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”?
Review of what we have learned so far
### Notation

\[
R(f) = \Pr[Y \neq f(X)]
\]

\[
\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} 1\{Y_i \neq f(X_i)\}
\]

\[
R_n^*(f) = \inf_{f: \mathcal{X} \to \mathbb{R}} R(f)
\]

\[
\hat{R}_n^*, \mathcal{F} = \inf_{f \in \mathcal{F}} \hat{R}_n(f)
\]

\[
R_n^*, \mathcal{F} = \inf_{f \in \mathcal{F}} R(f)
\]

\[
f_n^*, \mathcal{F} = \arg \min_{f \in \mathcal{F}} \hat{R}_n(f)
\]

This is what the learning algorithm produces

---

**We will need these definitions, please copy it!**

\[
R(f) = \text{Risk} \quad R^* = \text{Bayes risk}
\]

\[
\hat{R}_n(f) = \text{Empirical risk} \quad f^* = \text{Bayes classifier}
\]

\[
f_n^*, \mathcal{F} = \text{the classifier that the learning algorithm produces}
\]
Big Picture

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

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

Risk of the classifier \( f_n^* \)

\[ R(f_n^*) - R^* = \underbrace{R(f_n^*) - R_{\mathcal{F}}^*}_{\text{Bayes risk}} + \underbrace{R_{\mathcal{F}}^* - R^*}_{\text{Best classifier in } \mathcal{F}} \]

Estimation error

Approximation error

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

\( \{ f : \mathcal{X} \to \{0, 1\} \} \)
**Goal of Learning:**

For a fixed $\mathcal{F}$, make the $|R(f_n^*) - R(f_{\mathcal{F}}^*)|$ estimation error small.
Learning Theory
From Hoeffding’s inequality, we have seen that

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

\[
\begin{align*}
\Pr \left( \sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| > \varepsilon \right) &\leq 2N \exp \left( -2n\varepsilon^2 \right) \\
\Pr \left( \sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| < \sqrt{\frac{\log(N) + \log(2/\delta)}{2n}} \right) &\geq 1 - \delta
\end{align*}
\]

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
From Hoeffding’s inequality, we have seen that

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

\[
\begin{align*}
\Pr \left( \sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| > \varepsilon \right) &\leq 2N \exp \left( -2n\varepsilon^2 \right) \\
\Pr \left( \sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| < \sqrt{\frac{\log(N) + \log(2/\delta)}{2n}} \right) &\geq 1 - \delta
\end{align*}
\]

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

\[|R(f_n^*) - R(f_{\mathcal{F}}^*)| \leq 2 \sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| = 2\varepsilon\]

This is what the learning algorithm produces and its true risk
Theorem: Let $\mathcal{F} = \{f : \mathcal{X} \to \{0, 1\}\}$, and $|\mathcal{F}| \leq N$

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

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

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

Observation!

It does not matter how many elements $\mathcal{F}$ has. All that matters in the union bound is how many elements

$$\{[f(X_1), \ldots, f(X_n)] \mid f \in \mathcal{F}\}$$

has. (The effective size of $\mathcal{F}$). It can’t even be more than $2^n$. 

10
McDiarmid’s Bounded Difference Inequality

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

$$\sup_{x_1, x_2, \ldots, x_n, \tilde{x}_i} |f(x_1, x_2, \ldots, x_n) - f(x_1, x_2, \ldots, x_{i-1}, \tilde{x}_i, x_{i+1}, \ldots, x_n)| \leq c_i$$

for $1 \leq i \leq n$ (Bounded Difference Assumption: replacing the $i$-th coordinate $x_i$ changes the value of $f$ by at most $c_i$.)

It follows that

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

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

$$\Pr \{ |E[f(X_1, X_2, \ldots, X_n)] - f(X_1, X_2, \ldots, X_n)| \geq \varepsilon \} \leq 2 \exp \left( -\frac{2\varepsilon^2}{\sum_{i=1}^{n} c_i^2} \right).$$
Our main goal is to bound \( \sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| \)

**Lemma:**

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

**Proof:**

Let \( g \) denote the following function:

\[
\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} 1\{f(X_i) \neq Y_i\}
\]

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

**Observation:**

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

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

\( \Rightarrow \) **McDiarmid can be applied for** \( g \)!
The "bounded difference condition" is satisfied for \( \sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| \)

**Corollary:**

\[
\Pr \left\{ g - \mathbb{E}[g] \geq \varepsilon \right\} \leq \exp \left( -\frac{2\varepsilon^2}{\sum_{i=1}^{n} c_i^2} \right) \quad \text{for } g = \sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)|
\]

\[ c_i = 1/n \]

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

\[ \Rightarrow \sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| \text{ is concentrated around its mean!} \]

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

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

\[ Z_n = \sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| \]
Our main goal is to bound $\sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)|$

We already know:

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

Vapnik-Chervonenkis inequality:

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

Corollary: Vapnik-Chervonenkis theorem:

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

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?

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

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

**Definition**

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

($=5$ in this example)

**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
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 \( \{x_1, x_3\} \)
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. \( f(x_i) = y_i \) \( (1 \leq i \leq n) \)

- 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 measures how rich \( \mathcal{F} \) is.

If the VC dimension is high, e.g. \( \infty \), then it is easy to overfit!
Examples
What’s the VC dim. of decision stumps in 2d?

There is a placement of 3 pts that can be shattered \(\Rightarrow\) VC dim \(\geq 3\)
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.

3 collinear

1 in convex hull of other 3

quadrilateral
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 \( \Rightarrow \) VC dim \( \geq 3 \)
There is a placement of 4 pts that can be shattered \( \Rightarrow \) VC dim \( \geq 4 \)
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.

4 collinear

2 in convex hull

1 in convex hull

pentagon
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:

\[
|\mathcal{F}| = 7
\]

\[
\begin{array}{c|ccc}
  & x_1 & x_2 & x_3 \\
\hline
 f_1 & 0 & 0 & 0 \\
 f_2 & 0 & 1 & 0 \\
 f_3 & 1 & 1 & 1 \\
 f_4 & 1 & 0 & 0 \\
 f_5 & 0 & 1 & 0 \\
 f_6 & 1 & 1 & 1 \\
 f_7 & 0 & 1 & 1 \\
\end{array}
\]

\[
|\mathcal{F}| = 7
\]

\[
\begin{array}{c|ccc}
  & x_1 & x_2 & x_3 \\
\hline
 f_1 & 0 & 0 & 0 \\
 f_2 & 0 & 1 & 0 \\
 f_3 & 1 & 1 & 1 \\
 f_4 & 1 & 0 & 0 \\
 f_5 & 0 & 1 & 0 \\
 f_6 & 1 & 1 & 1 \\
 f_7 & 0 & 1 & 1 \\
\end{array}
\]

32
Proof of Sauer’s Lemma

We will prove that

\[ |\mathcal{F}| = 7 \]

\[
\begin{array}{ccc}
  f_1 & 0 & 0 & 0 \\
  f_2 & 0 & 1 & 0 \\
  f_3 & 1 & 1 & 1 \\
  f_4 & 1 & 0 & 0 \\
  f_7 & 0 & 1 & 1 \\
\end{array}
\]

\[ = A \]

Shattered subsets of columns:

\[ \{\emptyset, \{x_1\}, \{x_2\}, \{x_3\}, \{x_1, x_2\}, \{x_1, x_3\} \]
Proof of Sauer’s Lemma

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

In this example: $6 \leq 1+3+3=7$

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

for any binary matrix with no repeated rows.

In this example: $5 \leq 6$
Proof of Lemma 1

Shattered subsets of columns:

\[ \{\emptyset\}, \{x_1\}, \{x_2\}, \{x_3\}, \{x_1, x_2\}, \{x_1, x_3\} \]

In this example: \(6 \leq 1+3+3 = 7\)

**Lemma 1**

\[ \# \text{ shattered subsets of columns of } A \leq \sum_{k=0}^{\text{VC}_F} \binom{n}{k} \]

Proof

\(\text{VC}_F\) is the size of largest imaginable shattered sample. \(\text{VC}_F = \max\{n : S_F(n) = 2^n\}\)

If a shattered subsets of columns has \(d\) elements, then \(\text{VC}_F \geq d\)

For example if \(\{x_1, x_3\}\) are shattered in \(A\), then \(\text{VC}_F \geq 2\).
Proof of Lemma 2

Lemma 2

\[ \# \text{ rows}(A) \leq \# \text{ shattered subsets of columns of } A \]

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:

- \( A = (0) \Rightarrow 1 \leq 1 \) \hspace{1cm} \text{shattered subsets of columns: } \{\emptyset\}
- \( A = (1) \Rightarrow 1 \leq 1 \) \hspace{1cm} \text{shattered subsets of columns: } \{\emptyset\}
- \( A = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \Rightarrow 2 \leq 2 \) \hspace{1cm} \text{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 occur 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,

\[
\# \text{ rows}(A) = \# \text{ rows}(B) + \# \text{ rows}(C) \\
\leq \# \text{ shattered subsets of columns of } (B) \\
\quad + \# \text{ shattered subsets of columns of } (C)
\]

By induction (less columns)
Proof of Lemma 2

\[ \{\emptyset\} \leq \# \text{ shattered subsets of columns of } (C) \]
\[ \{\emptyset\}, \{x_1\}, \{x_2\}, \{x_3\}, \{x_1, x_2\}, \{x_1, x_3\} \]

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 \( S \cup x_m \).
Vapnik-Chervonenkis inequality

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

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

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

From Sauer’s lemma:

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

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

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

We already know that
\[
\mathbb{E}[|R(f^n_*) - R(f^*_F)|] \leq 4\sqrt{\frac{V C_F \log(n + 1) + \log 2}{n}}
\]

For linear classifiers in dimension when \( X = \mathbb{R}^d \): \( V C_F = d + 1 \).

\[
\Rightarrow \mathbb{E}[|R(f^n_*) - R(f^*_F)|] \leq 4\sqrt{\frac{(d + 1) \log(n + 1) + \log 2}{n}}
\]

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

\[
\Rightarrow \mathbb{E}[|R(f^n_*) - R(f^*_F)|] \leq 4\sqrt{\frac{(d' + 1) \log(n + 1) + \log 2}{n}}
\]
We already know from McDiarmid:

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

Vapnik-Chervonenkis inequality:

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

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

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

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

Hoeffding + Union bound for finite function class:

When \(|\mathcal{F}| = N < \infty\), \(\Rightarrow \Pr \left( \sup_{f \in \mathcal{F}} \left| \hat{R}_n(f) - R(f) \right| > t \right) \leq 2N \exp \left( -2nt^2 \right)\)
PAC Bound for the Estimation Error

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

Inversion: \[
8S_\mathcal{F}(n) \exp(-nt^2/32) \leq \delta \quad \Rightarrow \quad t^2 \geq \frac{32}{n} \log \left( \frac{8S_\mathcal{F}(n)}{\delta} \right)
\]

\[
\Rightarrow \Pr \left( \sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| \leq 8 \sqrt{\frac{\log(S_\mathcal{F}(n)) + \log \left( \frac{8}{\delta} \right)}{2n}} \right) \geq 1 - \delta
\]

\[
S_\mathcal{F}(n) \leq \left( \frac{ne}{V C_\mathcal{F}} \right)^{V C_\mathcal{F}} \Rightarrow \Pr \left( \sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| \leq 8 \sqrt{\frac{V C_\mathcal{F} \log \left( \frac{ne}{V C_\mathcal{F}} \right) + \log \left( \frac{8}{\delta} \right)}{2n}} \right) \geq 1 - \delta
\]

Don’t forget that \[
|R(f_n^*) - R(f_\mathcal{F}^*)| \leq 2 \sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)|
\]

Estimation error \[
\Rightarrow \Pr \left( |R(f_n^*) - R(f_\mathcal{F}^*)| \leq 16 \sqrt{\frac{\log(V C_\mathcal{F} \log \left( \frac{ne}{V C_\mathcal{F}} \right) + \log \left( \frac{8}{\delta} \right))}{2n}} \right) \geq 1 - \delta
\]
Structural Risk Minimization

Risk of the classifier $f_n^*$
\[ R(f_n^*) - R^* = R(f_n^*) - R_{\mathcal{F}}^* + R_{\mathcal{F}}^* - R^* \]

Estimation error
Approximation error
Bayes risk

Ultimate goal:
\[ R(f_n^*) - R^* = 0 \]

So far we studied when estimation error $\rightarrow 0$, but we also want approximation error $\rightarrow 0$

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 😊