To avoid overfitting, we previously fixed a predictor class $\mathcal{F}$, and considered ERM <u>within</u> $\mathcal{F}$. $\longrightarrow$ inductive bias

If the learning algorithm returns $\bar{f} \in \mathcal{F}$, then

$$R(\bar{f}) - R^* = \underbrace{R(\bar{f}) - \inf_{f \in \mathcal{F}} R(f)}_{\text{excess risk}} + \underbrace{\inf_{f \in \mathcal{F}} R(f) - R^*}_{\text{approximation error}}$$

where $R^* = \inf_{f : \mathcal{X} \to \mathcal{Y}} R(f)$ is the Bayes risk.

In the last two chapters, we analyzed the <u>excess risk</u>.

Now, we analyze the <u>approximation error</u>.

—— // ——

<u>Classical setup</u>

- compete with continuous predictors.
- all models of some fixed neural net. architecture.

$$\sup_{f^* \text{ cont.}} \inf_{f \in \mathcal{F}} R(f) - R(f^*).$$

<u>PROP</u> (Good function approximator $\Rightarrow$ good test error)

(i) $\mathcal{Y} = \mathbb{R}$, $\mathcal{X} \subset \mathbb{R}^d$. (constrained regression)

Let $z \mapsto \ell(z, y)$ be L-Lipschitz for all $y \in \mathcal{Y}$. Then,

$$R(f) - R(h) = \mathbb{E}\left[\ell(f(x), y) - \ell(h(x), y)\right]$$
$$\leq L \cdot \int_{\mathcal{X}} |f(x) - h(x)| \, dx \leq L \cdot \sup_{x \in \mathcal{X}} |f(x) - h(x)|.$$

(ii) $\mathcal{Y} = \{-1, 1\}$, $\mathcal{X} \subset \mathbb{R}^d$.

Let $\ell(f(x), y) = \ell_0(f(x) \cdot y)$ for L-Lipschitz ℓ_0. Then,

$$R(f) - R(h) \leq L \cdot \int_{\mathcal{X}} |f(x) - h(x)| \, dx \leq L \cdot \|f - h\|_\infty .$$

A very simple start: grid-based approximation.

THEOREM 1 | (Grid-based approximation: univariate)

$h : \mathbb{R} \to \mathbb{R}$ is L-Lipschitz. For any $\varepsilon > 0$, there exists a 1-hidden-layer neural network with $m = \left\lceil \frac{L}{\varepsilon} \right\rceil$ neurons and $\sigma(z) = \mathbb{1}\{z \geq 0\}$ activation s.t.

$$\sup_{x \in [0,1]} |f(x) - h(x)| \leq \varepsilon.$$

Proof: Let $c_0 \triangleq h(0)$,

$$c_i \triangleq h\left(\frac{i \cdot \varepsilon}{L}\right) - h\left(\frac{i-1}{L} \cdot \varepsilon\right), \quad i = 1, \ldots, m$$

Then, let

$$f(x) = \sum_{i=0}^{m-1} c_i \, \sigma\left(x - \frac{i \cdot \varepsilon}{L}\right) = \sum_{i < m} c_i \, \mathbb{1}\left\{x \geq \frac{i \cdot \varepsilon}{L}\right\}.$$

For any $x \in [0,1]$, let $k = \max\left\{ i \in \{1, \ldots, m\} : x \geq \frac{i \cdot \varepsilon}{L} \right\}.$

Then,

$$|f(x) - h(x)| = \left| \underbrace{f(x) - f\left(\frac{k\varepsilon}{L}\right)}_{=0} + f\left(\frac{k\varepsilon}{L}\right) - h\left(\frac{k\varepsilon}{L}\right) + h\left(\frac{k\varepsilon}{L}\right) - h(x) \right|$$

$$\leq \left| f\left(\frac{k\varepsilon}{L}\right) - h\left(\frac{k\varepsilon}{L}\right) \right| + \left| h\left(\frac{k\varepsilon}{L}\right) - h(x) \right|$$

$$\leq \left| \sum_{i=0}^{k} c_i - h\left(\frac{k\varepsilon}{L}\right) \right| + L \cdot \overbrace{\left| x - \frac{k\varepsilon}{L} \right|}^{\leq \left(\frac{k+1}{\varepsilon} - \frac{k}{\varepsilon}\right).}$$

$$\leq \left| h(0) + \sum_{i=0}^{k} \left(h\left(\frac{i\varepsilon}{L}\right) - h\left(\frac{(i-1) \cdot \varepsilon}{L}\right) \right) - h\left(\frac{k\varepsilon}{L}\right) \right| + \varepsilon$$

$$= \varepsilon. \quad \blacksquare$$

Grids worked well in $\mathbb{R}$. How about $\mathbb{R}^d$?

THEOREM 2 | (Grid-based approximation : multivariate)

$h : [0,1]^d \to \mathbb{R}$ continuous, $\varepsilon > 0$ given.

Choose $\delta > 0$ s.t. $\|x - x'\|_\infty < \delta \Rightarrow |h(x) - h(x')| \leq \varepsilon$.

Then, there exist a ReLU neural network f with two hidden layers

s.t. $m = \Omega(1/\delta^d)$ and $\displaystyle\int_{[0,1]^d} |f(x) - h(x)|\, dx \leq 2\varepsilon$.

Pf: Constructive. See (Telgarsky, 2021 ; Theorem 2.1).

Remarks : Note the width $m = \Omega\left(\frac{1}{\delta^d}\right)$, which grows exponentially with d in the exponent.

This terrible dependence in d is called curse of dimensionality in approximation.