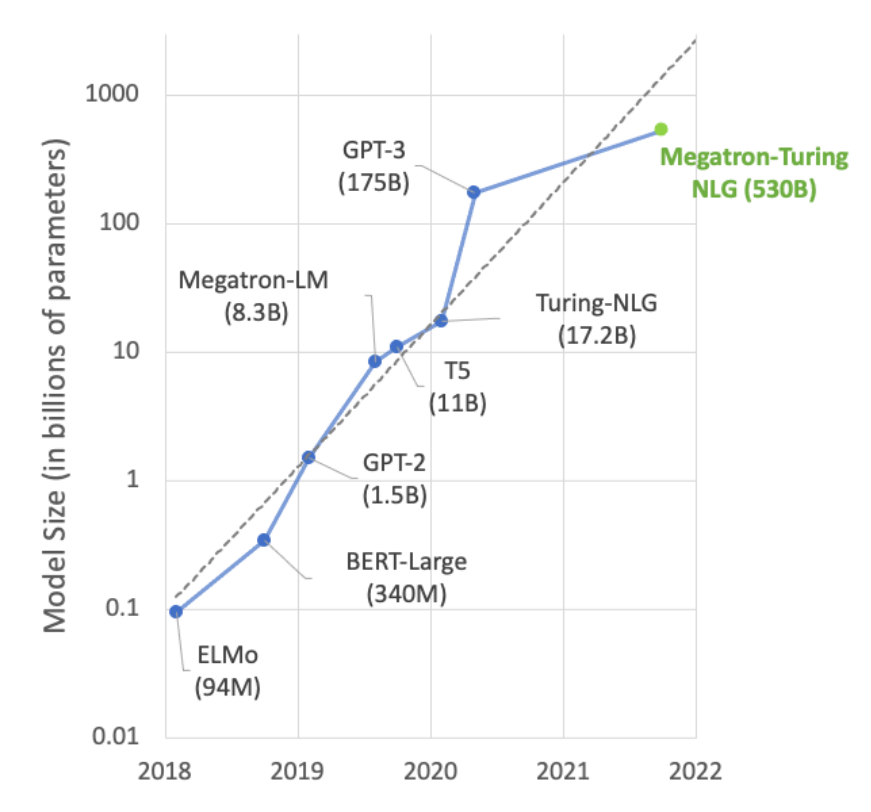
MATHEMATICAL TOPICS IN MACHINE LEARNING

(LECTURE 3 – GENERALIZATION BOUNDS)

Professor Gavin Brown

OUR QUESTION

"Are bigger models always better models?"



TODAY:

A way to estimate how bad our testing error might get, based on…

- the training error,
- the "capacity" of the model,
- the size of your data.

KEY POINTS FROM LAST WEEK

$$\hat{R}(f, \mathcal{S}_n) := \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{y}_i, f(\boldsymbol{x}_i)).$$

$$f_{erm} := \operatorname*{arginf}_{f \in \mathcal{F}} \hat{R}(f, \mathcal{S}_n).$$

The **estimated** risk, using an IID sample of size *n*.

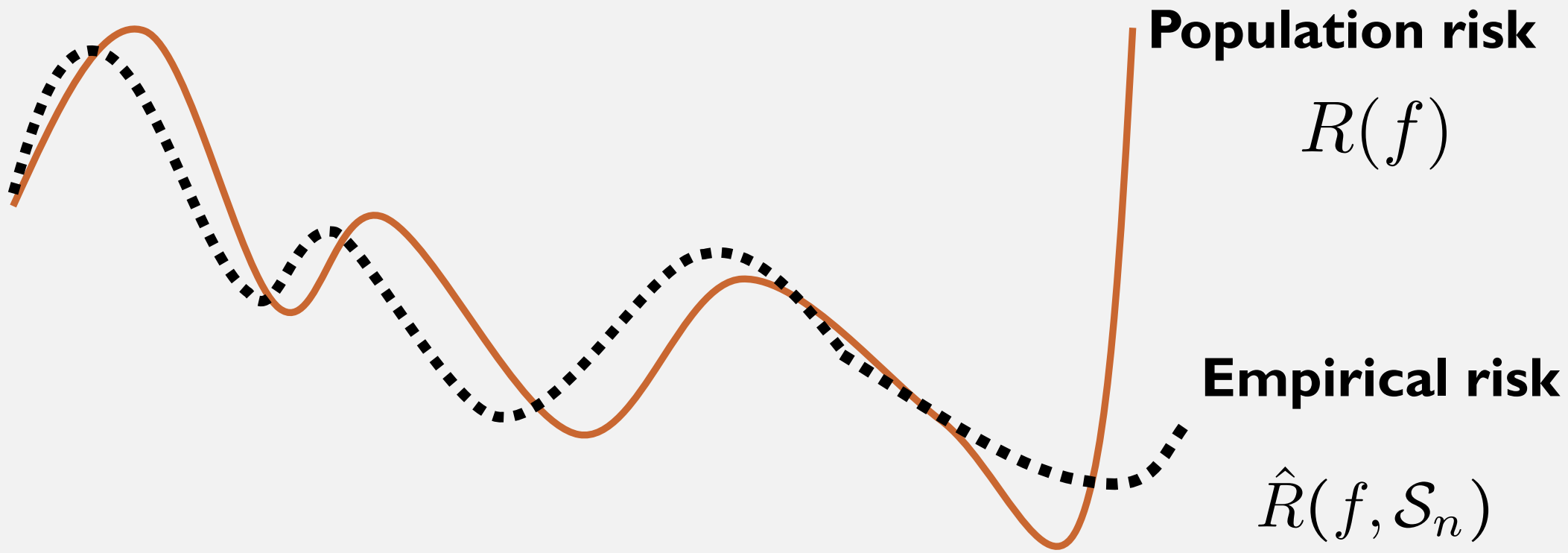
$$R(f) := \mathbb{E}_{(\boldsymbol{x}, \boldsymbol{y}) \sim \mathcal{D}} \Big[\ell(\boldsymbol{y}, f(\boldsymbol{x})) \Big]$$

$$f^* := \operatorname*{arginf}_{f \in \mathcal{F}} R(f).$$

The **true** risk, using all possible data we may ever encounter.

*(**NOT** the same thing as testing data!)*

KEY POINTS FROM LAST WEEK



KEY POINTS FROM LAST WEEK

$$R(f_{erm}) - R(y^*) = \underbrace{R(f_{erm}) - R(f^*)}_{\text{Estimation}} + \underbrace{R(f^*) - R(y^*)}_{\text{Approximation}}$$

Approximation error … error due to having a restricted model family, unable to represent the Bayes model.

Estimation error … error due to having a small sample, where empirical risk is a poor estimate of population risk.

HOW CLOSE CAN WE GET?

$$\hat{R}(f, \mathcal{S}_n) \quad \approx \quad R(f)$$

WEAK LAW OF LARGE NUMBERS

$$\hat{R}(f, \mathcal{S}_n) \quad \rightarrow \quad R(f)$$

$$\text{as } n \rightarrow \infty$$

The empirical risk "converges in probability" to the population risk.

But this says nothing about how FAST it converges.

A TOY EXAMPLE

Assume $x = \{0, 1\}$ is a random variable, following dist'n *p(x)*.

The expected (i.e. true) value is…

$$\mathbb{E}[X] = \Big(0 \times p(x = 0)\Big) + \Big(1 \times p(x = 1)\Big)$$

Our estimate is…

$$\frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\frac{1}{n} \sum_{i=1}^{n} x_i \to \mathbb{E}[X]$$

$$\text{as } n \to \infty$$

LET'S SEE HOW 'FAST' THEY CONVERGE

$R(f)$ $\hat{R}(f, \mathcal{S}_n)$

$$\left| \mathbb{E}[X] - \frac{1}{n} \sum_{i=1}^{n} x_i \right| \geq \epsilon$$

abs. value of difference "deviation" $\geq$ some small value

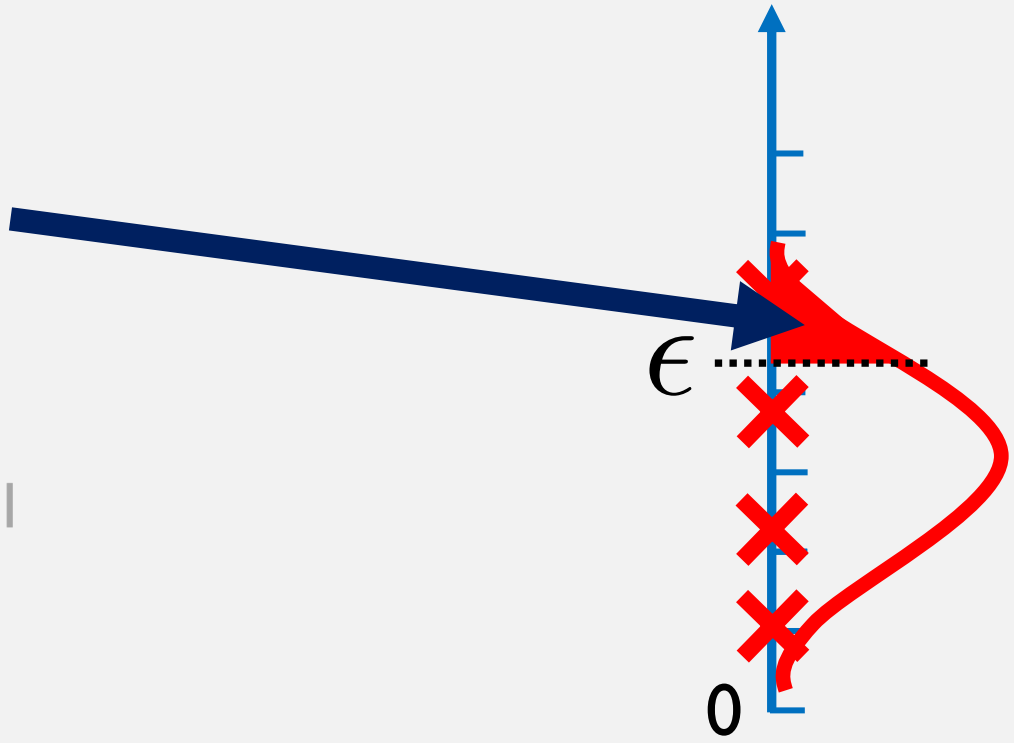
LET'S SEE HOW 'FAST' THEY CONVERGE

Probability of deviation $\geq \epsilon$

$$p\left(\left| \mathbb{E}[X] - \frac{1}{n}\sum_{i=1}^{n} x_i \right| \geq \epsilon \right)$$

abs. value of difference "deviation" $\geq$ some small value

$$\left| \mathbb{E}[X] - \frac{1}{n}\sum_{i=1}^{n} x_i \right|$$



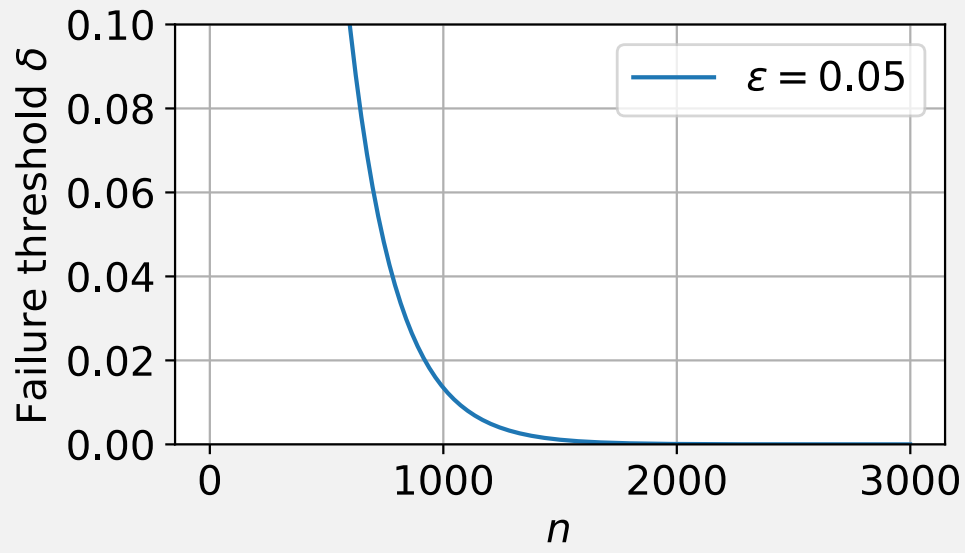
HOEFFDING'S INEQUALITY

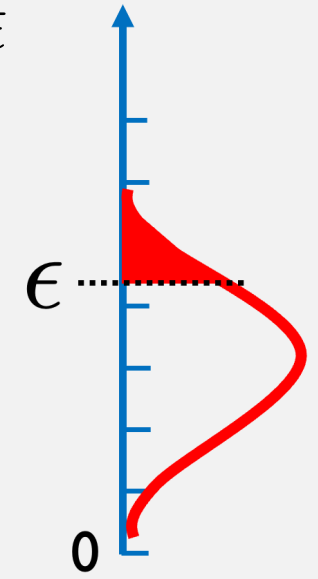
Proof is beyond the scope of the module.

Probability of deviation $\geq \epsilon$ is, at most

$$p\left(\left| \mathbb{E}[X] - \frac{1}{n}\sum_{i=1}^{n} x_i \right| \geq \epsilon \right) \leq \delta = 2\exp(-2n\epsilon^2)$$

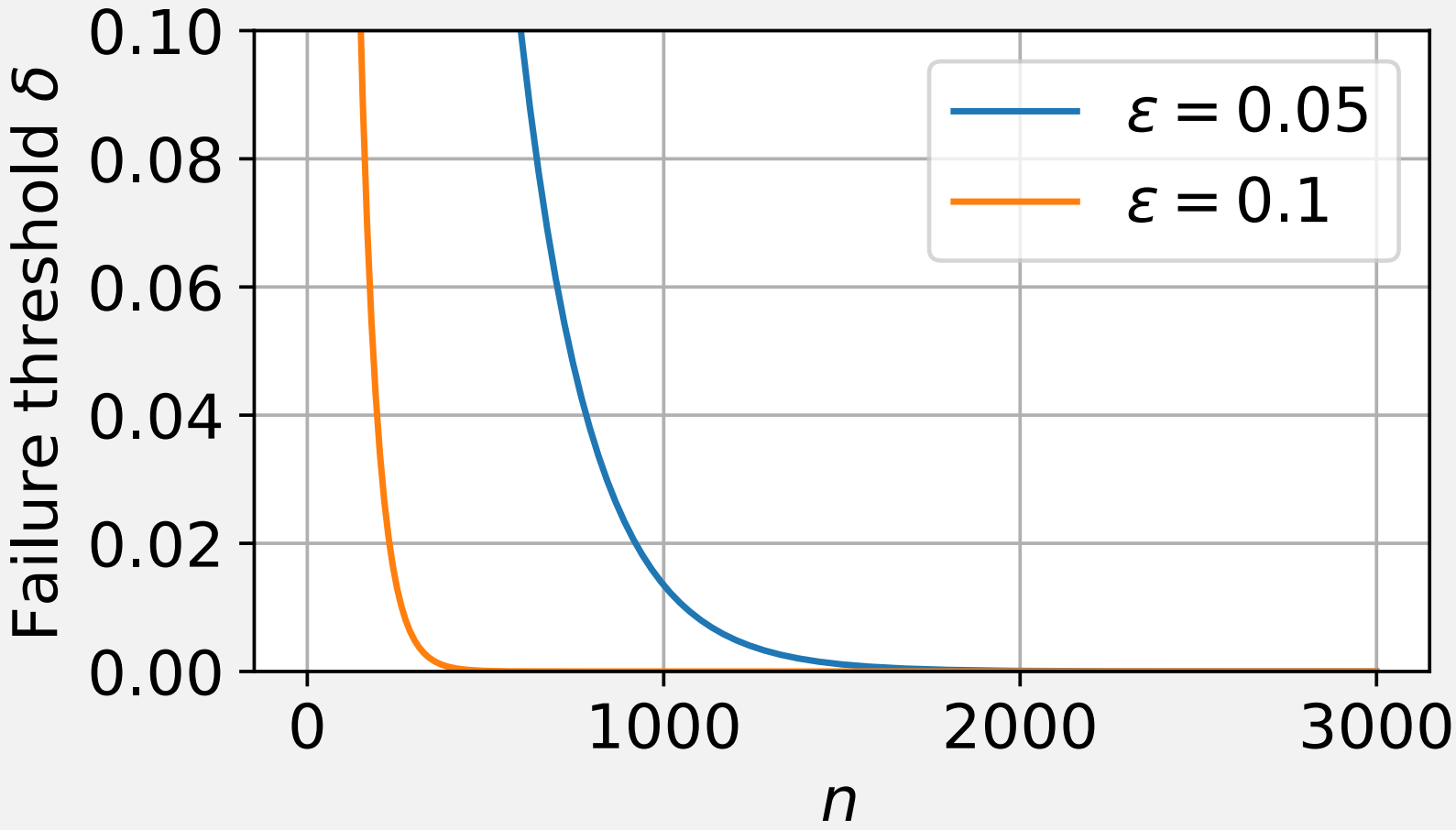
Function of n, ϵ

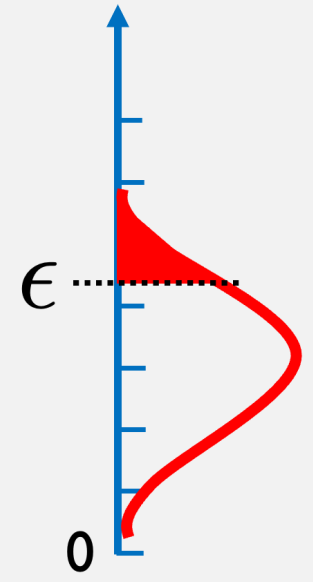




HOEFFDING'S INEQUALITY

$$p\left(\left|\mathbb{E}[X] - \frac{1}{n}\sum_{i=1}^{n} x_i\right| \geq \epsilon\right) \leq \delta \quad = \quad 2\exp(-2n\epsilon^2)$$

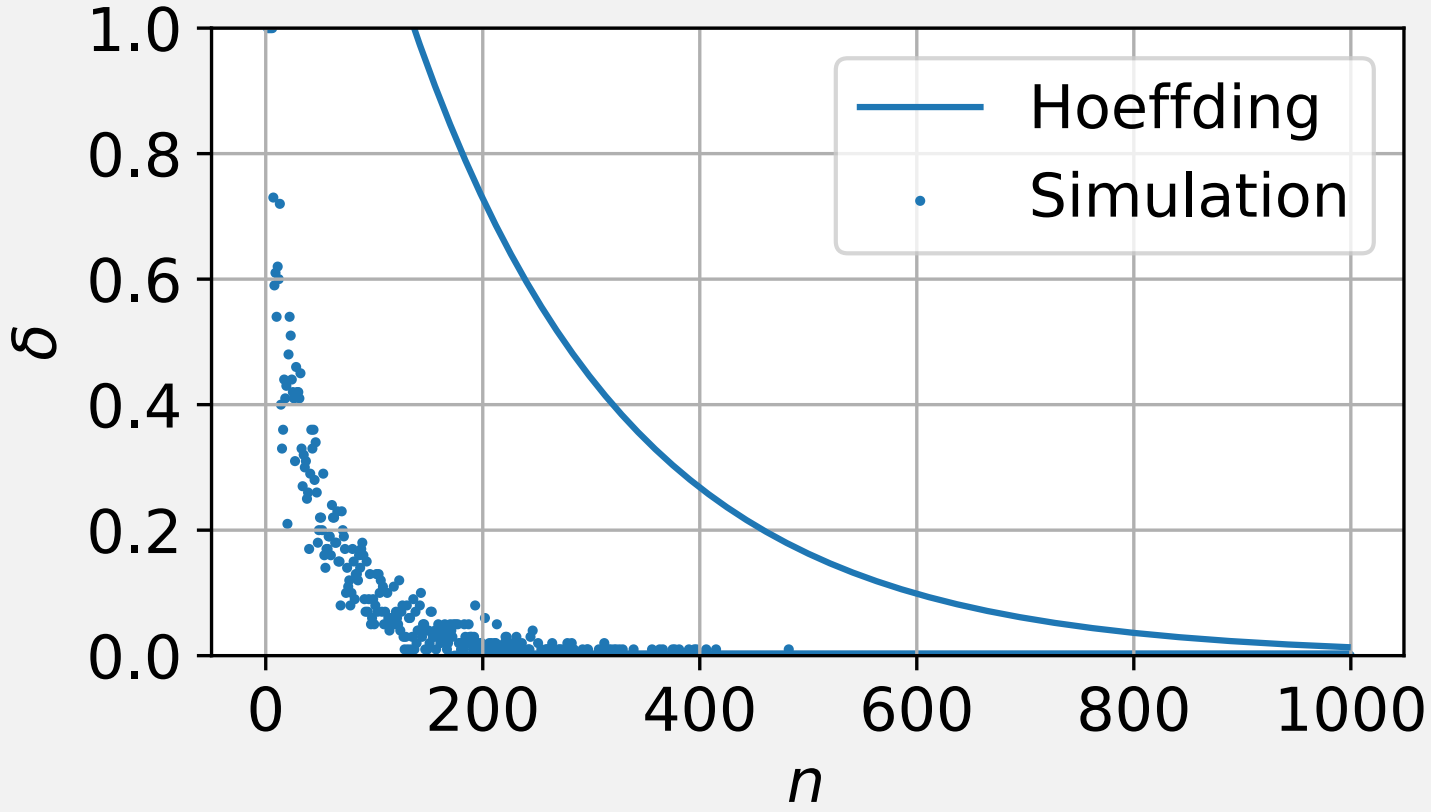


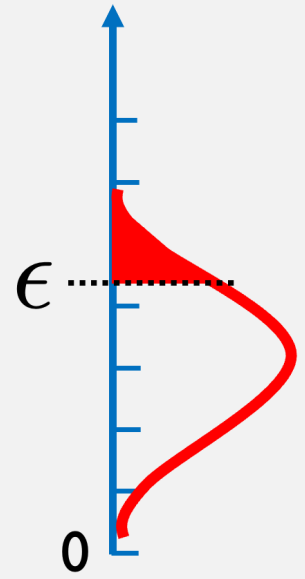


HOEFFDING'S INEQUALITY

$$p\left(\left|\mathbb{E}[X] - \frac{1}{n}\sum_{i=1}^{n} x_i\right| \geq \epsilon\right) \leq \delta \quad = \quad 2\exp(-2n\epsilon^2)$$

The bound is loose.
See notes & simulation



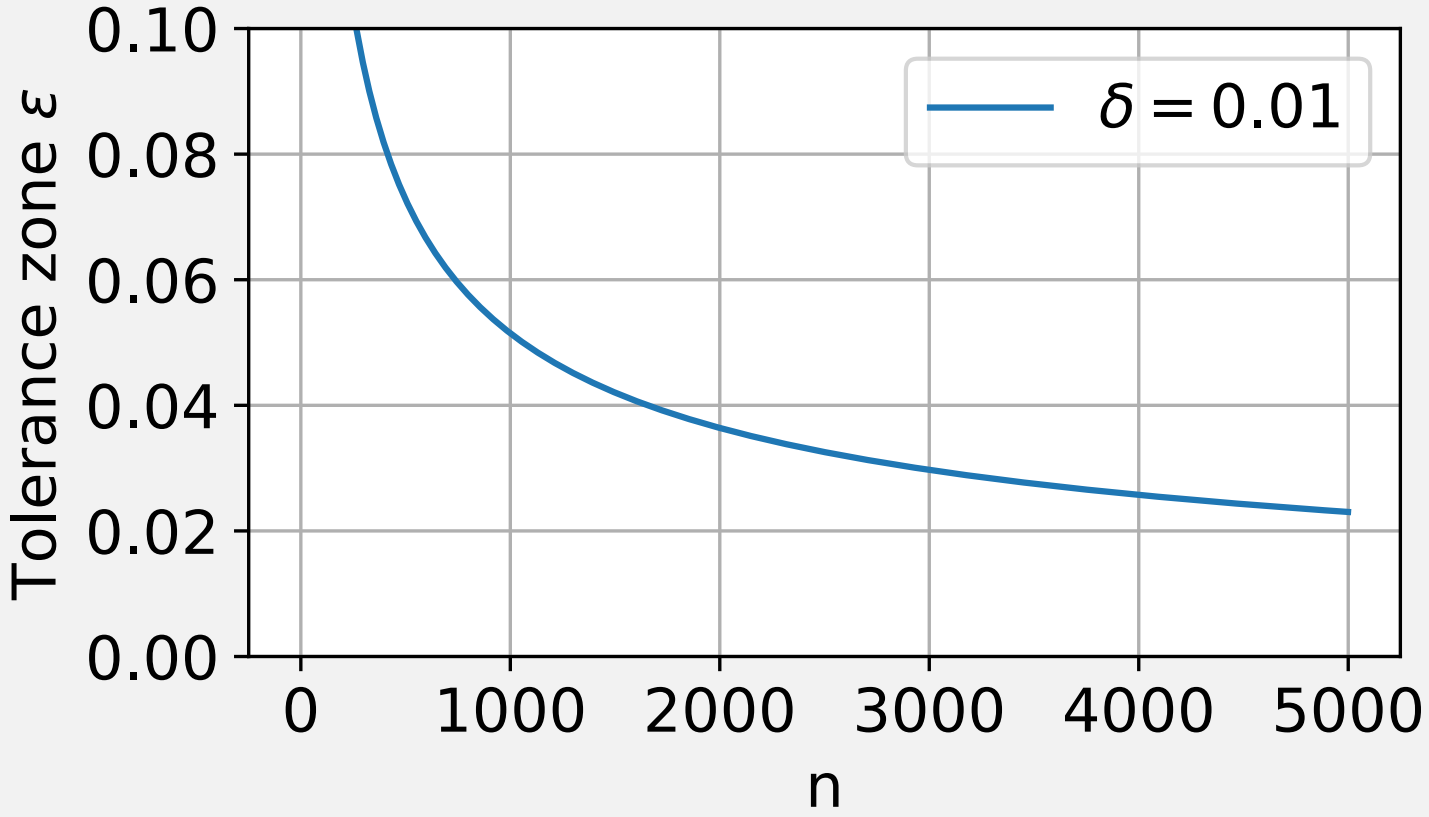


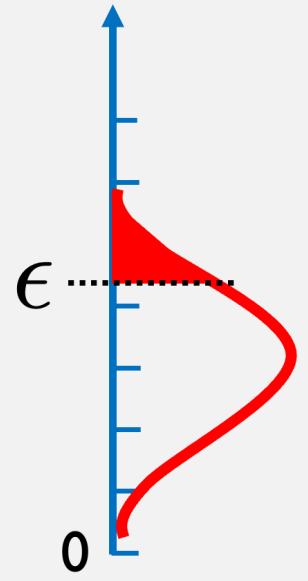
HOEFFDING'S INEQUALITY, REARRANGED

$$\delta = 2\exp(-2n\epsilon^2) \quad \xrightarrow[\text{solve}]{} \quad \epsilon = \sqrt{\frac{\ln(2/\delta)}{2n}}$$

$$\delta = 0.01$$

… we want 99% confidence

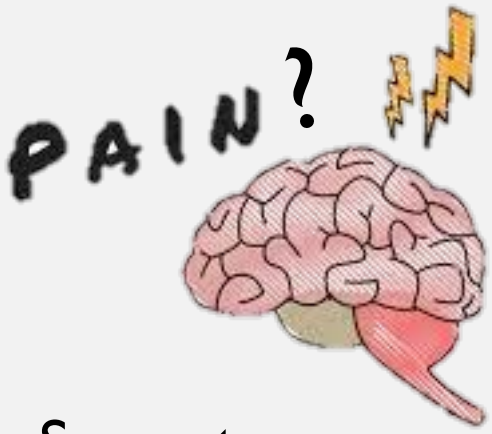




TAKE A SHORT BREAK

$$p\left(\left|\mathbb{E}[X] - \frac{1}{n}\sum_{i=1}^{n} x_i\right| \geq \epsilon\right) \leq \delta \quad = \quad 2\exp(-2n\epsilon^2)$$

$$p\left(\left|R(f) - \hat{R}(f, \mathcal{S}_n)\right| \geq \epsilon\right) \leq \delta = 2\exp(-2n\epsilon^2)$$

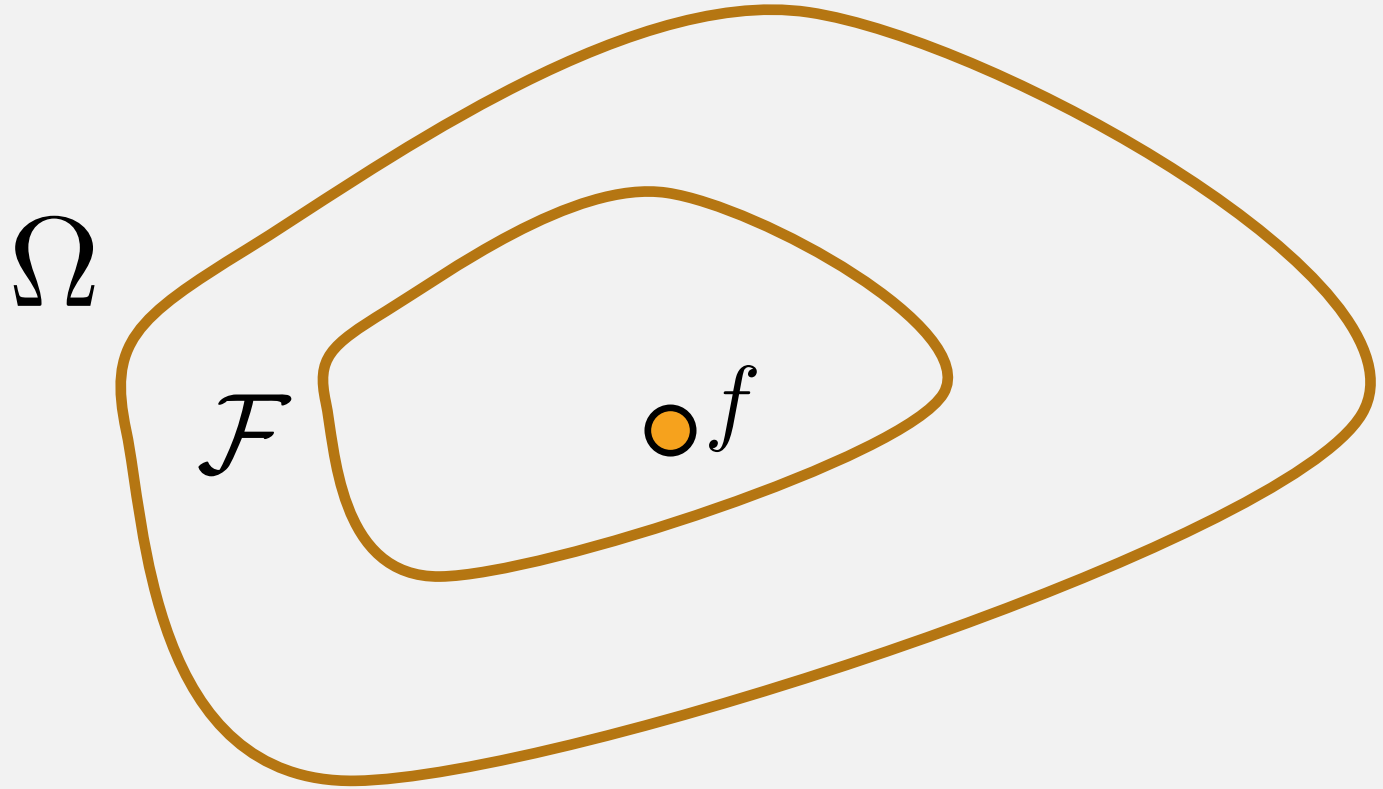


We assumed x = {0,1} but this applies more widely to any bounded loss. See notes.

THIS ONLY HOLDS FOR A FIXED FUNCTION…

$$p\left(\left|R(f) - \hat{R}(f, \mathcal{S}_n)\right| \geq \epsilon\right) \leq \delta = 2\exp(-2n\epsilon^2)$$

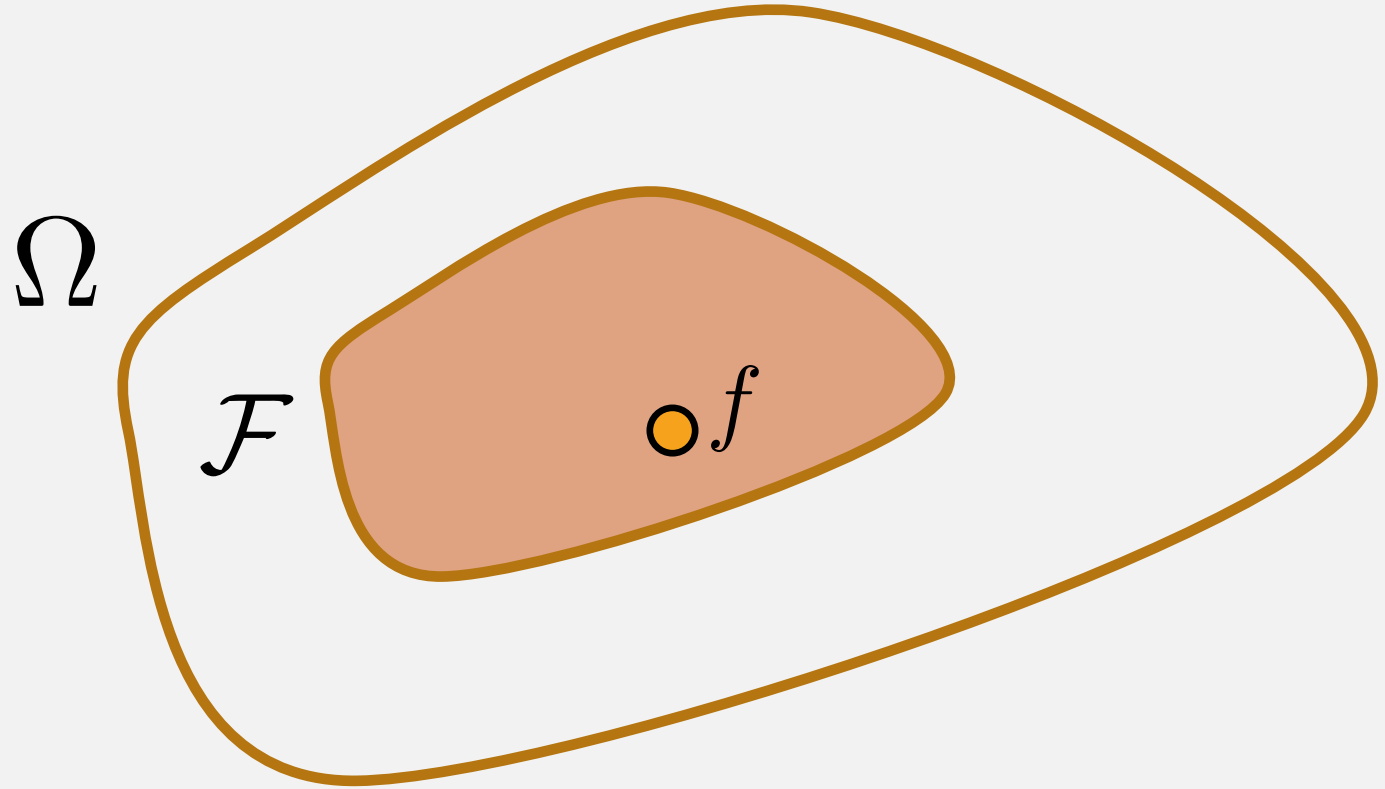
Chosen in advance.



BUT WE WANT IT TO HOLD FOR ALL FUNCTIONS

$$p\left(\left|R(f) - \hat{R}(f, \mathcal{S}_n)\right| \geq \epsilon\right) \leq \delta = 2\exp(-2n\epsilon^2)$$

?

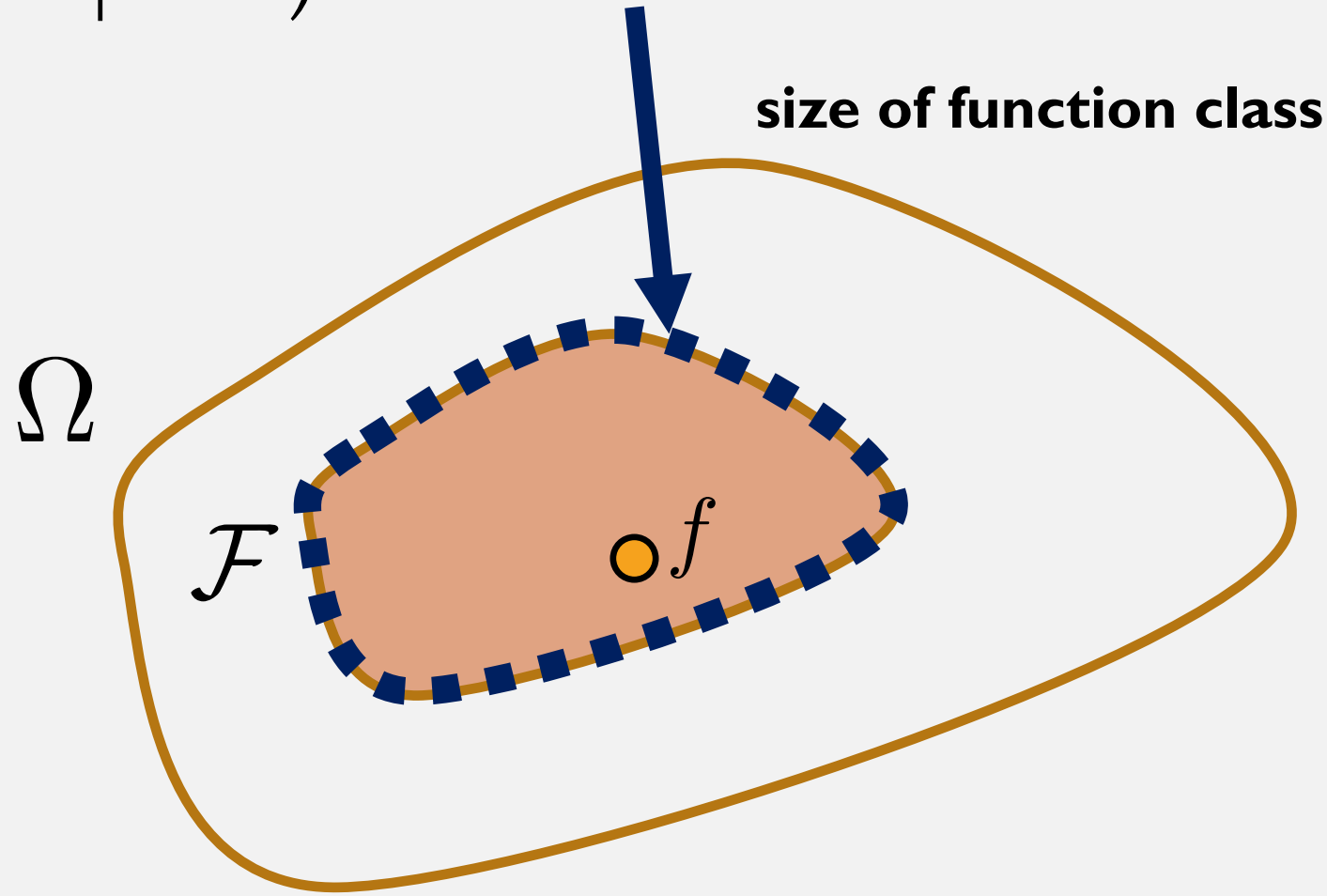


THE "UNIFORM DEVIATION BOUND"

$$p\Big(\exists f \in \mathcal{F},\ \Big|R(f) - \hat{R}(f, \mathcal{S}_n)\Big| \ \geq \ \epsilon\Big) \leq \delta = 2|\mathcal{F}|\exp(-2n\epsilon^2)$$

"there exists"

size of function class



The probability that there exists ***<u>at least one</u>*** *model violating our threshold,* δ

Proof in notes.

TAKING ONLY THE UPPER TAIL OF THE DISTRIBUTION

We bounded the probability of the event….

$$|R(f) - \hat{R}(f, \mathcal{S}_n)| > \epsilon$$

i.e. the absolute deviation. But we really only care if the true error is GREATER than our estimate…

$$R(f) - \hat{R}(f, \mathcal{S}_n) > \epsilon$$

Bounding that, we get a slight variation (details in notes) …

$$\delta = |\mathcal{F}| \exp(-2n\epsilon^2) \quad \rightarrow \quad \epsilon = \sqrt{\frac{\ln(|\mathcal{F}|) + \ln(1/\delta)}{2n}}$$

GENERALISATION BOUND

With probability $1 - \delta$ ...

size of function class

$$R(f) \leq \hat{R}(f, \mathcal{S}_n) + \sqrt{\frac{\ln(|\mathcal{F}|) + \ln(1/\delta)}{2n}}$$

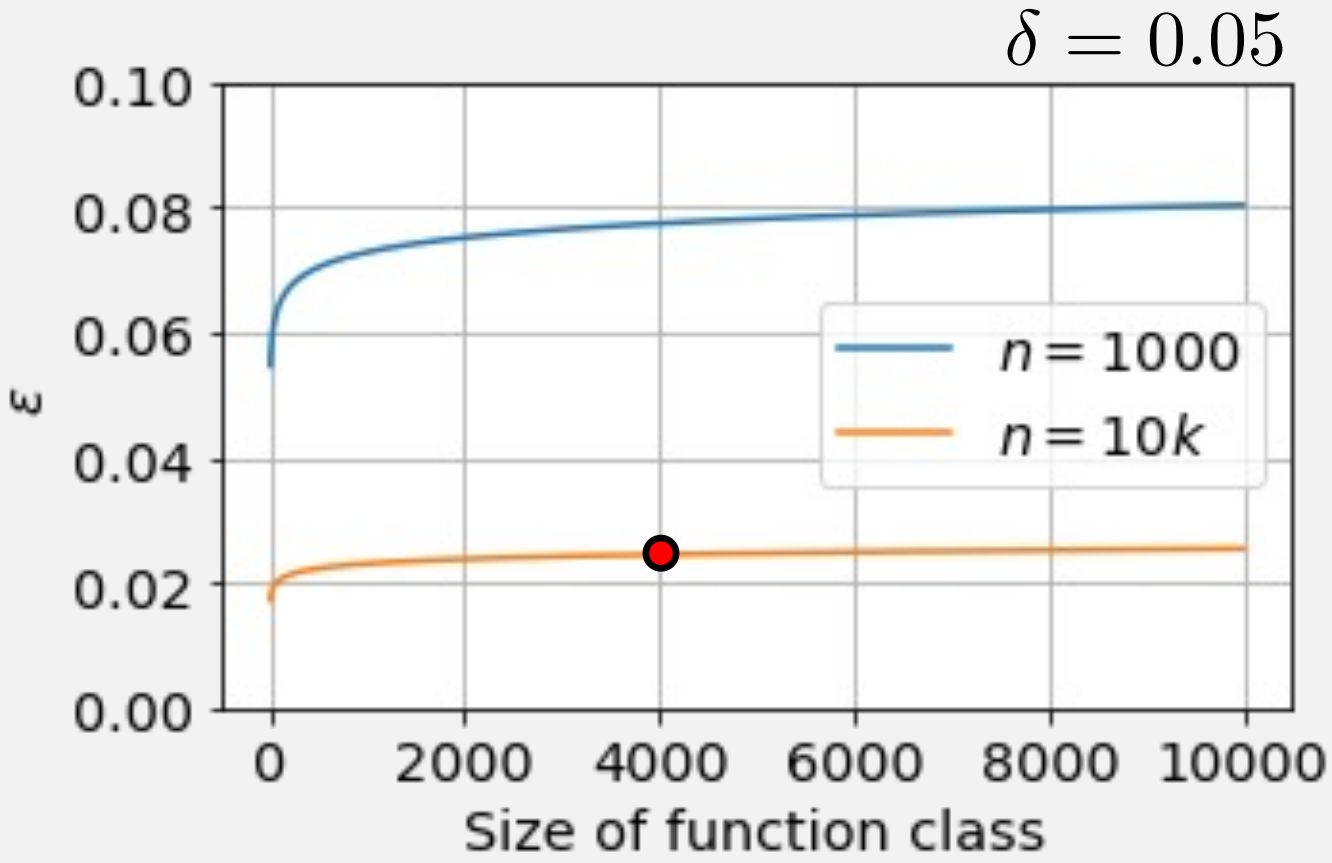
Population risk is not more than

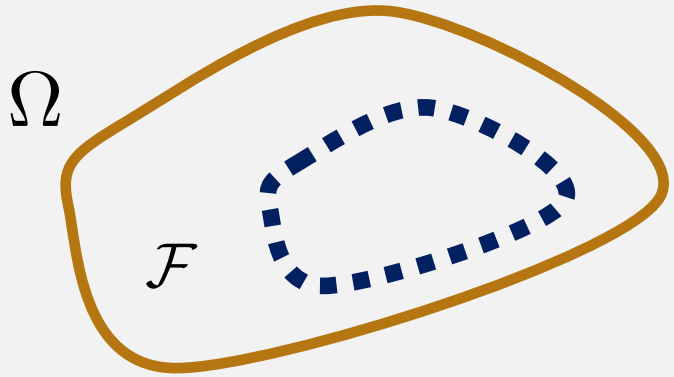
Empirical risk

plus this small value, ϵ

GROWING OUR FUNCTION SPACE

$$R(f) \leq \hat{R}(f, \mathcal{S}_n) + \sqrt{\frac{\ln(|\mathcal{F}|) + \ln(1/\delta)}{2n}}$$





We can be 95% confident that…

with n=10k datapoints
and a function class of size 4000

the population risk
is at most
the empirical risk
plus 0.025.

TODAY WE DID…

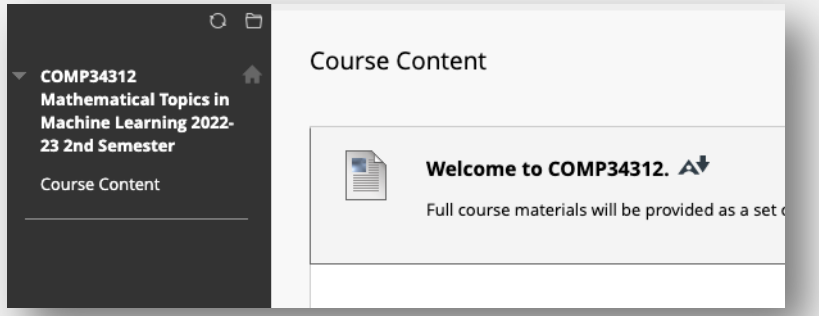
- Hoeffding's inequality
- uniform deviation bound
- a generalisation bound for finite function classes

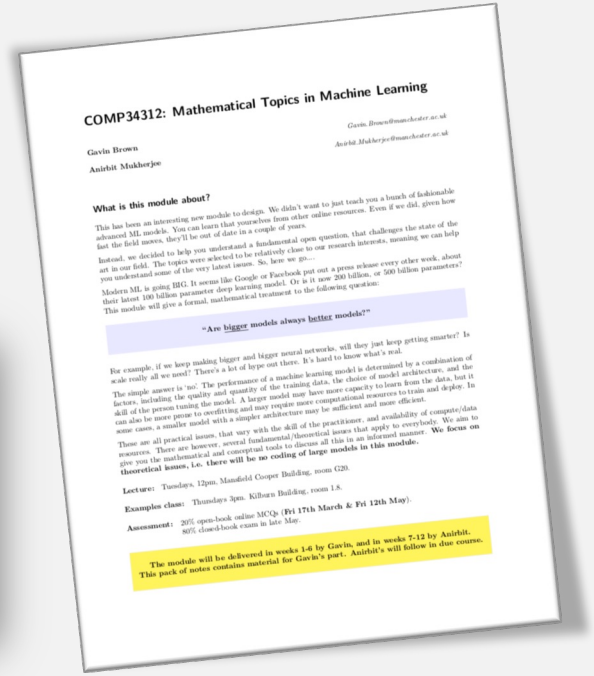
This week :

…Read the notes, thoroughly.

…Go over the proof for the uniform deviation bound.

…Play with the simulation code.




COMP34312: Mathematical Topics in Machine Learning

Gavin Brown

Anirbit Mukherjee

What is this module about?

"Are bigger models always better models?"

Lecture: Tuesdays, 12pm, Mansfield Cooper Building, room G20.

Examples class: Thursdays 3pm. Kilburn Building, room 1.8.

Assessment: 20% open-book online MCQs (Fri 17th March & Fri 12th May).
80% closed-book exam in late May.

The module will be delivered in weeks 1-6 by Gavin, and in weeks 7-12 by Anirbit.
This pack of notes contains material for Gavin's part. Anirbit's will follow in due course.

SEE YOU IN THE STUDY SESSIONS