PROBABLY APPROXIMATELY CORRECT LEARNING

FRANÇOIS FLEURET

EPFL – CVLAB

Statistical Learning for Classification

The usual setting for learning in a context of classification

- ➠ A training set
- ➠ A family of classifiers
- ➠ A test set

Choose a classifier according to its performances on the **training set** to get good performances on the **test set**.

TOPIC OF THIS TALK

The goal of this talk is to give an intuitive understanding of the Probably Approximately Correct learning (PAC learning for short) theory.

- Concentration inequalities
- Basic PAC results
- Relation with Occam's principle
- Relation to Vapnik-Chervonenkis dimension

NOTATION

$\mathcal{X}$ the space of the objects to classify (for instance images)

$\mathcal{C}$ the family of classifiers

$S = ((X_1, Y_1), \ldots, (X_{2N}, Y_{2N}))$ a random variable on $(\mathcal{X} \times \{0, 1\})^{2N}$ standing for the samples (both training and testing)

F a random variable on $\mathcal{C}$ standing for the learned classifier (which can be a deterministic function of S or not)

REMARKS

- ➠ The set $\mathcal{C}$ contains all the classifiers obtainable with the learning algorithm. For an ANN for instance, there is one element of $\mathcal{C}$ for every single configuration of the synaptic weights.
- ➠ The variable S is not **one** sample, but a family of $2N$ samples with their labels. It contains both the training and the test set.

For every $f \in \mathcal{C}$, we denote by $\xi(f, S)$ the difference between the training and the test errors of f estimated on S

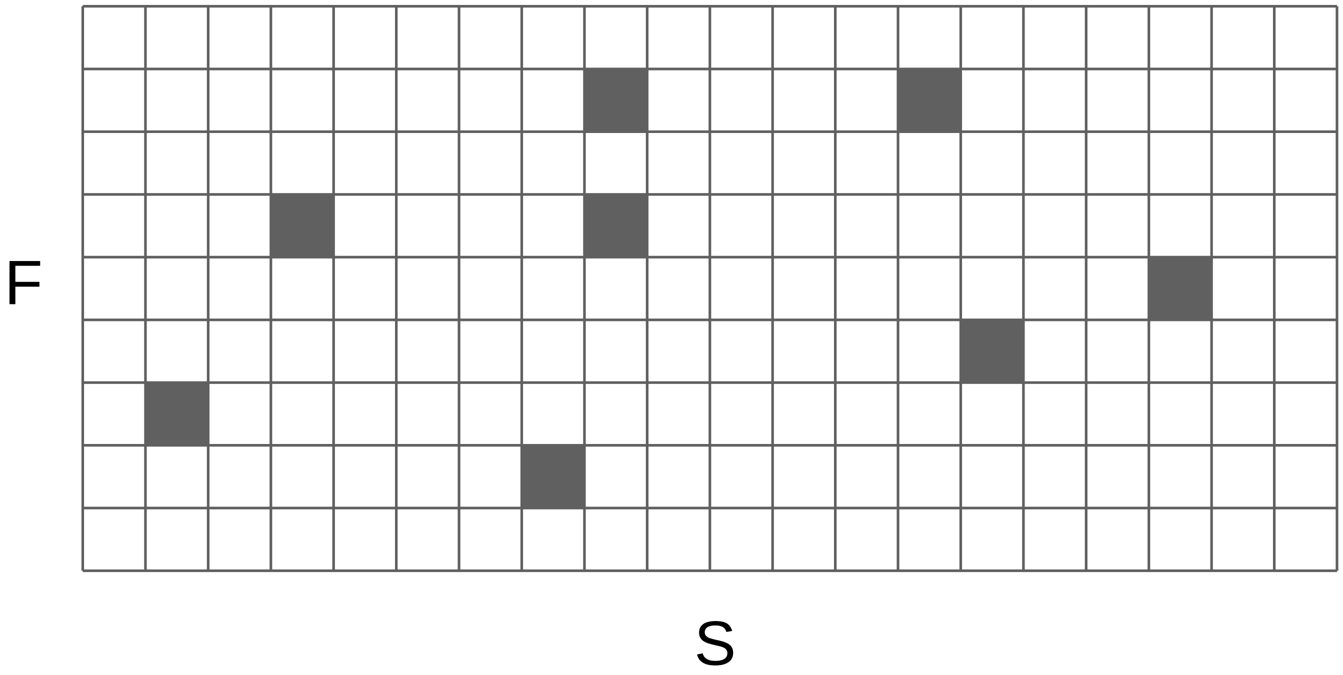
$$\xi(f, S) \quad = \quad \underbrace{\frac{1}{N} \sum_{i=1}^{N} 1\{f(X_{N+i}) \neq Y_{N+i}\}}_{\text{test error}} - \underbrace{\frac{1}{N} \sum_{i=1}^{N} 1\{f(X_i) \neq Y_i\}}_{\text{training error}}$$

Where $1\{t\}$ is equal to 1 if t is true, and 0 otherwise. Since S is random, this is a random quantity.

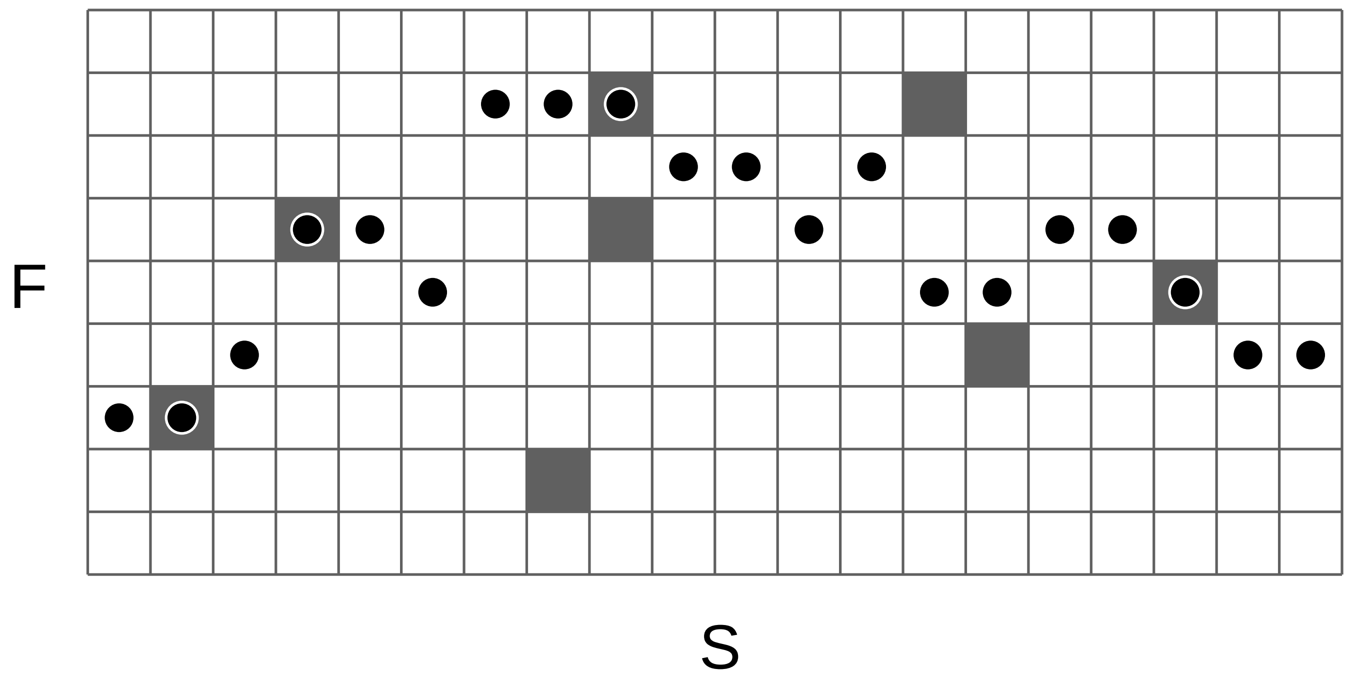
Given η, we want to bound the probability that the test error is less than the training error plus η

$$P\left(\xi(F, S) \leq \eta\right) \quad \geq \quad \textbf{?}$$

F is not constant and depends on the $X_1, \ldots, X_{2N}$ and the $Y_1, \ldots, Y_N$.



Gray squares correspond to the (S, F) for which $\xi(F, S) \geq \eta$.



A training algorithm associates an F to every S, here shown with dots. We want to bound the number of dots on gray cells.

CONCENTRATION INEQUALITY

How we see that for any fixed f, the test and training errors are likely to be similar . . .

HŒFFDING'S INEQUALITY (1963)

Given a family of independent random variables $Z_1, \dots, Z_N$, bounded $\forall i, Z_i \in [a_i, b_i]$, if we let S denote $\sum_i Z_i$, we have

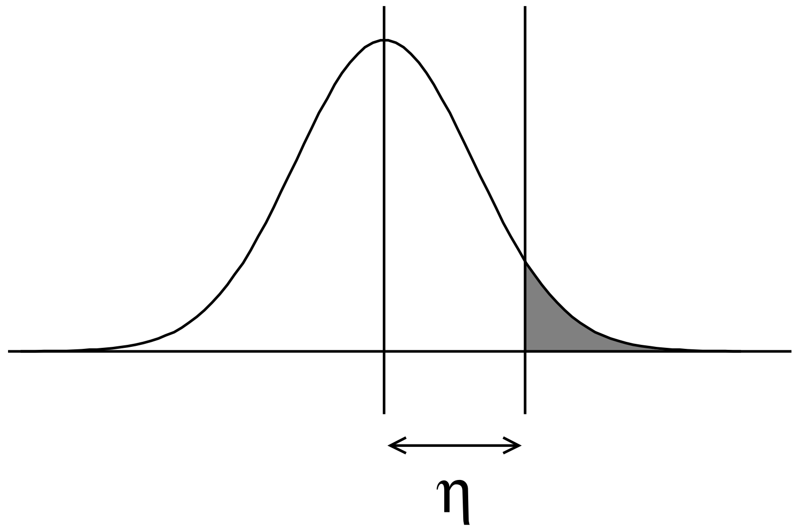
$$P(S - E(S) \geq t) \quad \leq \quad \exp\left(-\frac{2t^2}{\sum_i (b_i - a_i)^2}\right)$$

Note that the $1\{f(X_i) \neq Y_i\}$ are i.i.d Bernoulli, and we have

$$\begin{aligned} \xi(f, S) &= \frac{1}{N} \sum_{i=1}^{N} 1\{f(X_{N+i}) \neq Y_{N+i}\} - \frac{1}{N} \sum_{i=1}^{N} 1\{f(X_i) \neq Y_i\} \\ &= \frac{1}{N} \sum_{i=1}^{N} \underbrace{1\{f(X_{N+i}) \neq Y_{N+i}\} - 1\{f(X_i) \neq Y_i\}}_{\Delta_i} \end{aligned}$$

Thus ξ is the averaged sum of the Δ_i, which are i.i.d random variables on $\{-1, 0, 1\}$ of zero mean.

When f is fixed $\xi(f,\, S)$ is with high probability around 0, and we have (Hœffding)



$$\forall f,\ \forall \eta,\ \ P\left(\xi(f,\, S) \geq \eta\right) \quad \leq \quad \exp\left(-\frac{1}{2}\,\eta^2\, N\right)$$

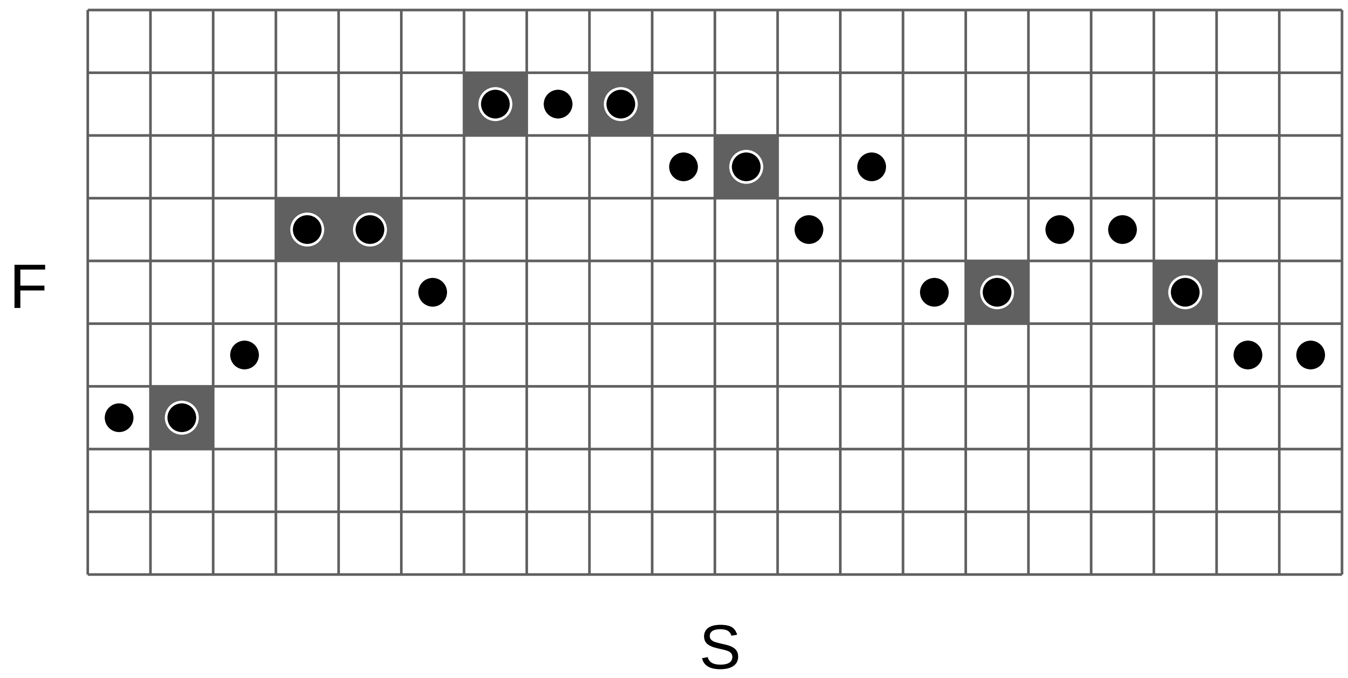
Hence, we have an upper bound on the number of gray cells per row.

UNION BOUND

How we see that the probability the chosen F fails is lower than the probability that there exists a f that fails . . .

We have

$$
\begin{aligned}
P(\xi(F,\, S) \geq \eta) &= \sum_f P(F = f,\, \xi(F,\, S) \geq \eta) \\
&= \sum_f P(F = f,\, \xi(f,\, S) \geq \eta) \\
&\leq \sum_f P(\xi(f,\, S) \geq \eta) \\
&\leq ||\mathcal{C}|| \, \exp\left(-\frac{1}{2}\,\eta^2\, N\right)
\end{aligned}
$$



We can see that graphically as a situation when the dots meet all the gray squares.

Since

$$P(\xi(F, S) \geq \eta) \quad \leq \quad ||\mathcal{C}|| \exp\left(-\frac{1}{2}\,\eta^2\,N\right)$$

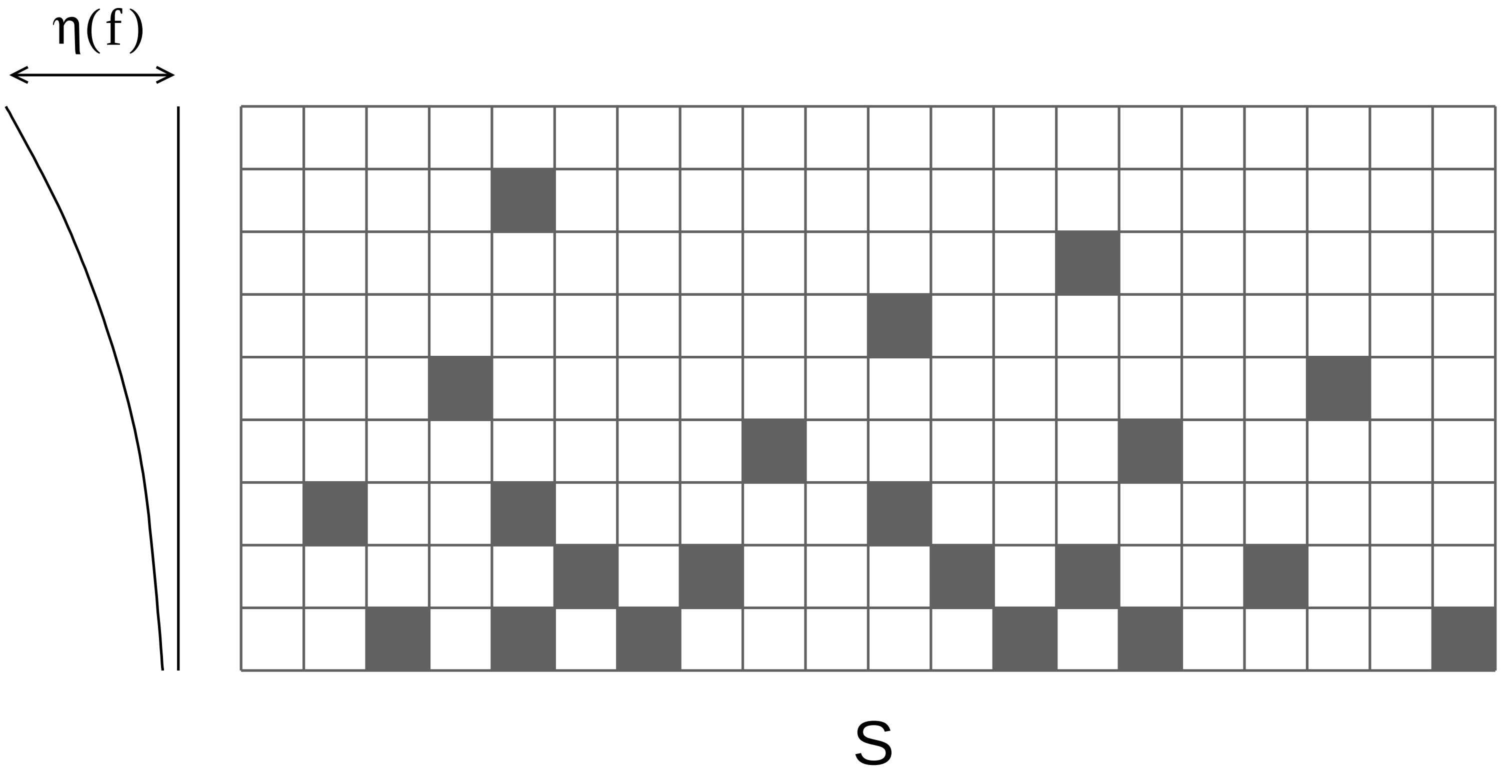
we have

$$P\left(\xi(F, S) \geq \sqrt{2\,\frac{\log||\mathcal{C}|| + \log\frac{1}{\epsilon^\star}}{N}}\right) \quad \leq \quad \epsilon^\star$$

Thus, the margin between the training and test errors η which is verified for a fixed probability $\epsilon^\star$ grows like the square root of the log of the number of classifiers $||\mathcal{C}||$.

PRIOR ON $\mathcal{C}$

How we see weird results when we arbitrarily distribute allowed errors on the fs before looking at the training data . . .



If the margin η depends on F, the proportion of gray squares is not the same on every row.

Let $\epsilon(f)$ denote the (bound on the) probability that the constraint is not verified for f

$$
\begin{aligned}
P(\xi(F, S) \geq \eta(F)) &\leq P\left(\exists f \in \mathcal{C}, \xi(f, S) \geq \eta(f)\right) \\
&\leq \sum_f P(\xi(f, S) \geq \eta(f)) \\
&\leq \sum_f \epsilon(f)
\end{aligned}
$$

and we have

$$
\forall f,\ \eta(f) = \sqrt{2 \frac{\log \frac{1}{\epsilon(f)}}{N}}
$$

Let define $\epsilon^\star = \sum_f \epsilon(f)$ and $\rho(f) = \frac{\epsilon(f)}{\epsilon^\star}$. The later is a distribution on $\mathcal{C}$.

Note that both can be fixed **arbitrarily**, and we have

$$\forall f,\ \eta(f) \quad = \quad \sqrt{2\,\frac{\log\frac{1}{\rho(f)} + \log\frac{1}{\epsilon^\star}}{N}}$$

We can see $\log \frac{1}{\rho(f)}$ as the optimal description length of f. From that point of view, $\eta(f)$ is consistent with the principle of parsimony of William Occam (1280 – 1349)

Entities should not be multiplied unnecessarily.

Picking a classifier with a long description leads to a bad control on the test error.

EXCHANGEABLE SELECTION

How we see that the family of classifiers can be a function of both the training and the test X*s . . .*

VARIABLE FAMILY OF CLASSIFIERS

Consider a family of classifiers which are functions of the sample $\{X_1, \ldots, X_{2N}\}$ in an exchangeable way. For instance with Xs in $\mathbb{R}^k$, one could rank the X_i according to the lexicographic order, and make the f functions of the ordered Xs.

Under such a constraint, the Δ_i remains i.i.d. with the same law, and all our results hold.

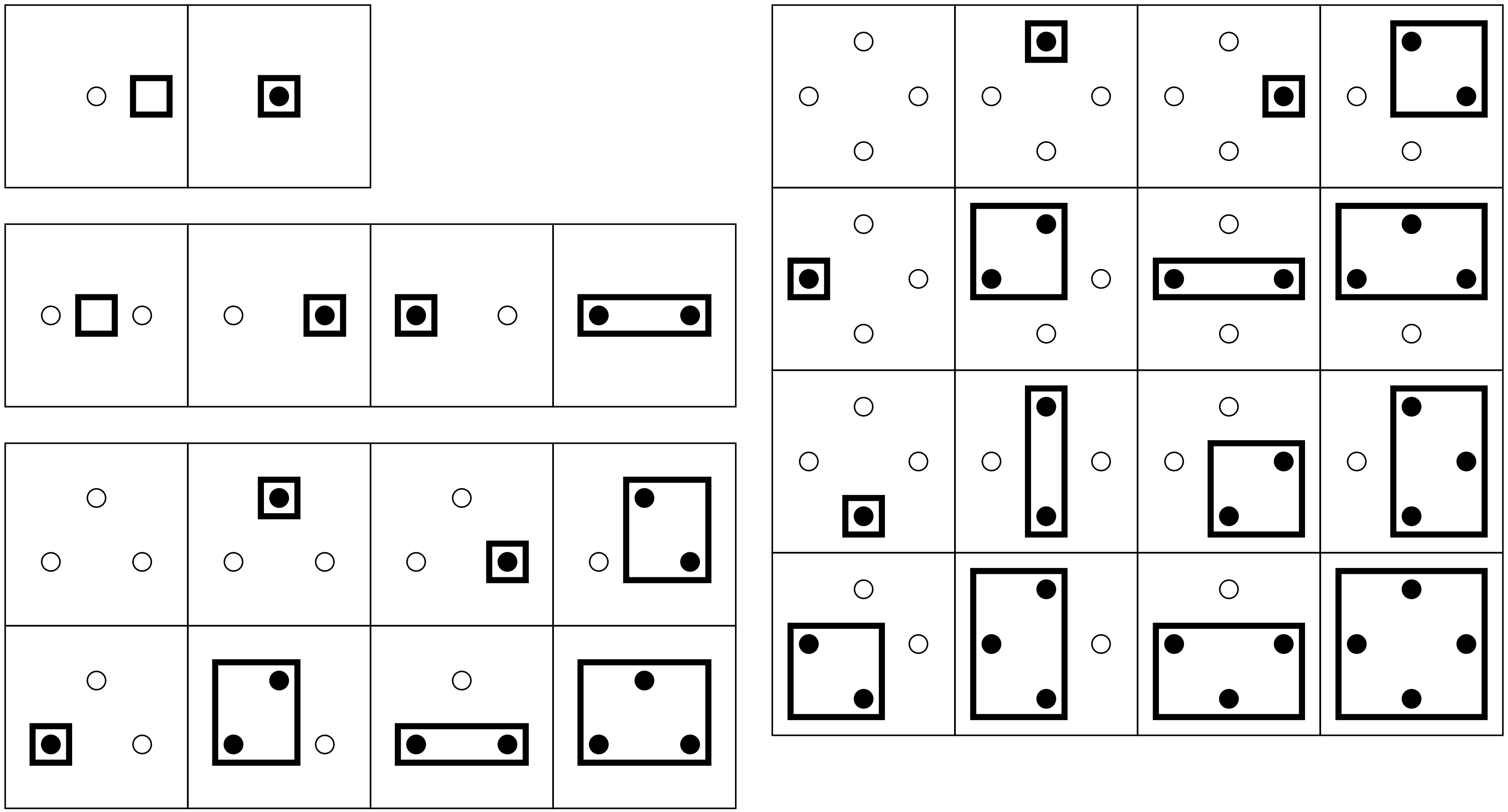
VAPNIK-CHERVONENKIS

How we realize that our classifier sets are not as rich as we though . . .

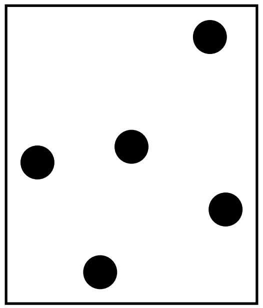
DEFINITION

The Vapnik-Chervonenkis dimension of $\mathcal{C}$ is the largest D so that exists a family $x_1, \ldots, x_D \in \mathcal{X}^D$ which can be arbitrarily labeled with a classifier from $\mathcal{C}$.

Consider for $\mathcal{C}$ the characteristic functions of rectangles. We can find families of 1, 2, 3 or 4 points which can be labelled arbitrarily:



However, given a family of 5 points, if the four *external* points are labelled 1 and the center point labelled 0, than no function from $\mathcal{C}$ can predict that labelling. Hence here $D = 4$.



The VC-dimension is mainly useful because we can compute from it a bound on the number of possible labellings of a family of N points.

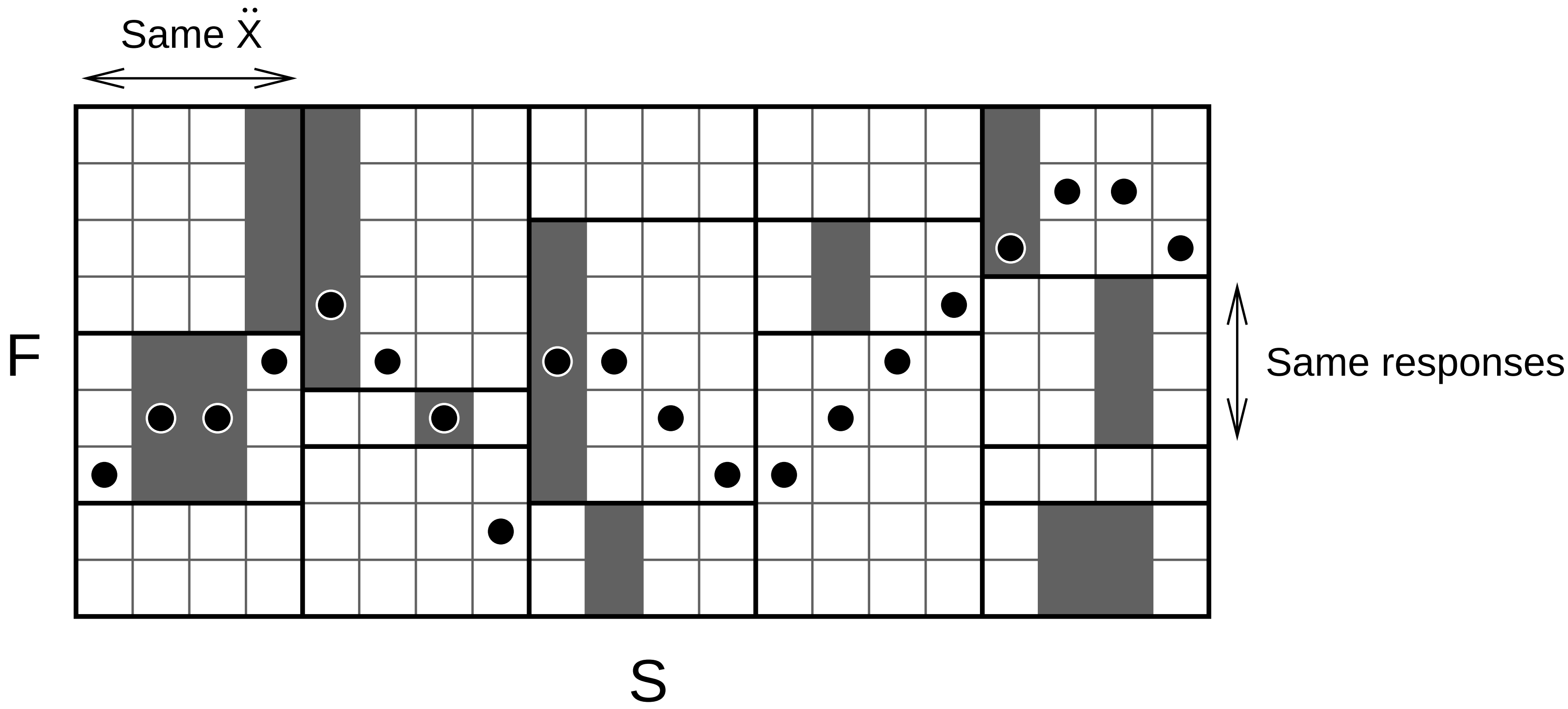
Let $S_{\mathcal{C}}(N)$ be this bound. We have (*Sauer's lemma*)

$$S_{\mathcal{C}}(N) \quad \leq \quad (n+1)^D$$

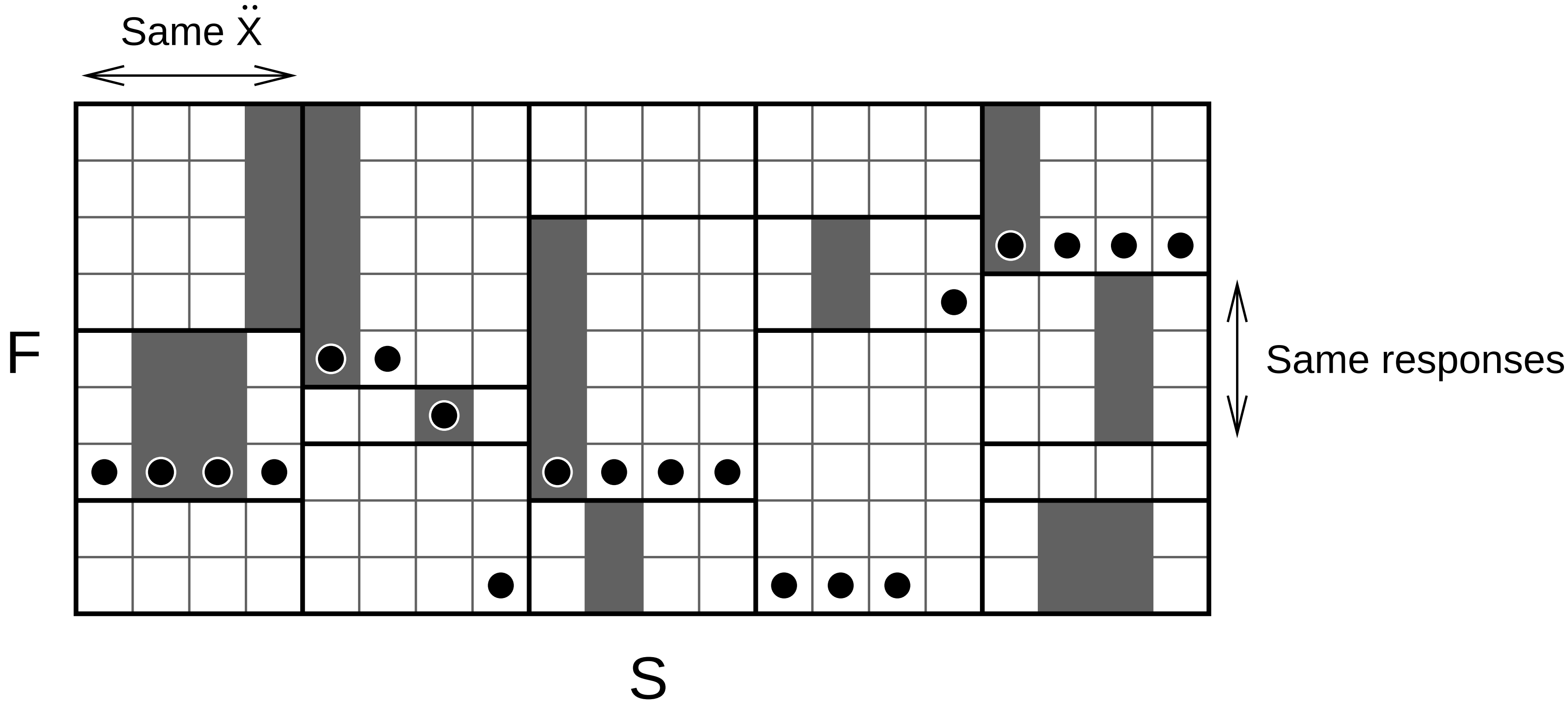
This is far smaller than the number of arbitrary labelings 2^N.

We let $\ddot{X}$ denote the **non-ordered** set $\{X_1, \ldots, X_{2N}\}$ and for $\alpha \subset \mathcal{X}$, let $\mathcal{C}_{|\alpha}$ denote a subset of $\mathcal{C}$ so that two elements of $\mathcal{C}_{|\alpha}$ are not equal when restrained to α. We have:

$$
\begin{aligned}
P(\xi(F,\, S) \geq \eta) &= \sum_{\alpha} P(\xi(F,\, S) \geq \eta \,|\, \ddot{X} = \alpha)\, P(\ddot{X} = \alpha) \\
&= \sum_{\alpha} \sum_{f \in \mathcal{C}_{|\alpha}} P(F_{|\alpha} = f_{|\alpha},\, \xi(F,\, S) \geq \eta \,|\, \ddot{X} = \alpha)\, P(\ddot{X} = \alpha) \\
&\leq \sum_{\alpha} \sum_{f \in \mathcal{C}_{|\alpha}} P(\xi(f,\, S) \geq \eta \,|\, \ddot{X} = \alpha)\, P(\ddot{X} = \alpha) \\
&\leq \sum_{\alpha} S_{\mathcal{C}}(2N) \exp\left(-\frac{1}{2}\,\eta^2\, N\right) P(\ddot{X} = \alpha) \\
&= S_{\mathcal{C}}(2N) \exp\left(-\frac{1}{2}\,\eta^2\, N\right)
\end{aligned}
$$



We group the Ss and fs into blocks of constant $\ddot{X}$ and fs. The bound on the number of gray cells holds in a piece of line in such a block, and we can bound the the number of such blocks for every given S by $S_{\mathcal{C}}(2N)$.



The training algorithm meets as many gray cells as another one which lives in the lowest rows of the blocks.

Contact

François Fleuret
EPFL – CVLAB
francois.fleuret@epfl.ch
http://cvlab.epfl.ch/~fleuret