9. Bias-Variance Tradeoff

Sunday, January 23, 2022 5:08 PM

Error Decomposition aka Bias-Variance Tradeoff §5.2 [SS]

Let $h_S \in ERM_{\mathcal{H}}(S)$

Total Error (true risk) is

$$L_D(h_S) = \underbrace{\overset{\varepsilon_{est.}}{L_D(h_S) - \min_{h \in \mathcal{H}} L_D(h)}}_{\text{estimation error}} + \underbrace{\overset{\varepsilon_{approx.}}{\min_{h\in\mathcal{H}} L_D(h)}}_{\text{approximation error}}$$

or $$\underbrace{L_D(h_S) - \underbrace{\min_{\text{all } h} L_D(h)}_{\text{Bayes risk, } \varepsilon_{Bayes}}}_{\text{"excess risk"}} = \varepsilon_{est} + \tilde{\varepsilon}_{approx}$$

$\tilde{\varepsilon}_{approx} = \varepsilon_{approx} - \varepsilon_{Bayes}$

ε_{est}: error due to using $\hat{L}_S$ instead of L_D. This is what our agnostic PAC bounds cover.

aka generalization. Results like $m = O\left(\frac{\log(|\mathcal{H}|/\delta)}{\varepsilon^2}\right)$ ...

Smaller $|\mathcal{H}|$ is "good": better sample complexity, generalization, lower variance

in fact, soon we'll see metrics to deal with $|\mathcal{H}| = \infty$

The intuition is low complexity $\mathcal{H}$ is good

ε_{approx}: we've ignored so far. This is more classical, or unknowable a priori.

Ex papayas

we're looking to approximate h_{true} by $h \in \mathcal{H}$

Ex $\mathcal{H} = \{$ all polynomials of degree $\leq 100 \}$

h_{true} is an arbitrary function (maybe even continuous)

ε_{approx} unlikely 0

ε_{Bayes} accounts for inherent noise in labels

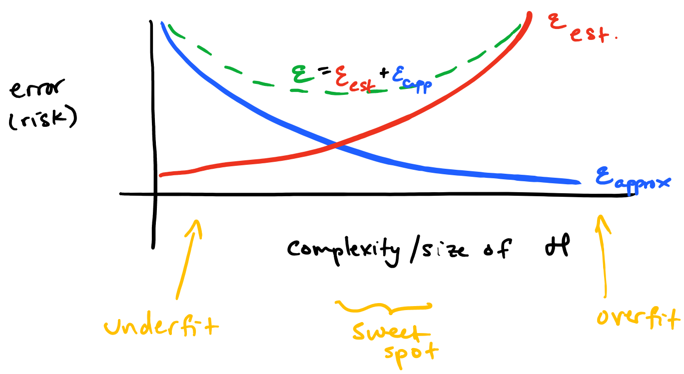
eg, (pure) PAC: $y = f(x)$ so $\min_h L_D(h) = L_D(f) = 0$

eg, agnostic PAC: $y \sim D(y|x)$

eg $y = f(x) + z$, $z \sim N(0, \sigma^2)$

not our focus

(textbook) bias-variance tradeoff

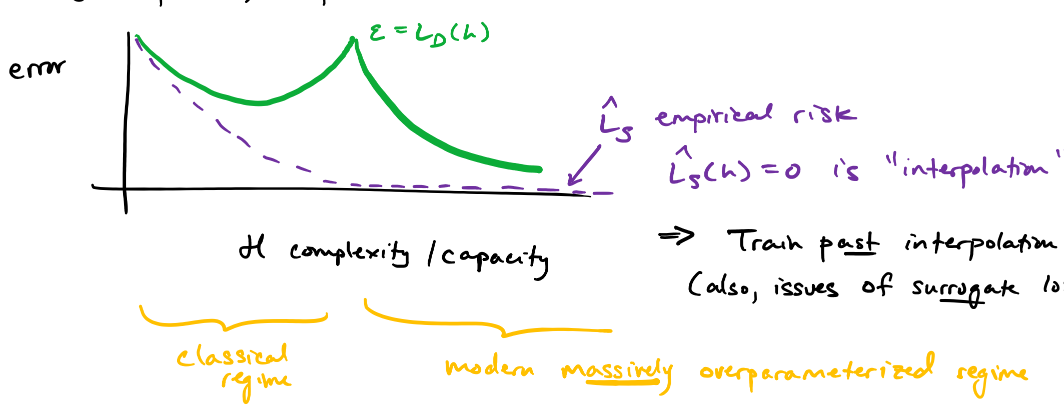


Point: both terms matter, so $|\mathcal{H}|$ small not always good

ε_{approx} hard to predict a priori

Double-descent

Belkin, Hsu, Ma, Mandal PNAS '19



$\Rightarrow$ Train past interpolation!

(also, issues of surrogate loss)

Classical: "A model with 0 training error is overfit to the training data and will typically generalize poorly"

(p 221, Hastie, Tibsirani, Friedman "The Elements of Statistical Learning" 2001)

That can be true, but need not be

- Deep Learning (2013+) empirical results strongly show best results obtained on massively overparameterized ($\hat{L}_S(h) = 0$)

neural networks

(caveat: how you train matters, eg., $\arg\min_h \hat{L}_S(h)$ is a big set, and not all ERM sol'n generalize... but the ones we find via SGD work well)

- Double descent can show up
  Though I'd disagree that it always does

We don't fully understand deep learning yet

many ideas in lit.

That paper suggests

- our notions of complexity not a good measure
- smoothness of h might be better
  (already sometimes exploited)
- $\mathcal{H}$ larger might allow for smoother functions
  (exactly at interpolation threshold is likely not smooth)
- or, w/ many sol'n to choose from, a least-squares (or other crit.) works well