Introduction to Machine Learning CMU-10701

11. Learning Theory

Barnabás Póczos

MACHINE LEARNING DEPARTMENT
Carnegie Mellon.
School of Computer Science

Learning Theory

We have explored many ways of learning from data

But…

- How good is our classifier, really?
- How much data do we need to make it “good enough”?

Please ask ***Questions***
and give us ***Feedbacks***!

Review of what we have learned so far

Notation

$R(f) = \Pr[Y \neq f(X)]$	$R^* = R(f^*) = \inf_{f:\mathcal{X}\to\mathbb{R}} R(f)$	$f^* = \arg \inf_{f:\mathcal{X}\to\mathbb{R}} R(f)$
	$R^*_{\mathcal{F}} = R(f^*_{\mathcal{F}}) = \inf_{f\in\mathcal{F}} R(f)$	$f^*_{\mathcal{F}} = \arg \inf_{f\in\mathcal{F}} R(f)$
$\widehat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Y_i \neq f(X_i)\}}$	$\widehat{R}^*_{n,\mathcal{F}} = \inf_{f\in\mathcal{F}} \widehat{R}_n(f)$	$f^*_{n,\mathcal{F}} = \arg \min_{f\in\mathcal{F}} \widehat{R}_n(f)$

This is what the learning algorithm produces

We will need these definitions, please copy it!

$R(f)$ = Risk

R^* = Bayes risk

$\widehat{R}_n(f)$ = Empricial risk

f^* = Bayes classifier

$f^*_{n,\mathcal{F}}$ = the classifier that the learning algorithm produces

Big Picture

Ultimate goal: $R(f_n^*) - R^* = 0$

ERM: $f_n^* = f_{n,\mathcal{F}}^* = \arg\min_{f \in \mathcal{F}} \widehat{R}_n(f) = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i))$

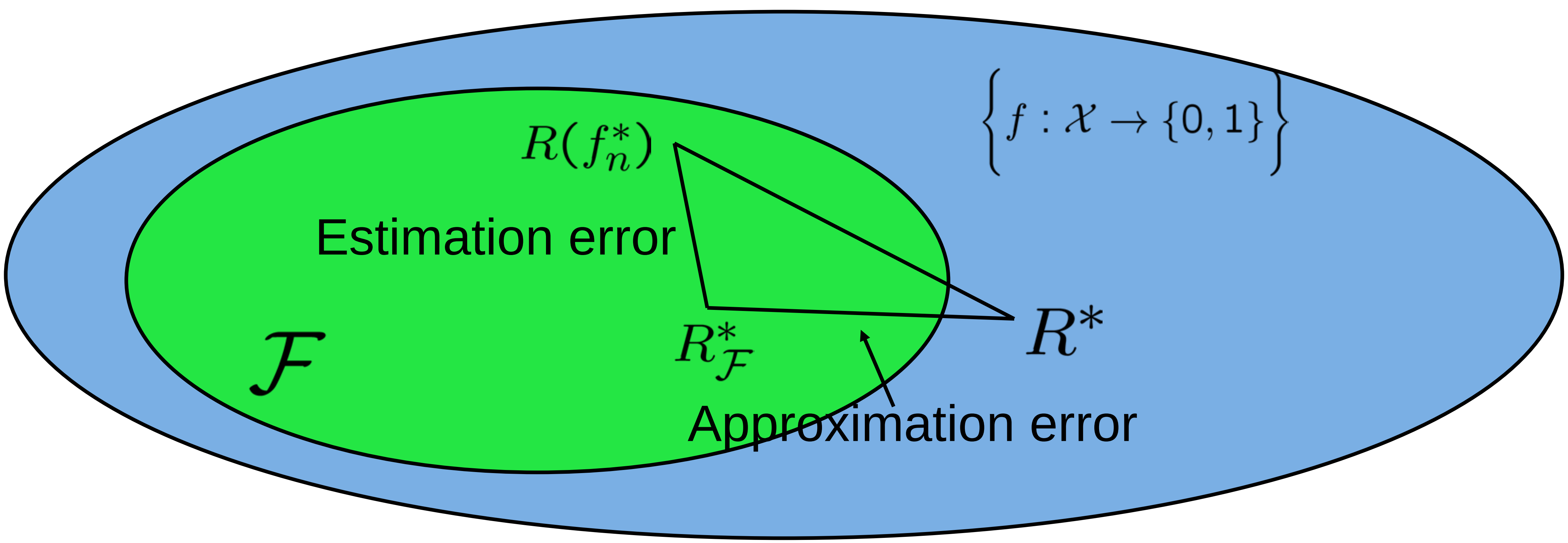
Risk of the classifier f_n^*

$$R(f_n^*) - R^* = \overbrace{R(f_n^*) - R_{\mathcal{F}}^*}^{\text{Estimation error}} + \overbrace{R_{\mathcal{F}}^* - R^*}^{\text{Approximation error}}$$

Bayes risk

Bayes risk

$R_{\mathcal{F}}^* = \inf_{g \in \mathcal{F}} R(g)$ Best classifier in $\mathcal{F}$



Big Picture

Risk of the classifier f_n^*

$$\underbrace{R(f_n^*)}_{\text{Risk of the classifier } f_n^*} - \underbrace{R^*}_{\text{Bayes risk}} = \overbrace{R(f_n^*) - R^*_{\mathcal{F}}}^{\text{Estimation error}} + \overbrace{R^*_{\mathcal{F}} - R^*}^{\text{Approximation error}}$$

Bayes risk

$R^*_{\mathcal{F}} = \inf_{g\in\mathcal{F}} R(g)$ Best classifier in $\mathcal{F}$

$\mathcal{F}$ is too big for a fixed n $\Rightarrow$

- $R^*_{\mathcal{F}}$ is small, close to R^*
- Approximation error is small.
- Estimation error is big.
- Overfitting
- $R(f_n^*)$ is big
- $\widehat{R}_n(f_n^*)$ is small, close to 0.

Big Picture

Risk of the classifier f_n^*

$$\underbrace{R(f_n^*)}_{\text{Risk of the classifier } f_n^*} - \underbrace{R^*}_{\text{Bayes risk}} = \overbrace{R(f_n^*) - R^*_{\mathcal{F}}}^{\text{Estimation error}} + \overbrace{R^*_{\mathcal{F}} - \underbrace{R^*}_{\text{Bayes risk}}}^{\text{Approximation error}}$$

$R^*_{\mathcal{F}} = \inf_{g \in \mathcal{F}} R(g)$ Best classifier in $\mathcal{F}$

$$\left.\begin{array}{l} \mathcal{F} \text{ is too small} \\ \text{e.g. } |\mathcal{F}| < \infty \\ n \text{ is large enough} \end{array}\right\} \Rightarrow$$

- $R^*_{\mathcal{F}}$ is big, far from R^* (usually)
- Approximation error is big.
- The estimation error is small.
- $\widehat{R}_n(f_n^*)$ is close to $R(f_n^*)$ (if n is large enough)
- $R(f_n^*)$ is close to $R^*_{\mathcal{F}}$
- Underfitting: $R^*_{\mathcal{F}}, R(f_n^*)$, $\widehat{R}_n(f_n^*)$ are big.

Big Picture: Illustration of Risks

$$|\widehat{R}_n(f_n^*) - R(f_n^*)| \le \sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)| = \varepsilon$$

$$|R(f_n^*) - R(f_{\mathcal{F}}^*)| \le 2\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)| = 2\varepsilon$$

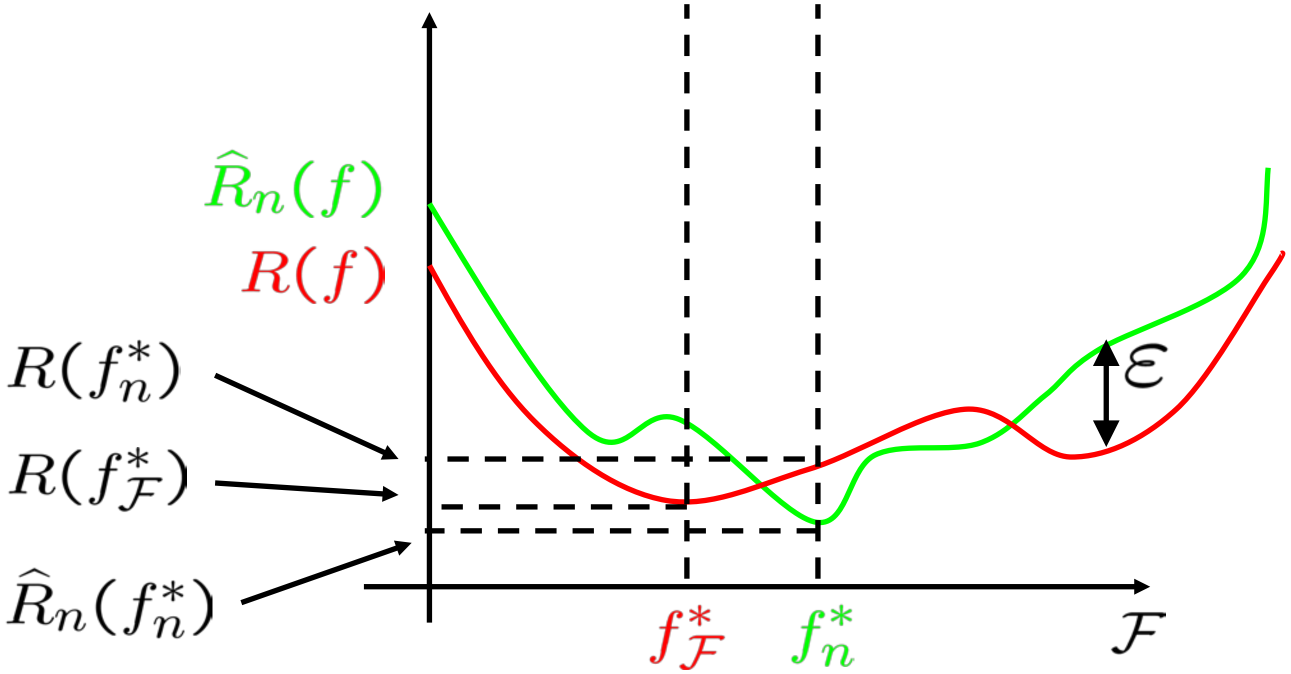
$$|\widehat{R}_n(f_n^*) - R(f_{\mathcal{F}}^*)| \le 3\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)| = 3\varepsilon$$

Upper bound

$\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|$

Goal of Learning:

For a fixed $\mathcal{F}$, make the $|R(f_n^*) - R(f_{\mathcal{F}}^*)|$ estimation error small



11. Learning Theory

Outline

From Hoeffding's inequality, we have seen that

Theorem: Let $\mathcal{F} = \{f : \mathcal{X} \to \{0,1\}\}$, and $|\mathcal{F}| \leq N$

$$\Rightarrow \begin{cases} \Pr\left(\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)| > \varepsilon\right) \leq 2N \exp\left(-2n\varepsilon^2\right) \\ \Pr\left(|\widehat{R}_n(f) - R(f)| < \sqrt{\frac{\log(N) + \log(2/\delta)}{2n}}\right) \geq 1 - \delta \end{cases}$$

These results are useless if N is big, or infinite. (e.g. all possible hyper-planes)

Today we will see how to fix this with the Shattering coefficient and VC dimension

Outline

From Hoeffding's inequality, we have seen that

Theorem: Let $\mathcal{F} = \{f : \mathcal{X} \to \{0,1\}\}$, and $|\mathcal{F}| \le N$

$$\Rightarrow \begin{cases} \Pr\left(\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)| > \varepsilon\right) \le 2N \exp\left(-2n\varepsilon^2\right) \\ \Pr\left(|\widehat{R}_n(f) - R(f)| < \sqrt{\dfrac{\log(N) + \log(2/\delta)}{2n}}\right) \ge 1 - \delta \end{cases}$$

After this fix, we can say something meaningful about this too:

$$|R(f_n^*) - R(f_{\mathcal{F}}^*)| \le 2 \sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)| = 2\varepsilon$$

Best true risk in $\mathcal{F}$

This is what the learning algorithm produces and its true risk

Hoeffding inequality

Theorem: Let $\mathcal{F} = \{f : \mathcal{X} \to \{0,1\}\}$, and $|\mathcal{F}| \leq N$

$$\Rightarrow \Pr\left(\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)| > \varepsilon\right) \leq 2N \exp\left(-2n\varepsilon^2\right)$$

$$\Pr\left(|\widehat{R}_n(f) - R(f)| < \sqrt{\frac{\log(N) + \log(2/\delta)}{2n}}\right) \geq 1 - \delta$$

$$\widehat{R}_n(f) = \arg\min_{f\in\mathcal{F}} \tfrac{1}{n} \textstyle\sum_{i=1}^n 1_{\{Y_i \neq f(X_i)\}}$$

Observation:

It does not matter how many elements $\mathcal{F}$ has. All that matters is how many different behaviours $[f(x_1), \ldots, f(x_n)]$ $f \in \mathcal{F}$ has. (The effective size of $\mathcal{F}$). It can't even be more than 2^n.

McDiarmid's Bounded Difference Inequality

Suppose $X_1, X_2, \ldots, X_n$ are independent and assume that

$$\sup_{x_1,x_2,\ldots,x_n,\widehat{x}_i} |f(x_1,x_2,\ldots,x_n) - f(x_1,x_2,\ldots,x_{i-1},\widehat{x}_i,x_{i+1},\ldots,x_n)| \le c_i \quad \text{for } 1 \le i \le n$$

(**Bounded Difference Assumption**: replacing the i-th coordinate x_i changes the value of f by at most c_i.)

It follows that

$$\Pr\{f(X_1,X_2,\ldots,X_n) - E[f(X_1,X_2,\ldots,X_n)] \ge \varepsilon\} \le \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

$$\Pr\{E[f(X_1,X_2,\ldots,X_n)] - f(X_1,X_2,\ldots,X_n) \ge \varepsilon\} \le \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

$$\Pr\{|E[f(X_1,X_2,\ldots,X_n)] - f(X_1,X_2,\ldots,X_n)| \ge \varepsilon\} \le 2\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

Bounded Difference Condition

Our main goal is to bound $\sup_{f \in \mathcal{F}} |\widehat{R}_n(f) - R(f)|$

Lemma:

The **"bounded difference condition"** is satisfied for $\sup_{f \in \mathcal{F}} |\widehat{R}_n(f) - R(f)|$

Proof:

$$\widehat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{f(X_i) \neq Y_i\}}$$

Let *g* denote the following function:

$$g(Z_1, \ldots, Z_n) = g((X_1, Y_1), \ldots, (X_n, Y_n)) = \sup_{f \in \mathcal{F}} |\widehat{R}_n(f) - R(f)|$$

Observation:

If we change $Z_i = (X_i, Y_i)$, then g can change $c_i = 1/n$ at most.

(Look at how much $\widehat{R}_n(f)$ can change if we change either X_i or Y_i!)

) McDiarmid can be applied for *g!*

Bounded Difference Condition

The **"bounded difference condition"** is satisfied for $\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|$

Corollary:

$$\Pr\{g - \mathbb{E}[g] \geq \varepsilon\} \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

for $g = \sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|$

$c_i = 1/n$

$$\Pr\left\{\left|\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)| - \mathbb{E}[\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|]\right| \geq \varepsilon\right\} \leq 2\exp\left(-2\varepsilon^2 n\right)$$

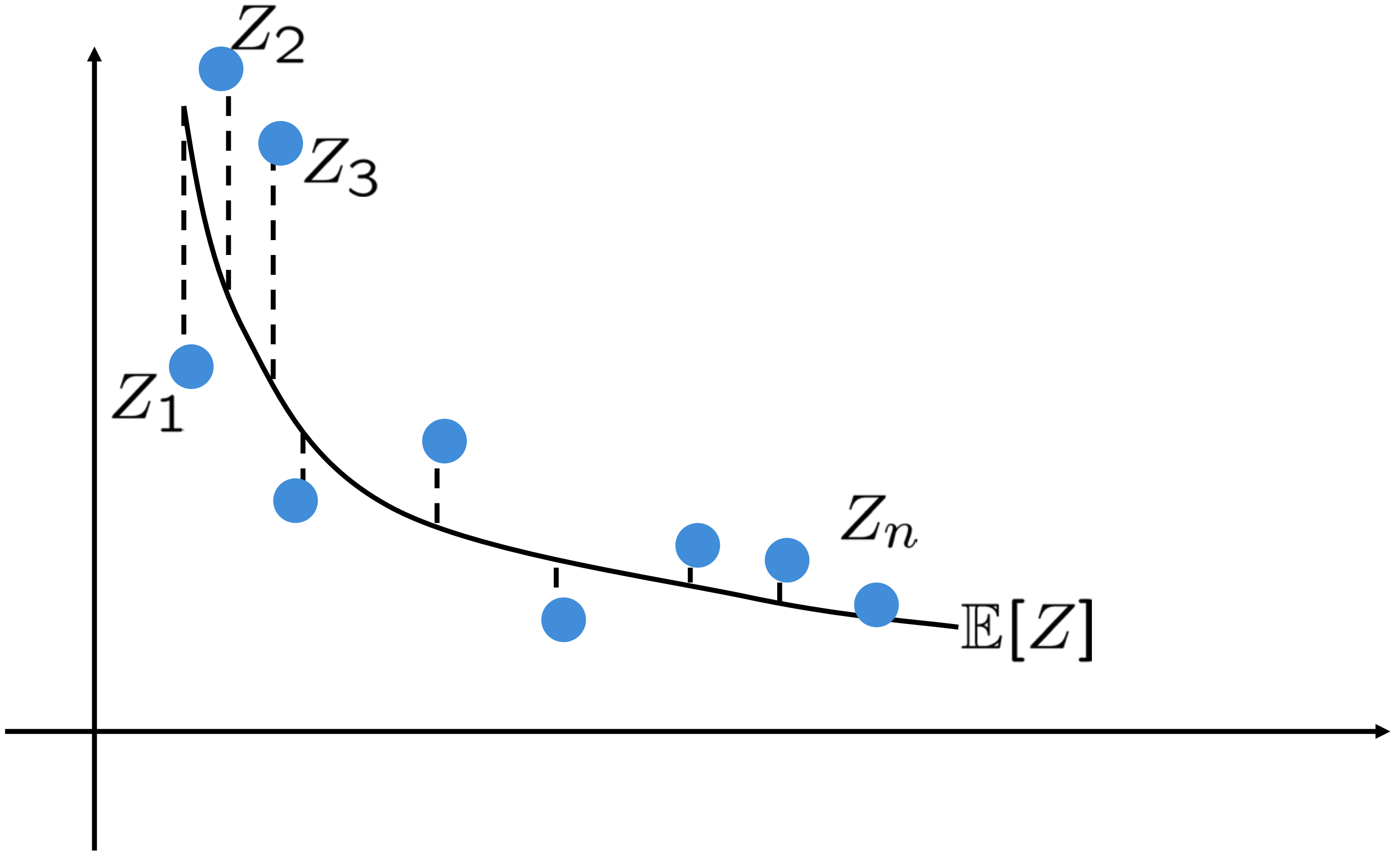
$\Rightarrow |\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|$ is concentrated around its mean!

Therefore, it is enough to study how $\mathbb{E}[\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|]$ behaves.

The Vapnik-Chervonenkis inequality does that with the ***shatter coefficient*** (and ***VC dimension)!***

Concentration and Expected Value

$$Z_n = |\sup_{f \in \mathcal{F}} |\widehat{R}_n(f) - R(f)|$$



Vapnik-Chervonenkis inequality

Our main goal is to bound $\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|$

We already know:

$$\Pr\left\{ \left| \sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)| - \mathbb{E}[\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|] \right| \geq \varepsilon \right\} \leq 2\exp\left(-2\varepsilon^2 n\right)$$

Vapnik-Chervonenkis inequality:

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|\right] \leq 2\sqrt{\frac{\log(2S_{\mathcal{F}}(n))}{n}}$$

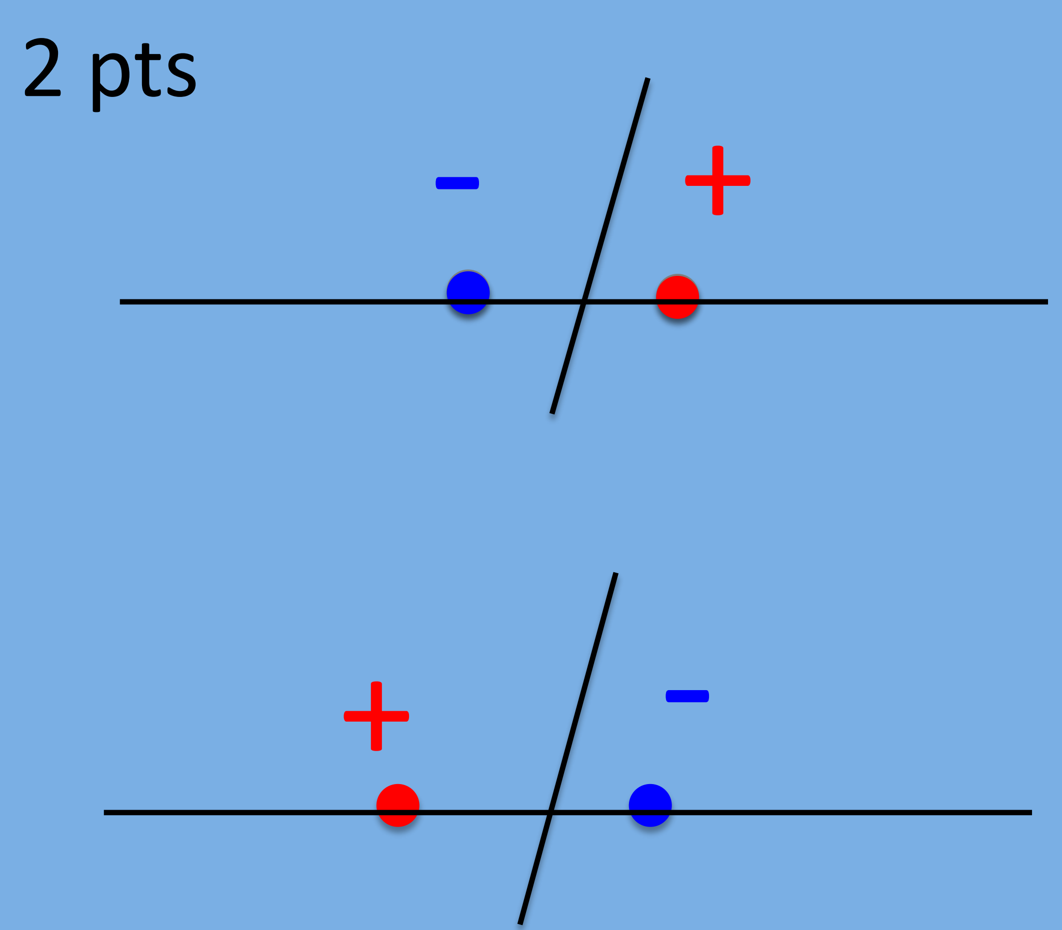
Corollary: Vapnik-Chervonenkis theorem:

$$\Pr\left(\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)| > t\right) \leq 4S_{\mathcal{F}}^2(n)\exp(-nt^2/8)$$

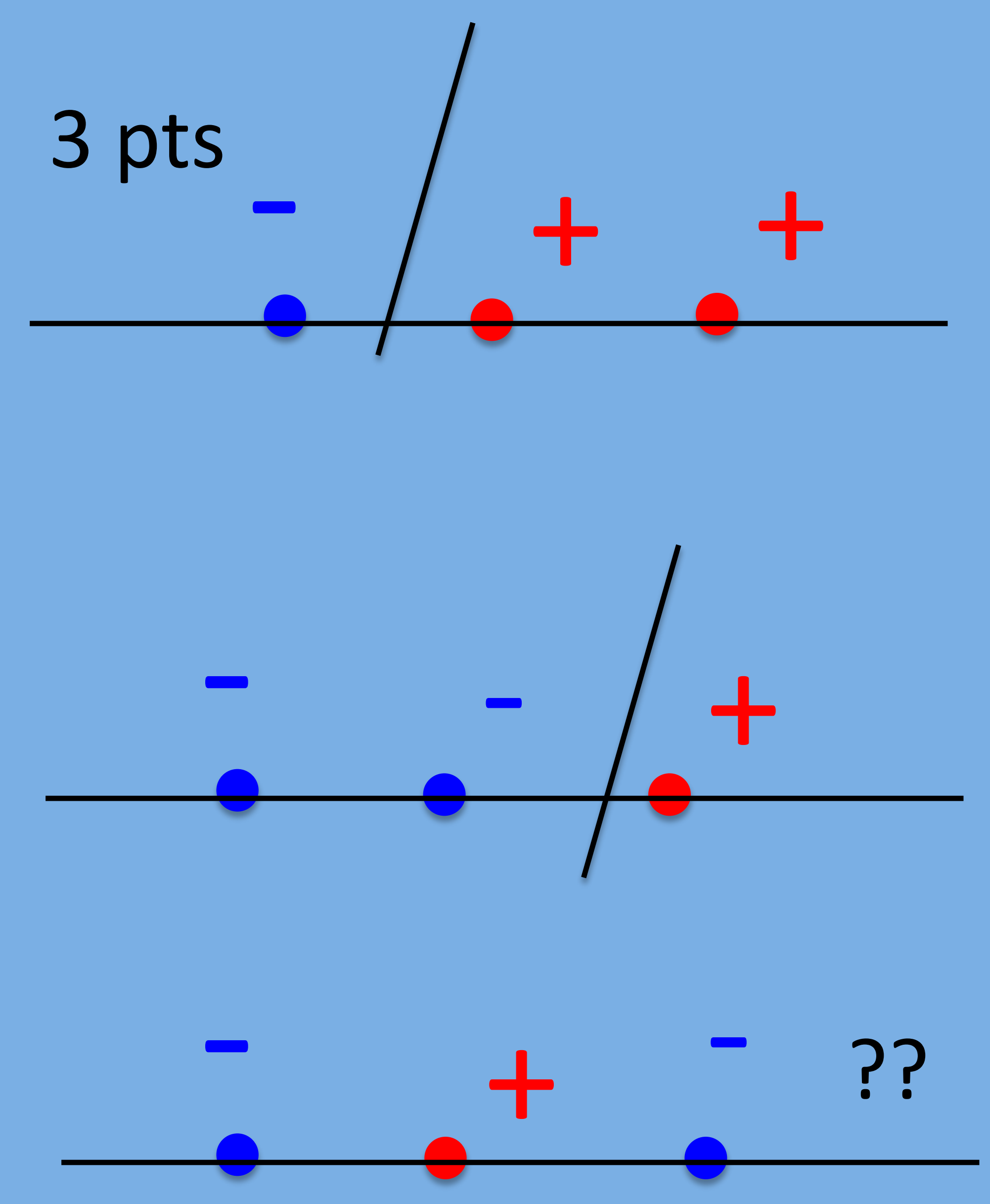
We will define $S_{\mathcal{F}}(n)$ later.

Shattering

How many points can a linear boundary classify exactly in 1D?



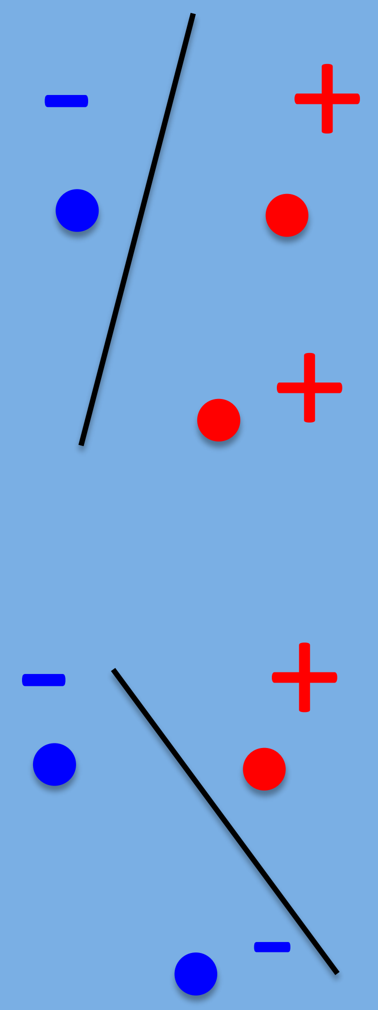
There exists placement s.t.
all labelings can be classified



The answer is 2

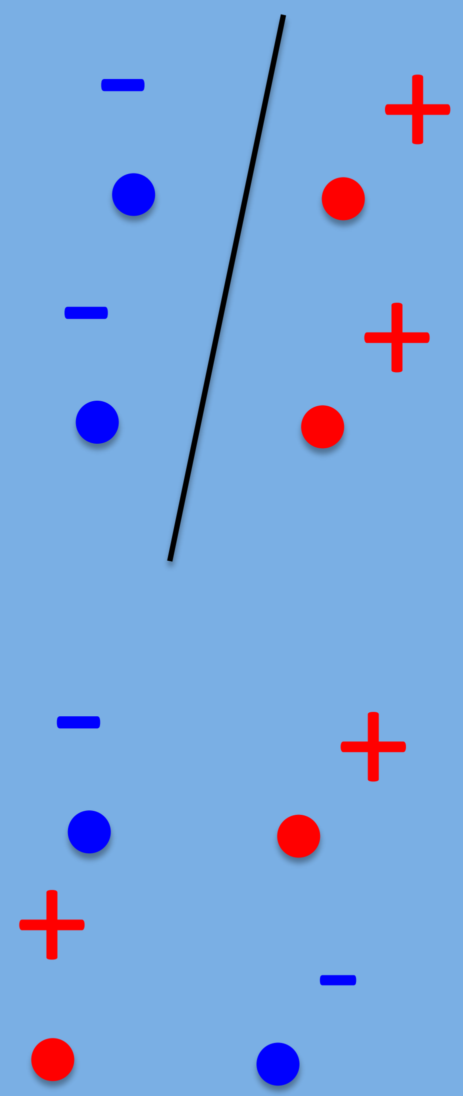
How many points can a linear boundary classify exactly in 2D?

3 pts



There exists placement s.t.
all labelings can be classified

4 pts

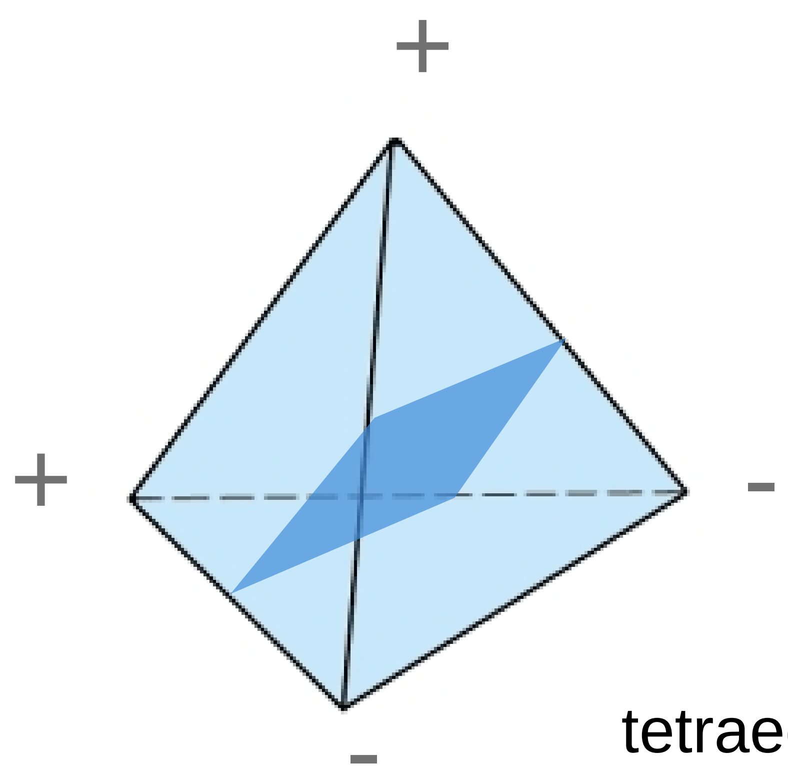


??

The answer is 3

How many points can a linear boundary classify exactly in 3D?

The answer is 4



How many points can a linear boundary classify exactly in d-dim?

The answer is d+1

Growth function, Shatter coefficient

Let $\mathcal{F} = \mathcal{X} \to \{0, 1\}$

How many different behaviour can we get with $[f(x_1), \dots, f(x_n)]$, $f \in \mathcal{F}$?

$|\mathcal{F}| = 7$

	x_1	x_2	x_3
f_1	0	0	0
f_2	0	1	0
f_3	1	1	1
f_4	1	0	0
f_5	0	1	1
f_6	0	1	0
f_7	1	1	1

Definition

$$S_{\mathcal{F}}(x_1, \dots, x_n) = |\{f(x_1), \dots, f(x_n)\}; f \in \mathcal{F}|$$

(=5 in this example)

Growth function, Shatter coefficient

$$S_{\mathcal{F}}(n) = \max_{x_1, \dots, x_n} |\{f(x_1), \dots, f(x_n)\}; f \in \mathcal{F}|$$

maximum number of behaviors on *n* points

Growth function, Shatter coefficient

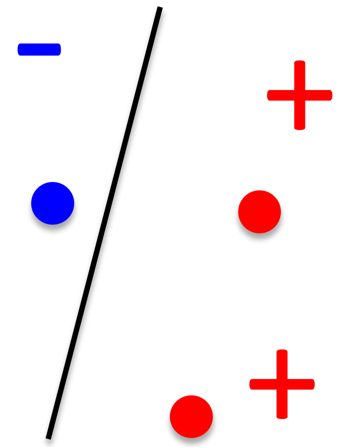
Definition

$$S_{\mathcal{F}}(x_1, \ldots, x_n) = |\{f(x_1), \ldots, f(x_n)\}; f \in \mathcal{F}|$$

Growth function, Shatter coefficient

$$S_{\mathcal{F}}(n) = \max_{x_1, \ldots, x_n} |\{f(x_1), \ldots, f(x_n)\}; f \in \mathcal{F}|$$

maximum number of behaviors on *n* points

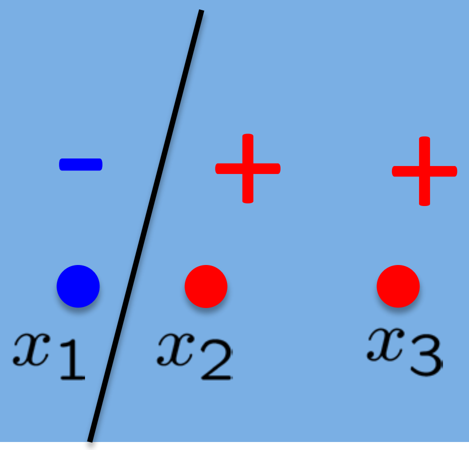


Example: Half spaces in 2D $\Rightarrow S_{\mathcal{F}}(3) = 2^3 = 8$

(Although $\exists x_1, x_2, x_3$ such that $S_{\mathcal{F}}(x_1, x_2, x_3) = 6 < 8$)

$\{\emptyset\}, \{x_1\}, \{x_3\}, \{x_1, x_2\}, \{x_2, x_3\}, \{x_1, x_2, x_3\}$

We can't get $\{x_2\}$ and $\{x_1, x_3\}$



VC-dimension

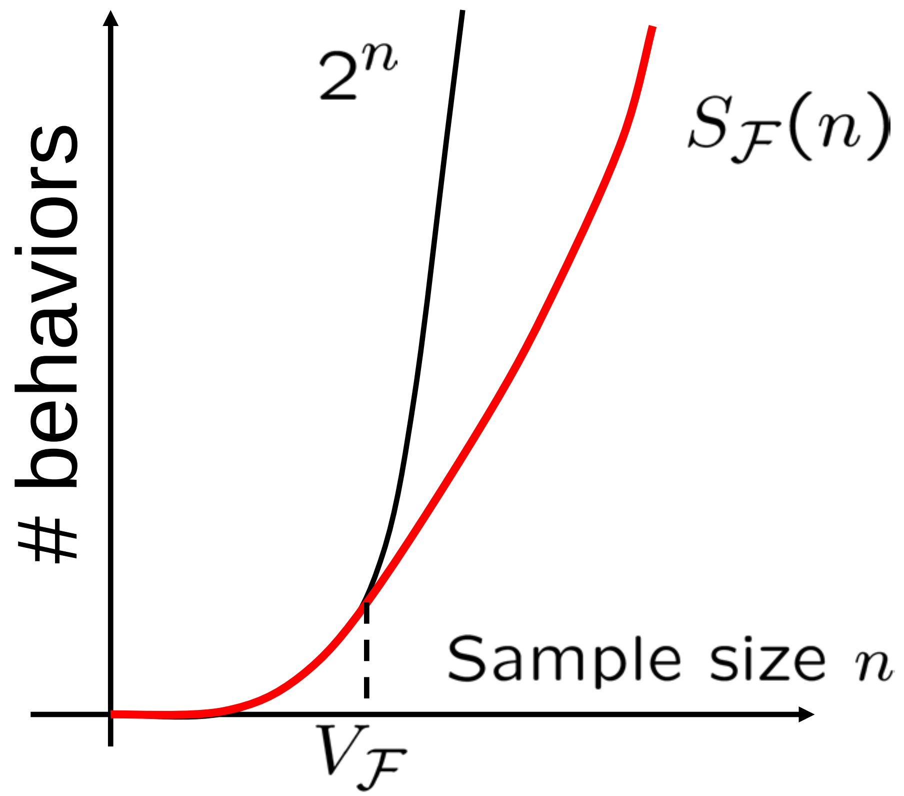
Definition

$$S_{\mathcal{F}}(x_1, \ldots, x_n) = |\{f(x_1), \ldots, f(x_n)\}; f \in \mathcal{F}|$$

Growth function, Shatter coefficient

$$S_{\mathcal{F}}(n) = \max_{x_1, \ldots, x_n} |\{f(x_1), \ldots, f(x_n)\}; f \in \mathcal{F}|$$

maximum number of behaviors on *n* points



Definition: VC-dimension

$$V_{\mathcal{F}} = \max\{n : S_{\mathcal{F}}(n) = 2^n\}$$

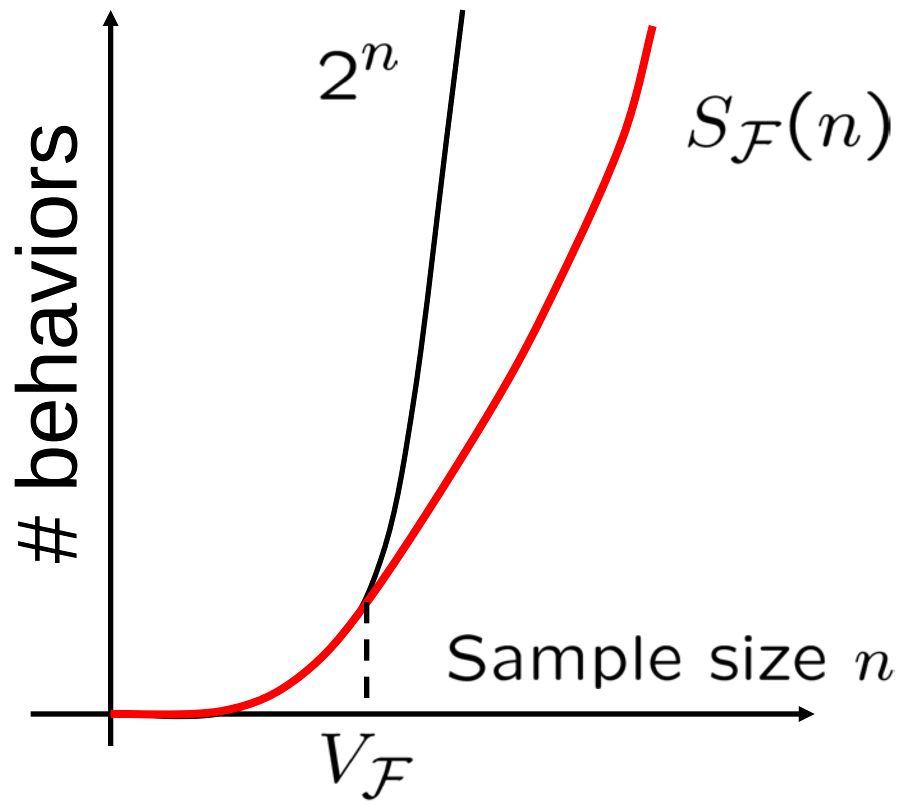
Definition: Shattering

$\mathcal{F}$ shatters the sample $x_1, \ldots, x_n$ iff $\mathcal{F}$ has all the 2^n behaviors on the sample.

Note: $V_{\mathcal{F}}$ is the size of largest shattered sample

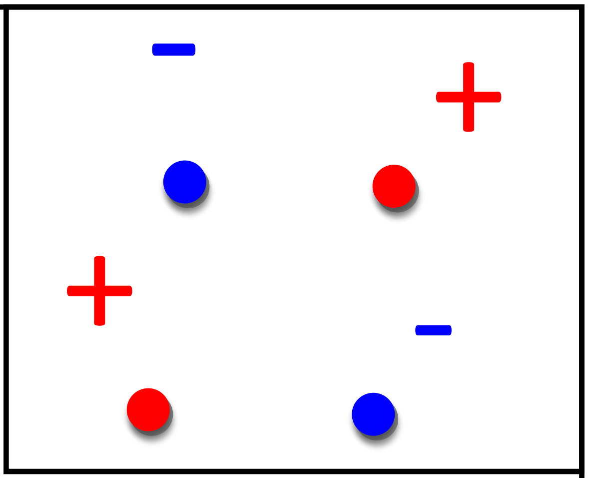
VC-dimension

Definition $V_{\mathcal{F}} = \max\{n : S_{\mathcal{F}}(n) = 2^n\}$



- If the VC dimension is n, then we can find n points that can be shattered, i.e. show 2^n behaviours.
- $n + 1$ points never show 2^{n+1} behaviours.

VC-dimension



- You pick set of points $x_1, \ldots, x_n$
- Adversary assigns labels $y_1, \ldots, y_n$
- If $VC_{\mathcal{F}} \geq n$, then you find a hypothesis f in $\mathcal{F}$ consistent with the labels, i.e. $f(x_i) = y_i \quad (1 \leq i \leq n)$
- If $VC_{\mathcal{F}} = n$, then for any $n+1$ points, there exists a labeling that cannot be shattered (can't find a hypothesis f in $\mathcal{F}$ consistent with it)

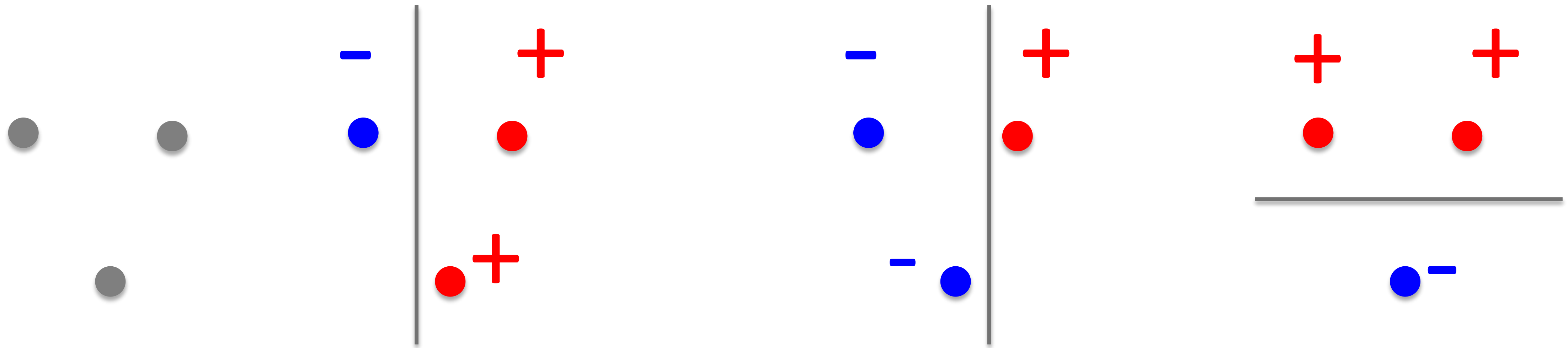
The VC dimension measure how rich $\mathcal{F}$ is.

If the VC dimension is high, e.g. ∞, then it is easy to overfit!

Examples

VC dim of decision stumps (axis aligned linear separator) in 2d

What's the VC dim. of decision stumps in 2d?

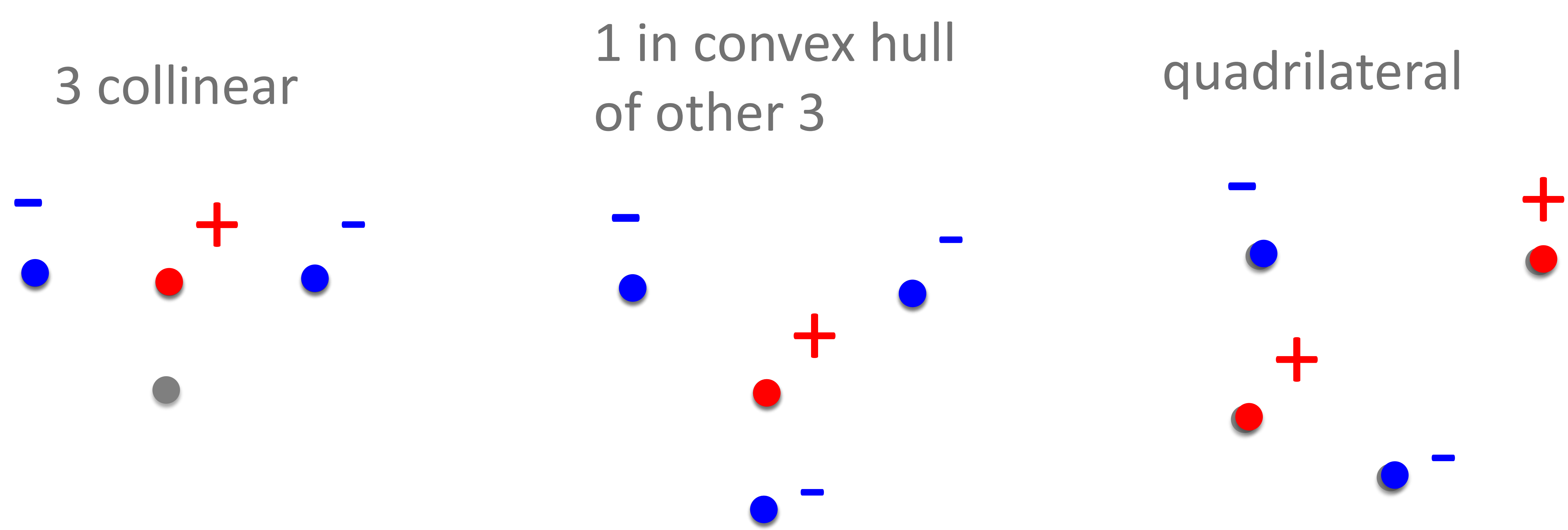


There is a placement of 3 pts that can be shattered) VC dim ≥ 3

VC dim of decision stumps (axis aligned linear separator) in 2d

What's the VC dim. of decision stumps in 2d?

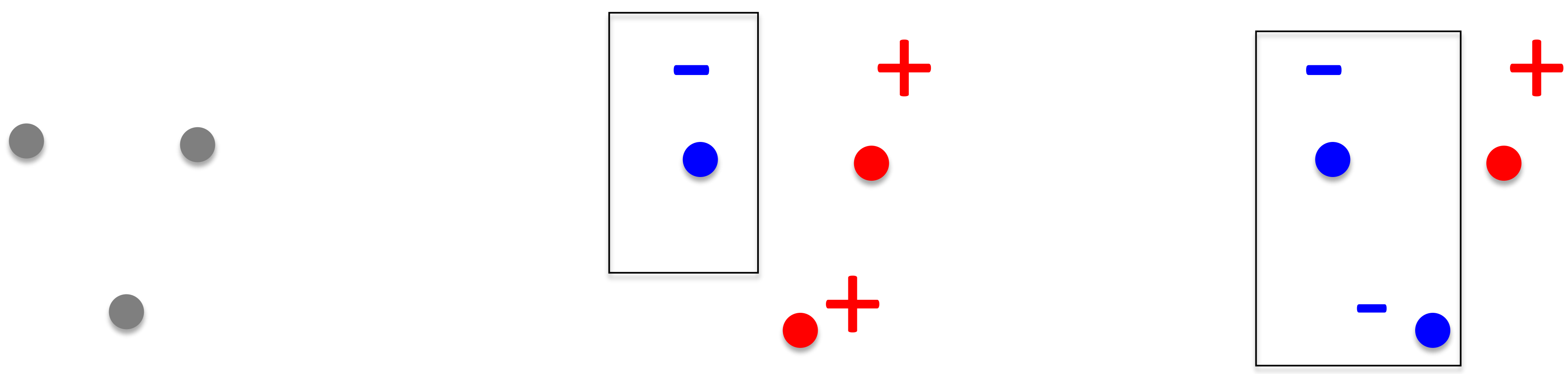
If VC dim = 3, then for all placements of 4 pts, there exists a labeling that can't be shattered



VC dim. of axis parallel rectangles in 2d

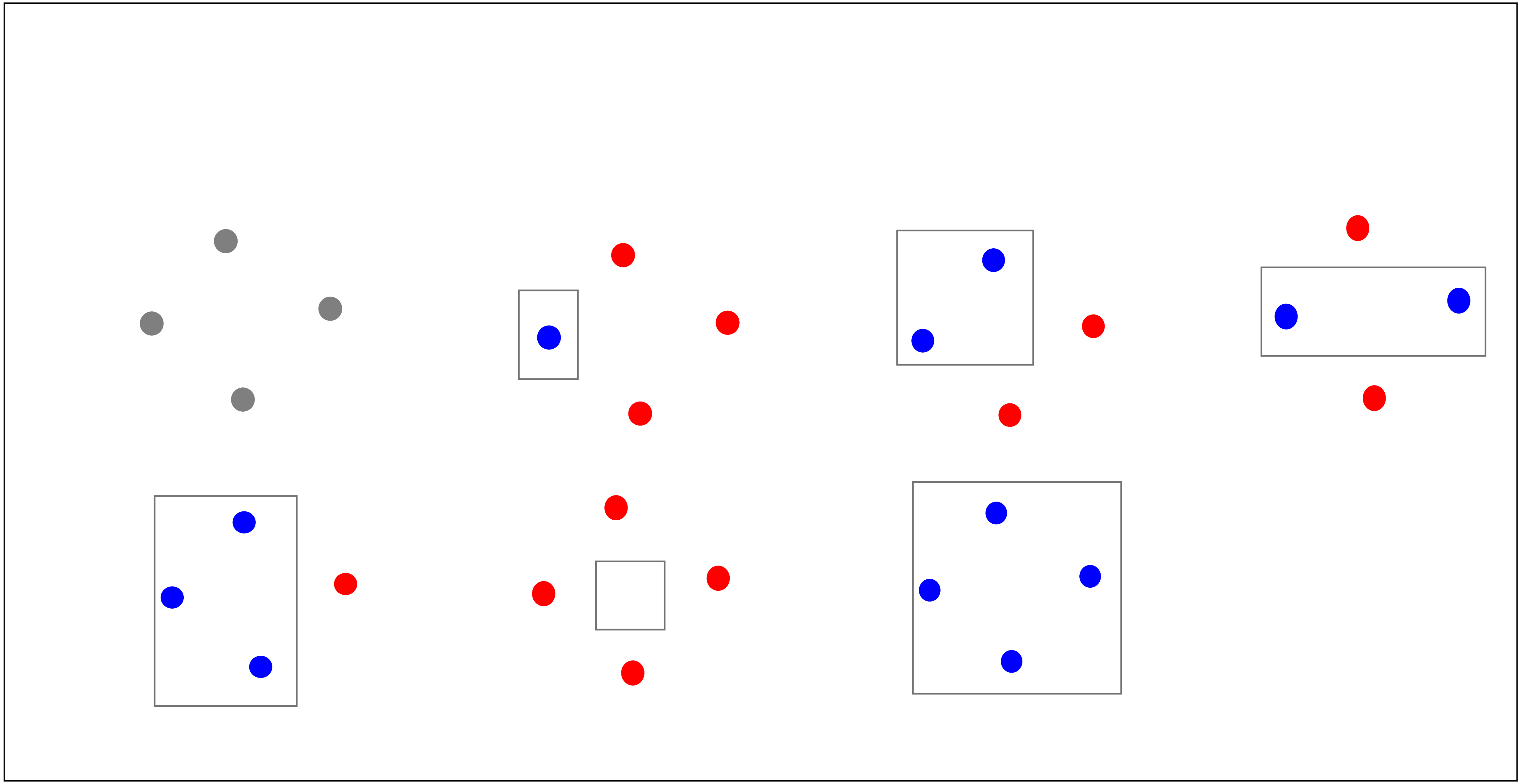
What's the VC dim. of axis parallel rectangles in 2d?

$$f(x) = \text{sign}(1 - 2 \cdot 1_{\{x \in \text{ rectangle}\}})$$



There is a placement of 3 pts that can be shattered) VC dim ≥ 3

VC dim. of axis parallel rectangles in 2d



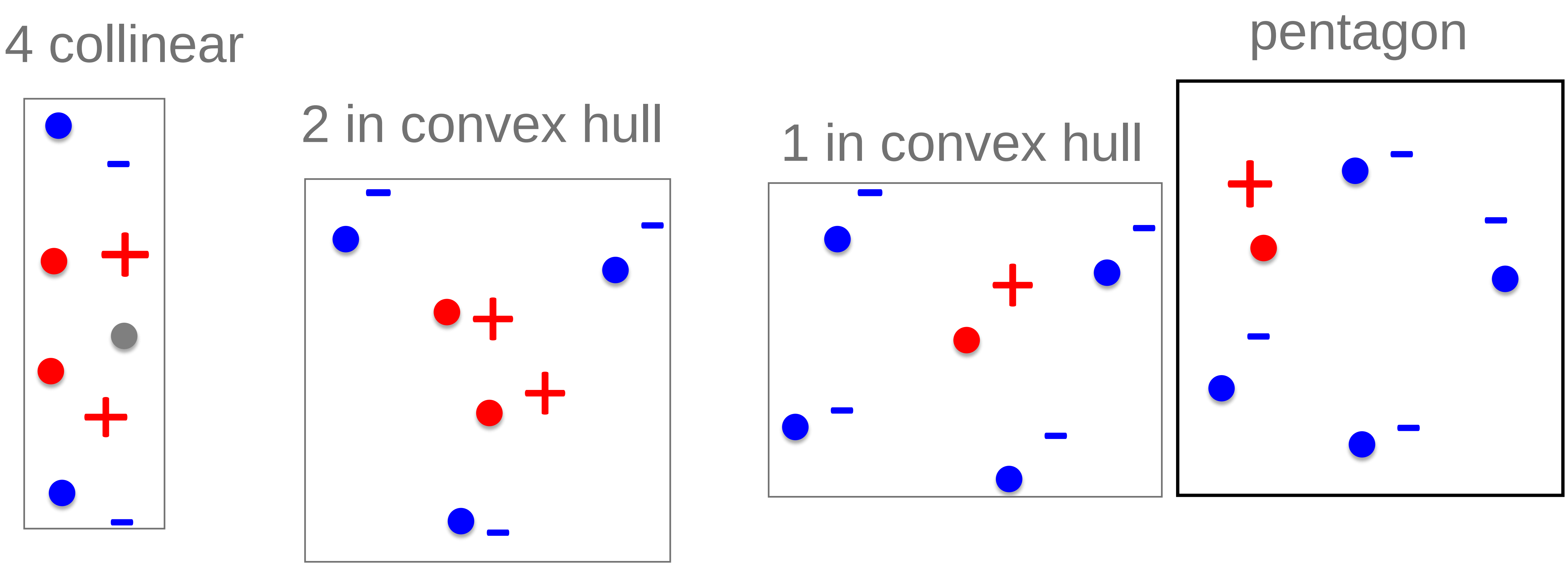
There is a placement of 4 pts that can be shattered) VC dim ≥ 4

VC dim. of axis parallel rectangles in 2d

What's the VC dim. of axis parallel rectangles in 2d?

$$f(x) = \text{sign}(1 - 2 \cdot 1_{\{x \in \text{ rectangle}\}})$$

If VC dim = 4, then for all placements of 5 pts, there exists a labeling that can't be shattered



Sauer's Lemma

We already know that $S_{\mathcal{F}}(n) \leq 2^n$ [Exponential in *n*]

Sauer's lemma:

$$S_{\mathcal{F}}(n) \leq \sum_{k=0}^{VC_{\mathcal{F}}} \binom{n}{k}$$

The VC dimension can be used to upper bound the shattering coefficient.

Corollary:

$$S_{\mathcal{F}(n)} \leq (n+1)^{VC_{\mathcal{F}}}$$

[Polynomial in *n*]

$$S_{\mathcal{F}}(n) \leq \left(\frac{ne}{VC_{\mathcal{F}}}\right)^{VC_{\mathcal{F}}}$$

Proof of Sauer's Lemma

Write all different behaviors on a sample $(x_1, x_2, \ldots x_n)$ in a matrix:

$\lvert\mathcal{F}\rvert = 7$	x_1	x_2	x_3
f_1	0	0	0
f_2	0	1	0
f_3	1	1	1
f_4	1	0	0
f_5	0	1	0
f_6	1	1	1
f_7	0	1	1

$\lvert\mathcal{F}\rvert = 7$	x_1	x_2	x_3
f_1	0	0	0
f_2	0	1	0
f_3	1	1	1
f_4	1	0	0
f_7	0	1	1

Proof of Sauer's Lemma

$$|\mathcal{F}| = 7 \quad \begin{array}{c|ccc} & x_1 & x_2 & x_3 \\ \hline f_1 & 0 & 0 & 0 \\ f_2 & 0 & 1 & 0 \\ f_3 & 1 & 1 & 1 \\ f_4 & 1 & 0 & 0 \\ f_7 & 0 & 1 & 1 \end{array} = A$$

Shattered subsets of columns:

$$\{\emptyset\}, \{x_1\}, \{x_2\}, \{x_3\}, \{x_1, x_2\}, \{x_1, x_3\}$$

We will prove that

$$S_{\mathcal{F}}(x_1, \ldots, x_n) = \#\ \text{rows}(A) \leq \#\ \text{shattered subsets of columns of } A \leq \sum_{k=0}^{VC_{\mathcal{F}}} \binom{n}{k}$$

Therefore,

$$S_{\mathcal{F}(n)} = \max_{x_1, \ldots, x_n} S_{\mathcal{F}}(x_1, \ldots, x_n) \leq \sum_{k=0}^{VC_{\mathcal{F}}} \binom{n}{k}$$

Proof of Sauer's Lemma

$$|\mathcal{F}| = 7 \quad \begin{array}{c|ccc|} & x_1 & x_2 & x_3 \\ \hline f_1 & 0 & 0 & 0 \\ f_2 & 0 & 1 & 0 \\ f_3 & 1 & 1 & 1 \\ f_4 & 1 & 0 & 0 \\ f_7 & 0 & 1 & 1 \\ \hline \end{array} = A$$

Shattered subsets of columns:

$$\{\emptyset\}, \{x_1\}, \{x_2\}, \{x_3\}, \{x_1, x_2\}, \{x_1, x_3\}$$

Lemma 1 $\#$ shattered subsets of columns of $A \leq \sum_{k=0}^{VC_{\mathcal{F}}} \binom{n}{k}$

In this example: 6· 1+3+3=7

Lemma 2 $\#$ rows$(A) \leq \#$ shattered subsets of columns of A

for any binary matrix with no repeated rows.

In this example: 5· 6

Proof of Lemma 1

$$|\mathcal{F}| = 7 \quad \begin{array}{c|ccc|} & x_1 & x_2 & x_3 \\ \hline f_1 & 0 & 0 & 0 \\ f_2 & 0 & 1 & 0 \\ f_3 & 1 & 1 & 1 \\ f_4 & 1 & 0 & 0 \\ f_7 & 0 & 1 & 1 \\ \hline \end{array} = A$$

Shattered subsets of columns:

$$\{\emptyset\}, \{x_1\}, \{x_2\}, \{x_3\}, \{x_1, x_2\}, \{x_1, x_3\}$$

In this example: 6· 1+3+3=7

Lemma 1 $\#$ shattered subsets of columns of $A \leq \sum_{k=0}^{VC_{\mathcal{F}}} \binom{n}{k}$

Proof

$VC_{\mathcal{F}}$ is the size of largest imaginable shattered sample. $VC_{\mathcal{F}} = \max\{n : S_{\mathcal{F}}(n) = 2^n\}$

If a shattered subsets of columns has d elements, then $VC_{\mathcal{F}} \geq d$

For example if $\{x_1, x_3\}$ are shattered in A, then $VC_{\mathcal{F}} \geq 2$.

Proof of Lemma 2

Lemma 2 $\#\ \mathsf{rows}(A) \leq \#\ \mathsf{shattered\ subsets\ of\ columns\ of}\ A$

for any binary matrix with no repeated rows.

Proof Induction on the number of columns

Base case: A has one column. There are three cases:

$A = (0)$) 1 · 1 shattered subsets of columns: $\{\emptyset\}$

$A = (1)$) 1 · 1 shattered subsets of columns: $\{\emptyset\}$

$A = \binom{0}{1}$) 2 · 2 shattered subsets of columns: $\{\emptyset\}, \{x_1\}$

Proof of Lemma 2

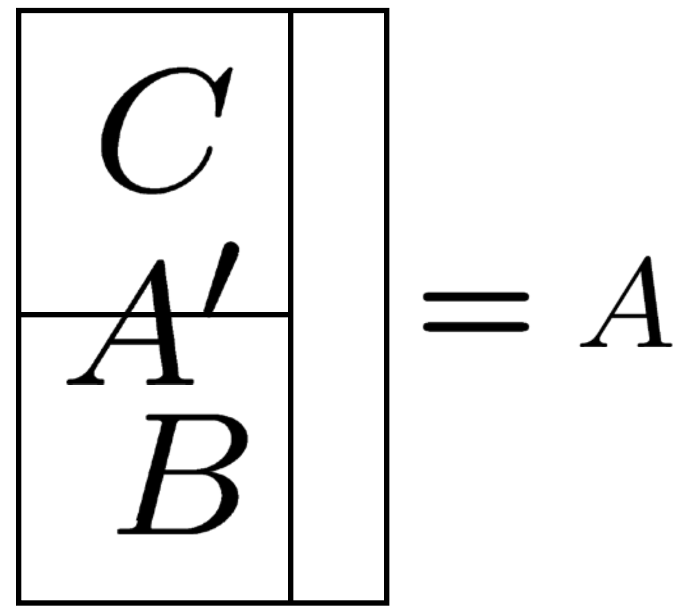
Inductive case: A has at least two columns.

Let A' be A minus its last column x_m removed

In A' each row can occure once or twice.

If "twice" $\Rightarrow$ move one of them to B the other to C

If "once" $\Rightarrow$ move them to C



We have,

$$\# \text{ rows}(A) = \# \text{ rows}(B) + \# \text{ rows}(C)$$

$$\leq \# \text{ shattered subsets of columns of } (B)$$

$$+ \# \text{ shattered subsets of columns of } (C)$$

By induction (less columns)

0	0	0
0	1	0
1	1	1
1	0	0
0	1	1

Proof of Lemma 2

$\{\emptyset\}$ $\qquad\qquad$ $\{\emptyset\}, \{x_1\}, \{x_2\}\{x_1, x_2\}$

$\#$ shattered subsets of columns of (B) + $\#$ shattered subsets of columns of (C)

$\leq$ $\#$ shattered subsets of columns of (A)

$\{\emptyset\}, \{x_1\}, \{x_2\}, \{x_3\}, \{x_1, x_2\}, \{x_1, x_3\}$

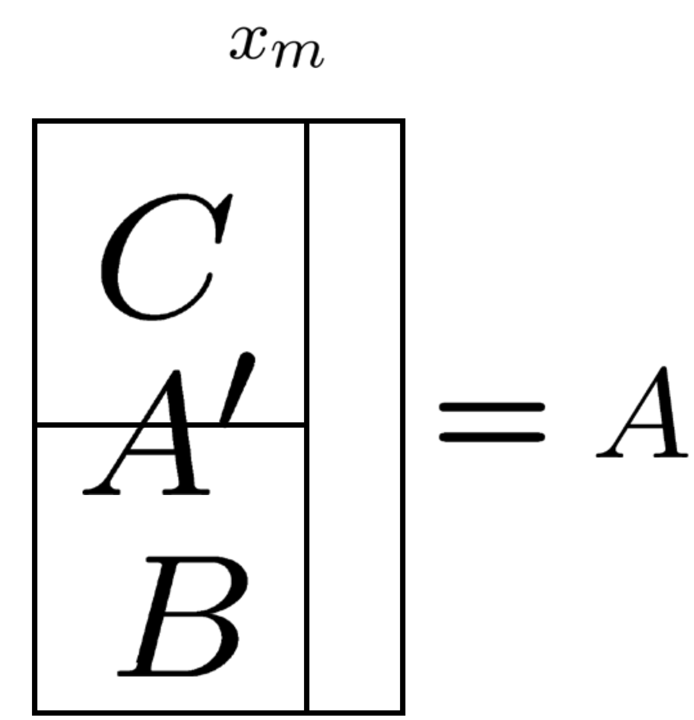
because

"once" $\Rightarrow$ move them to C

Therefore, if C shatters S e.g. $\{x_1, x_2\}$, then A shatters S.

"twice" $\Rightarrow$ move one of them to B the other to C

Therefore, if B shatters S, then A shatters $S \cup x_m$.



0	0	0
0	1	0
1	1	1
1	0	0
0	1	1

Vapnik-Chervonenkis inequality

When $|\mathcal{F}| = N < \infty$, we already know $\mathbb{E}\left[\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|\right] \le \sqrt{\frac{\log(2N)}{2n}}$

Vapnik-Chervonenkis inequality: [We don't prove this]

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|\right] \le 2\sqrt{\frac{\log(2S_{\mathcal{F}}(n))}{n}}$$

From Sauer's lemma:

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|\right] \le 2\sqrt{\frac{\log(2S_{\mathcal{F}}(n))}{n}} \le 2\sqrt{\frac{VC_{\mathcal{F}}\log(n+1) + \log 2}{n}}$$

Since $|R(f_n^*) - R(f_{\mathcal{F}}^*)| \le 2\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|$

Therefore, $\mathbb{E}[|R(f_n^*) - R(f_{\mathcal{F}}^*)|] \le 4\sqrt{\frac{VC_{\mathcal{F}}\log(n+1) + \log 2}{n}}$

Yayyy! ☺

Estimation error

Linear (hyperplane) classifiers

We already know that

$$\mathbb{E}[|R(f_n^*) - R(f_{\mathcal{F}}^*)|] \leq 4\sqrt{\frac{VC_{\mathcal{F}} \log(n+1) + \log 2}{n}}$$

Estimation error

For linear classifiers in dimension when $\mathcal{X} = \mathbb{R}^d$: $VC_{\mathcal{F}} = d + 1$.

$$\Rightarrow \mathbb{E}[|R(f_n^*) - R(f_{\mathcal{F}}^*)|] \leq 4\sqrt{\frac{(d+1) \log(n+1) + \log 2}{n}}$$

Estimation error

If we do feature map first, $x = \phi(x) \in \mathbb{R}^{d'}$, then linear separation (SVM) $\Rightarrow VC_{\mathcal{F}} = d' + 1$.

Estimation error

$$\Rightarrow \mathbb{E}[|R(f_n^*) - R(f_{\mathcal{F}}^*)|] \leq 4\sqrt{\frac{(d'+1) \log(n+1) + \log 2}{n}}$$

Vapnik-Chervonenkis Theorem

We already know from McDiarmid:

$$\Pr\left\{\left|\sup_{f\in\mathcal{F}}|\widehat{R}_n(f)-R(f)|-\mathbb{E}[\sup_{f\in\mathcal{F}}|\widehat{R}_n(f)-R(f)|]\right|\geq\varepsilon\right\}\leq 2\exp\left(-2\varepsilon^2 n\right)$$

Vapnik-Chervonenkis inequality: $\mathbb{E}\left[\sup_{f\in\mathcal{F}}|\widehat{R}_n(f)-R(f)|\right]\leq 2\sqrt{\frac{\log(2S_{\mathcal{F}}(n))}{n}}$

Corollary: Vapnik-Chervonenkis theorem: [We don't prove them]

$$\Pr\left(\sup_{f\in\mathcal{F}}|\widehat{R}_n(f)-R(f)|>t\right)\leq 4S_{\mathcal{F}}(2n)\exp(-nt^2/8)$$

$$\Pr\left(\sup_{f\in\mathcal{F}}|\widehat{R}_n(f)-R(f)|>t\right)\leq 8S_{\mathcal{F}}(n)\exp(-nt^2/32)$$

Hoeffding + Union bound for finite function class:

When $|\mathcal{F}|=N<\infty,\quad\Rightarrow\Pr\left(\sup_{f\in\mathcal{F}}|\widehat{R}_n(f)-R(f)|>t\right)\leq 2N\exp\left(-2nt^2\right)$

PAC Bound for the Estimation Error

VC theorem: $$\Pr\left(\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)| > t\right) \le 8S_{\mathcal{F}}(n)\exp(-nt^2/32)$$

Inversion: $$8S_{\mathcal{F}}(n)\exp(-nt^2/32) \le \delta \qquad \Rightarrow t^2 \ge \frac{32}{n}\log\left(\frac{8S_{\mathcal{F}}(n)}{\delta}\right)$$

$$\Rightarrow \Pr\left(\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)| \le 8\sqrt{\frac{\log(S_{\mathcal{F}}(n)) + \log\left(\frac{8}{\delta}\right)}{2n}}\right) \ge 1-\delta$$

$$S_{\mathcal{F}}(n) \le \left(\frac{ne}{VC_{\mathcal{F}}}\right)^{VC_{\mathcal{F}}} \qquad \Rightarrow \Pr\left(\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)| \le 8\sqrt{\frac{VC_{\mathcal{F}}\log\left(\frac{ne}{VC_{\mathcal{F}}}\right) + \log\left(\frac{8}{\delta}\right)}{2n}}\right) \ge 1-\delta$$

Don't forget that $|R(f_n^*) - R(f_{\mathcal{F}}^*)| \le 2\sup_{f\in\mathcal{F}} |\widehat{R}_n(f) - R(f)|$

Estimation error

$$\Rightarrow \Pr\left(|R(f_n^*) - R(f_{\mathcal{F}}^*)| \le 16\sqrt{\frac{\log(VC_{\mathcal{F}}\log\left(\frac{ne}{VC_{\mathcal{F}}}\right) + \log\left(\frac{8}{\delta}\right)}{2n}}\right) \ge 1-\delta$$

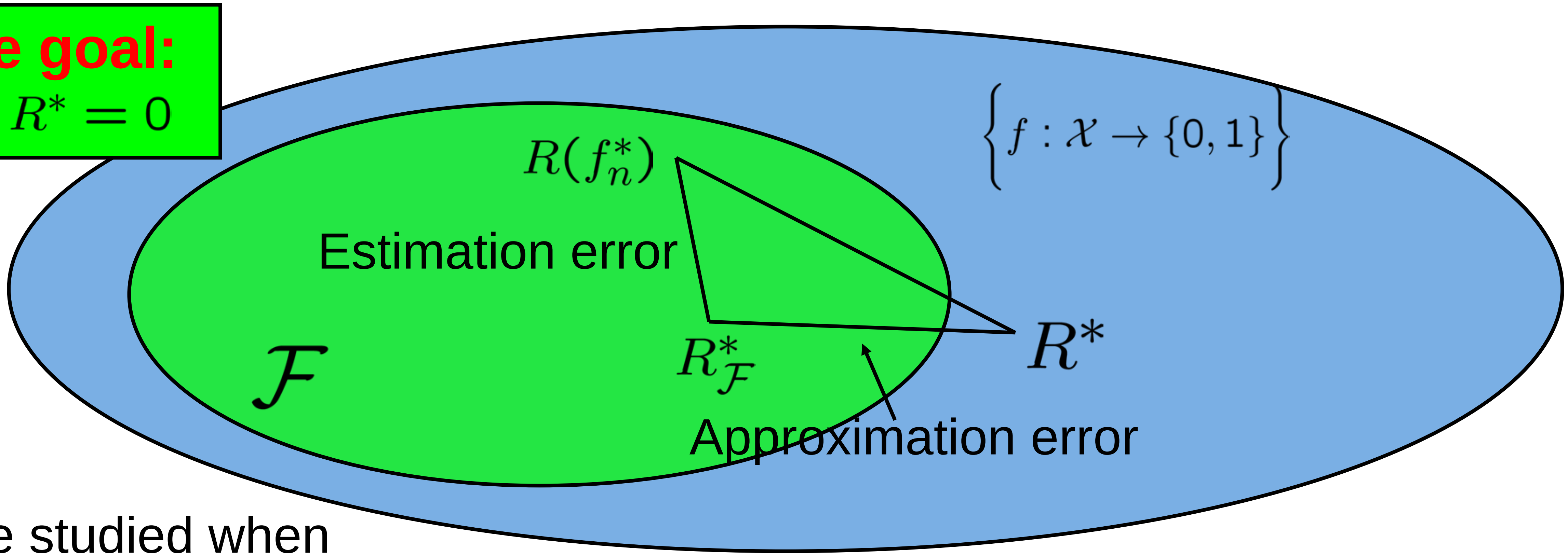
Structoral Risk Minimization

Risk of the classifier f_n^*

$$R(f_n^*) - R^* = \overbrace{R(f_n^*) - R^*_{\mathcal{F}}}^{\text{Estimation error}} + \overbrace{R^*_{\mathcal{F}} - R^*}^{\text{Approximation error}}$$

Bayes risk

Ultimate goal: $R(f_n^*) - R^* = 0$



So far we studied when estimation error ! 0, but we also want approximation error ! 0

Let $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \ldots \subseteq \mathcal{F}_n \subseteq \ldots$ such that $VC_{\mathcal{F}_1} \leq VC_{\mathcal{F}_2} \leq \ldots \leq VC_{\mathcal{F}_n} \leq \ldots$

Many different variants…

penalize too complex models to avoid overfitting

What you need to know

Complexity of the classifier depends on number of points that can be classified exactly

Finite case – Number of hypothesis
Infinite case – Shattering coefficient, VC dimension

PAC bounds on true error in terms of empirical/training error and complexity of hypothesis space

Empirical and Structural Risk Minimization

Thanks for your attention ☺