

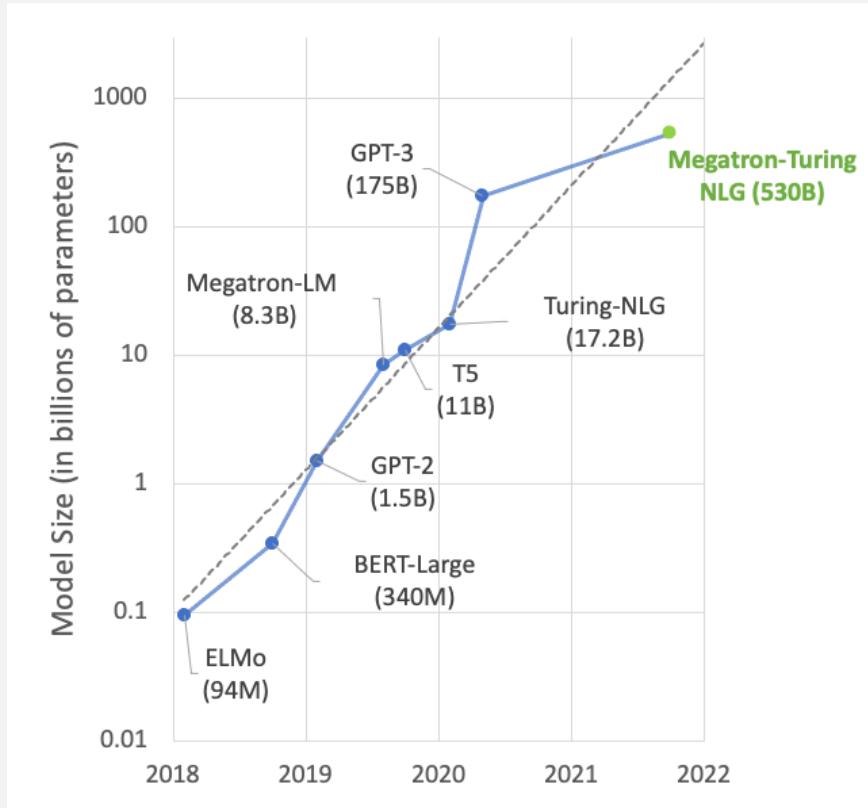
MATHEMATICAL TOPICS IN MACHINE LEARNING

(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

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

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

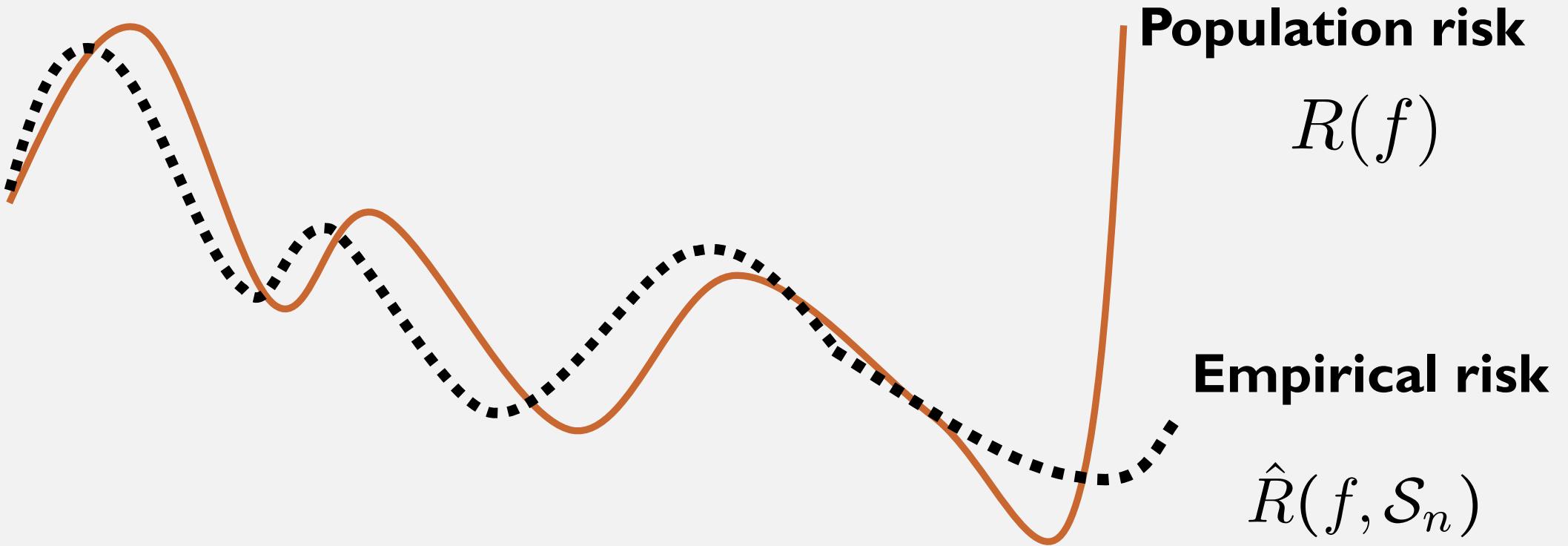
The **estimated** risk, using
an IID sample of size n .

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

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

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

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

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?

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

WEAK LAW OF LARGE NUMBERS

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

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...

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

Our estimate is...

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

$$\frac{1}{n} \sum_{i=1}^n x_i \rightarrow \mathbb{E}[X] \quad \text{as } n \rightarrow \infty$$

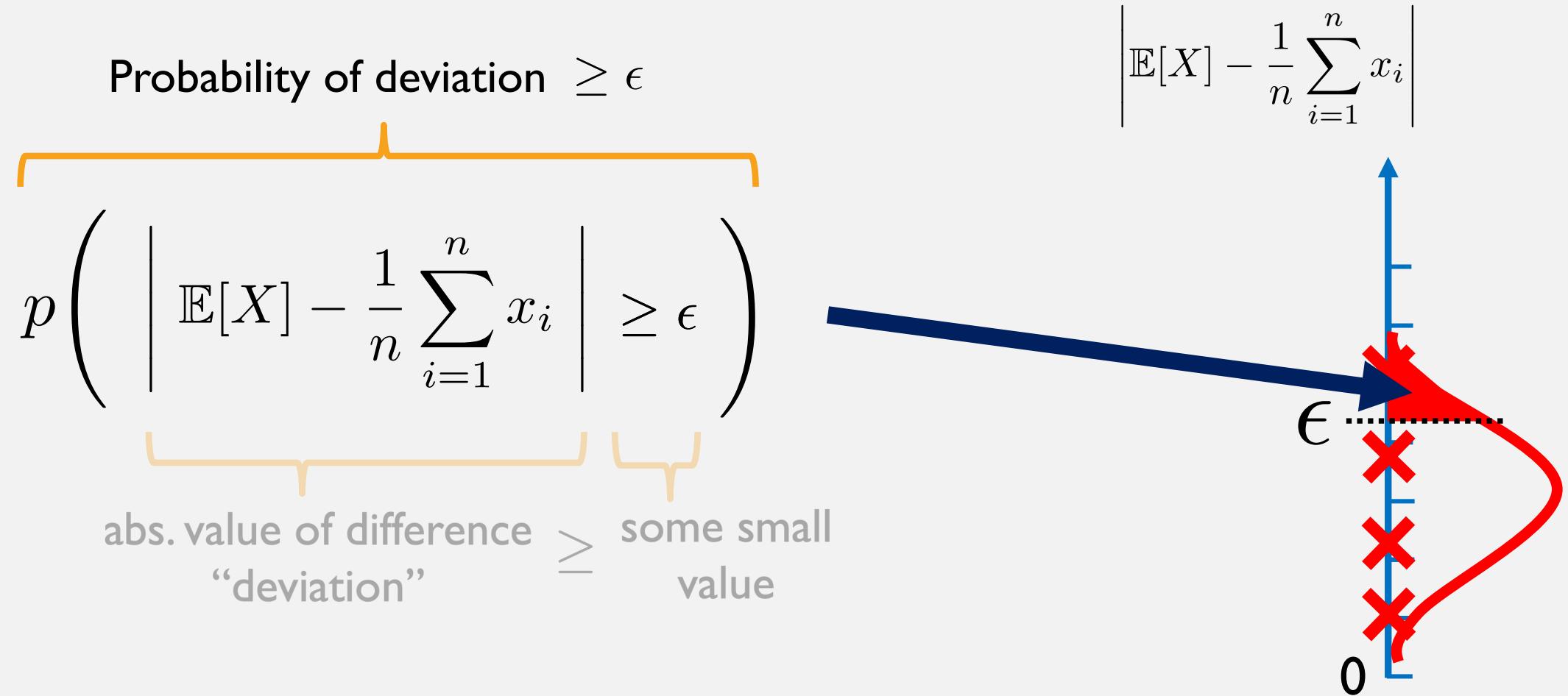
LET'S SEE HOW 'FAST' THEY CONVERGE

$$R(f) \quad \hat{R}(f, \mathcal{S}_n)$$
$$\left| \mathbb{E}[X] - \frac{1}{n} \sum_{i=1}^n x_i \right| \geq \epsilon$$



abs. value of difference \geq some small
“deviation” value

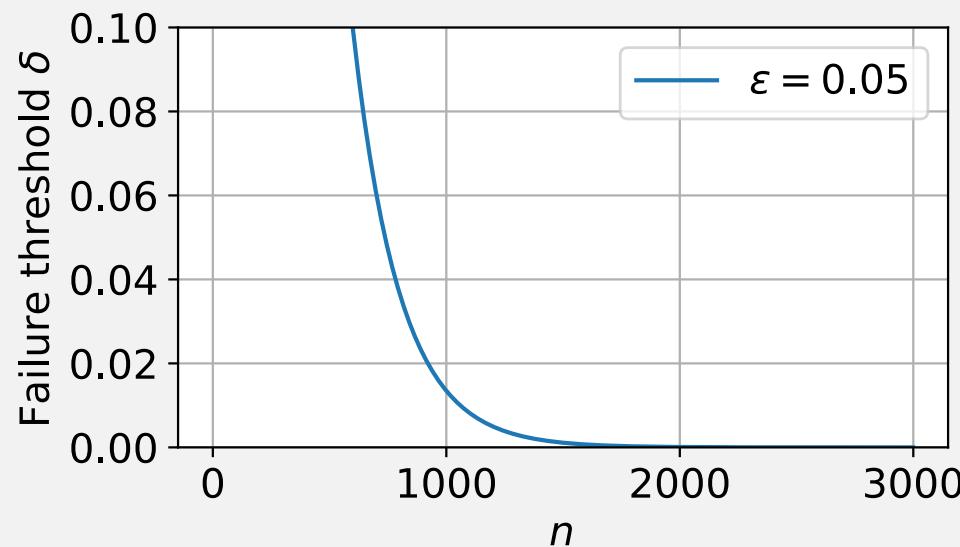
LET'S SEE HOW 'FAST' THEY CONVERGE



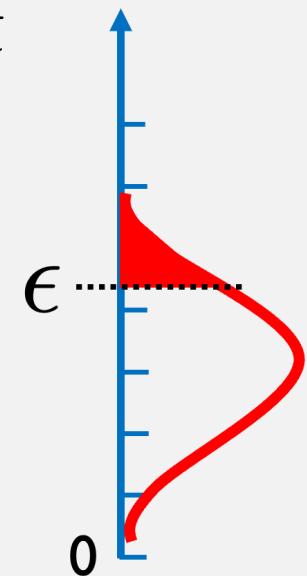
HOEFFDING'S INEQUALITY

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

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



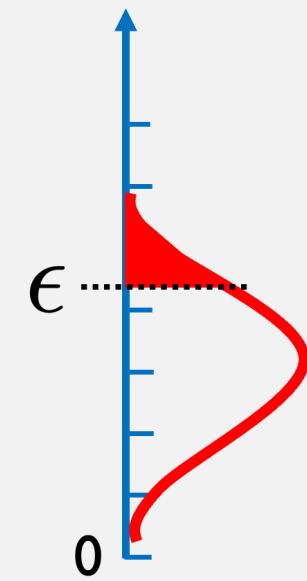
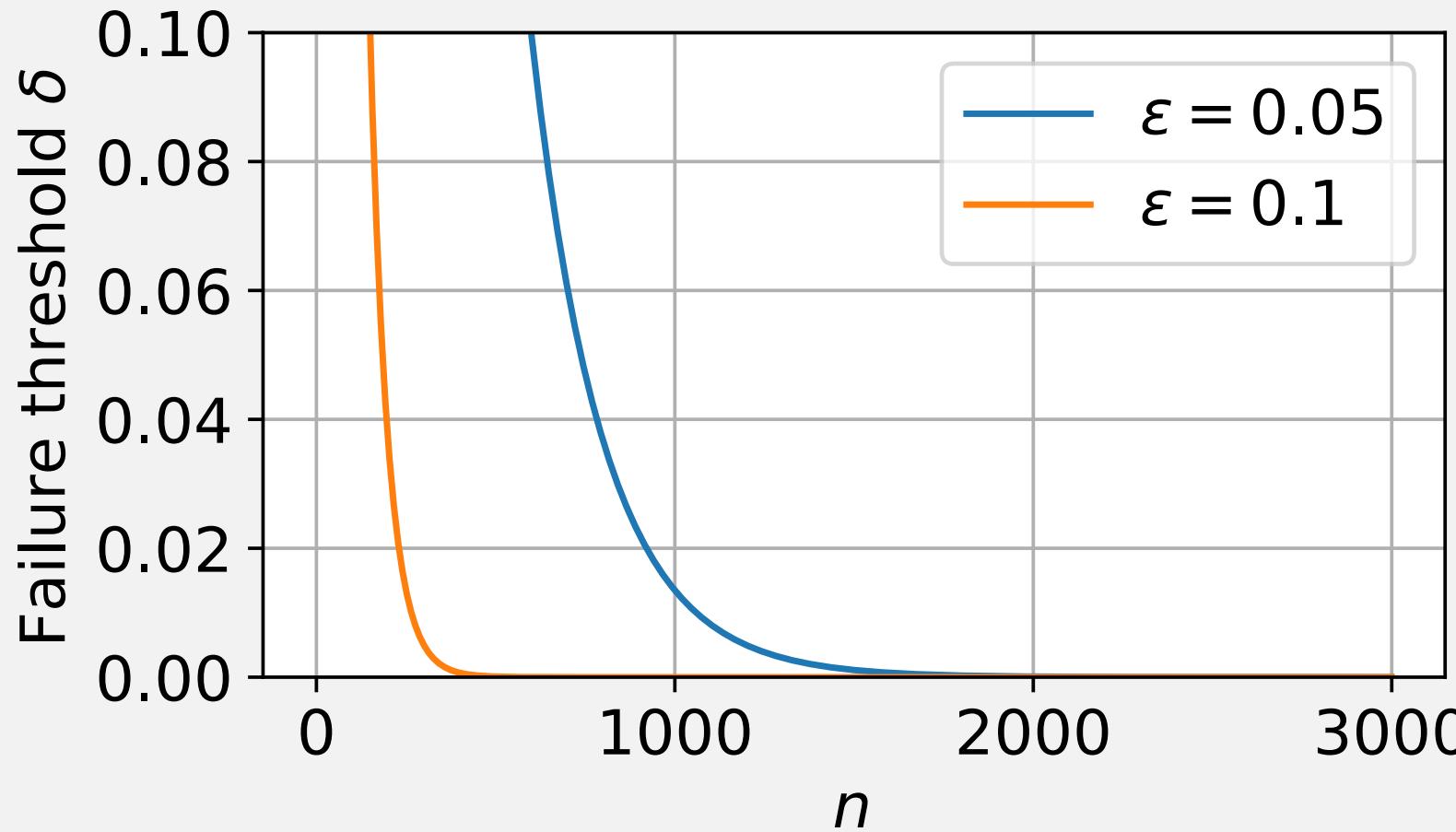
Function of n, ϵ



Proof is beyond the scope of the module.

HOEFFDING'S INEQUALITY

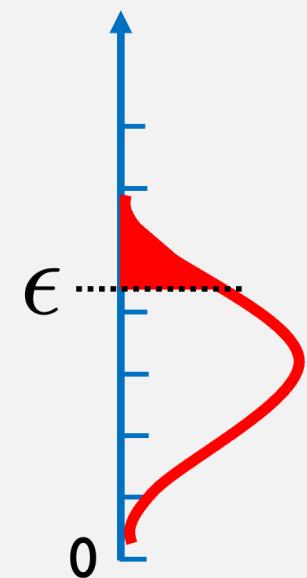
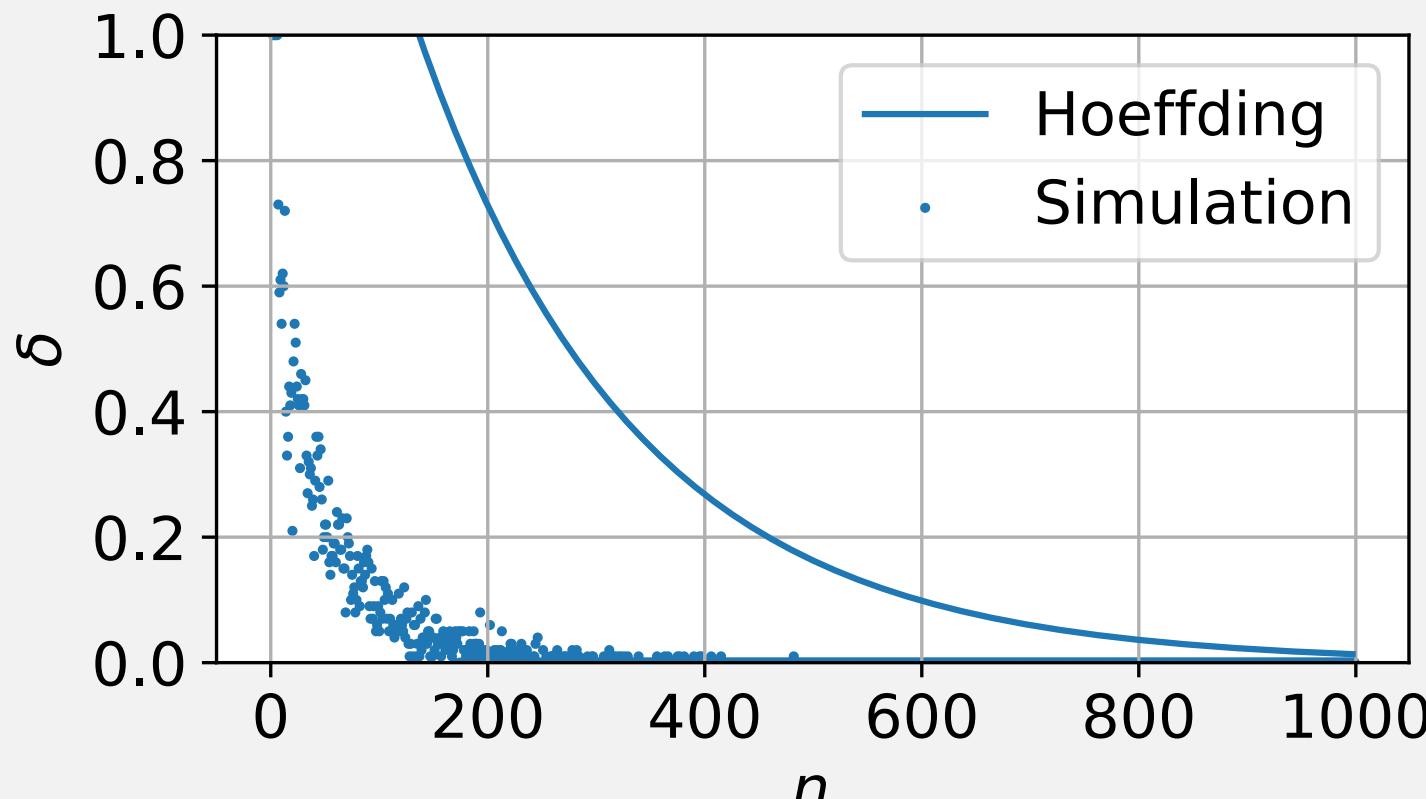
$$p \left(\left| \mathbb{E}[X] - \frac{1}{n} \sum_{i=1}^n x_i \right| \geq \epsilon \right) \leq \delta = 2\exp(-2n\epsilon^2)$$



HOEFFDING'S INEQUALITY

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

**The bound is loose.
See notes & simulation**



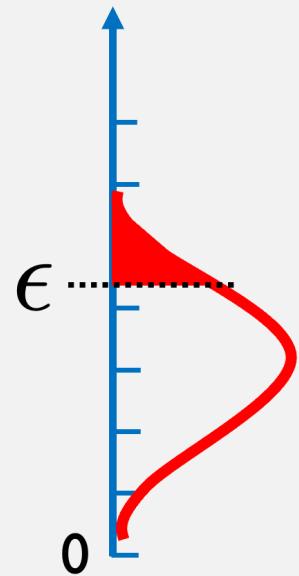
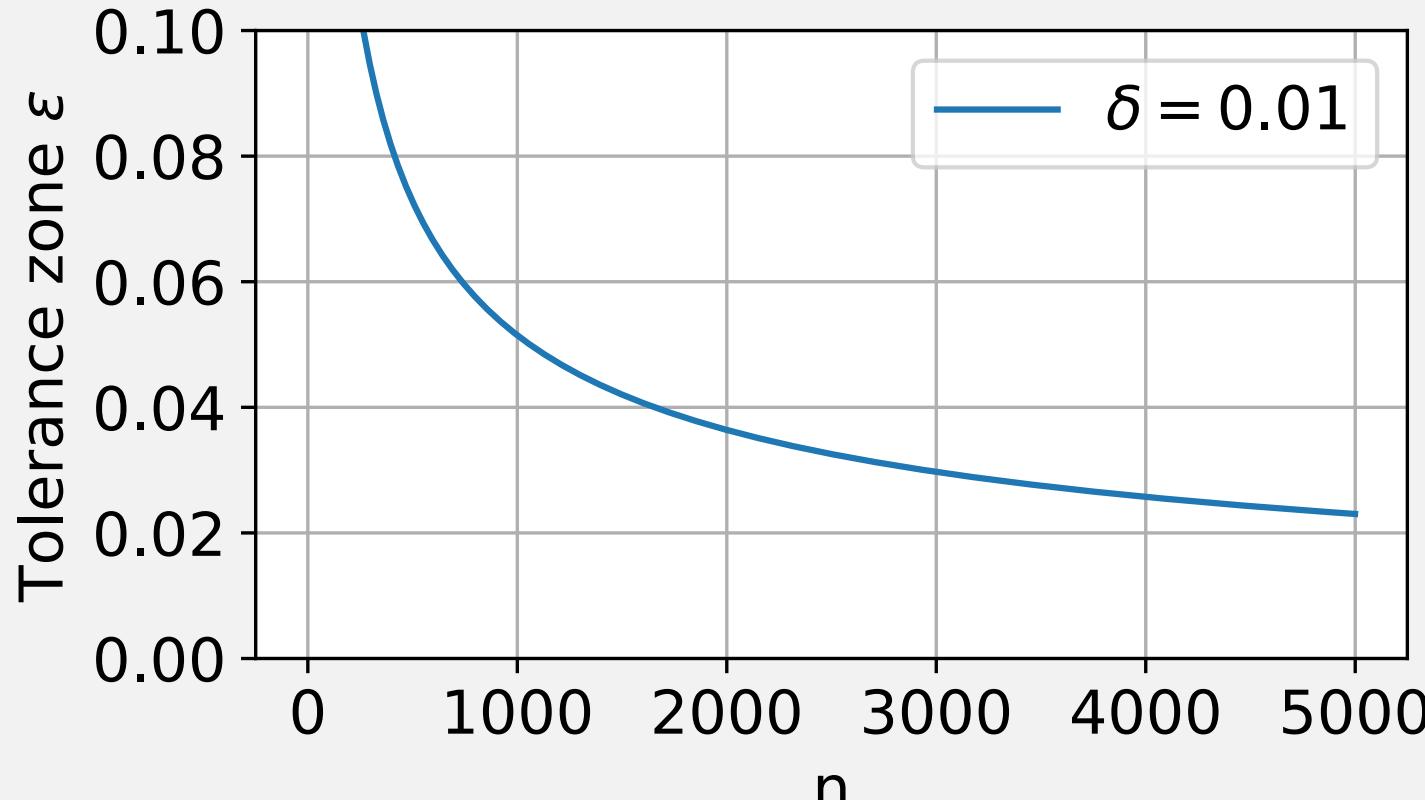
HOEFFDING'S INEQUALITY, REARRANGED

$$\delta = 2\exp(-2n\epsilon^2) \rightarrow \epsilon = \sqrt{\frac{\ln(2/\delta)}{2n}}$$

solve

$$\delta = 0.01$$

... we want 99% confidence



TAKE A SHORT BREAK

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



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

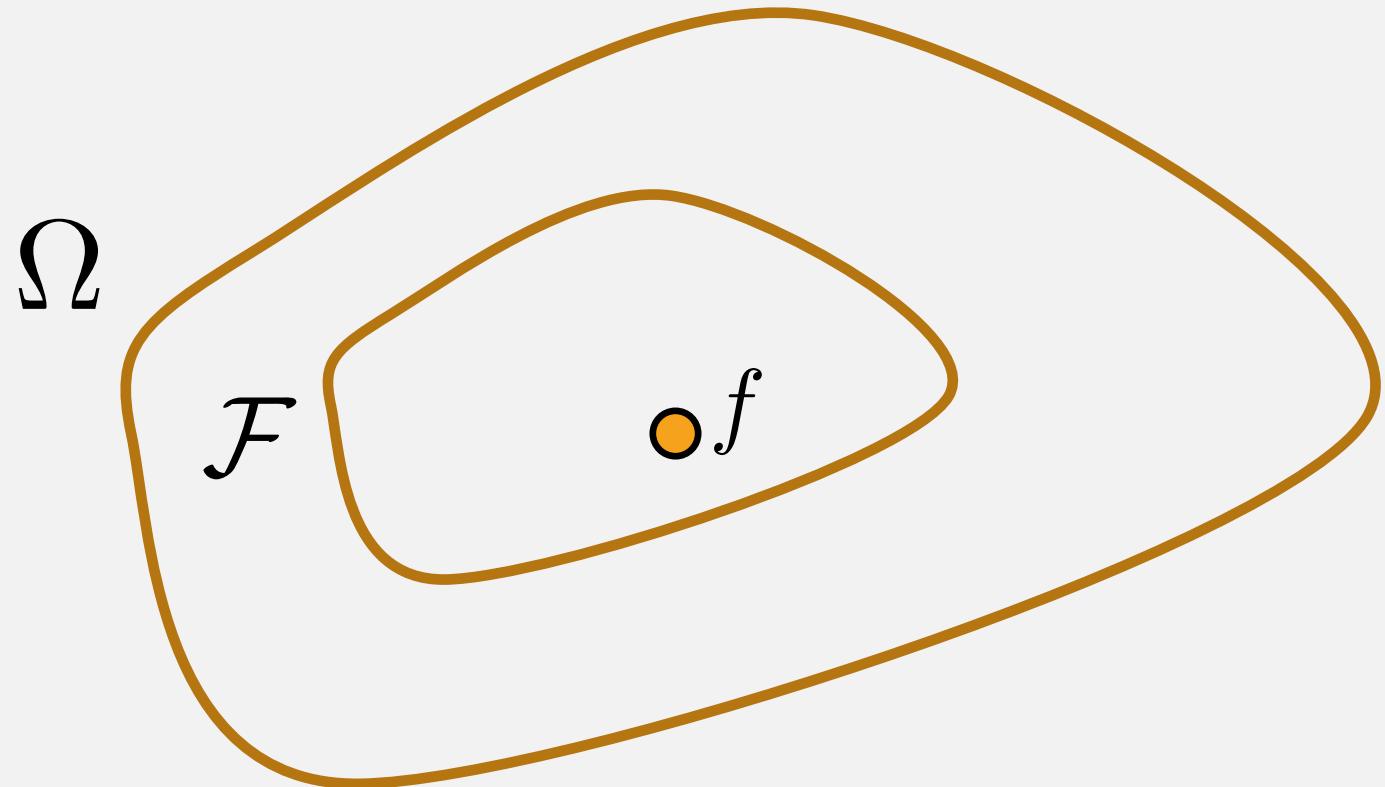


We assumed $x = \{0, 1\}$ but this applies more widely to any bounded loss. See notes.

THIS ONLY HOLDS FOR A FIXED FUNCTION...

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

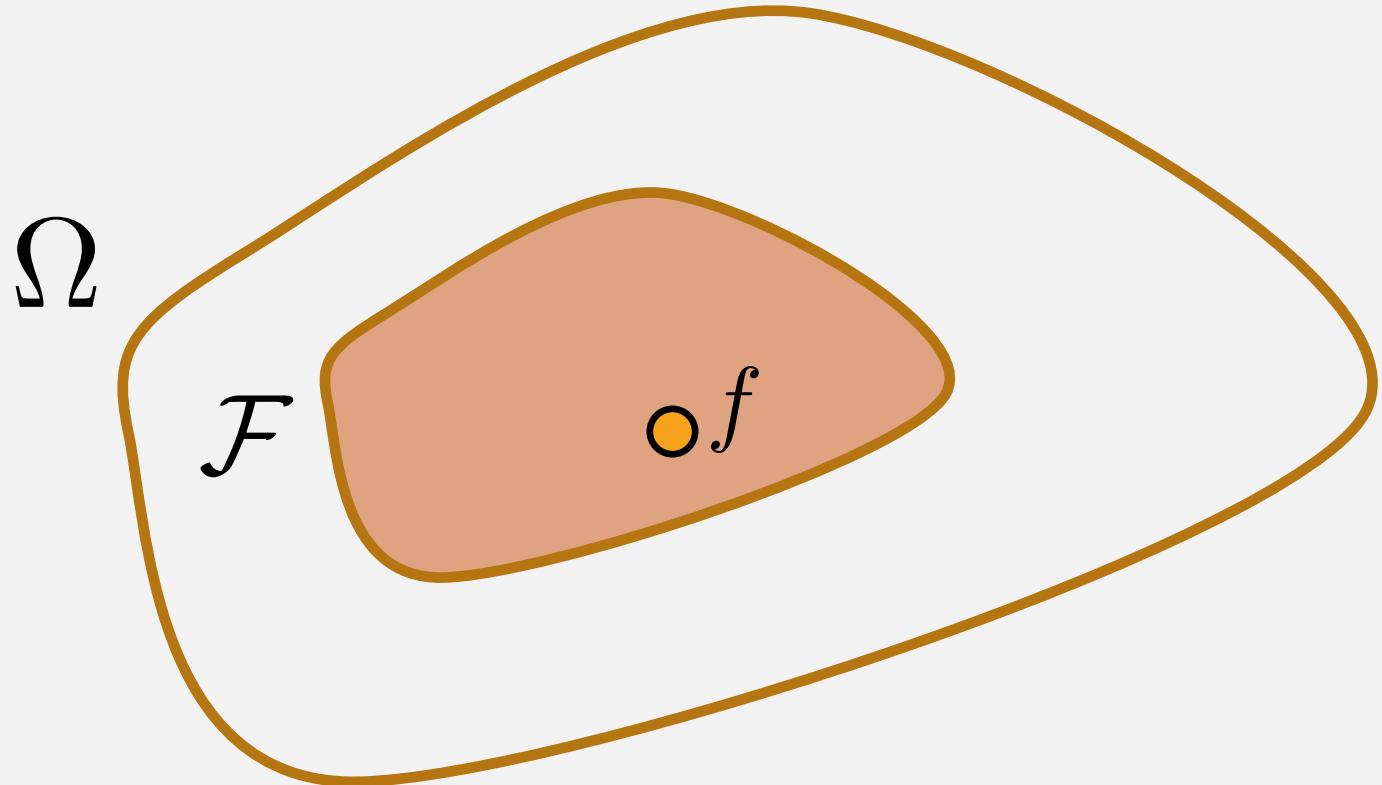
Chosen in advance.



BUT WE WANT IT TO HOLD FOR ALL FUNCTIONS

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

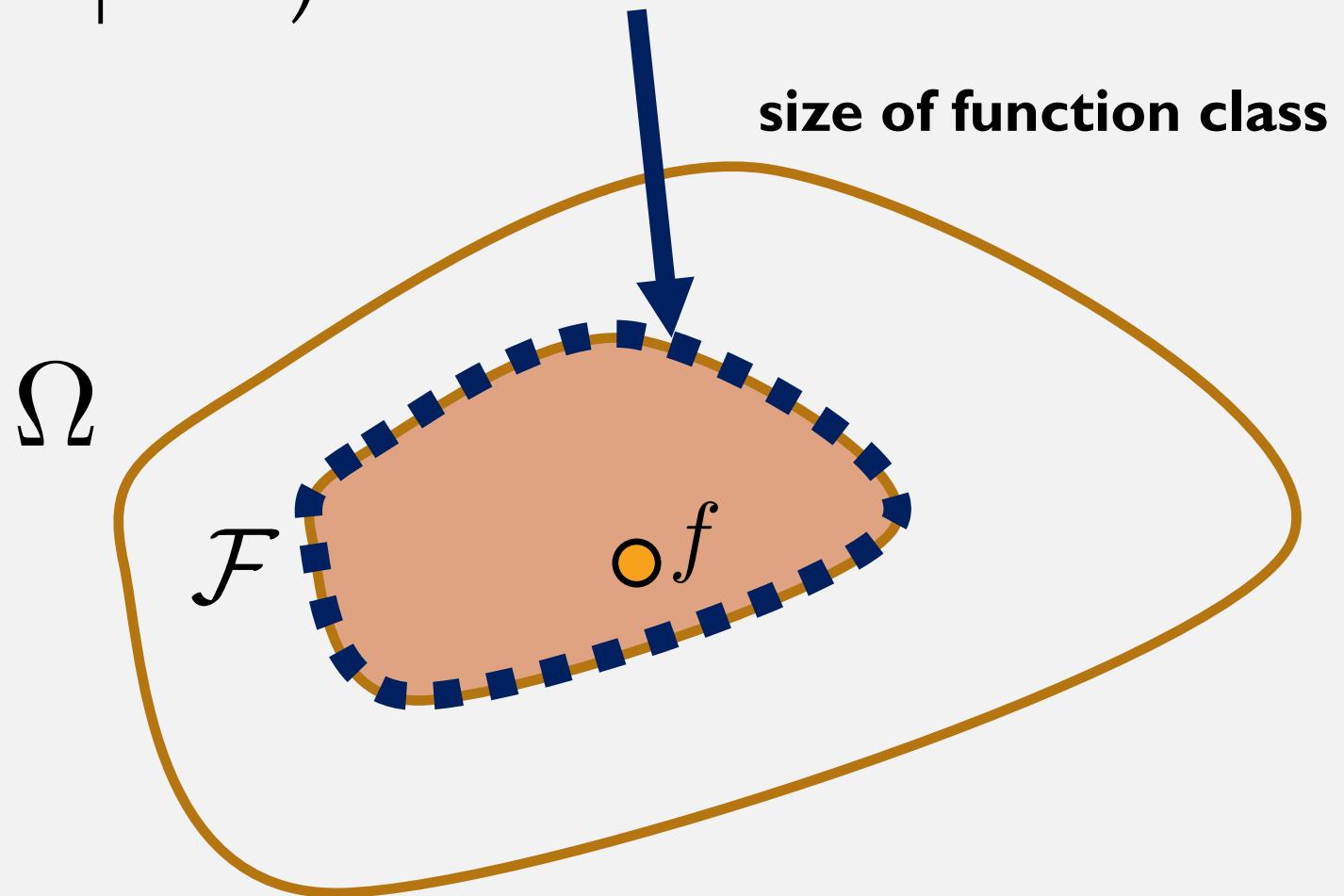
?



THE “UNIFORM DEVIATION BOUND”

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

“there exists”



The probability that there exists
at least one 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 $1 - \delta$...

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

size of function class

Population risk
is not more than

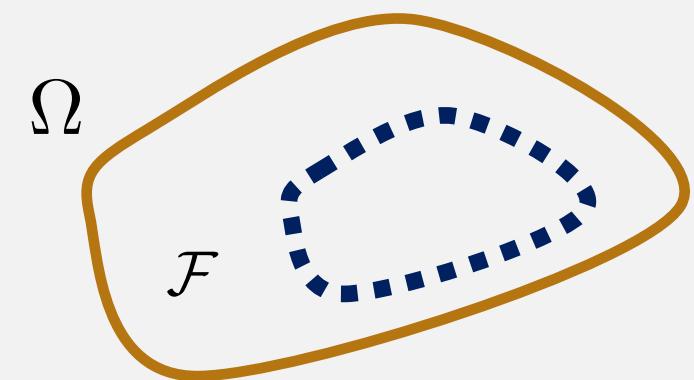
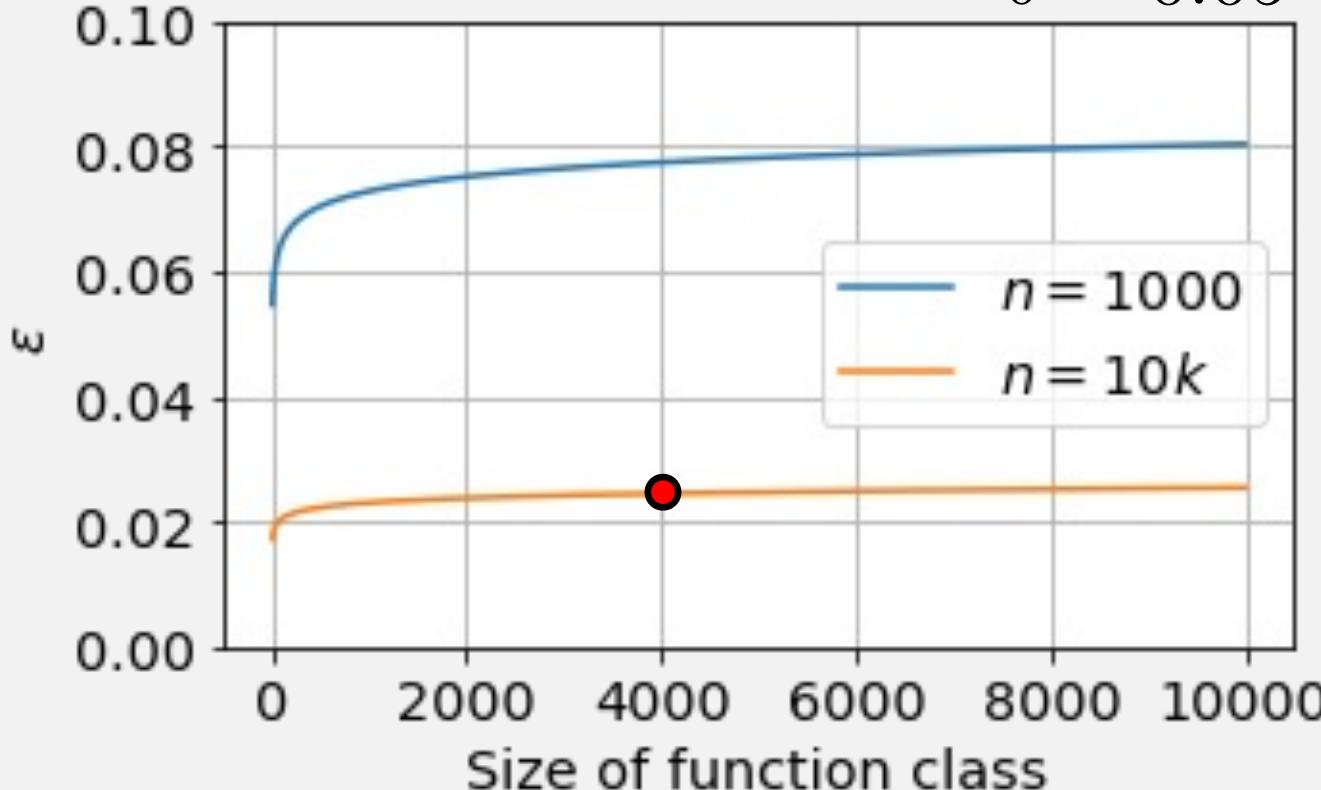
Empirical
risk

plus this small value, ϵ

GROWING OUR FUNCTION SPACE

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

$$\delta = 0.05$$



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.

The screenshot shows a course page with the following details:

- COMP34312**
Mathematical Topics in
Machine Learning 2022-
23 2nd Semester
- Course Content**
- Welcome to COMP34312.** A small icon of a document with a downward arrow is next to the text.
- Full course materials will be provided as a set of notes.**

COMP34312: Mathematical Topics in Machine Learning

Gavin.Brown@manchester.ac.uk
Anishit.Mukherjee@manchester.ac.uk

What is this module about?

This has been an interesting new module to design. We didn't want to just teach you a bunch of fashionable advanced ML models. You can learn that yourself from other online resources. Even if we did, given how fast the field moves, they'll be out of date in a couple of years.

Instead, we decided to help you understand a fundamental open question, that challenges the state-of-the-art in our field. The topic was selected to be relatively close to our research interests, meaning we can help you understand some of the very latest issues. So, what is this question? What is the future of machine learning? Is it going IIGC. It seems like Google or Facebook put out a press release every other week, about their latest 100 billion parameter neural learning model. Or is it now 200, or 500 billion parameters? This module will give a formal, mathematical treatment to the following question:

"Are bigger models always better models?"

For example, if we keep making bigger and bigger neural networks, will they just keep getting smarter? Is scale really all we need? There's a lot of hype out there. It's hard to know what's real.

The simple answer is 'no'. The performance of a machine learning model is determined by a combination of its architecture, the quality and quantity of training data, the choice of model architecture, and the skill of the person tuning the model. A larger model may have more capacity to learn from the data, but it can also be more difficult to train and may require more computational resources to train and deploy. In some cases, a smaller model with a simpler architecture may be sufficient and more efficient.

These are all practical issues, that vary with the skill of the practitioner, and availability of compute/data resources. There are however some fundamental theoretical issues that apply to everybody. We focus on these theoretical issues, i.e. there will be no coding of large models in this module.

Lecture: Tuesdays, 12pm, Mansfield Cooper Building, room G20.
Example class: Thursdays 1pm, Kilburn Building, room 1.8.
Assessment: 20% open-book online MCQ (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 Anishit. This pack of notes contains material for Gavin's part. Anishit's will follow in due course.

SEE YOU IN THE STUDY SESSIONS