Measures of Hypothesis Complexity: 9.520

Class 9, 10 March 2002

Alex Rakhlin and Sayan Mukherjee

Plan

- Measuring the complexity of function spaces.
- Definitions of VC dimension and scale sensitive versions.
- Necessary and sufficient conditions for uniform convergence.

Uniform convergence for classification

Our loss function is now $V(f(x), y) = \Theta(-yf(x))$ and our RKHS is $||f||_K^2 \leq M$.

Our goal is to bound the following

$$P\left\{ \sup_{f \in \mathcal{H}: \|f\|_K^2 \leq M} |I[f] - I_S[f]| > \epsilon \right\}.$$

For one function we could use the Chernoff bound

$$P\{|I[f] - I_S[f]| > \epsilon\} < 2 \exp(-2\epsilon^2 \ell).$$

Uniform convergence for classification (cont)

We then would want to use the union bound over the number of "essential" functions in the class which we already determined. We have seen how to relate the ϵ in the bound with the r covering radius for square loss.

What about if $V(f(x), y) = \Theta(-yf(x))$?

Classification is scale insensitive

The key result in computing $r(\epsilon)$ was showing that if

$$||f_1(x) - f_2(x)||_\infty < r(\epsilon)$$

then

$$|V(f_1(x), y) - V(f_2(x), y)| \leq \epsilon \qquad \forall x, y.$$

For the classification loss function $\epsilon = 1$ and varying $r(\epsilon)$ has no effect.

Counting classification functions

Given ℓ points $\{(x_1, y_1), ..., (x_\ell, y_\ell)\}$, for every $f \in \mathcal{H}(M)$ we get different "labelings" $\{\Theta(-y_1 f(x_1)), ..., \Theta(-y_\ell f(x_\ell))\}$ (or, alternatively, different vertices of the $[0,1]^\ell$ cube are spanned).

We define the random VC entropy as the number of labelings that can be implemented over $f \in \mathcal{H}(M)$ written as

$$\mathcal{N}^{\mathcal{H}(M)}((x_1, y_1), ..., (x_\ell, y_\ell)).$$

An obvious property of $\mathcal{N}^{\mathcal{H}(M)}((x_1, y_1), ..., (x_\ell, y_\ell))$ is:

$$\mathcal{N}^{\mathcal{H}(M)}((x_1, y_1), ..., (x_\ell, y_\ell)) \leq 2^\ell.$$

Counting classification functions

Notice that

$$\mathcal{N}^{\mathcal{H}(M)}((x_1, y_1), ..., (x_\ell, y_\ell)).$$

depends on data so we need to take the expectation to use it

$$\overline{\mathcal{N}} = \mathbb{E}_{x_1,y_1,...,x_\ell,y_\ell} \mathcal{N}^{\mathcal{H}(M)}((x_1, y_1), ..., (x_\ell, y_\ell)).$$

We can use the following bound

$$\mathbb{P}\left\{ \sup_{f \in \mathcal{H}: \|f\|_K^2 \leq M} |I[f] - I_S[f]| > \epsilon \right\} < 2\overline{\mathcal{N}} \exp(-2\epsilon^2 \ell).$$

A necessary and sufficient condition

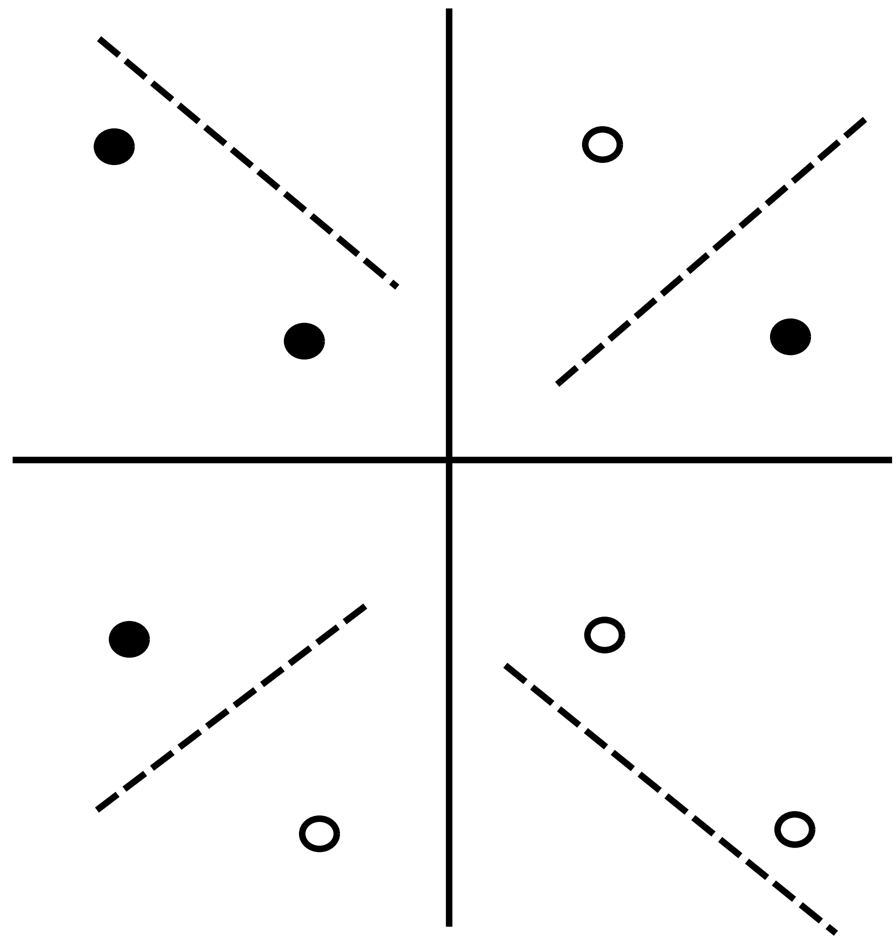
Iff

$$\lim_{\ell\to\infty} \frac{\log \overline{\mathcal{N}}}{\ell} \to 0,$$

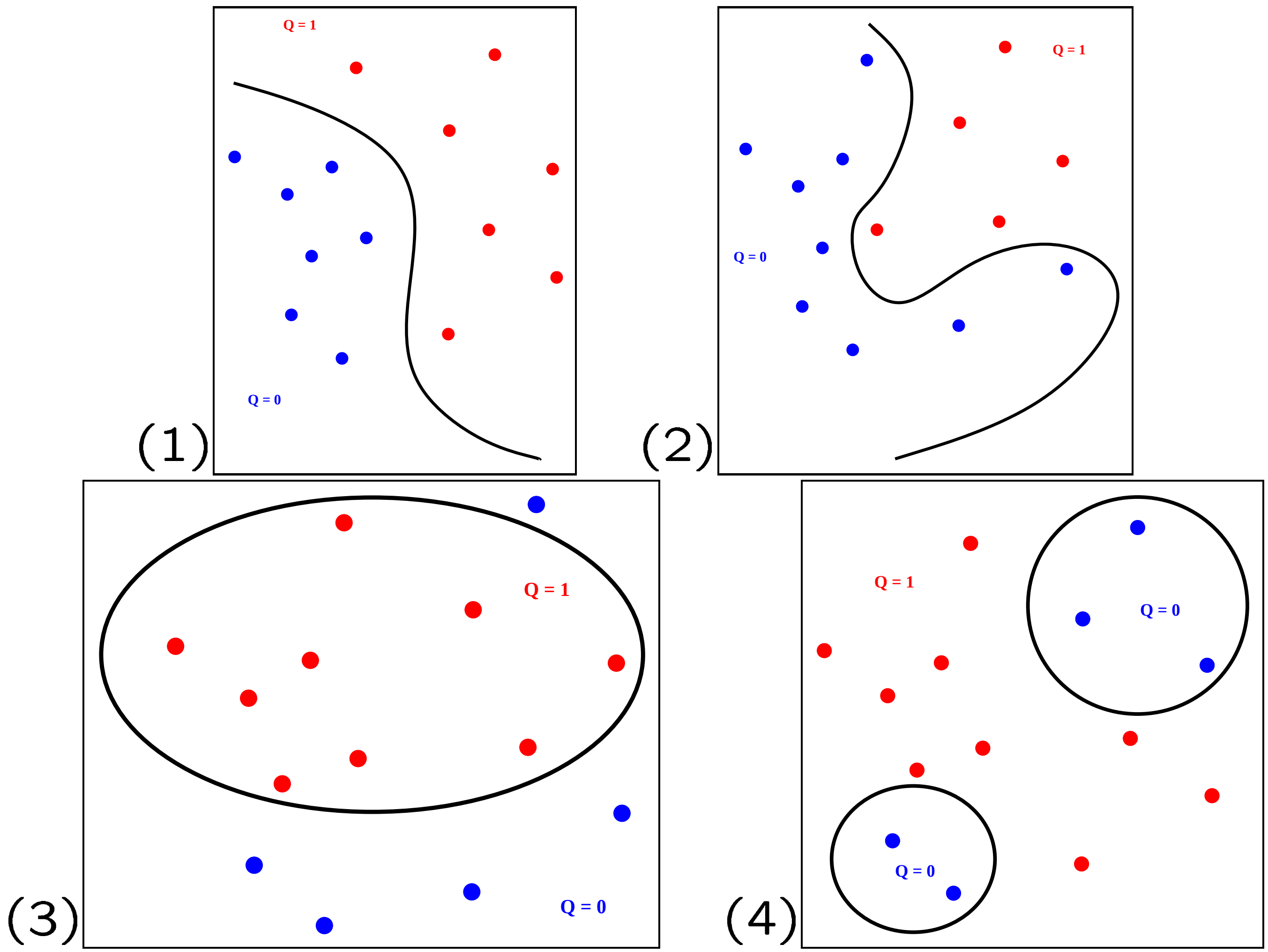
do we get uniform convergence in probability.

So the capacity can increase polynomially in ℓ but not exponentially.

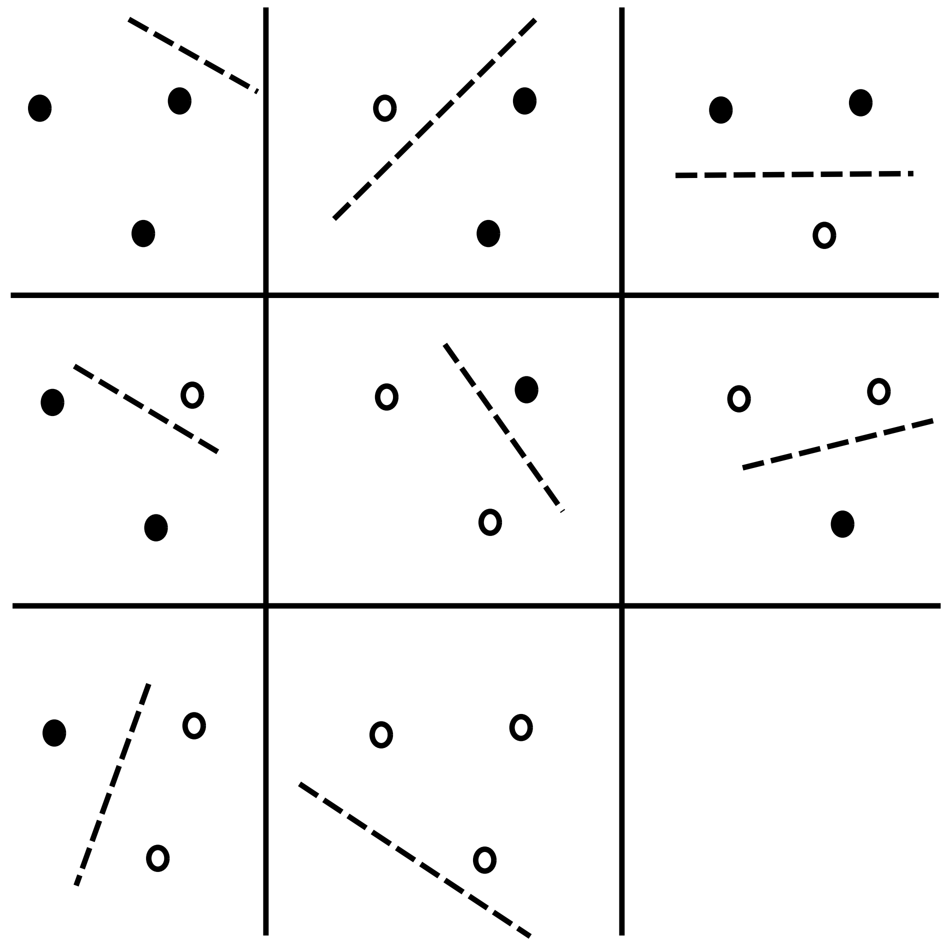
Implementation of different labelings



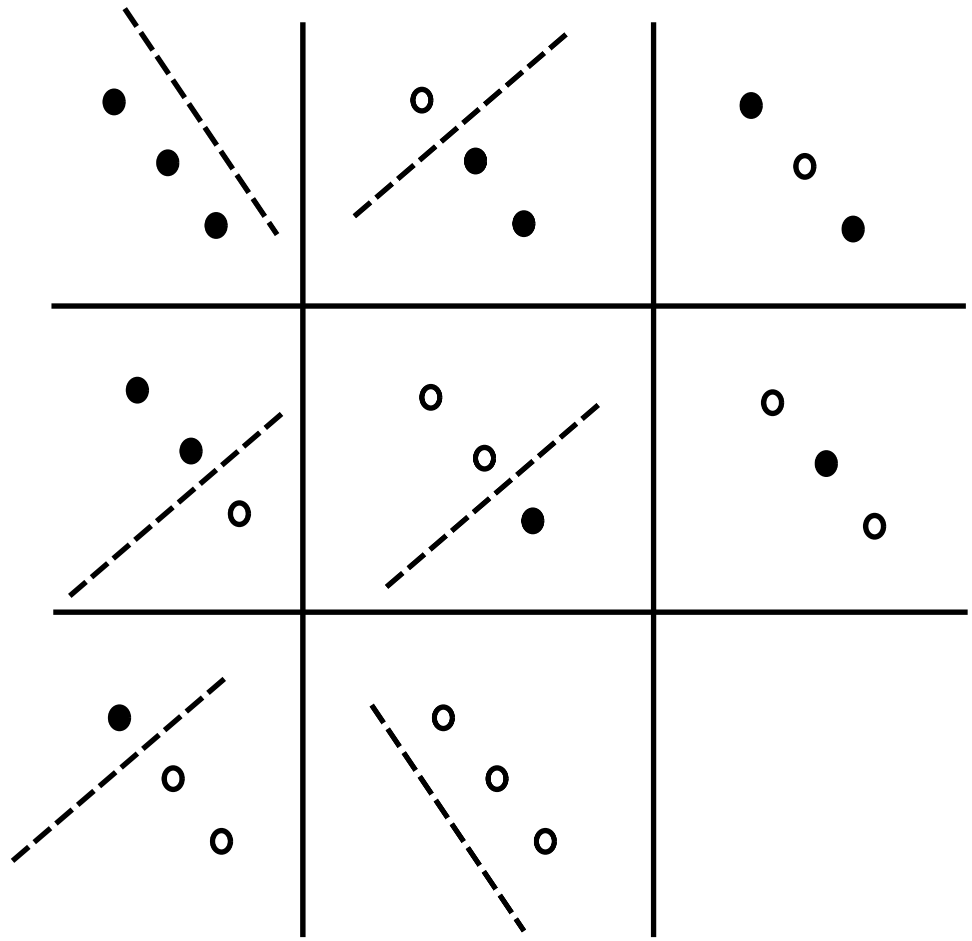
Implementation of different labelings



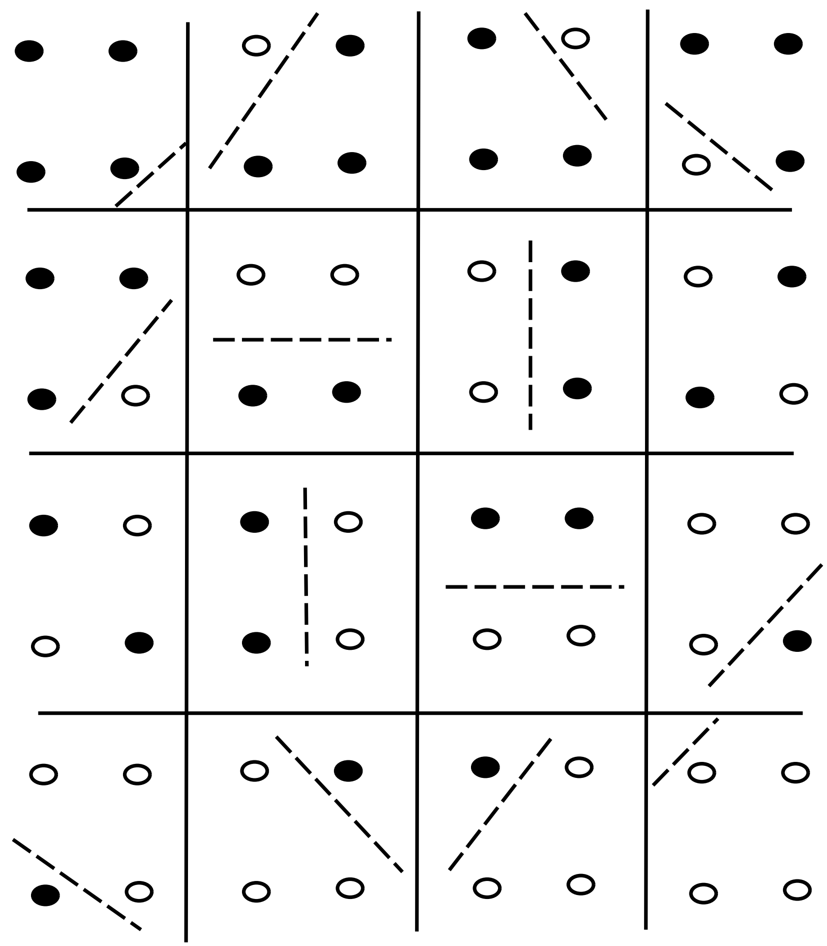
The 8 possible labelings of 3 points in 2D



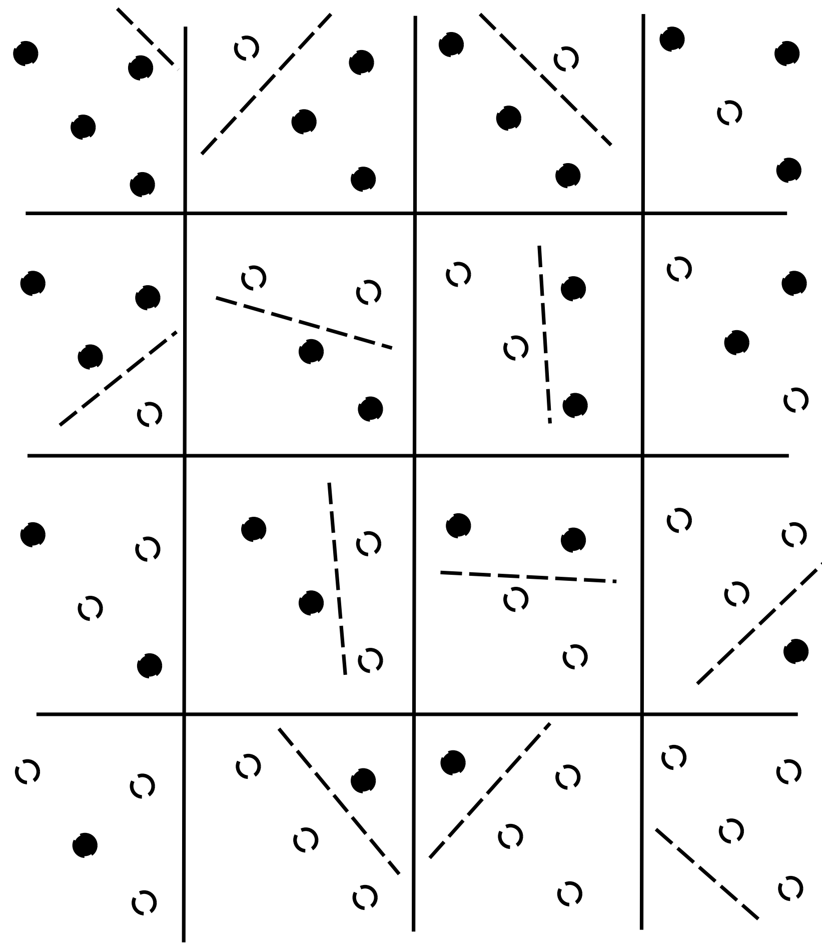
Example



Example



Example



How Many Labelings?
Sauer's Lemma

If the hypothesis space can separate h points in all possible (2^h ways), then $\ell > h$ points can be labeled in

$$\sum_{i=1}^{h} \binom{\ell}{i} < \left(\frac{e\ell}{h}\right)^h$$

possible ways and

$$\sum_{i=1}^{h} \binom{\ell}{i} < 2^\ell.$$

VC-dimension

The VC-dimension of a set of binary functions is h if and only if

- There is **at least one set of** h **points** that can be labeled in all possible ways;

- there is **no set of** $h + 1$ **points** that can be labeled in all possible ways;

Classification

The finiteness of the VC-dimension of the set of functions $f \in \mathcal{H}(M)$ for the classification loss is a **necessary and sufficient** for uniform convergence of Ivanov regularization (empirical risk minimization in a bounded function class) for arbitrary probability distributions with a fast rate of convergence.

$$\overline{\mathcal{N}} \leq \left(\frac{e\ell}{h}\right)^{h}.$$

VC-bound

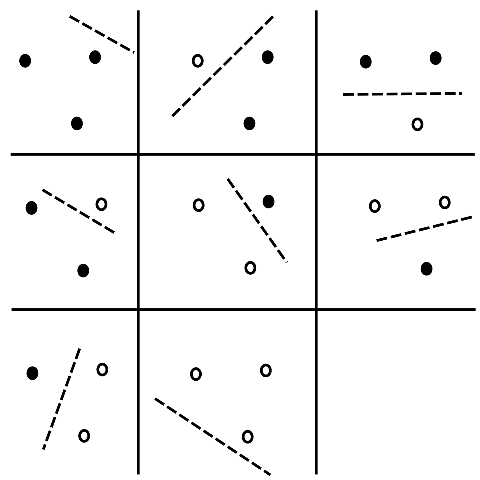
We can now bound the defect in the case of classification

$$P\left\{\sup_{f\in\mathcal{H}:\|f\|_K^2\leq M}|I[f]-I_S[f]|>\epsilon\right\}<2\left(\frac{e\ell}{h}\right)^h\exp(-2\epsilon^2\ell).$$

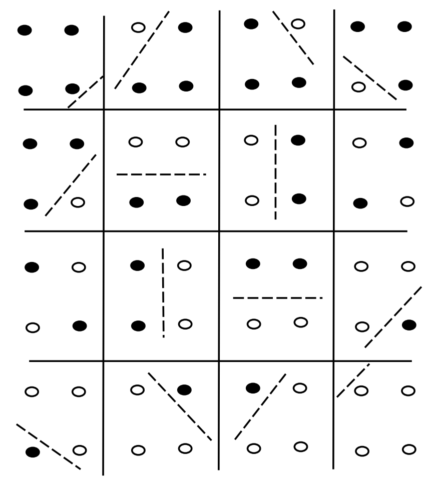
Which allows us to state that with probability $1-\delta$

$$I[f]\leq I_S[f]+\sqrt{\frac{h\ln(e\ell/h)-\ln(\delta/2)}{\ell}}.$$

VC dimension of hyperplanes



all the possible labelings



not all the possible labelings

VC-dimension = 3

VC dimension for RKHS

For hyperplanes in $\mathbb{R}^d$ the VC-dimension is $d + 1$. For a RKHS with dimensionality N the VC-dimension is $N + 1$ independent of the restriction on the norm.

What happens in the case of Gaussian kernels ?

"Dear Tommy, it may be infinite"—— V. Vapnik 1999.

VC-dimension and free parameters

The VC-dimension is proportional, but not necessarily equal, to the number of parameters.

- For Multilayer Perceptrons with hard thresholds $h \propto n \ln n$ (Maass, 1994);

- For Multilayer perceptrons with standard sigmoid thresholds $h \propto n^2$ (Koiran and Sontag, 1995);

- For classification functions of the form $\theta(-y \sin(\alpha x))$ the VC-dimension is infinite;

Empirical covering numbers

Instead of using the sup norm as the metric of our cover we can use

$$d_{x_\ell}(f_1, f_2) = \max_{x_i} |f_1(x_i) - f_2(x_i)|.$$

The **empirical covering number** $\mathcal{N}(\mathcal{H}, r, d_{x_\ell})$ is the minimal $m \in \mathbb{N}$ such that there exists m disks in $\mathcal{H}$ with radius r covering function values at ℓ points.

Empirical covering numbers

Notice that

$$\mathcal{N}(\mathcal{H}, r, d_{x_\ell}).$$

depends on data so we need to take the expectation to use it

$$\overline{\mathcal{N}} = \mathbb{E}_S \mathcal{N}(\mathcal{H}, r, d_{x_\ell}).$$

A necessary and sufficient condition

Iff for any given $r > 0$

$$\lim_{\ell \to \infty} \frac{\log \overline{\mathcal{N}}}{\ell} \to 0,$$

do we get uniform convergence in probability.

So the capacity can increase polynomially in ℓ but not exponentially at any scale.

Is there a number like VC dimension for classification that can be used to bound the empirical cover ?

V_γ dimension and shattering

The V_γ-dimension of $\mathcal{F}_{\mathcal{H},V}$ is defined as the the maximum number h of vectors $\{(x_1, y_1), \ldots, (x_h, y_h)\}$ that can be separated into two classes in all 2^h possible ways using rules:

$$\text{class 1 if: } V(y_i, f(x_i)) \geq s + \gamma$$
$$\text{class 0 if: } V(y_i, f(x_i)) \leq s - \gamma$$

for some $s \geq 0$. If, for any number N, it is possible to find N points that can be separated in all possible ways, the V_γ-dimension is infinite.

Key result (Alon et al. 93)

Finiteness of the V_γ dimension for every $\gamma > 0$ is a **necessary and sufficient** condition for distribution independent uniform convergence of the ERM method for real-valued functions.

V_γ dimension

The expectation of the cover is bounded by the V_γ dimension

$$\mathbb{E}_S \mathcal{N}(\mathcal{H}, r, d_{x_\ell}) \leq 2 \left(\frac{4\ell}{r^2}\right)^{h \log(2e\ell/(hr))}.$$

For the square loss bounded with the same constants as we saw in last class we get

$$\mathbb{P}\left\{\sup_{f \in \mathcal{H}} |I[f] - I_S[f]| \leq \epsilon\right\} \leq 1 - 4\left(\frac{4\ell}{(\epsilon/8B')^2}\right)^{h \log(2e\ell/(h(\epsilon/8B')))} \exp(-\epsilon^2 \ell / B^2).$$