted $P(f_n)$) be close to the minimum possible expected risk for all $f \in \mathcal{F}$?

Formally, will we have

$$P(f_n) \xrightarrow{P} \inf_{f \in \mathcal{F}} P(f)$$

as $n \to \infty$?

We can consider the following two general learning problems:
Learning Problem 1: Does there exist a learning algorithm that satisfies:

$$\lim_{n\to\infty} \sup_P Pr\left(|P(f_n) - \inf_{f\in\mathcal{F}} P(f)| > \epsilon\right) = 0$$

A related problem involves empirical risk: For distribution P, and loss function $f \in \mathcal{F}$, the empirical risk is defined as $P_n(f) := \frac{1}{n}\sum_{i=1}^{n} f(x_i)$.
Learning Problem 2: Does there exist a learning algorithm that satisfies:

$$\lim_{n\to\infty} \sup_P Pr\left(|P_n(f_n) - \inf_{f\in\mathcal{F}} P(f)| > \epsilon\right) = 0$$

The focus of this presentation will be characterizing learnability in the probabilistic concept model. Probabilistic concept learning has the following setup:

> Let each data point x be a pair (y, z), with $y \in \mathbf{Y}, z \in \{0, 1\}$, drawn i.i.d. from unknown distribution P on $\mathbf{Y} \times \{0, 1\}$.
> Goal: minimize mean squared error.
> Loss functions will have the form $f(y, z) := (z - g(y))^2$,
> for $g \in \mathcal{G}$, a class of estimators of z given y.

For pattern recognition problems, $\mathcal{F}$ is a class of indicator functions. For regression problems, $f \in \mathcal{F}$ will be real-valued. We'll restrict to the probabilistic concept learning setting throughout the talk, and with loss functions $f \in \mathcal{F}$ taking values in $[0, 1]$. Most of the proofs, however, hold in greater generality.

In section 4, Alon et al. show that learnability in the probabilistic concept model is equivalent to the uniform Glivenko-Cantelli property. A few weeks ago, Emin presented an "easy to check" characterization of uniform Glivenko-Cantelli classes of *indicator* functions. He argued that a class of indicator functions $\mathcal{F}$ is uniform Glivenko-Cantelli iff the corresponding growth function $H_n(\mathcal{F})$ grew slowly as the sample size n increased (i.e. $H_n(\mathcal{F})$ was $o(n)$). Then Emin gave an "easy to check" purely combinatorial criterion for deciding if the growth function was $o(n)$: $\mathcal{F}$ having finite VC-dimension. This is great if one is concerned with pattern recognition, where the class of loss functions corresponding to a class $\mathcal{G}$ of estimators takes values in $\{0,1\}$. However, for problems such as regression we'll need to deal with more general loss functions. In this presentation we use the probabilistic concept learning model described above, which involves classes $\mathcal{F}$ of real-valued loss functions.

Alon et al. extend the results for indicator functions to a more general scenario, where the functions $f \in \mathcal{F}$ take on real values. They assume that the functions are uniformly bounded, i.e. there exists a constant C s.t. $\forall f \in \mathcal{F}, \forall x \in X, |f(x)| \leq C$. For simplicity, let's assume the range of all our functions is the interval $[0,1]$. Unfortunately, we cannot directly apply the above machinery used on classes of indicator functions, since the definitions of growth function and VC-dimension only make sense for indicator functions. We need to generalize these concepts and find new proofs to characterize probabilistic concept learning for classes of $[0,1]$-valued functions. That is the focus of the Alon, Ben-David, Cesa-Bianchi and Haussler paper [1997].

The two problems to tackle are:
1. Generalize the definition of growth function $H_n(\mathcal{F})$ to classes of real-valued functions, and show it grows "slowly" (i.e. it is $o(n)$) iff $\mathcal{F}$ is a uniform Glivenko-Cantelli class.
2. Generalize the idea of VC-dimension to real-valued functions and show it is finite iff the growth function is $o(n)$.
At the end, we'll even get bounds on the rate of uniform 2-sided convergence of empirical means $(P_n(f) := \frac{1}{n}\sum_{i=1}^{n} f(x_i))$ to actual means $P(f) := \int f dP$ in terms of a scale parameter γ.

3. Probabilistic Concept Learning, and a Refresher on uniform Glivenko-Cantelli classes, growth functions, and VC-dimension for indicator functions

Let's look at how probabilistic concept learnability, uniform Glivenko-Cantelli classes, growth functions, and VC-dimension are defined for indicator functions, since we'll soon be generalizing these ideas to real-valued functions.

3.1 Probabilistic Concept Learning
A class of functions is called learnable in the probabilistic concept model if there exists a single learning algorithm that is uniformly "good" over all probability distributions.

Definition 1 *$\mathcal{F}$ is ϵ-learnable in the probabilistic-concept model*[1] *if there exists a learning algorithm that maps finite samples $x_1, ..., x_n$ onto loss functions $f_n \in \mathcal{F}$ and such that*

$$\lim_{n\to\infty} \sup_P Pr\left(P(f_n) > \inf_{f\in\mathcal{F}} P(f) + \epsilon\right) = 0.$$

The outer supremum is taken with respect to all probability measures P defined on a suitable σ-algebra $\mathcal{A}$. $P(f)$ denotes the expectation of f with respect to the distribution P. We say that $\mathcal{F}$ is learnable for the probabilistic concept model if it is ϵ-learnable for all $\epsilon > 0$.

3.2 Uniform Glivenko-Cantelli classes

Uniform Glivenko-Cantelli classes capture the notion of 2-sided convergence of empirical means to actual means, uniformly over all probability distributions and all functions in a class $\mathcal{F}$.

Definition 2 *We call a class of functions $\mathcal{F}$ an ϵ-uniform Glivenko-Cantelli class if*[1]

$$\lim_{n\to\infty} \sup_P Pr\left(\sup_{m\geq n} \sup_{f\in\mathcal{F}} |P_m(f) - P(f)| > \epsilon\right) = 0.$$

The outer supremum is taken with respect to all probability measures P defined on $\mathcal{A}$. The empirical mean of f, denoted $P_m(f)$, is defined as $\frac{1}{m}\sum_{i=1}^{m} f(x_i)$. We say that $\mathcal{F}$ is a uniform Glivenko-Cantelli class if it is an ϵ-uniform Glivenko-Cantelli class for all $\epsilon > 0$.

3.3 Growth Function for indicator functions

For a sequence of n points $\vec{x}^n := (x_1, x_2, ..., x_n) \subseteq X$, let $\mathcal{N}(\mathcal{F}, \vec{x}^n)$ be the number of distinct vectors of the form $(f(x_1), ..., f(x_n))$ for $f \in \mathcal{F}$. In other words, $\mathcal{N}(\mathcal{F}, \vec{x}^n)$ is the number of distinct separations of this sample that can be obtained using functions from $\mathcal{F}$. Now, we can define the growth function for $\mathcal{F}$:

Definition 3 *Growth Function for classes $\mathcal{F}$ of indicator functions.*
The growth function $H_n(\mathcal{F}) := \log \sup_{\vec{x}^n} \mathcal{N}(\mathcal{F}, \vec{x}^n)$.

3.4 Vapnik-Chervonenkis Dimension for indicator functions

Vapnik-Chervonenkis Dimension (VC-dimension) measures the "capacity" of a class of functions.

[1] Measurability concerns are ignored in the Alon et al. paper. I will treat measurability issues in footnotes to the relevant sections. If, for example, we assume that $\mathcal{F}$ is image-admissible Suslin, then the quantities we are examining are all measurable. [see Dudley 1999 5.3.5 pp. 188].

Definition 4 *VC-dimension for classes $\mathcal{F}$ of indicator functions.*
We say that a class $\mathcal{F}$ of indicator functions with domain X shatters a finite set $A \subseteq X$ if for every $E \subseteq A$, there is an $f \in \mathcal{F}$ such that $f(x) = 1$ for $x \in E$ and $f(x) = 0$ for $x \in A \setminus E$. The VC-dimension of such a class is the maximal cardinality of a set $A \subseteq X$ that is shattered by $\mathcal{F}$. (If $\mathcal{F}$ shatters sets of unbounded finite sizes, then let its VC-dimension $= \infty$.)

Example 1
To get some intuition for how these quantities are related, let's look at an example. Take our probability space to be $\mathbf{Y} \times \mathbf{Z} := \mathbf{N} \times \{0, 1\}$. Let our class of estimators $\mathcal{G}$ be the class of all indicator functions g of the form:

$$g(y) := \begin{cases} b & \text{if } y < t \\ 1 - b & \text{if } y \geq t \end{cases}$$

for some $t \leq \infty$, and some $b \in \{0, 1\}$. Let $\mathcal{F}$ be the class of loss functions associated with $\mathcal{G}$, where $\mathcal{F} := \{(z - g(y))^2 | g \in \mathcal{G}\}$.

It is easy to see that $\mathcal{F}$ has VC-dimension equal to two. One can shatter any two elements $(y_1, z_1), (y_2, z_2)$ with $y_1 \neq y_2$ with the functions $\mathcal{F}$, but it is impossible to shatter any set of three or more elements since no function in $\mathcal{G}$ changes values from b to $1 - b$ then back again.

Next, the growth function $H_n(\mathcal{F})$ is not hard to calculate exactly. Since for any set $\vec{x}^n := \{(y_1, z_1), ..., (y_n, z_n)\}$ with distinct y_i, we get $\mathcal{N}(\mathcal{F}, \vec{x}^n) = 2n$, we have $H_n(\mathcal{F}) = \log 2n$. Note that $H_n(\mathcal{F})/n \to 0$ as $n \to \infty$. So $H_n(\mathcal{F})$ is $o(n)$.

Thirdly, by the theorems we will prove shortly, we have that $\mathcal{F}$ is a uniform Glivenko-Cantelli class, and $\mathcal{F}$ is learnable in the probabilistic concept model. But it is not obvious how to show this directly, even for this relatively simple problem. Thus, it is convenient to have an "easy to check" criterion, such as finite VC-dimension, when analyzing learnability for a particular class of functions.

Example 2
Let's look at another example. We'll use the same setup as in example 1, except this time we let $\mathcal{G}$ be the class of *all* indicator functions g with domain $\mathbf{N}$.

Since we can shatter finite sets of arbitrary sizes, $\mathcal{F}$ has infinite VC-dimension.

Next, since for any set $\vec{x}^n := \{(y_1, z_1), ..., (y_n, z_n)\}$ with distinct y_i, we get $\mathcal{N}(\mathcal{F}, \vec{x}^n) = 2^n$, we have $H_n(\mathcal{F}) = n \log 2$. Note that $H_n(\mathcal{F})/n \to \log 2$ as $n \to \infty$. So $H_n(\mathcal{F})$ is not $o(n)$.

Lastly, by the theorems we will prove shortly, we have that $\mathcal{F}$ is NOT a uniform

Glivenko-Cantelli class. Also we have that $\mathcal{F}$ is NOT learnable in the probabilistic concept model.

Now we're ready to generalize probabilistic concept learnability, uniform Glivenko-Cantelli classes, growth functions, and VC-dimension to classes $\mathcal{F}$ of real-valued functions. Since the first two definitions generalize without any modification, we'll start by generalizing growth functions, and next look at generalizing VC-dimension.

4. Scale Sensitive Packing Numbers and Metric Entropy

Covering and Packing Numbers; Metric Entropy
For a sequence of n points $\vec{x}^n := (x_1, x_2, ..., x_n) \subseteq X$, and $f, g \in \mathcal{F}$, define the $l^\infty_{\vec{x}^n}$-pseudo-metric by $l^\infty_{\vec{x}^n}(f,g) := \max_{1\le i\le n} |f(x_i) - g(x_i)|$. Let $\mathcal{N}(\epsilon, \mathcal{F}, \vec{x}^n)$ be the ϵ-covering number of $\mathcal{F}$ with respect to $l^\infty_{\vec{x}^n}$, that is, the minimum cardinality of a set $G \subseteq \mathcal{F}$ such that for all $f \in \mathcal{F}$ there is a $g \in G$ where $l^\infty_{\vec{x}^n}(f,g) < \epsilon$. Such a finite set G always exists, since the set of points $\vec{x}^n \in [0,1]^n$ is totally bounded for the metric $l'(\vec{x}^n, \vec{y}^n) := \max_{1\le i\le n} |x_i - y_i|$. We call a minimal such set G an ϵ-net for $\mathcal{F}$. Define the metric entropy $H_n(\epsilon, \mathcal{F}) := \log \sup_{\vec{x}^n} \mathcal{N}(\epsilon, \mathcal{F}, \vec{x}^n)$.

A related quantity is the ϵ-packing number, denoted $\mathcal{M}(\epsilon, \mathcal{F}, \vec{x}^n)$. Call a set $H \subseteq \mathcal{F}$ ϵ-separated if for all distinct $f, g \in H$, $l^\infty_{\vec{x}^n}(f,g) \ge \epsilon$. The ϵ-packing number $\mathcal{M}(\epsilon, \mathcal{F}, \vec{x}^n)$ is defined to be the maximum cardinality of a set $H \subseteq \mathcal{F}$ that is ϵ-separated. Later, we will use a consequence of the following easy to verify fact from Dudley [1.2.1. pp. 11] in the proof of Lemma 2 of the main theorem: $\mathcal{N}(\epsilon, \mathcal{F}, \vec{x}^n) \le \mathcal{M}(\epsilon, \mathcal{F}, \vec{x}^n)$.

The following theorem, proved earlier by Dudley et al., is a consequence of the main theorem proved in the Alon et al. paper:

Theorem [Dudley et al. 1991, Theorem 6, pp. 500]. Let $\mathcal{F}$ be a class of functions from X into $[0,1]$. Then [2]
(1) $\mathcal{F}$ is a uniform Glivenko-Cantelli class if and only if for all $\epsilon > 0$, $\lim_{n\to\infty} H_n(\epsilon, \mathcal{F})/n = 0$.
(2) For all $\epsilon > 0$, if $\lim_{n\to\infty} H_n(\epsilon, \mathcal{F})/n = 0$, then $\mathcal{F}$ is an (8ϵ)-uniform Glivenko-Cantelli class.

Example 3
Let $\mathcal{F}$ be the class of all $[0,1]$-valued functions f defined on $\mathbf{N}$ such that $f(n) \le 2^{-n}$ for all $n \in \mathbf{N}$. Assume $\epsilon < 1$.

[2] Under suitable measurability conditions, for example that $\mathcal{F}$ is image-admissible Suslin.

First, we can bound the ϵ-covering number for $\mathcal{F}$ from above. Note that for $n > \log \frac{1}{2\epsilon}$, we have $\forall f \in \mathcal{F} : f(n) < 2\epsilon$. Using this fact, we can show that the ϵ-covering number is at most $(\frac{2}{\epsilon})^{(\log \frac{1}{2\epsilon})+1}$. Note that this upper bound is independent of the number of sample points n. So for all $\epsilon > 0$, $\lim_{n\to\infty} H_n(\epsilon, \mathcal{F})/n = 0$ and we can apply the above theorem to conclude that $\mathcal{F}$ is a uniform Glivenko-Cantelli class.

5. Scale sensitive γ-dimension

One way to generalize VC-dimension to real-valued functions is the following definition due to Pollard [1990]:

Definition 5 *Pollard-dimension for classes of* $[0,1]$*-valued functions* $\mathcal{F}$.
We say that a class of $[0,1]$*-valued* functions $\mathcal{F}$ *on* X *shatters a finite set* $A \subseteq X$ *if there exists a function* $s : A \to \mathbf{R}$ *such that for every* $E \subseteq A$*, there is an* $f \in \mathcal{F}$ *where:* $\forall x \in A \backslash E, f(x) < s(x)$ *and* $\forall x \in E, f(x) \geq s(x)$*. The Pollard-dimension of* $\mathcal{F}$ *is the maximal cardinality of a set* $A \subseteq X$ *shattered by* $\mathcal{F}$*. (If* $\mathcal{F}$ *shatters sets of unbounded finite sizes, then let its Pollard-dimension* $= \infty$*.)*

Unfortunately, this extension of VC-dimension to classes of functions with range $[0,1]$ does not provide a characterization of the uniform Glivenko-Cantelli property for such classes of functions; I show in *Example 4* below that it provides only a sufficient condition for the uniform Glivenko-Cantelli property.

A slight modification of Definition 4 will yield a characterization of the uniform Glivenko-Cantelli property for classes of functions with range $[0,1]$. Let $\gamma > 0$. γ will be called a scale parameter, and we can define a notion of dimension with respect to this parameter:

Definition 6 *(Scale-Sensitive)* γ*-dimension for classes of* $[0,1]$*-valued functions* $\mathcal{F}$.
We say that a class of $[0,1]$*-valued* functions $\mathcal{F}$ *on* X γ*-shatters a finite set* $A \subseteq X$ *if there exists a function* $s : A \to \mathbf{R}$ *such that for every* $E \subseteq A$*, there is an* $f \in \mathcal{F}$ *where:* $\forall x \in A \backslash E, f(x) \leq s(x) - \gamma$ *and* $\forall x \in E, f(x) \geq s(x) + \gamma$*. In this case, we say* $\mathcal{F}$ γ*-shatters* A *according to* s*. The* γ*-dimension of* $\mathcal{F}$ *is the maximal cardinality of a set* $A \subseteq X$ γ*-shattered by* $\mathcal{F}$*. (If* $\mathcal{F}$ γ*-shatters sets of unbounded finite sizes, then let its* γ*-dimension* $= \infty$*.)*

We can now calculate the Pollard-dimension and γ-dimension of the class of functions from example 3:

Example 4
As in example 3, let $\mathcal{F}$ be the class of all $[0,1]$-valued functions f defined on $\mathbf{N}$ such that $f(n) \leq 2^{-n}$ for all $n \in \mathbf{N}$ and $f \in \mathcal{F}$.

First, $\mathcal{F}$ has infinite Pollard-dimension, since for any $n > 0$, setting $A := \{1, 2, .., n\}$, and $s(x) := \frac{2^{-n}}{2}$ for $x \in A$, we can find functions $f \in \mathcal{F}$ that shatter A. However, for all $\gamma > 0$, the γ-dimension of $\mathcal{F}$ is $\lfloor \log(\frac{1}{2\gamma}) \rfloor < \infty$.

According to the main theorem, to be proved below, $\mathcal{F}$ is a uniform Glivenko-Cantelli class. Thus the Pollard-dimension does not directly characterize the uniform Glivenko-Cantelli property; i.e. finite Pollard-dimension is not a necessary condition for a class of real-valued functions to be a uniform Glivenko-Cantelli class. It is, however, a sufficient condition, as follows from Theorem 1; we need only note that for all $\gamma > 0$, the γ-dimension of $\mathcal{F}$ is no more than the Pollard-dimension of $\mathcal{F}$.

6. The Main Theorem

The main theorem is a chain of equivalences, linking probabilistic concept learnability to finite γ-dimension.

Theorem 1 *Let $\mathcal{F}$ be a class of $[0,1]$-valued functions on X. Then*[3] *the following conditions are equivalent:*

(a) $\mathcal{F}$ is learnable in the probabilistic-concept model.

(b) $\mathcal{F}$ is a uniform Glivenko-Cantelli class.

(c) $\forall \epsilon > 0$, we have for the metric entropy: $\lim_{n\to\infty} H_n(\epsilon, \mathcal{F})/n = 0$.

(d) $\forall \gamma > 0$, $\mathcal{F}$ has finite γ-dimension.

Proving the above equivalence takes lots of work. Alon et al. show **d** $\Rightarrow$ **c**, **b** $\Rightarrow$ **d**, and cite a result of Vapnik to show **c** $\Rightarrow$ **b**, closing the circle. They then prove **a** $\iff$ **b**. Their main contribution is a combinatorial lemma (Claim 3' below) generalizing Sauer's Lemma [4], which is used to show **d** $\Rightarrow$ **c**. This is what I will present.

[3] Under suitable measurability conditions: see Dudley [1999 6.6 pp. 223-226].
[4] Sauer's lemma is proved by Dudley [1999 4.1.2, pg. 135].

7. Proof of $\mathbf{d} \Rightarrow \mathbf{c}$: How to Bound the 4γ-covering number by the γ-dimension

Our aim is to prove the following lemma, which gives $\mathbf{d} \Rightarrow \mathbf{c}$. The full proof is in the Alon et al. paper. I will give an outline, highlighting the main ideas.

Lemma 1 *For all $\gamma \in (0, \frac{1}{4})$, for $d :=$ the γ-dimension of $\mathcal{F}$, $d < \infty$, we have:*

$$\sup_{(x_1,\ldots,x_n)} \mathcal{N}(4\gamma, \mathcal{F}, \vec{x}^n) \leq 2\left(\frac{9n}{16\gamma^2}\right)^{1+d\log(3en/(4d\gamma))}$$

Proof Outline The proof idea is to first discretize the functions in $\mathcal{F}$, reducing the problem to one involving functions that take only a finite set of values. Next we solve this discrete problem using combinatorial tools.

It will be convenient to work with packing numbers $\mathcal{M}$ instead of covering numbers $\mathcal{N}$. It is not difficult to show *Claim 1:*

$$\sup_{(x_1,\ldots,x_n)} \mathcal{N}(4\gamma, \mathcal{F}, \vec{x}^n) \quad \leq \quad \sup_{(x_1,\ldots,x_n)} \mathcal{M}(4\gamma, \mathcal{F}, \vec{x}^n).$$

We now discretize our class of functions at "coarseness" ρ. For a function $f : X \to [0,1]$, and real number $\rho > 0$, the ρ-discretization of f, denoted f^ρ, is defined to be $f^\rho(x) := \lfloor f(x)/\rho \rfloor$. For a class of $[0,1]$-valued functions $\mathcal{F}$, let $\mathcal{F}^\rho := \{f^\rho : f \in \mathcal{F}\}$.

We'll first need the following two-part claim, which will let us reduce the problem of bounding the packing number of $\mathcal{F}$ to bounding the packing number of $\mathcal{F}^\rho$.

Claim 2: for every $\beta \geq 2\rho$ and every $\vec{x}^n \in X^n$, $\mathcal{M}(\beta, \mathcal{F}, \vec{x}^n) \leq \mathcal{M}(2, \mathcal{F}^\rho, \vec{x}^n)$.

Proof of Claim 2:
By definition of $\mathcal{F}^\rho$, for all $f, g \in \mathcal{F}$, and $x \in X$, $|f(x) - g(x)| \geq 2\rho$ implies $|f^\rho(x) - g^\rho(x)| \geq 2$. Thus, if for some $G \subseteq \mathcal{F}$, we have $\forall f, g \in G$, $\sup_{x \in \vec{x}^n} |f(x) - g(x)| \geq 2\rho$, then $\forall f', g' \in G^\rho$, $\sup_{x \in \vec{x}^n} |f'(x) - g'(x)| \geq 2$, completing the proof of the claim.

We can now continue the chain of inequalities from above using Claim 2:

$$\sup_{(x_1,\ldots,x_n)} \mathcal{M}(4\gamma, \mathcal{F}, \vec{x}^n) \quad \leq \quad \sup_{(x_1,\ldots,x_n)} \mathcal{M}(2, \mathcal{F}^{2\gamma}, \vec{x}^n).$$

Now that we have discretized our class of functions, our task is to bound the packing number on the previous line. The following is an extension of Sauer's lemma [5], and is the main contribution in the Alon et al. paper.

Claim 3': If $\mathcal{H}$ is a class of functions from a finite domain X of cardinality n to a finite range $B := \{1, 2, ..., b\}$, and the 1-dimension of $\mathcal{H}$ is $c < \infty$, then $\mathcal{M}(2, \mathcal{H}, X) < 2(nb^2)^{\lceil \log y \rceil}$, where $y := \sum_{i=1}^{c} \binom{n}{i} b^i$.

The proof of Claim 3' is long but not difficult, and relies on repeated uses of the pigeon-hole principle.

If we take $b = 3$, we need only consider functions with range $\{1, 3\}$, which can be thought of instead as indicator functions. Here, the 1-dimension of $\mathcal{H}$ reduces to the VC-dimension of the corresponding class of indicator functions, and we have for the packing numbers: $\mathcal{M}(2, \mathcal{H}, X)$ reduces to $\mathcal{N}(\mathcal{H}, X)$. In this special case, Sauer's lemma bounds $\mathcal{N}(\mathcal{H}, X)$ in terms of the VC-dimension:

Sauer's Lemma: For $\mathcal{H}$ a set of indicator functions with finite VC-dimension c, we have $\mathcal{N}(\mathcal{H}, X) \leq \begin{pmatrix} n \\ 0 \end{pmatrix} + \begin{pmatrix} n \\ 1 \end{pmatrix} + ... + \begin{pmatrix} n \\ c \end{pmatrix}$.

Sauer's lemma was the key combinatorial result used in the proof that for classes of indicator functions $\mathcal{F}$, the growth function is $o(n)$ iff $\mathcal{F}$ has finite VC-dimension. Claim 3' generalizes this lemma to functions that take values in $\{1, ..., b\}$.

We need just one more claim, so that we can bound the 1-dimension of $\mathcal{F}^{2\gamma}$ in terms of the γ-dimension of $\mathcal{F}$.

Claim 4: For any $\rho > 0$, for every $\eta \leq \rho/2$, the 1-dimension of $\mathcal{F}^{\rho}$ is no more than the η-dimension of $\mathcal{F}$,

Proof of Claim 4:
We will show that any set 1-shattered by $\mathcal{F}^{\rho}$ is also $\rho/2$-shattered by $\mathcal{F}$. If $A \subseteq X$ is 1-shattered by $\mathcal{F}^{\rho}$, then by definition there exists a function $s : A \to \mathbf{R}$ such that for every $E \subseteq A$, there is an $f \in \mathcal{F}$ where: $\forall x \in A \backslash E, f^{\rho}(x) \leq s(x) - 1$ and $\forall x \in E, f^{\rho}(x) \geq s(x) + 1$. Define $s' : A \to \mathbf{R}$ by $s'(x) := (s(x) + 1/2)\rho$. Now, if $x \in A \backslash E$, then $f^{\rho}(x) \leq s(x) - 1$, so that $f(x) \leq \rho s(x) = s'(x) - \rho/2$. If $x \in E$, then $f^{\rho}(x) \geq s(x) + 1$, so that $f(x) \geq \rho s(x) + \rho = s'(x) + \rho/2$. Thus, A is $\rho/2$-shattered by $\mathcal{F}$, as evidenced by s', and so A is also η-shattered for all $\eta \leq \rho/2$.

Given that Claim 3' is true, we can combine it with Claim 4 to prove Lemma 1 above.

[5] [Sauer's lemma is proved in Dudley 1999 4.1.2 pp. 135]

By Claim 4, we have that the 1-dimension of $\mathcal{F}^{2\gamma}$ is no more than d, the γ-dimension of $\mathcal{F}$. Setting $b := \lfloor \frac{1}{2\gamma} \rfloor + 1$, and using Stirling's approximation and a proposition from Dudley [1999 4.1.5 pp. 136] to bound $y := \sum_{i=1}^{d} \binom{n}{i} b^i$, we can apply Claim 3' to get

$$\sup_{(x_1,\ldots,x_n)} \mathcal{M}(2, \mathcal{F}^{2\gamma}, \vec{x}^n) \leq 2\left(\frac{9n}{16\gamma^2}\right)^{1+d\log(3en/(4d\gamma))}$$

which yields Lemma 1.

8. For the Over-Enthusiastic: Proof of **b**$\Rightarrow$ **d**

It is not that difficult to show **b** $\Rightarrow$ **d**. The idea is that if the γ-dimension of $\mathcal{F}$ is infinite, we can find probability distributions that cause the worst case convergence of functions in $\mathcal{F}$ to occur arbitrarily slowly. This proof is very similar to the proof for $\mathcal{F}$ a class of indicator functions [Vapnik 1998. 4.9.3. pg. 149-150]. **b** $\Rightarrow$ **d** will follow directly from the following claim:

Claim 5: If the γ-dimension of $\mathcal{F}$ is infinite, then $\mathcal{F}$ is not a $(2\gamma-\tau)$-uniform Glivenko-Cantelli class for any $\tau > 0$.

Proof of Claim 5: If for some $\gamma > 0$, the γ-dimension of $\mathcal{F}$ is ∞, I will show that for any $\tau > 0$, $\mathcal{F}$ is not a $(2\gamma - \tau)$-uniform Glivenko-Cantelli class. For any sample size n, find a set $A \subseteq X$ with $\lceil \frac{n}{\tau} \rceil$ elements that is γ-shattered by $\mathcal{F}$. Then there exists a function $s : A \to [0, 1]$ such that for every $E \subseteq A$, there is an $f := f_E \in \mathcal{F}$ where: $\forall x \in A\backslash E, f(x) \leq s(x) - \gamma$ and $\forall x \in E, f(x) \geq s(x) + \gamma$. Take P to be the uniform distribution on A. We then have $(X, 2^X, P)$ is a probability triple making $\mathcal{F}$ 2^X-measurable. For this choice of P, and $\vec{x}^n := (x_1, x_2, ..., x_n)$ chosen according to the product measure P^n, at least one of the following events occurs:
$E_1 := \left\{\frac{\sum_{i=1}^n s(x_i)}{n} \geq \frac{\sum_{x\in A} s(x)}{|A|}\right\}$ or $E_2 := \left\{\frac{\sum_{i=1}^n s(x_i)}{n} \leq \frac{\sum_{x\in A} s(x)}{|A|}\right\}$.
We have from above that for $\vec{x}^n \subseteq A^n$, there is an f such that $\forall x \in A\backslash\{x_1, ..., x_n\}, f(x) \leq s(x) - \gamma$ and $\forall x \in \{x_1, ..., x_n\}$, $f(x) \geq s(x) + \gamma$. So, on the event E_1,

$$\begin{aligned}|P_n(f) - P(f)| &\geq P_n(f) - P(f)\\ &= \frac{1}{n}\sum_{i=1}^{n} f(x_i) - \frac{\sum_{x\in A} f(x)}{|A|}\end{aligned}$$

$$
\begin{aligned}
&\geq \frac{1}{n}\sum_{i=1}^{n}[s(x_i)+\gamma] - \frac{\sum_{x\in A\setminus\{x_1,\dots x_n\}}[s(x)-\gamma]+n}{|A|} \\
&\geq \frac{1}{n}\sum_{i=1}^{n} s(x_i) - \frac{\sum_{x\in A} s(x)}{|A|} + 2\gamma - \frac{n}{|A|} \\
&\geq 2\gamma - \tau.
\end{aligned}
$$

Similarly, on the event E_2, we can find $f' \in \mathcal{F}$ such that $|P_n(f') - P(f')| \geq 2\gamma - \tau$.

So for any n, and any sample $(x_1, ..., x_n)$, $\sup_{f\in\mathcal{F}} |P_n(f) - P(f)| \geq 2\gamma - \tau$, proving $\mathcal{F}$ is not a $(2\gamma-\tau)$-uniform Glivenko-Cantelli class. Thus we have demonstrated Claim 5.

9. Bounds on the rate of uniform convergence

If the γ-dimension of $\mathcal{F}$ is finite, we have

Theorem 2 *For all distributions P, and all $\epsilon, \delta > 0$,*
$Pr\left(\sup_{f\in\mathcal{F}} |P_n(f) - P(f)| > \epsilon\right) \leq \delta$
for $n = O\left(\frac{1}{\epsilon^2}\left(d\ln^2\frac{d}{\epsilon} + \ln\frac{1}{\delta}\right)\right)$, where d is the $\epsilon/24$-dimension of $\mathcal{F}$.

10. Conclusions

- Alon et al. characterize exactly which classes of real-valued loss functions $\mathcal{F}$ are learnable in the probabilistic-concept model. These are the uniform Glivenko-Cantelli classes.

- They characterize uniform Glivenko-Cantelli classes, via an intermediate step considering the metric entropy, by finiteness of the scale sensitive γ-dimension.

- The γ-dimension is a purely combinatorial quantity extending VC-dimension to real-valued functions.

11. References

Alon, N., Ben-David, S., Cesa-Bianchi, N., and Haussler, D., 1997. Scale Sensitive Dimensions, Uniform Convergence, and Learnability. *J. ACM* **44**, 615-631.

Assouad, P. and Dudley, R. 1989. Minimax nonparametric estimation over classes of sets. Preprint.

Dudley, R. 1999. *Uniform Central Limit Theorems.* Cambridge University Press.

Dudley, R., Giné, E., and Zinn, J. 1991. Uniform and universal Glivenko-Cantelli classes. *J. Theoretical Probability* **4**. 485-510.

Pollard, D. 1990. *Empirical Processes: theory and Applications*, Volumne 2 of *NSF-CBMS Regional Conference Series in Probability and Statistics.* Institute of Mathematical Statistics and American Statistical Association.

Vapnik, V. 1982. *Estimation of Dependencies Based on Empirical Data.* Springer-Verlag, New York.

Vapnik, V. 1998. *Statistical Learning Theory.* John Wiley and Sons, Inc, New York.

Vapnik, V. 1999. An Overview of Statistical Learning Theory. *IEEE Transactions on Neural Networks.* Vol. 10. No. 5, 988-999.

Vapnik, V., and Chervonenkis, A. 1971. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Prob. and its Applic. 16* 2, 264-280.