PAC learning theory

Machine Learning @ UWr 2020

Lecture 11

Simplified PAC theory

- The PAC (Probably Approximately Correct) model:
- The data distribution is stationary:
  - $x, y \sim P(x, y)$
- The training samples are drawn i.i.d. (independently, identically distributed)
- The hypothesis space $\mathcal{H}$ is finite and has size $|\mathcal{H}|$
- The error rate (error probability on a random sample) of an $h \in \mathcal{H}$ is:
  - $error(h) = \sum_{all\ x,y}[y \neq h(x)]P(x, y) = \mathbb{E}_{P(x,y)}[y \neq h(x)]$
- We learn by choosing a $h_O \in \mathcal{H}$ that agrees with all training data
- What is the probability that h_O has a low error rate: $error(h) < \epsilon$?

PAC intuition

- We can find a seriously wrong hypothesis by testing it against N examples.
  If we can't say h is bad after sufficiently many tests, it is unlikely that h is seriously wrong.
- We will then say it is probably approximately correct.

Our learning algorithm

Assume a hypothesis space $\mathcal{H}$

Test each hypothesis on N samples, return first 100% accurate

```
for h in ℋ:
    for i in 1..N:
        if y^(i) ≠ h(x^(i)):
            next h
    return h          # hypothesis h is consistent with N samples
return None           # no hypothesis in H was found
```

A (simplified) PAC bound

$\mathcal{H}_{good} = \{h \in \mathcal{H}: error(h) < \epsilon\}$
$\mathcal{H}_{bad} = \mathcal{H} \setminus \mathcal{H}_{good}$

What is the prob. of not rejecting an $h_b \in \mathcal{H}_{bad}$?

$error(h_b) > \epsilon$
$P(h_b \text{ correct on } N \text{ samples}) \leq (1-\epsilon)^N$

What is the prob. of selecting $h_b \in \mathcal{H}_{bad}$ tested to be consistent with N samples?

$P(\text{selecting } h_b \in \mathcal{H}_{bad}) \leq |\mathcal{H}_{bad}|(1-\epsilon)^N \leq |\mathcal{H}|(1-\epsilon)^N \leq |\mathcal{H}|e^{-N\epsilon}$

A PAC bound

The prob. That our learning algo fails (gives us a bad hyp.) is

$P(\text{selecting } h_b \in \mathcal{H}_{bad}) \leq |\mathcal{H}|e^{-N\epsilon}$

We want to ensure this is less than δ.
Solve for N:

$$|\mathcal{H}|e^{-N\epsilon} < \delta$$

$$N \geq \frac{1}{\epsilon}\left(\ln\frac{1}{\delta} + \ln|\mathcal{H}|\right)$$

The space of all Boolean functions

- There are 2^{2^n} Boolean functions of n variables
- Therefore, to learn a hypothesis from the space of all Boolean functions of n variables we need to see $O(2^n)$ examples, or nearly all of them :(
- To learn from smaller number of examples we need to constrain our hypothesis space – e.g., consider only simple functions.

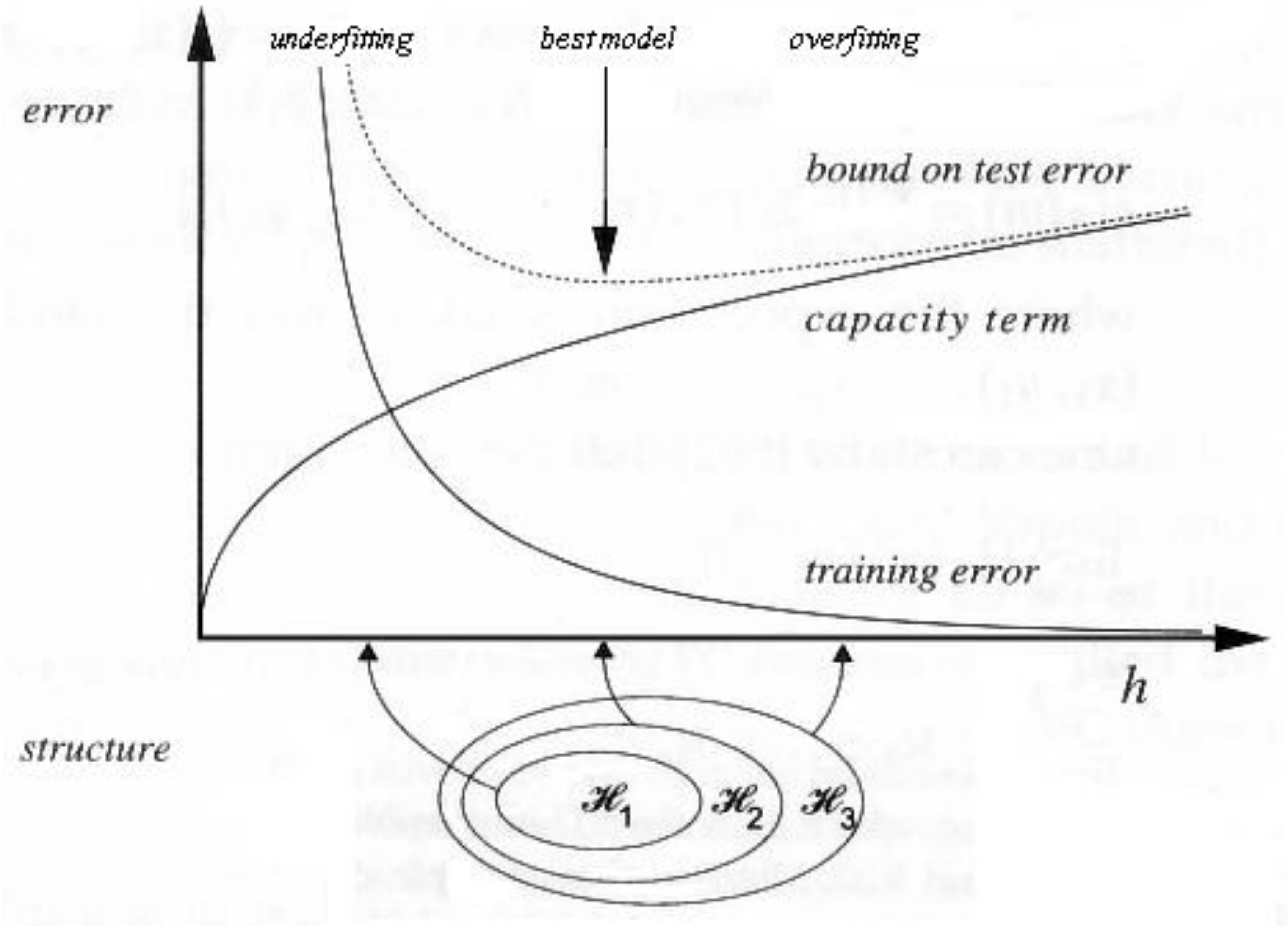
What about infinite $\mathcal{H}$?

- A naïve approach assumes that in a PC we never get an infinite number of models (floats have limited precision)
- The truly infinite case is solved by the Statistical Learning Theory (or the Vapnik-Chervonenkis, VC-theory)
- It introduces a measure of hypothesis complexity called VC-dimension
- PAC and VC theory are consistent
- If you are interested, see the book “Statistical Learning Theory” by Vladimir Vapnik.

How is regularization related to PAC

- Intuitively, less parameters means smaller $|\mathcal{H}|$.
- For infinite models, the VC dimension measures the hypothesis complexity.
- The more regularized a model, the smaller its VC dimension.
- Models with low VC dimension underfit, while those with a large VC dimension overfit.
- Need to optimally regularize (find optimal VC dim) (This is called structural risk minimization)

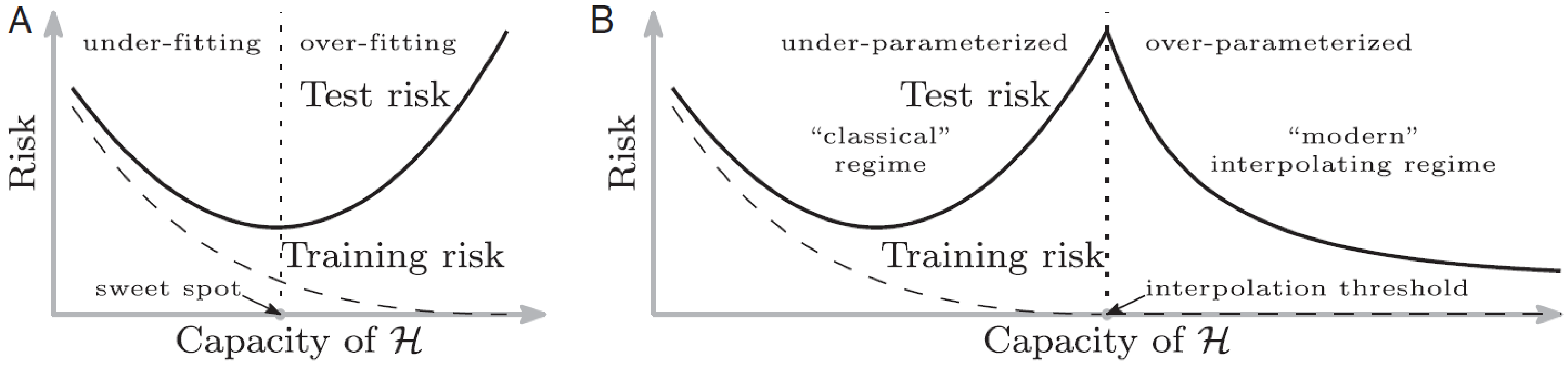
Classical theory: Structural Risk Minimization



Trouble with Boosting, again

- Modern models can be hugely overparameterized.
  They have more tunable parameters than needed to fit the data!
- They don’t overfit (as much as we could expect):
  - Boosted ensembles tend not to overfit
  - Modern deep learning uses massive models (billions of tunable parameters!!!)

Modern theory: Double Descent Hypothesis



Near the threshold

Very few models fit the training data, (think about a polynomial of degree d+1 interpolating d points).

Regularization can't help, as there are no models to choose from ☹

Extremely overparameterized regime

There are many models that fit the training data!

Regularization can help to choose e.g. a wide-margin model!

Profit!!!!!

Belkin et al.: Reconciling modern machine-learning practice and the classical bias–variance trade-off, PNAS 2019
OpenAI: https://openai.com/blog/deep-double-descent/