Empirical Risk Minimization

→ What is SUPERVISED LEARNING ?

List of INPUTS :
$x_1, \dots, x_n \in \mathcal{X}$

List of OUTPUTS :
$y_1, \dots, y_n \in \mathcal{Y}$

GOAL: Given a new input x, predict the corresponding y.

Two large families :

CLASSIFICATION $\mathcal{Y} = \{a, b, c, \dots\}$

REGRESSION $\mathcal{Y} = \mathbb{R}$

Examples :

INPUTS		OUTPUTS
Pictures	⟶	Objects
Movie reviews	⟶	Review rating
Patient	⟶	Is the patient cured ?

1) Risks and losses :

⚠ There is no single good notion to quantify the quality of a prediction.

Def : A loss function is a function
$$\ell : \mathcal{Y} \times \mathcal{Y} \longrightarrow [0, \infty).$$

Examples: 1) Binary classification $\mathcal{Y} = \{-1, +1\}$

- Choice 1 : $\ell(y, y') = \begin{cases} 1 & \text{if } y \neq y' \\ 0 & \text{otherwise} \end{cases}$

- Choice 2 : $\ell(y, y') = \begin{cases} a & \text{if } y = 1 \text{ and } y' = -1 \\ b & \text{if } y = -1 \text{ and } y' = 1 \\ 0 & \text{if } y = y' \end{cases}$

⇝ Choice 2 makes sense (for instance) in medical settings, where it is a more serious mistake to predict that a patient is not sick ($y' = -1$) whereas they are ($y = +1$) than the opposite.

2. Regression $\mathcal{Y} = \mathbb{R}^D$ $\quad y = (y_1, \dots, y_D)$ $\quad y' = (y'_1, \dots, y'_D)$

- L_∞ norm $\quad \|y' - y\|_\infty = \max_{i=1 \dots D} |y_i - y'_i|$
- L_p norm $\quad \|y' - y\|_p = \left(\sum_{i=1}^{D} |y_i - y'_i|^p \right)^{1/p}$
- Weighted L_p norm $\quad \left(\sum_{i=1}^{D} w_i |y_i - y'_i|^p \right)^{1/p}$

GOAL: Find a predictor $f : \mathcal{X} \to \mathcal{Y}$ such that, $\ell(f(x), y)$ is small on NEW samples $(x'_1, y'_1) \dots (x'_m, y'_m)$.

TRAINING SAMPLES
(x_1, y_1)
$\dots$ (x_n, y_n)

$\rightsquigarrow$ Find a predictor f.

NEW SAMPLES
$(x'_1, y'_1) \dots (x'_m, y'_m)$

We want $\ell(f(x'_i), y'_i)$ to be small on average.
(i.e. good predictions on new samples)

Example: Electric consumption forecasting

Training samples : y_i = Electric consumption on Day i

x_i = characteristics of Day i (weather, day of the week, etc.)

TRAIN A MODEL

$\Downarrow$

Testing sample : Predict tomorrow's consumption based on tomorrow's weather, etc.

ASSUMPTION: The observations $(X_1, Y_1) \dots (X_n, Y_n)$ are i. i. d. with law P.

independant identically distributed.

Def Given a function $f : \mathcal{X} \to \mathcal{Y}$, the P-risk of f is defined as:

$$R_P(f) := E_P[\ell(f(X), Y)]$$

Theorem: The P-risk is minimized for the Bayes predictor f_P^* defined by

$$f_P^*(x) \in \underset{z \in \mathcal{Y}}{\operatorname{argmin}} \; E_P[\ell(Y, z) \mid X = x].$$

proof: Let $\Psi(x,z) = \mathbb{E}_P[\ell(Y,z) \mid X=x]$

$\rightsquigarrow \quad \Psi(x,z) \geq \Psi(x, f_P^*(x)) \quad$ (by definition)

$$R_P(f) = \mathbb{E}_P[\ell(Y,f(X))] \overset{(*)}{=} \mathbb{E}_P[\Psi(X,f(X))]$$
$$\geq \mathbb{E}_P[\Psi(X,f_P^*(X))]$$
$$\overset{(*)}{=} \mathbb{E}_P[\ell(Y,f_P^*(X))] = R_P(f_P^*)$$

(*) LAW OF TOTAL EXPECTATION

$$\mathbb{E}[A] = \mathbb{E}[\mathbb{E}[A \mid X]]$$

$\blacksquare$

Examples:

① Binary classification:

$$\ell(y,y') = \begin{cases} 1 & \text{if } y \neq y' \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbb{E}_P[\ell(Y,z) \mid X=x] = P(Y \neq z \mid X=x)$$

$z = -1$ or $+1$ $\quad = 1 - P(Y=z \mid X=x)$

→ if $\eta(x) = P(Y=1 \mid X=x) \geq \frac{1}{2}$,

then $f_P^*(x) = 1$.

Otherwise, $f_P^*(x) = -1$.

② Quadratic loss $\mathcal{Y} = \mathbb{R}$ $\ell(y,y') = (y-y')^2$

Question : A random variable.
What is the minimum of

$$a \mapsto \mathbb{E}[(A-a)^2] = \mathbb{E}[A^2] - 2\mathbb{E}[A]a + a^2$$

$$\partial_a = 2(a - \mathbb{E}[A]) \rightsquigarrow a = \mathbb{E}[A]$$

$$\longrightarrow f_P^*(x) = E_P[Y \mid X = x]$$

② Empirical risk

P is unknown $\Longrightarrow$ f_P^* is unknown

GOAL : Approximate f_P^* using the observations
$((X_1, Y_1), \ldots, (X_n, Y_n))$

Def : The empirical risk of a function f is

$$\left[R_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i) \right]$$

LAW OF LARGE NUMBERS

$$\mathcal{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i)$$

$$\Big\downarrow{\scriptstyle n \to +\infty}$$

$$\mathcal{R}_P(f) = \mathbb{E}_P[\ell(f(X), Y)]$$

$\rightsquigarrow$ Minimizing $\mathcal{R}_n(f)$ $\approx$ Minimizing $\mathcal{R}_P(f)$

Def: Let $\mathcal{F}$ be a set of functions from $\mathcal{X} \to \mathcal{Y}$.

The empirical risk minimizer of $\mathcal{F}$ is

$$\left[\ \hat{f}_{\mathcal{F}} \in \underset{f \in \mathcal{F}}{\operatorname{argmin}}\ \mathcal{R}_n(f)\ \right]$$

Examples:

1. Linear Regression: $\mathcal{X} = \mathbb{R}^d \quad \mathcal{Y} = \mathbb{R}$

$$\ell(y, y') = (y - y')^2$$

$$\mathcal{F}_{lin} = \{ f_\theta : x \mapsto \langle x, \theta \rangle : \theta \in \mathbb{R}^d \}$$

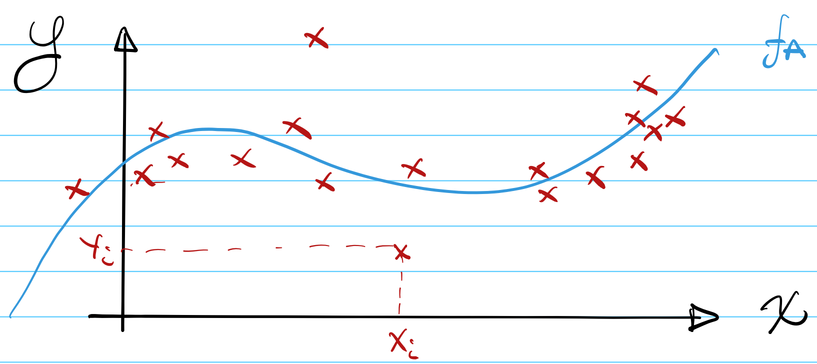
$$\left.\begin{aligned} R_n(f_\theta) &= \frac{1}{n}\sum_{i=1}^{n}(y_i - f_\theta(x_i))^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \langle x_i, \theta\rangle)^2 \\ &= \frac{1}{n}\|\mathbb{Y} - \mathbb{X}\theta\|^2 \end{aligned}\right\} \text{LINEAR REGRESSION}$$

$$\mathbb{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \qquad \mathbb{X} = \underbrace{\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}}_{d} \Big\updownarrow n$$

② Polynomial Regression

$$\mathcal{X} = \mathbb{R} \qquad \mathcal{Y} = \mathbb{R} \qquad \ell(y, y') = |y - y'|^2$$

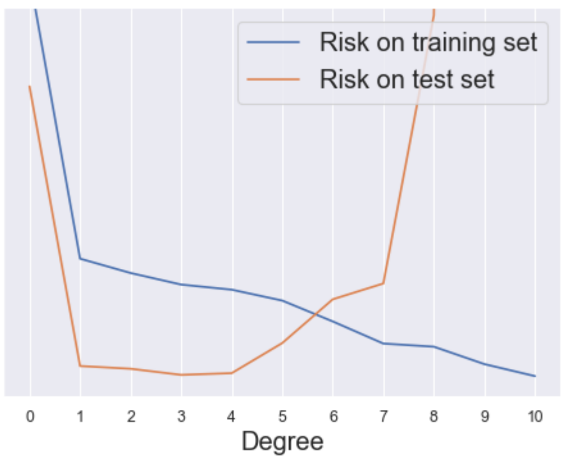
$$\mathcal{F}_d = \Big\{ \underbrace{x \longmapsto \sum_{i=0}^{d} a_i x^i}_{= f_A} \; : \; \underbrace{a_0 \ldots a_d}_{= A\, \in \mathbb{R}^{d+1}} \in \mathbb{R} \Big\}$$



— $d = 1$

— $d = 2$

— $d = 8$



⇝ Let us explain this phenomenon.

⚠ Training set $(X_1, Y_1) \dots (X_n, Y_n)$

⟹ Find prediction $\hat{f}_{\mathcal{F}}$ DEPENDING on the Training set.

⟹ We can NOT apply the Law of Large numbers to say that $R_P(\hat{f}_{\mathcal{F}}) \approx R_n(\hat{f}_{\mathcal{F}})$

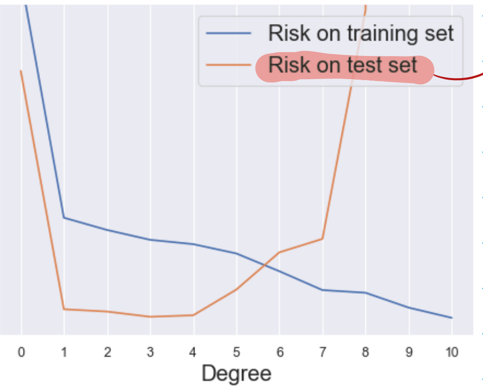
$$= \frac{1}{n} \sum_{i=1}^{n} \underbrace{\ell(\hat{f}_{\mathcal{F}}(X_i), Y_i)}_{\text{not indep !!}}$$

⟹ On the testing sample $(X'_1, Y'_1) \dots (X'_m, Y'_m)$ INDEPENDENT from the training sample,

LLN CONDITIONALLY ON THE TRAINING SAMPLE

⇓

$$R_P(\hat{f}_{\mathcal{F}}) \approx \frac{1}{m} \sum_{i=1}^{m} \underbrace{\ell(\hat{f}_{\mathcal{F}}(X'_i), Y'_i)}_{\text{conditionally independent.}}$$



Good Approximation of $R_P(\hat{f}_{\mathcal{F}})$

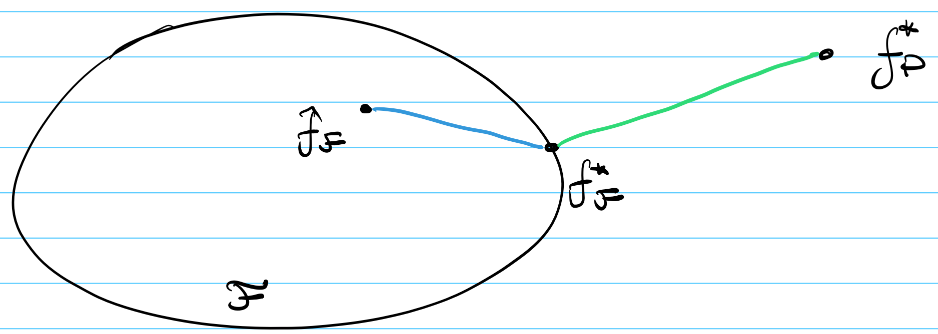
Decomposition of the Empirical Risk

Minimal Risk: $R_P^* = R_P(f_P^*) = \min_f R_P(f)$

Excess Risk:

$$R_P(\hat f_{\mathcal F}) - R_P^*$$

$$= \underbrace{\left(R_P(\hat f_{\mathcal F}) - \inf_{f\in\mathcal F} R_P(f)\right)}_{\text{Estimation Error } \geq 0} + \underbrace{\left(\overbrace{\inf_{f\in\mathcal F} R_P(f)}^{= R_P(f^*_{\mathcal F})} - R_P^*\right)}_{\text{Approximation Error } \geq 0}$$



Approximation Error: "How far is the model $\mathcal F$ from the truth?"

Estimation Error: "How good can we estimate the best predictor in $\mathcal F$?"

Harder if $\mathcal F$ is "large"

Bound on the Estimation Error:

$$R_P(\hat{f}_{\mathcal{F}}) - \inf_{f \in \mathcal{F}} R_P(f) = R_P(\hat{f}_{\mathcal{F}}) - R_P(f^*_{\mathcal{F}})$$

$$= \left(R_P(\hat{f}_{\mathcal{F}}) - R_n(\hat{f}_{\mathcal{F}})\right)$$

$$\left[+ \left(R_n(\hat{f}_{\mathcal{F}}) - R_n(f^*_{\mathcal{F}})\right) \right] \leq 0$$

$$+ \left(R_n(f^*_{\mathcal{F}}) - R_P(f^*_{\mathcal{F}})\right)$$

$$\leq \underbrace{\sup_{f \in \mathcal{F}} \left(R_P(f) - R_n(f)\right)}_{\text{uniform the deviation between empirical risk and P-risk.}} + \left(R_n(f^*_{\mathcal{F}}) - R_P(f^*_{\mathcal{F}})\right)$$

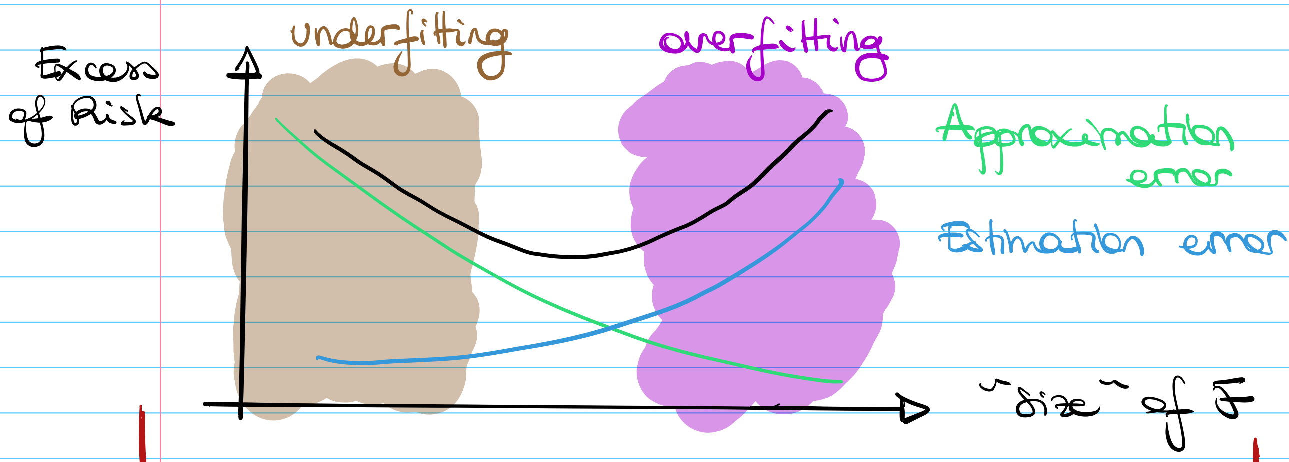
→ CENTRAL LIMIT THEOREM :

$$R_P(f) - R_n(f) = \frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{E}[\ell(f(X), Y)] - \ell(f(X_i), Y_i) \right)$$

$$\approx \frac{1}{\sqrt{n}} \times \text{Gaussian}$$

↳ But what about $\sup_{f \in \mathcal{F}} \left(R_P(f) - R_n(f)\right)$?

The Picture



③ Bound on the estimation error in Binary Classification

$\mathcal{Y} = \{-1, 1\}$ $\quad\quad$ $\ell(y, y') = \mathbb{1}\{y \neq y'\}$

GOAL : Bound $\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left(R_p(f) - R_n(f)\right)\right]$

CASE 1 : Finite Number of Predictors

$\mathcal{F} = \{f_1 \, --- \, f_k\}$

THEOREM (Maximal Inequality)

Let $Z_1, \dots, Z_k$ be random variables with

$\forall \lambda > 0,\quad \mathbb{E}[e^{\lambda Z_j}] \le e^{\lambda^2 \sigma^2 / 2}$ (subgaussianity condition)

Then,

$$\boxed{\mathbb{E}\Big[\max_{j=1 \dots k} Z_j\Big] \le \sigma\sqrt{2 \log k}}$$

proof:

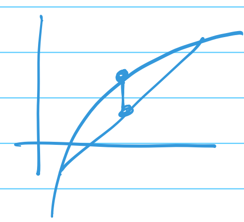
$$\max_{j=1 \dots k} e^{\lambda a_j} \le \sum_{j=1}^{k} e^{\lambda a_j} \qquad (*)$$

$$\Longrightarrow \mathbb{E}\Big[\max_{j=1 \dots k} Z_j\Big] = \frac{1}{\lambda}\, \mathbb{E}\Big[\log\Big(\max_{j=1 \dots k} e^{\lambda Z_j}\Big)\Big]$$

$\frac{1}{\lambda}\log(e^{\lambda x}) = x$

JENSEN

$$\le \frac{1}{\lambda} \log\Big(\mathbb{E}\Big[\max_{j=1 \dots k} e^{\lambda Z_j}\Big]\Big)$$



$$\overset{(*)}{\le} \frac{1}{\lambda} \log\Big(\underbrace{\mathbb{E}\Big[\sum_{j=1}^{k} e^{\lambda Z_j}\Big]}_{\sum_{j=1}^{k} \mathbb{E}[e^{\lambda Z_j}] \,\le\, k e^{\lambda^2 \sigma^2 / 2}}\Big)$$

$$\le \frac{\log k}{\lambda} + \frac{1}{\cancel{\lambda}} \frac{\lambda^2 \sigma^2}{2}$$

$\rightsquigarrow$ Choose $\lambda = \frac{\sqrt{2 \log k}}{\sigma}$. $\blacksquare$

Recall

$$\mathbb{E}\left[R_P(\hat{f}_{\mathcal{F}}) - \inf_{f \in \mathcal{F}} R_P(f)\right]$$

$$\leq \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left(R_P(f) - R_n(f)\right)\right] + \mathbb{E}\left[\underbrace{\left(R_n(f^*_{\mathcal{F}}) - R_P(f^*_{\mathcal{F}})\right)}\right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{f^*_{\mathcal{F}}(X_i) \neq Y_i\} - \mathbb{E}_P\left[\mathbb{1}\{f^*_{\mathcal{F}}(X) \neq Y\}\right]$$

$$\Rightarrow \mathbb{E}\left[\quad\right] = 0$$

See notes: $R_P(f) - R_n(f)$ is $\frac{1}{\sqrt{n}}$ - subgaussian.

BOUND ON THE ESTIMATION ERROR

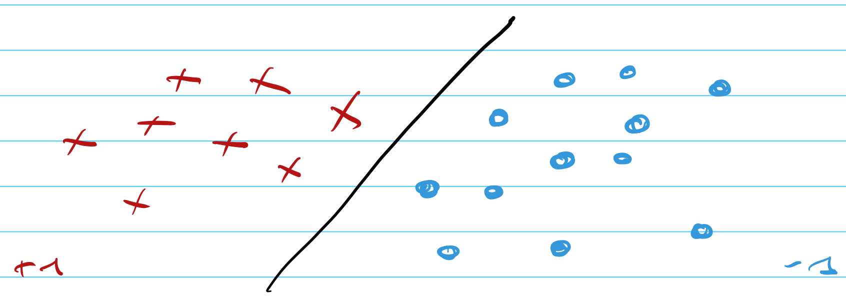
$$\Longrightarrow \boxed{\mathbb{E}\left[R_P(\hat{f}_{\mathcal{F}}) - \inf_{f \in \mathcal{F}} R_P(f)\right] \leq \sqrt{\frac{2\log k}{n}}}$$

size of $\mathcal{F}$.

CASE 2 : With VC-dimension

What if $\mathcal{F}$ is infinite?

ex: $\mathcal{F}_{lin} = \{\text{hyperplane classifiers}\}$



Even if $\mathcal{F}$ is infinite, there is only a finite number of classifications.

$$\mathcal{C}_{\mathcal{F}}(x_1 \ldots x_n) = \{ (f(x_1), \ldots, f(x_n)) : f \in \mathcal{F} \}$$
$$\subseteq \{-1, +1\}^n$$

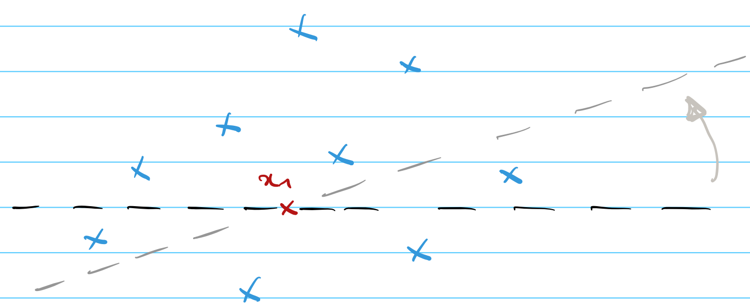
of size $\mathcal{N}_{\mathcal{F}}(x_1, \ldots, x_n) \leq 2^n$

$\Rightarrow$ Actually the size of $\mathcal{F}$ can be replaced by $\mathcal{N}_{\mathcal{F}}(x_1 \ldots x_n)$!

THEOREM :

$$\boxed{E\left[\sup_{f \in \mathcal{F}} \left(R_P(f) - R_n(f)\right)\right] \leq 2\, E\left[\sqrt{\frac{2 \log \mathcal{N}_{\mathcal{F}}(x_1 \ldots x_n)}{n}}\right]}$$

Example: $\mathcal{F} = \mathcal{F}_{lin}$ in $\mathbb{R}^2$



Rotation around x_1 : $\leq n$ different classifications
Rotation around x_2 : $\leq n$ different classifications
$\vdots$
Rotation around x_n : $\leq n$ different classifications

$\Rightarrow$ At most n^2 different classifications!

$\mathcal{N}_{\mathcal{F}}(x_1, \dots, x_n) \leq n^2$

$\Rightarrow$ $\mathbb{E}[\text{estimation error}] \leq 4 \sqrt{\frac{\log n}{n}}$

And for a general $\mathcal{F}$?

- We say that $\mathcal{F}$ **shatters** $(x_1, ---, x_n)$ if $$\mathcal{N}_{\mathcal{F}}(x_1, --- x_n) = 2^n.$$

Vapnik– Chervonenkis

- The **VC dimension of $\mathcal{F}$** is the longest n such that there exists a configuration $(x_1 --- x_n)$ of n points being shattered by $\mathcal{F}$.

$\rightsquigarrow$ $VC(\mathcal{F})$

For $n \leq VC(\mathcal{F})$, we cannot bound $\mathcal{N}_{\mathcal{F}}(x_1 --- x_n)$ meaningfully ☹

What if $n > VC(\mathcal{F})$?

$\Downarrow$ MAGICAL RESULT

SAUER LEMMA: if $n > VC(\mathcal{F})$,

$$\log \mathcal{N}_{\mathcal{F}}(x_1, ---, x_n) \leq VC(\mathcal{F}) \log\left(\frac{en}{VC(\mathcal{F})}\right)$$

$\rightsquigarrow$ if $n > 2VC(\mathcal{F})$,

$$\mathcal{N}_{\mathcal{F}}(---) \lesssim n^{VC(\mathcal{F})} \lll 2^n \;!!$$

THEOREM: if $n > 2VC(\mathcal{F})$

$$\left[\; \underbrace{\mathbb{E}\left[R_P(\hat{f}_{\mathcal{F}}) - \inf_{f \in \mathcal{F}} R_P(f) \right]}_{\text{Estimation error in binary classification}} \leq 2\sqrt{\frac{2VC(\mathcal{F})}{n} \log\left(\frac{en}{VC(\mathcal{F})}\right)} \;\right]$$

$$\lesssim \sqrt{\frac{VC(\mathcal{F})}{n}}$$

TAKE-HOME MESSAGES :

- The quality of a prediction is measured by a loss ℓ.

$$P\text{-Risk} = \mathbb{E}_P[\text{loss}]$$

- The best theoretical predictor is the Bayes predictor.
  $\rightsquigarrow$ cannot be computed

- A strategy to find a good predictor is to minimize the empirical risk on a model $\mathcal{F}$.

- The "size" of $\mathcal{F}$ has to be properly chosen to avoid

UNDERFITTING and OVERFITTING

UNDERFITTING	OVERFITTING
Large approximation error ↳ The model $\mathcal{F}$ is too simple to capture the complexity of the dataset	Large estimation error ↳ in binary classification, bounded by $\approx \sqrt{\frac{VC(\mathcal{F})}{n}}$ Large if $\mathcal{F}$ is too complex.