More Learning Theory: Characterizing Uniform Glivenko-Cantelli Classes using Scale-Sensitive Dimension

Michael Rosenblum

March 22, 2001

Outline

1. Learning Theory Motivation: Probabilistic Concept Learning
2. Review of Ideas for Indicator Functions
3. Preview of Main Theorem
4. Generalizing the Growth Function and VC-dimension to Real Valued Functions
5. Proof of Main Theorem (Outlined)
6. Rates of Convergence
7. Conclusions

General Learning Problem

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.

Will g_n generalize well? In other words, is the expected risk of f_n (denoted $P(f_n)$) close to the minimum possible expected risk for all $f \in \mathcal{F}$?

Formally, will we have as $n \to \infty$,

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

Uniform Learning and Empirical Risk

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:

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

Probabilistic Concept Learning

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, with loss functions $f \in \mathcal{F}$ taking values in $[0, 1]$. Most of the proofs, however, hold in greater generality.

Definitions

For a class of functions $\mathcal{F}$ with domain X, and $\epsilon > 0$,

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

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

2. $\mathcal{F}$ is an ϵ-uniform Glivenko-Cantelli class if

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

Growth Function and VC-dimension for indicator functions

3. For $\vec{x}^n := (x_1, x_2, ..., x_n) \subseteq X$, let $\mathcal{N}(\mathcal{F}, \vec{x}^n)$ be the number of distinct vectors of the form $(f(x_1), ..., f(x_n))$ for $f \in \mathcal{F}$.

The growth function $H_n(\mathcal{F}) := \log \sup_{\vec{x}^n} \mathcal{N}(\mathcal{F}, \vec{x}^n)$.

4. 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 $\mathcal{F}$.

Theorem for indicator loss functions

Theorem:

Let $\mathcal{F}$ be a class of $\{0,1\}$-valued functions on X. 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) we have for the growth function, $\lim_{n\to\infty} H_n(\mathcal{F})/n = 0$.

(d) $\mathcal{F}$ has finite VC-dimension.

Goals of Alon, Ben-David, Cesa-Bianchi, and Haussler

1. Generalize the definition of growth function $H_n(\mathcal{F})$ to classes of real-valued functions.

   Prove 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 classes of real-valued functions.

   Prove it is finite iff the growth function is $o(n)$.

Preview: Main Theorem

Main Theorem:

Let $\mathcal{F}$ be a class of $[0,1]$-valued functions on X. 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.

Scale Sensitive Covering Numbers

For $\vec{x}^n := (x_1, x_2, ..., x_n) \subseteq X$, and $f, g \in \mathcal{F}$,

Define the $l^\infty_{\vec{x}^n}$-pseudo-metric by

$$l^\infty_{\vec{x}^n}(f, g) := \max_{1 \leq i \leq n} |f(x_i) - g(x_i)|.$$

- The ϵ-covering number $\mathcal{N}(\epsilon, \mathcal{F}, \vec{x}^n)$ is defined to be:

  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 $l^\infty_{\vec{x}^n}(f, g) < \epsilon$.

- The ϵ-packing number $\mathcal{M}(\epsilon, \mathcal{F}, \vec{x}^n)$ is defined to be:

  the maximum cardinality of a set $H \subseteq \mathcal{F}$ such that for all distinct $f, g \in H$, $l^\infty_{\vec{x}^n}(f, g) \geq \epsilon$.

Claim 1

Claim 1: For all possible values of $\vec{x}^n$,

$$\mathcal{N}(\epsilon, \mathcal{F}, \vec{x}^n) \quad \leq \quad \mathcal{M}(\epsilon, \mathcal{F}, \vec{x}^n).$$

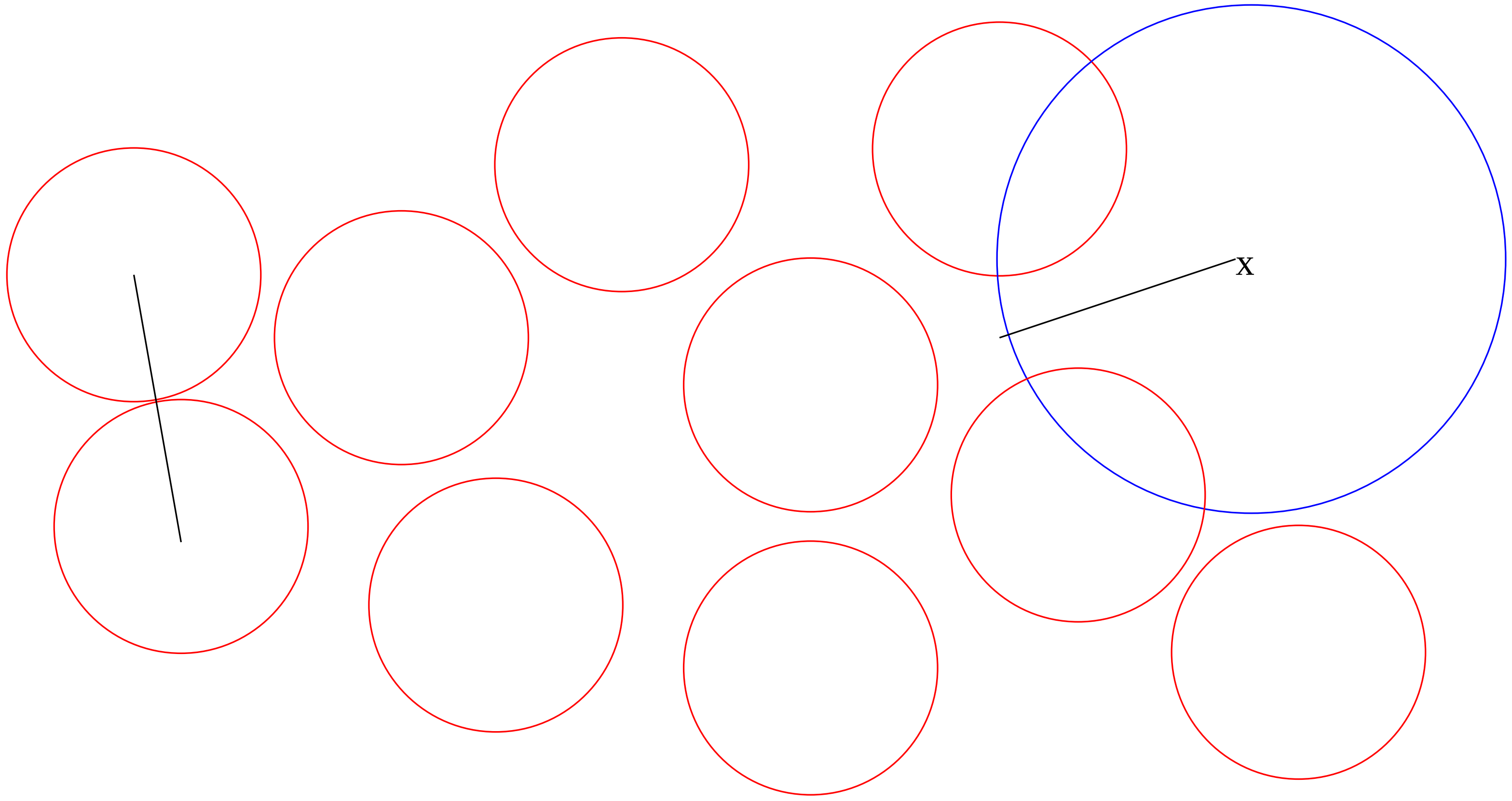
Proof: It is enough to show that any maximal ϵ-packing for $\mathcal{F}$ is an ϵ-covering for $\mathcal{F}$.

Suppose to the contrary, that we have a maximal ϵ-packing $H \subseteq \mathcal{F}$, which was not an ϵ-covering.

Then there is a function $f \in \mathcal{F}$ with $l^{\infty}_{\vec{x}^n}$ distance at least ϵ from all $h \in H$.

But then we could add f to H and still have an ϵ-packing, contradicting H being maximal.

Picture Proof of Claim 1



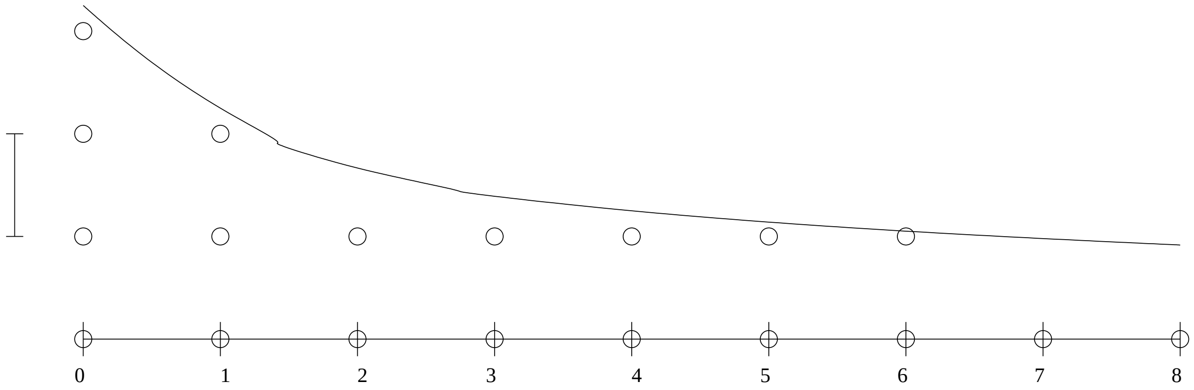
Metric Entropy and an Example

Define the metric entropy $H_n(\epsilon, \mathcal{F}) := \log \sup_{\vec{x}^n} \mathcal{N}(\epsilon, \mathcal{F}, \vec{x}^n)$.

Example: 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 $n \in \mathbf{N}$.

In this example, how does the metric entropy behave for large n?

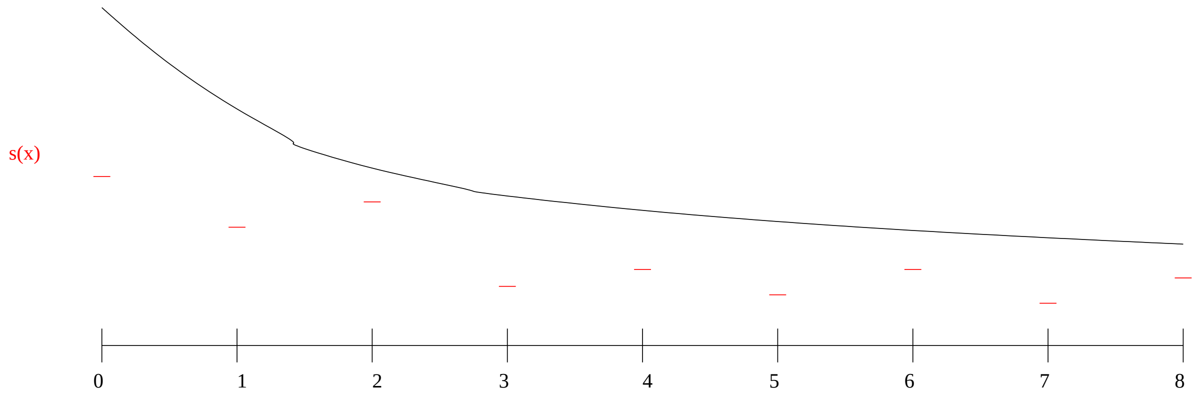
- For all $\epsilon > 0$, the ϵ-covering number is finite, and can be bounded *independent of n.*
- Thus, for all $\epsilon > 0$, $\lim_{n\to\infty} H_n(\epsilon, \mathcal{F})/n = 0$.



Pollard-dimension

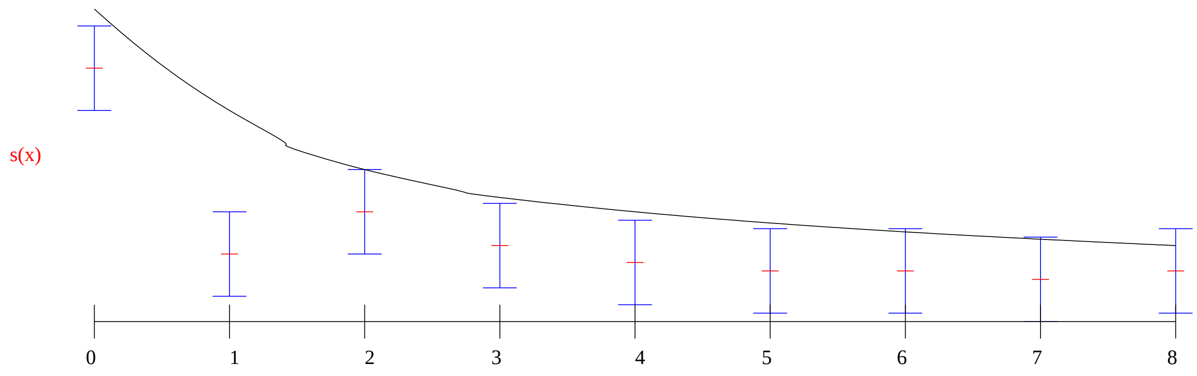
A first attempt to generalize VC-dimension to real-valued functions: We say that a class of $[0,1]$-valued functions $\mathcal{F}$ on X *shatters* a finite set $A \subseteq X$ if there exists a function $s : A \to \mathbf{R}$ such that for every $E \subseteq A$, there is an $f \in \mathcal{F}$ where: $\forall x \in A \backslash E, f(x) < s(x)$ 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 $\mathcal{F}$. Example:



Scale sensitive γ-dimension

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 $\mathcal{F}$.
Example:



Proving $\mathbf{d} \Rightarrow \mathbf{c}$

Recall the **Main Theorem:**

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

(a) $\forall \epsilon > 0$, $\mathcal{F}$ is ϵ-learnable in the probabilistic-concept model.

(b) $\forall \epsilon > 0$, $\mathcal{F}$ is an ϵ-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.

Let's show $\mathbf{d} \Rightarrow \mathbf{c}$.

Proving $\mathbf{d} \Rightarrow \mathbf{c}$

Goal: prove $\forall \gamma > 0$ if $\mathcal{F}$ has finite γ-dimension, then $\lim_{n\to\infty} H_n(4\gamma, \mathcal{F})/n = 0$.

Enough to prove the following lemma:

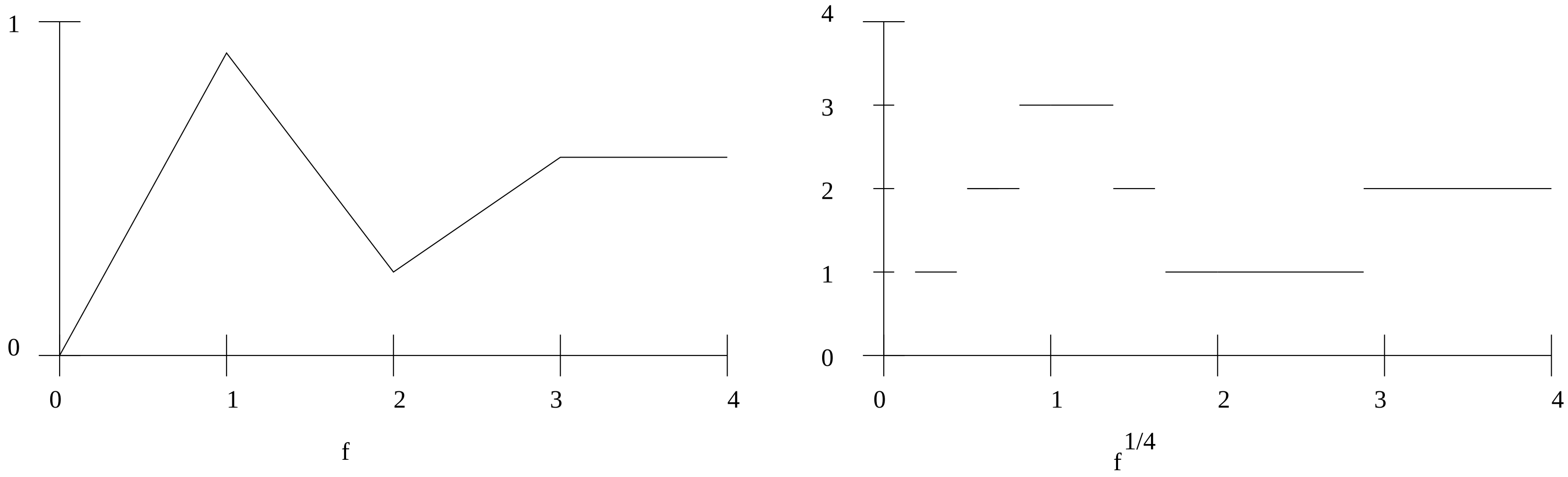
Lemma 1: For all $\gamma \in (0, \frac{1}{4})$, for $d :=$ the γ-dimension of $\mathcal{F}$, $d < \infty$, we have:
$\sup_{(x_1,\ldots,x_n)} \mathcal{N}(4\gamma, \mathcal{F}, \vec{x}^n) \le 2\left(\frac{9n}{16\gamma^2}\right)^{1+d\log(3en/(4d\gamma))}$

Proving **d** ⇒ **c**: Discretizing $\mathcal{F}$

First, discretize the functions in $\mathcal{F}$.

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 $\mathcal{F}^\rho := \{f^\rho : f \in \mathcal{F}\}$.
Example: $\rho = 1/4$,



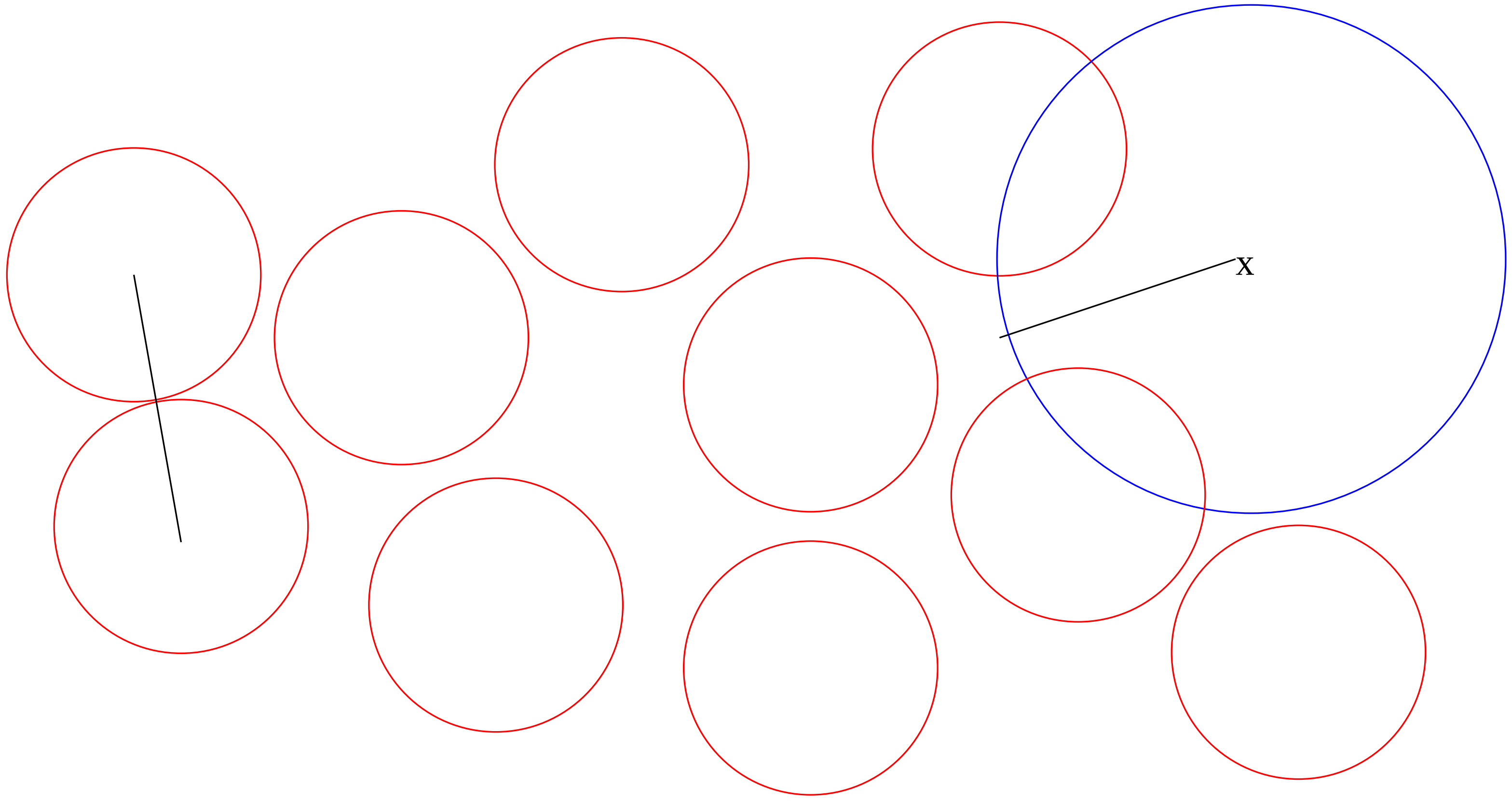
Lemma 1: Proof Overview

To prove Lemma 1, we'd like to bound $\sup_{(x_1,\ldots,x_n)} \mathcal{N}(4\gamma, \mathcal{F}, \vec{x}^n)$.

I will show the following inequalities hold for $d < \infty$ the γ-dimension of $\mathcal{F}$ and for c the 1-dimension of the discretized class $\mathcal{F}^{2\gamma}$.

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

Picture Proof of Claim 1



Proof of Claim 2

We'd like to show the following fact about packing numbers:
$\sup_{(x_1,\ldots,x_n)} \mathcal{M}(4\gamma, \mathcal{F}, \vec{x}^n) \leq \sup_{(x_1,\ldots,x_n)} \mathcal{M}(2, \mathcal{F}^{2\gamma}, \vec{x}^n)$.

It is enough to show that for any choice of $\vec{x}^n$,
$\mathcal{M}(4\gamma, \mathcal{F}, \vec{x}^n) \leq \mathcal{M}(2, \mathcal{F}^{2\gamma}, \vec{x}^n)$

We'll show for any $\rho > 0$, that every 2-ρ packing of $\mathcal{F}$ corresponds to a 2 packing of $\mathcal{F}^\rho$ of the same size:

> By definition of $\mathcal{F}^\rho$, for all $f, g \in \mathcal{F}$, and $x \in \{x_1, ..., x_n\}$,
> $|f(x) - g(x)| \geq 2\rho$ implies $|f^\rho(x) - g^\rho(x)| \geq 2$.
>
> Thus, if for some $G \subseteq \mathcal{F}$,
> we have for all distinct $f, g \in G$, $\sup_{x \in \vec{x}^n} |f(x) - g(x)| \geq 2\rho$,
> then for all distinct $f', g' \in G^\rho$, $\sup_{x \in \vec{x}^n} |f'(x) - g'(x)| \geq 2$.

This completes the proof of Claim 2.

Proof of Claim 3

We first examine the main combinatorial result:

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 the packing number $\mathcal{M}(2, \mathcal{H}, X) < 2(nb^2)^{\lceil \log y \rceil}$, where $y := \sum_{i=1}^{c} \binom{n}{i} b^i$.

Setting $\mathcal{H} := \mathcal{F}^{2\gamma}$, and $b := \lfloor \frac{1}{2\gamma} \rfloor + 1$, 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 get from Claim 3' that

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

which is **Claim 3**.

Sauer's Lemma: a special case of Claim 3'

Let's look at a special case of Claim 3', setting $b = 3$:

- We need only consider functions with range $\{1, 3\}$ (i.e. indicator functions).
- The 1-dimension of $\mathcal{H}$ reduces to the VC-dimension of the corresponding class of indicator functions.
- The packing number $\mathcal{M}(2, \mathcal{H}, X)$ reduces to $\mathcal{N}(\mathcal{H}, X)$.

In the special case $b = 3$ of Claim 3', 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 $\mathcal{N}(\mathcal{H}, X) \leq \binom{n}{0} + ... + \binom{n}{c}$.

Proof of Claim 4

Claim 4: The 1-dimension of $\mathcal{F}^\rho$ is no more than the $\rho/2$-dimension of $\mathcal{F}$.

Proof:

- Enough to show any set $A \subseteq X$ 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 $s'(x) := (s(x) + 1/2)\rho$

Proof of Claim 4 continued

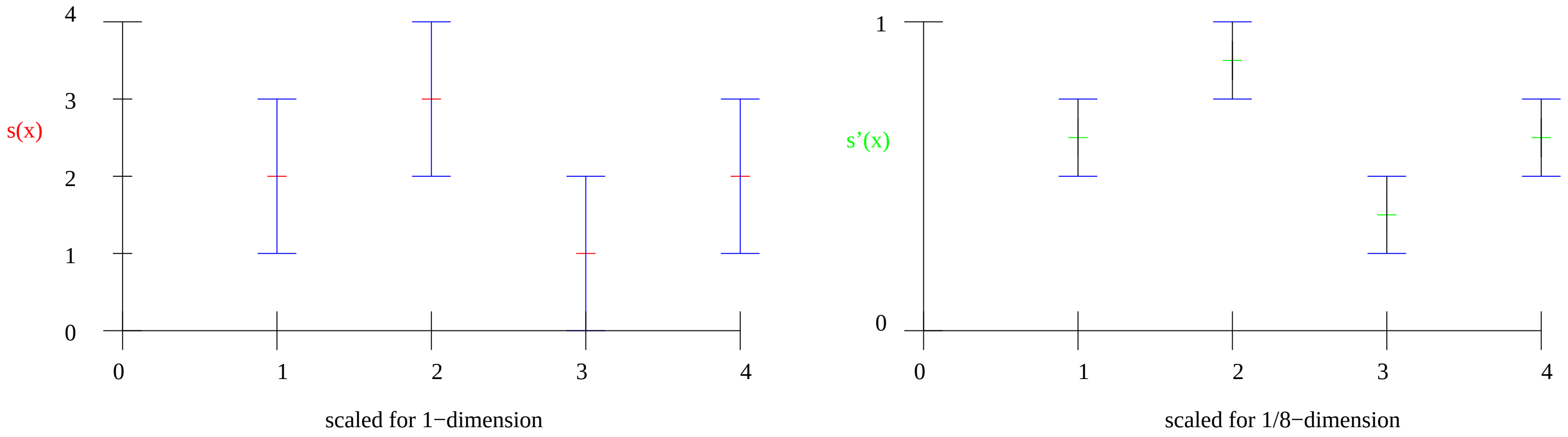
Define $s' : A \to \mathbf{R}$ by $s'(x) := (s(x) + 1/2)\rho$

If $x \in A \backslash E$, then $f^\rho(x) \leq s(x) - 1$, so $f(x) \leq \rho s(x) = s'(x) - \rho/2$.

If $x \in E$, then $f^\rho(x) \geq s(x) + 1$, so $f(x) \geq \rho s(x) + \rho = s'(x) + \rho/2$

Thus A is $\rho/2$-shattered by $\mathcal{F}$, and we are done.

Example: $\rho = 1/4$, $A = \{1, 2, 3, 4\}$:



The Big Picture

Recall the **Main Theorem:**

> Let $\mathcal{F}$ be a class of $[0,1]$-valued functions on X. 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.

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**.

Next, let's take one slide to get the intuition behind **c** $\Rightarrow$ **b**.

Showing $\mathbf{c} \Rightarrow \mathbf{b}$

We want to show if $\forall\epsilon > 0$, we have $\lim_{n\to\infty} H_n(\epsilon, \mathcal{F})/n = 0$, then $\forall\epsilon > 0, \mathcal{F}$ is an ϵ-uniform Glivenko-Cantelli class.

This follows from:

Lemma 2: For all $\epsilon > 0$, for all probability measures P, we have, for $\vec{x}^{2n} := (x_1, x_2, ..., x_{2n})$ chosen independently at random according to P, and
$P_{n'}(f) := \frac{1}{n}\sum_{i=1}^{n} f(x_i)$
$P_{n''}(f) := \frac{1}{n}\sum_{i=n+1}^{2n} f(x_i)$:
we have

$$Pr(\sup_{f\in F} |P_{n'}(f) - P_{n''}(f)| > \epsilon) \leq 5\sqrt{n}\mathbf{E}[\mathcal{N}(\epsilon/3, \mathcal{F}, \vec{x}^{2n})] \exp\left(-\epsilon^2 n/36\right)$$

The idea is that given $\vec{x}^{2n} := (x_1, x_2, ..., x_{2n})$, at precision ϵ there are at most $\mathcal{N}(\epsilon/3, \mathcal{F}, \vec{x}^{2n})$ "distinguishable" functions $f \in \mathcal{F}$.

Connecting learning to uniform Glivenko-Cantelli classes: **b** $\Rightarrow$ **a**

Finally, let's prove **b** $\Rightarrow$ **a**.
Given an ϵ-uniform Glivenko-Cantelli class $\mathcal{F}$, we'll show it's 3ϵ-learnable in the probabilistic concept model.

As our learning algorithm, we'll use empirical risk minimization.
That is, given data points $\vec{x}^n := (x_1, x_2, ..., x_n)$, our algorithm chooses the function f_n from $\mathcal{F}$ with minimal empirical risk:
$P_n(f) := \frac{1}{n}\sum_{i=1}^{n} f(x_i)$.

Proof of **b** $\Rightarrow$ **a** continued

One can show that for f_n the empirical risk minimizer, the event $\{P(f_n) > \inf_{f \in \mathcal{F}} P(f) + 3\epsilon\}$ is contained in the event $\left\{\sup_{f \in \mathcal{F}} |P_n(f) - P(f)| > \epsilon\right\}$.

Thus $\mathcal{F}$ an ϵ-uniform Glivenko-Cantelli class; i.e.

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

implies $\mathcal{F}$ is 3ϵ-learnable in the probabilistic-concept model; i.e.

$$\lim_{n \to \infty} \sup_{P} Pr\left(P(f_n) > \inf_{f \in \mathcal{F}} P(f) + 3\epsilon\right) = 0.$$

Bounds on the rate of uniform convergence

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

Theorem: For all distributions P, and all $\epsilon, \delta > 0$,
$Pr\left(\sup_{f\in F} |P_n(f) - P(f)| > \epsilon\right) \leq \delta$

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 $\mathcal{F}$.

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.