AI616 Problem Set 1.

20200130 Yujun Kim

#1.1

$g(x)=x^2$ with $\varepsilon=1>0$ works.

Any 2-layer ReLU network f is of the form

$$f(x)=v^T\sigma(wx+b)+c, \text{ where } v,w,b\in\mathbb{R}^m,\ c\in\mathbb{R},\ m\in\mathbb{N}$$

It is a piecewise linear fn with finite pieces. Let $a_1<\cdots<a_k$ be real nbrs so that f is linear on $(-\infty,a_1],[a_1,a_2],\cdots,[a_{k-1},a_k],[a_k,\infty)$.

Let $f(x)=px+q$ on $[a_k,\infty)$.

$$\lim_{x\to\infty} g(x)-f(x)=\lim_{x\to\infty} x^2-px+q=\infty.$$

Thus, $\exists x_0>a_k$ s.t. $g(x_0)-f(x_0)\geq\varepsilon$. $\Rightarrow \sup_{x\in\mathbb{R}}|f(x)-g(x)|\geq g(x_0)-f(x_0)\geq\varepsilon$.

#1.2

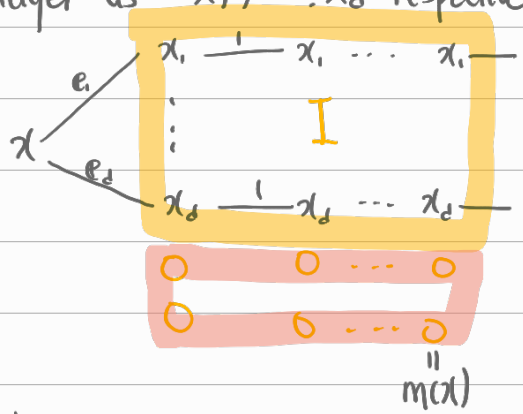
Let $\mathcal{P}$ be the collection of all polynomials in cpt domain $[0,1]^d$.

By Stone-Weierstrass, $\mathcal{P}$ is ρ_∞-dense in $C([0,1]^d,\mathbb{R})$.

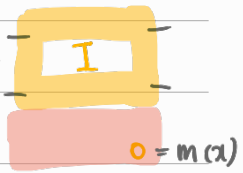
Hence, it suffices to show $\mathcal{P}\subseteq\mathcal{F}$.

① We show any monic monomial $m(x)$ is represented by width $d+2$ network.

Let $x=(x_1,\cdots,x_d)\in[0,1]^d$. We keep values in first d nodes of each layer as $x_1,\cdots,x_d$ respectively, denoted by I in figure.



Call this width $d+2$ network



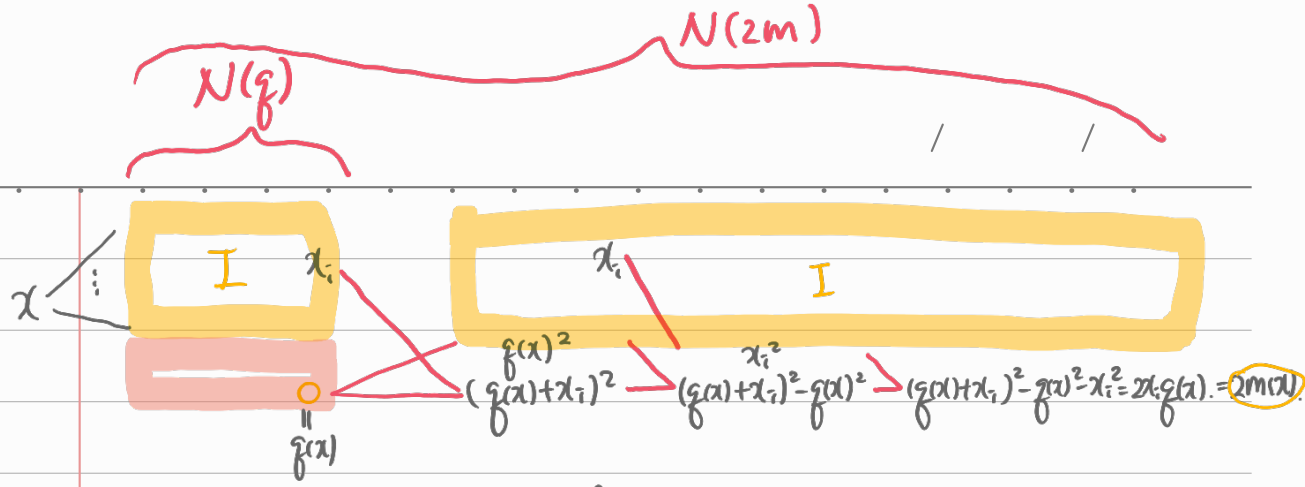
as $N(m)$, if it exists

This can be done by induction on $\deg m(x)$ to show $\exists N(m)$.

i) $\deg m(x)=0$ or 1 $\Rightarrow$ trivial.

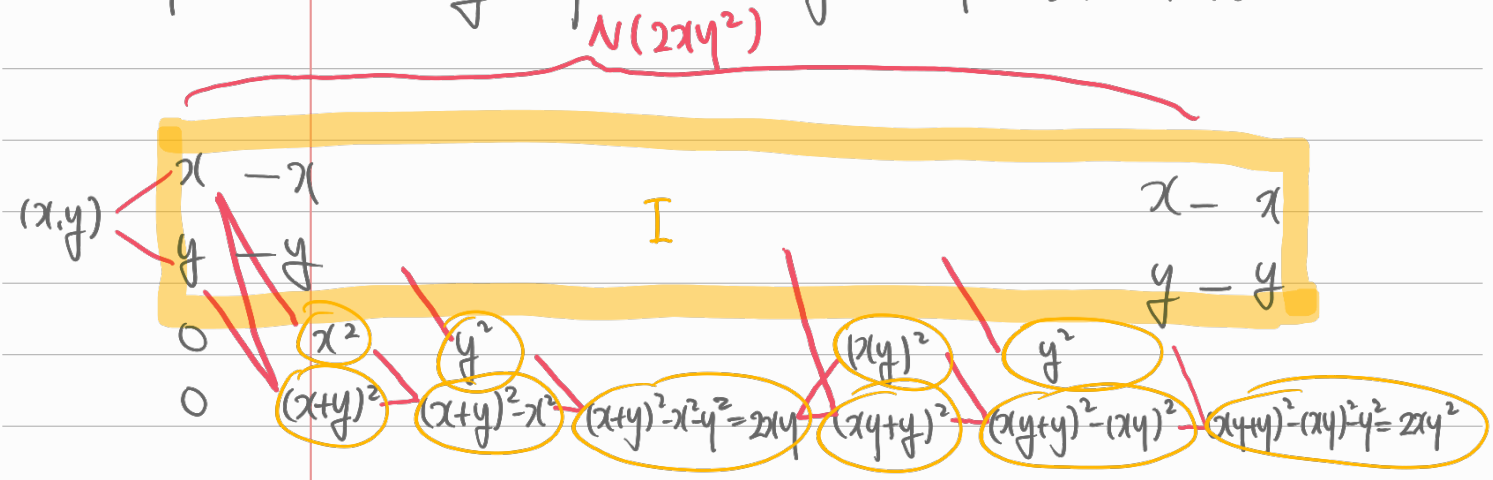
ii) If $\deg m(x)\geq 2$, let $m(x)=x_i q(x)$ for some i.

By induction hypothesis, $\exists N(q)$



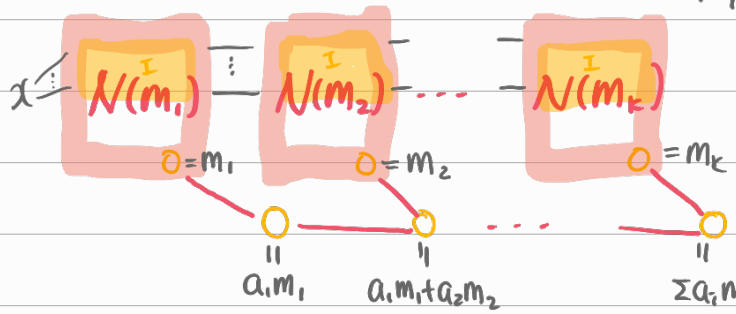
Thus, we created $N(2m)$ from $N(q)$, which infers that we can create $N(m)$ from $N(q/2)$.

Example Making $2xy^2$ from (x,y) on depth $2+2=4$ network.



② Now we show any polynomial is represented by width $d+3$ network. Let p be a polynomial s.t. $p=\sum_{i=1}^{k} a_i m_i$ for $a_i \in \mathbb{R}$, m_i: monic monomials.

Then,



gives width $d+3$ network that represent p. Here, $(d+3)$-th width is reserved to save added values.

③ By ②, we have $\mathcal{P} \subseteq \mathcal{F}$ and so we are done.

/ /

#1.3

By lemma 3.4 from the lecture,
$f:\mathbb{R}\to\mathbb{R}$, a ReLU network with d_1, d_2 hidden nodes has
output of each node in layer 2 has at most $2^2 d_1$ affine pieces.
$\Rightarrow f$ has at most $2^2 d_1 d_2 = 4 d_1 d_2$ affine pieces.

If $\max\{d_1, d_2\} \le C\sqrt{N}$, then $4 d_1 d_2 \le 4C^2 N$.
Let $C = \frac{1}{3}$, $N_0 = 18$ Then, for $N \ge N_0$, we have $4C^2 N = \frac{4}{9}N \le \frac{N}{2} - 1$ ($N \ge 18$)
Thus, nbr of affine pieces of f is less or equal to $\frac{N}{2} - 1$.
Consider $S = \{(n, y) \mid 0 \le n \le N-1,\ y = \begin{cases} 0 & \text{if } n \text{ is even} \\ 1 & n \text{ is odd} \end{cases}\}$

Lemma

Then, any piecewise affine g that memorize S should have at least one nonlinear point in $(0.2), (2,4), \cdots, (2(\lfloor \frac{N-1}{2} \rfloor - 1), 2\lfloor \frac{N-1}{2} \rfloor) \subseteq \mathbb{R}$ respectively.
i.e. If $k = N_A(g)$ so that g is affine in
$$(-\infty, a_1], [a_1, a_2], \cdots, [a_{k-2}, a_{k-1}], [a_{k-1}, \infty)$$
respectively, then each $(0.2), \cdots, (2(\lfloor \frac{N-1}{2} \rfloor - 1), 2\lfloor \frac{N-1}{2} \rfloor) \subseteq \mathbb{R}$ should contain at least one a_i respectively.

(pf of lemma)

$\because$ Otherwise, g gets affine on $[2(t-1), 2t]$ for some $t = 0, \cdots, \lfloor \frac{N-1}{2} \rfloor$ so that g cannot memorize all three points $(2(t-1), 0), (2t-1, 1), (2t, 0) \in S$

Corollary

From lemma, any g that memorize S has $N_A(g) \ge \lfloor \frac{N-1}{2} \rfloor + 1$, as $\lfloor \frac{N-1}{2} \rfloor$ nbr of disjoint interval each contains a nonlinear point of g.

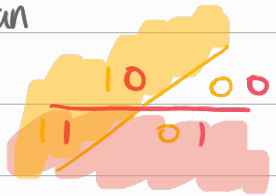
For any f in our fn class, $N_A(f) \le \frac{N}{2} - 1 < \frac{N-1}{2} \le \lfloor \frac{N-1}{2} \rfloor + 1$
By corollary, f cannot memorize S.

#1.4

Note that collection of outputs from nodes of first hidden layers fully characterize the output of the entire network. ···(*)

Also, each nodes in the first hidden layer output either 0 or 1 depending on whether an input is on certain side of a hyperplane. (**)

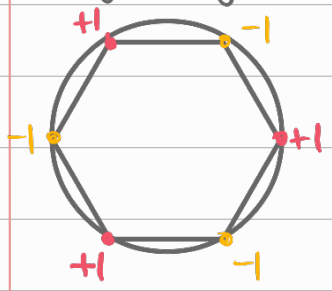
For example, ouput of two nodes u, v can vary as right figure depending on the input domain.



For even N, let $x_i = e^{\frac{2\pi i}{N}}$ for $i=1,\cdots,N$. Let $S=\{x_i\}_{i=1}^{N}$

Define $f: S \to \{\pm 1\}$ by $f(x_i) = \begin{cases} +1 & \text{if } i \text{ is even} \\ -1 & \text{if } i \text{ is odd.} \end{cases}$

6-gon (hexagon)



Let the network have k nodes in the first hidden layer.
Consider regular N-gon formed by points in S.

Let's say closed half plane $H = \{x \in \mathbb{R}^d : a^T x \geq b\}$ helps distinguishing $x, y \in \mathbb{R}^d$ if ($x \in H$ and $y \notin H$) or ($x \notin H$ and $y \in H$).

Refer(**) → Let $H_1, \cdots, H_k$ be hyperplanes given by nodes in the first hidden layer.

Refer (*) → For network to memorize f on S, each x_i, x_{i+1} should get help by some H_j to be distinguished. Observe that:

① There are N pair of points $(x_1,x_2),(x_2,x_3),\cdots,(x_{N-1},x_N),(x_N,x_1)$ that needs help to get distinguished by some H_j, $j=1,\cdots,k$

② Each hyperplane H_j can help distinguish at most two pair of points (x_i, x_{i+1}), $(x_{i'}, x_{i'+1})$ among aboves.

Hence, for N pairs of points to get at least one help each, we need

$2k \geq N \Rightarrow k \geq \frac{N}{2}$

Hence, we need at least $\frac{N}{2}$ nodes in the first hidden layer.