Lecture 2: basic norm-based apx.

* Apx goal; norm proxy; apx plan.
* univariate apx: L_∞, 1 layer, threshold
* $\mathbb{R}^d$ apx: L_1, 2 layers, ReLU
* "universal" apx: $\mathbb{R}^d$, L_∞, 1 layer, "any" activation

---

$R(f) = \mathbb{E}\, \ell(f(x), y)$; $\hat{R}(f) = \frac{1}{n}\sum_i \ell(f(x_i), y_i)$

$\hat{f} \in \mathcal{F}$ alg $\quad \bar{f} \in \mathcal{F}$ best in class

$R(\bar{f}) \approx \inf_{g \in \mathcal{F}} R(g)$

opt + gen

$R(\hat{f}) \approx R(\bar{f}) \approx \inf_{g \in \mathcal{F}} R(g)$ small

To argue $\inf_{g \in \mathcal{F}} R(g)$ is small,

lazy/standard/classical approach:

take a big class, e.g., $\mathcal{C}$ all cont. functions.

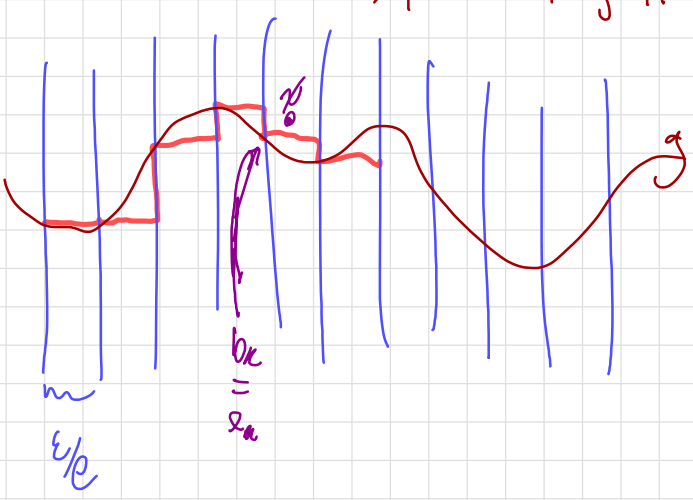
$\forall h \in \mathcal{C}$, $\exists g \in \mathcal{F}$ s.t. $\|g - h\|$ small (for some norm $\|\cdot\|$).

Proposition. Given c-Lipschitz $g: \mathbb{R} \to \mathbb{R}$, $\varepsilon > 0$

$\exists f(x) = \sum_{j=1}^{m} a_j \mathbb{1}[x \geq b_j]$, $m = \lceil \frac{c}{\varepsilon} \rceil$

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

Proof.



$b_0 := 0$, $a_0 := g(0)$; $b_j := \frac{j\varepsilon}{c}$, $a_j := g(b_j) - g(b_{j-1})$

Let $x \in [0,1)$ be given, pick largest $x_k := b_k \leq x$

$|g(x) - f(x)| \leq |g(x) - g(x_k)| + |g(x_k) - f(x_k)| + |f(x_k) - f(x)|$

$\leq c\left(\frac{\varepsilon}{c}\right) + |g(b_k) + (g(x_k) - g(x_{k-1}) + g(x_{k-1}) - \cdots$

$= \varepsilon$.

Remark.

* This is a non-adaptive; e.g., pays for flat regions.
  It seems DNNs are good because they effectively model natural phenomena.
* Polynomials pay for flat regions
* Better version in hwk 1; but over $\mathbb{R}^d$ open.

Theorem. $\forall$ cont $g: \mathbb{R}^d$ $\varepsilon > 0$ $\in [0,1]^d$
given $\delta > 0$ s.t. $\|x - x'\|_\infty \le \delta \Rightarrow |g(x) - g(x')| \le \varepsilon$

$\exists f: \mathbb{R}^d \to \mathbb{R}$ 2-hidden layer ReLU network with $\ge \left(\frac{1}{\delta}\right)^d$ nodes

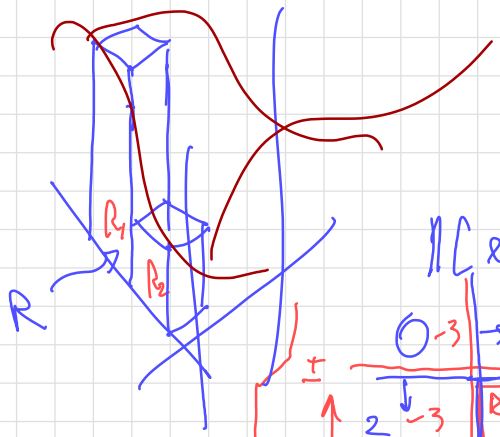
s.t. $\int_{[0,1]^d} |f(x) - g(x)| \, dx \le \varepsilon$.

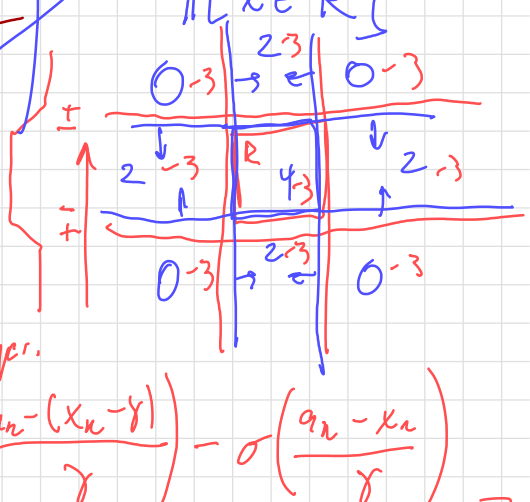
Remark.

* L_1 is bad; next theorem uses L_∞
* $1/\delta^d$ is a disaster; e.g., CIFAR dim $= 3072$.

Also matching lower bound.

Proof.





$R = \bigtimes_{k=1}^d [a_k, b_k)$

Use a second layer.

$$g_\gamma(x) = \sigma\left(\sum_{k=1}^d \left[\sigma\left(\frac{a_k - (x_k - \gamma)}{\gamma}\right) - \sigma\left(\frac{a_k - x_k}{\gamma}\right) - \sigma\left(\frac{b_k - x_k}{\gamma}\right) + \sigma\left(\frac{b_k - (x_k + \gamma)}{\gamma}\right)\right] - (d-1)\right)$$

Note $g_\gamma \to \mathbb{1}[\cdot \in R]$ pointwise as $\gamma \to 0$

Analysis fact: $\exists \, h(x) := \sum_{j=1}^m a_j \, \mathbb{1}[x \in R_j]$

s.t. $\sup_{x \in [0,1]^d} |h(x) - g(x)| \le \varepsilon/2$.

Define $f(x) := \sum_{j=1}^m a_j \, g_{\gamma,j}(x)$.

Then $\lim_{\gamma \downarrow 0} \int |f - g| \le \frac{\varepsilon}{2} + \lim_{\gamma \downarrow 0} \int |f - h|$

$\le \frac{\varepsilon}{2} + \sum_{j=1}^m |a_j| \lim_{\gamma \downarrow 0} \int_{[0,1]^d} |\mathbb{1}[\cdot \in R_j] - g_{\gamma,j}|$

$\to 0$.

* "universal approximation"
* Barron class / Fourier.
* NTK / "near initialization" ← related
* "Benefits of depth" (non-algorithmic)

Definition. $\mathcal{F}$ is a "universal approximator" if $\forall$ cont $g: [0,1]^d \to \mathbb{R}$, $\forall \varepsilon > 0$ $\exists f \in \mathcal{F}$, $\sup_{x \in [0,1]^d} |f(x) - g(x)| \le \varepsilon$.

Theorem. Let "sigmoidal" σ be given:
σ continuous, $\lim_{z \to -\infty} \sigma(z) = 0$, $\lim_{z \to \infty} \sigma(z) = 1$

Let $\mathcal{F}$ denote 1-hidden layer networks of unbounded width $\left(\sum_{j=1}^{M} a_j \sigma(w_j^T x + b_j) \ \forall M > 0, a_j, w_j, b_j \right)$
Then $\mathcal{F}$ is a universal approximator.

Remark:
* Found concurrently in ~1988
  Cybenko, Hornik-Stinchcombe-White, Funahashi, Barron. (Proofs differ)
* Still used!

Proof idea: (Hornik-Stinchcombe-White)

$$\mathbb{1}\left[x \in \sum_{k=1}^{d} [a_k, b_k)\right] = \prod_{k=1}^{d} \mathbb{1}\left[x_k \in [a_k, b_k)\right]$$

(left: apx; right: apx)

Weierstrass: polynomials of unbounded degree are universal approximators

Stone-Weierstrass: polynomial-like $\mathcal{F}$ are uni. apx

Why we use shallow networks.

Note $e^{w^T x + b} e^{u^T x + c} = e^{(w+u)^T x + (b+c)}$

Exp(·) activations are closed under products;
Arbitrary activations okay.