Lecture 2: shallow apx

Ann.
* project draft up
* lecture 2-4 typed by 11:59pm Monday

Today
* Shallow constructive apx
  * univariate apx
  * multivariate apx
  * universal apx
* infinite width
* initialization & overparameterization
* apx - opt - gen - other topics

$\boxed{1}$ $\mathcal{R}$ = abstract measure of future performance
$\mathcal{F}$ = deep network architecture
$\hat{f}$ = output of alg
$\bar{f}$ = good choice for $\mathcal{R}$ in $\mathcal{F}$

$\mathcal{R}(\hat{f}) \approx \mathcal{R}(\bar{f})$ (opt/gen)
now must $\mathcal{R}(\bar{f})$ small (apx)

Want to formalize "is $\inf_{f\in\mathcal{F}} \mathcal{R}(f)$ small"

(A) If: (i) know what future data looks like
(ii) believe in a perf criterion per datum, $\ell(f(x), y)$

$\Rightarrow$ ensure $\inf_{f\in\mathcal{F}} \mathbb{E}_{\text{future } x,y}\, \ell(f(x), y)$ small.

(B) If $\mathcal{R}$ satisfies "some regularity", we can say

$$\inf_{f\in\mathcal{F}} \mathcal{R}(f) \approx \inf_{\substack{g \\ \text{"every function"}}} \mathcal{R}(g)$$

if $\boxed{\forall g\ \exists f,\ \forall x\ |f(x) - g(x)| \le \varepsilon.}$

Proposition. $\forall$ ρ-Lip $g : \mathbb{R} \to \mathbb{R}$, $\varepsilon > 0$,

$\exists$ f 2-layer network with biases

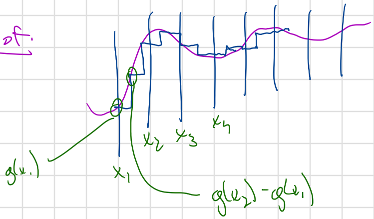
activation $\boxed{\sigma(z) = \mathbb{1}[z \geq 0]}$ $m = \lceil \frac{\rho}{\varepsilon} \rceil$

$\forall x \in [0,1]$, $|f(x) - g(x)| \leq \varepsilon$.

Remarks.

* pays for large flat regions, limitation of the proof.
* polynomials pay for the flat part

Proof.



$$b_j := \frac{(j-1)\varepsilon}{\rho}, \quad a_1 := g(b_1)$$
$$a_{j+1} := g(b_{j+1}) - g(b_j)$$

$$f(x) = \sum_j a_j \sigma(x - b_j)$$

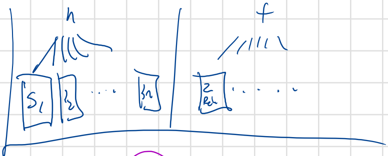
Let x given, $b_k \leq x$ be largest

$$|f(x) - g(x)| \leq |f(x) - f(b_k)| + |f(b_k) - g(b_k)| + |g(b_k) - g(x)|$$
$$\leq 0 + 0 + \rho\left(\frac{\varepsilon}{\rho}\right)$$

$$f(b_k) = \sum_{j=1}^{m} a_j \sigma(b_k - b_j) = \sum_{j=1}^{k} a_j$$
$$= g(b_1) + \sum_{j=2}^{k} \left(g(b_j) - g(b_{j-1})\right) = g(b_k).$$

Theorem. $\forall$ c-Lipschitz $g:\mathbb{R}^d \to \mathbb{R}$ & $\varepsilon > 0$ $\exists$ f 3-layer biased network with ($\sigma(z) = \mathrm{ReLU}(z)$, $m = \mathcal{O}(\lceil \frac{c}{\varepsilon} \rceil^d$)

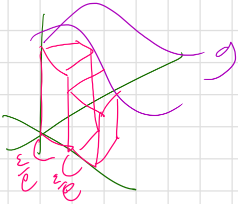
$$\int_{[0,1]^d} |f(x) - g(x)|\, dx \leq \varepsilon .$$



Proof.

Claim: done if p_j, a 2-layer ReLU net, with

$$\int_{[0,1]^d} |p_j(x) - \mathbb{1}[x \in S_j]| \leq \frac{\varepsilon}{\sum |g(a_j)|}.$$



Proof. Using $f(x) = \sum g(a_j)\, p_j(x)$,

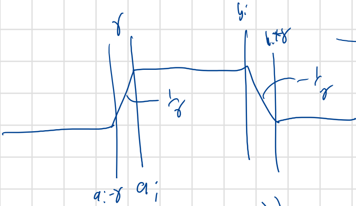
$$\int |h(x) - f(x)|\, dx \leq \sum_{j=1}^{m} |g(a_j)| \underbrace{\int |p_j(x) - \mathbb{1}[x \in S_j]|}$$

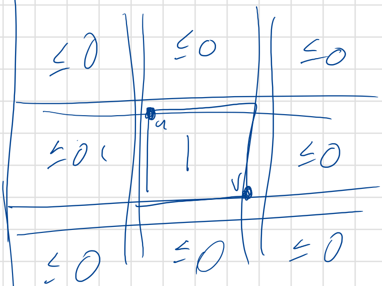
$\leq \varepsilon$. //

$(S_1 \cdots S_m)$ partition of $[0,1)^d$ into cubes, $S_j := \times_{i=1}^{d} [a_{j,i}, b_{j,i})$

$$h(x) = \sum_j g(a_j)\, \mathbb{1}[x \in S_j]$$

$$= \sum_j g(a_j) \underbrace{\prod_{i=1}^{d} \mathbb{1}[x_i \in [a_{j,i}, b_{j,i})]}$$





$$g_i(r) = \sigma\left(\frac{r - (a_i - \gamma)}{\gamma}\right) - \sigma\left(\frac{r - (a_i)}{\gamma}\right) - \sigma\left(\frac{r - b_i}{\gamma}\right) + \sigma\left(\frac{r - (b_i + \gamma)}{\gamma}\right).$$

$$x \mapsto \sigma\left(\left[\sum g_i(x_i)\right] - (d-1)\right)$$

Standard universal approximation.

Theorem. [Hornik–Stinchcombe–White '89, Leshno '93]

Suppose $\sigma : \mathbb{R} \to \mathbb{R}$ is continuous, but not a polynomial,

$\forall$ $g : \mathbb{R}^d \to \mathbb{R}$ continuous, $\forall \varepsilon$
$\exists$ $f : \mathbb{R}^d \to \mathbb{R}$ 2-layer biased σ network
$\forall x \in [0,1]^d$, $|f(x) - g(x)| \le \varepsilon$. [supremum norm / uniform norm]

Remarks.

* Not unique to neural networks (e.g., SVM + RBF).
* implicit exponentially large net.
* proofs.
* not polynomial is necessary

Proof plan.

* We'll invoke lemma for polynomial-like functions.

* Consider $\sigma(r) = \exp(r)$,

  where $\exp(r+s) = \exp(r)\exp(s)$.