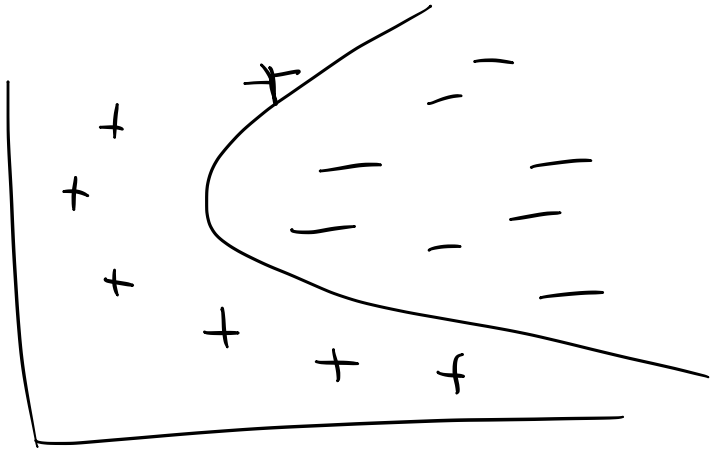
Statistical Learning Theory

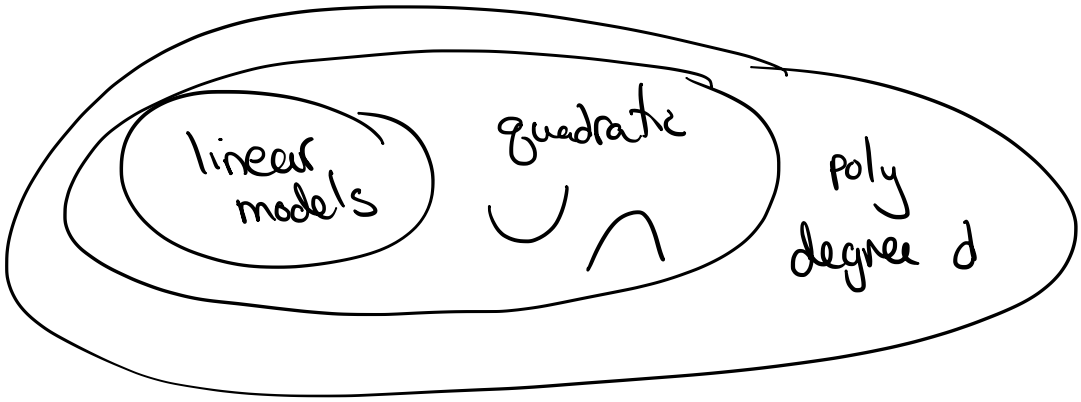
Generalization = Data + Knowledge

- cannot generalize from data alone. Need wise choice for function class.



- need "simplest" function class that describes the data well.

Occham's Razor - choose the simplest possible explanation to explain the phenomenon

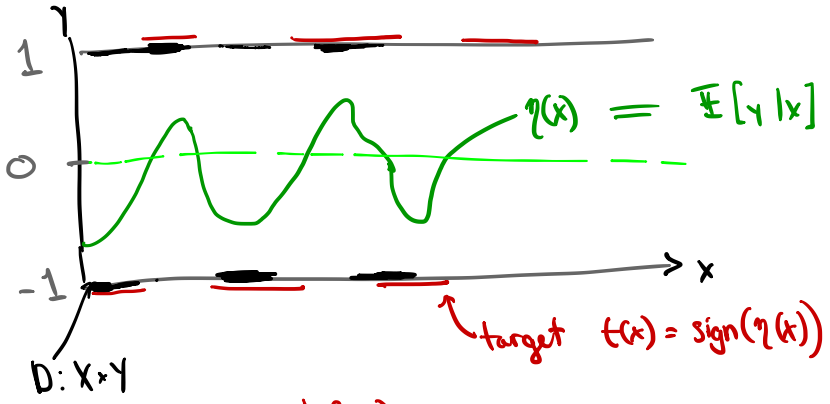


- can you measure the "complexity" of a function class by the number of parameters?

  answer: sometimes yes, sometimes no!

- how do we measure the "complexity" of a function class?

  answer: VC dim, Rademacher complexity, covering numbers



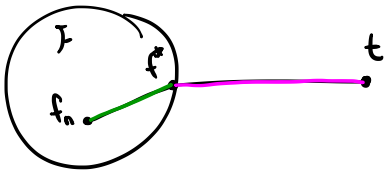
$$t \in \operatorname*{argmin}_{f \text{ (all functions)}} R^{true}(f), \quad R^{true}(f) = \mathbb{E}_{(x,y)\sim D} \mathbb{1}[f(x) \neq Y]$$

$$R^* = R^{true}(t)$$

$$f_n \in \operatorname*{argmin}_{f \in \mathcal{F}} R^{emp}(f) + C\|f\|^2$$

best-in-class is f^*

$$R^{true}(f^*) := \inf_{f \in \mathcal{F}} R^{true}(f)$$



How bad is f_n?

$$R^{true}(f_n) - R^* = \underbrace{\left[R^{true}(f^*) - R^*\right]}_{\text{approximation error}} + \underbrace{\left[R^{true}(f_n) - R^{true}(f^*)\right]}_{\text{estimation error}}$$

approximation error: need to know something about D

estimation error: want this to be as small as possible

Can't measure $R^{true}(f_n)$ what to do?

$$R^{true}(f_n) = \underbrace{R^{emp}(f_n)}_{ok} + \underbrace{\left[R^{true}(f_n) - R^{emp}(f_n)\right]}_{\leq \text{stuff}(n, \mathcal{F})}$$

Hoeffding's Inequality (one-sided)

$Z_1, \ldots, Z_n$ iid rv

$h(z) \in [a, b]$ bounded function

$\forall \varepsilon > 0$,

$$P_{Z\sim D^n}\left[\mathbb{E}_{z\sim D^n}[h(z)] - \frac{1}{n}\sum_{i=1}^{n} h(z_i) \geq \varepsilon\right] \leq \exp\left(\frac{-2n\varepsilon^2}{(b-a)^2}\right)$$

If I set $h(z) = \mathbb{1}_{[f(x)\neq y]}$ then I get

$$P_{Z\sim D^n}\left[R^{true}(f) - R^{emp}(f) \geq \varepsilon\right] \leq \exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right)$$

$b-a$ is 1-0

$$P_{Z\sim D^n}\left[R^{true}(f) \geq R^{emp}(f) + \varepsilon\right] \leq \underbrace{\exp\left(\frac{-2n\varepsilon^2}{(b-a)^2}\right)}_{\delta}$$

reparameterize

$$\delta = \exp\left(\frac{-2n\varepsilon^2}{(b-a)^2}\right)$$

$$\ln\delta = -\frac{2n\varepsilon^2}{(b-a)^2}$$

$$\varepsilon^2 = -\frac{\ln\delta}{2n}(b-a)^2 = \frac{\ln(1/\delta)}{2n}(b-a)^2 \Rightarrow \varepsilon = \sqrt{\frac{\ln(1/\delta)}{2n}}(b-a)$$

$$P_{Z\sim D^n}\left[R^{true}(f) > R^{emp}(f) + \sqrt{\frac{\ln(1/\delta)}{2n}}(b-a)\right] \leq \delta$$

"inversion"

with prob $\geq 1-\delta$,

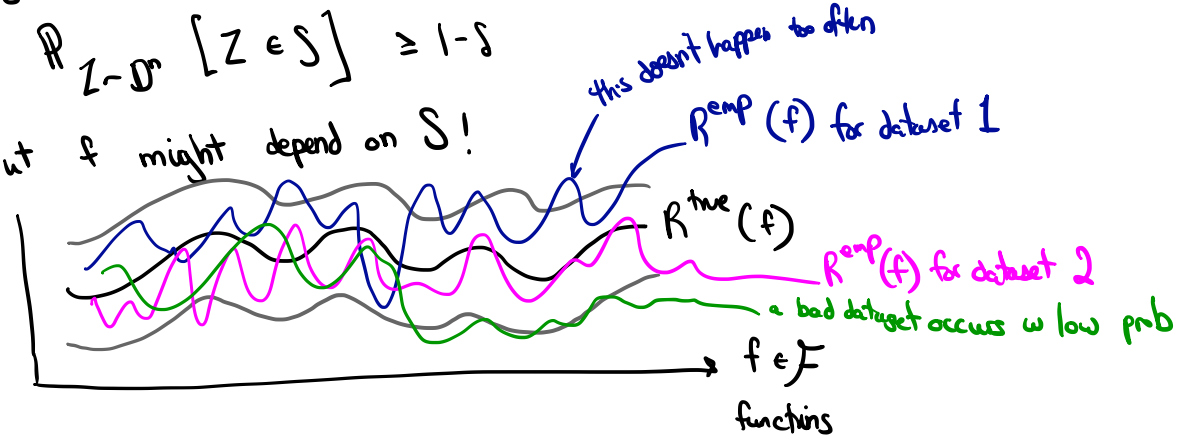
$$R^{true}(f) \leq R^{emp}(f) + \sqrt{\frac{\ln(1/\delta)}{2n}}(b-a)$$

← Bound for a single $f \in \mathcal{F}$

This bound says that for each f, there is a set of "good" datasets (S) where $R^{true}(f)$ is not too large and

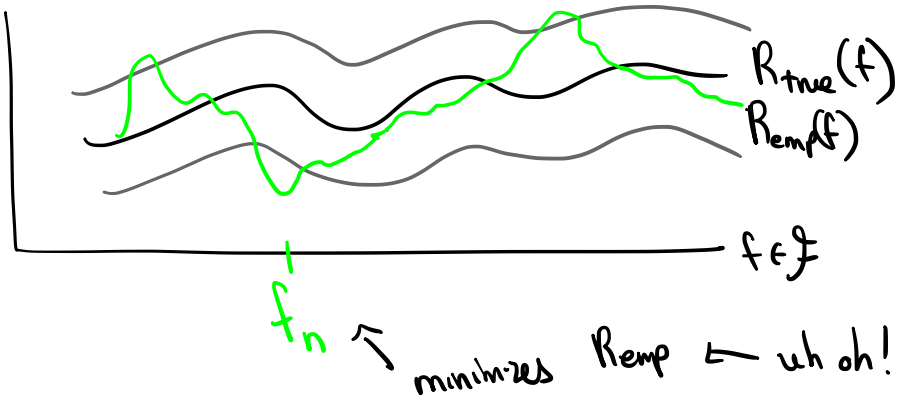
$$P_{Z\sim D^n}[Z \in S] \geq 1-\delta$$

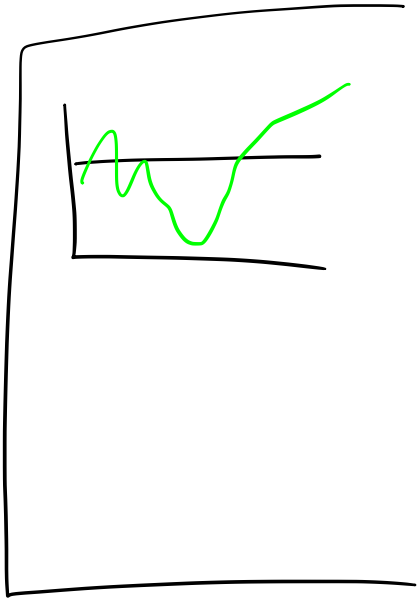
But f might depend on S!



Hoeffding's says that $R^{true}(f)$ & $R^{emp}(f)$ are close w.h.p. for any fixed f.

However, f is not fixed. The algorithm chooses f knowing the data. It tries to minimize R^{emp}.





- If $\mathcal{F}$ is really large, this problem gets worse. Oy!
- Want a bound that holds uniformly over all $f \in \mathcal{F}$

$$R^{true}(f_n) - R^{emp}(f_n) \leq \sup_{f\in\mathcal{F}}\left(R^{true}(f) - R^{emp}(f)\right)$$

Occham's Razor Bound (Hoeffding's + Union Bound)

$$P[C_1^{bad} \cup C_2^{bad}] \leq P(C_1^{bad}) + P(C_2^{bad})$$

↑ prob we hit a bad dataset for either f_1 or f_2

From Hoeffding's ineq $P(C_1^{bad}) \leq \delta$, $P(C_2^{bad}) \leq \delta$ so $P[C_1^{bad} \cup C_2^{bad}] \leq 2\delta$

$P[C_1^{bad} \cup \dots \cup C_M^{bad}] \leq M\delta$

Say $\mathcal{F}$ is finite $f_1 \dots f_m$.

$$P_{z\sim D^n}\left[\exists f \in \mathcal{F} : R^{true}(f) - R^{emp}(f) \geq \varepsilon\right] \leq$$

(↑ union)

$$\sum_{j=1}^{M} P_{z\sim D^n}\left[R^{true}(f_j) - R^{emp}(f_j) \geq \varepsilon\right] \leq \sum_{j=1}^{M} e^{-2n\varepsilon^2} = Me^{-2n\varepsilon^2}$$

reparameterize $\delta = Me^{-2n\varepsilon^2} \Rightarrow \varepsilon^2 = -\frac{1}{2n}\log\frac{\delta}{M}$

$$\varepsilon = \sqrt{\frac{\log(M) + \log(1/\delta)}{2n}}$$

Plug: $$P_{z\sim D^n}\left[\exists f \in \mathcal{F} : R^{true}(f) - R^{emp}(f) \geq \sqrt{\frac{\log(M) + \log(1/\delta)}{2n}}\right] \leq \delta$$

inverting: For all $\delta > 0$ w.p. at least $1-\delta$,

for all $f \in \mathcal{F} = \{f_1, f_2 \dots f_m\}$

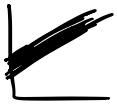
$$R^{true}(f) \leq R^{emp}(f) + \sqrt{\frac{\log(M) + \log(1/\delta)}{2n}}$$

Occham's Razor Bound

Summary: Our bound can be applied to f_n even though it depends on the data.

Extra log M term → says we want M bounds to hold simultaneously.

What if $\mathcal{F}$ is infinite? Bound is vacuous.



Recap:

- Generalization = data + knowledge
  restricting f to lie in restricted $\mathcal{F}$
  - for a fixed f, for most datasets
    $$R^{true}(f) - R^{emp}(f) \approx \frac{1}{\sqrt{n}}$$
  - for most datasets if $\mathcal{F}$ is finite, $|\mathcal{F}| = M$
    $$\sup_{f \in \mathcal{F}} \left[R^{true}(f) - R^{emp}(f) \right] \approx \sqrt{\frac{\log M}{n}}$$
  - Occham's Razor