Part 1

Question 1

Is it possible to define a function $f: \{\pm 1\}^d \to \{\pm 1\}$ that the lower bound on B is larger than 2^{d-1}?
(In order to let a shallow network with width B realizes f)

Answer: no! it's not possible.

Lemma: We can realize each f with $\leq 2^{d-1}$ AND gates that represent $\{x \mid f(x)=1\}$

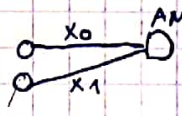
Proof

We will prove it by induction on d.

base: $d=2$

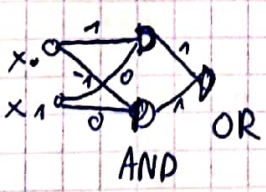
Let $f: \{\pm 1\}^2 \to \{\pm 1\}$. We will show that a network with width $2^{d-1}=2^{2-1}=2$ can realize f. Define $A := \{x \mid f(x)=1\}$

① If $|A|=j \leq 2$, then we can build network with width j, s.t each ~~AND~~ gate represents some $x \in A$

[if $X = X_0 X_1$ then 

will be the weights to this gate] and the weights after each AND gate will be 1. In that case, $N(x)=1$ if and only if one of the AND gates output 1 if and only if $x \in A$ by definition of the gates $\Rightarrow$ N realizes f with at most 2 width.

② if $|A|=4$, then $f \equiv 1$, the network N can be

N: 

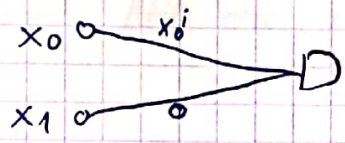
So, for every $x=X_0X_1 \in \{\pm 1\}^2$, $X_0=1$ or $X_0=-1$

So one of the AND gates outputs 1 as desired (then the OR gate outputs 1). Thus N realizes f with 2 width.

③ If $|A|=3$, then let's sign $A=\{y_1,y_2,y_3\}$,
$y_i = x_0^i x_1^i$.

By the pigenhole principle there must exist y_i, y_j
$i\neq j$ S.t $x_0^i = x_0^j$ (or $x_1^i = x_1^j$) and
$x_1^i \neq x_1^j$ (or $x_0^i \neq x_0^j$)

W.l.o.g we'll assume $x_0^i = x_0^j$ and $x_1^i \neq x_1^j$ then the AND gate which will be defined now, can represents y_i, y_j and only them:



Thus, we can build network with such AND gate and another AND gate for y_k ($k\neq i, k\neq j$, the left item on A) and as before the weights after the AND gates are 1.

So, we've built a network with 2 AND gates that realizes f.

Thus, the induction base for $d=2$ is correct.

Step:

Let's assume that we can build a network with width 2^{d-2} such that can realize any function $g:\{\pm1\}^{d-1}\to\{\pm1\}$.

Let $f:\{\pm1\}^d \to \{\pm1\}$ be a function.

We will prove that we can build a network with at most 2^{d-1} AND gates to realize f.

Let's sign, as in the induction base, $A := \{y \mid f(y)=1\}$

and define $B = \{y = x_2 x_2 \cdots x_d \mid y\in A \text{ and } x_1 = 1\}$

$C = \{y = x_1 x_2 \cdots x_d \mid y\in A \text{ and } x_1 = -1\}$

Of course, it holds $A = B \cup C$

Lemma

① We can build $\leq 2^{d-2}$ AND gates which represent B's items

② We can build $\leq 2^{d-2}$ AND gates which represent C's items.

If the lemma is true, then we can build a network with $\leq 2^{d-2}+2^{d-2} = 2^{d-1}$ AND gates represents $B \cup C = A$ as desired.

Proof of lemma

We'll prove ① and the proof for ② will be the same.

$B = \{y = x_1 x_2 \cdots x_d \mid y \in A, x_1 = 1\}$

define $g: \{\pm 1\}^{d-1} \to \{\pm 1\}$ $\quad g(x_2 \cdots x_d) = f(1 x_2 \cdots x_d)$

Then by the induction we can build $\leq 2^{d-2}$ AND gates that represent $\{x_2 \cdots x_d \mid g(x_2 \cdots x_d) = 1\}$, thus, if for each gate we'll add a weight of 1 on x_1 (in the sense of $f: \{\pm 1\}^d \to \{\pm 1\}$) then these $\leq 2^{d-2}$ AND gates will assure $x_2 \cdots x_d \in \{x_2 \cdots x_d \mid g(x_2 \cdots x_d) = 1\}$ and $x_1 = 1$ which means, by definition of g, they will assure and represent the items in $B = \{1 x_2 \cdots x_d \mid f(1 x_2 \cdots x_d) = g(x_2 \cdots x_d) = 1\}$ as desired.

□

Foundations of Deep Learning – Homework Assignment #2

Adi Album & Tomer Epshtein

Part 1: (2)

Question:

Prove that with polynomial width $\bar{B}$ the deep network cannot realize all possible functions. In particular derive an exponential in d lower bound on $\bar{B}$ required in order for its hypothesis space to be y^X.

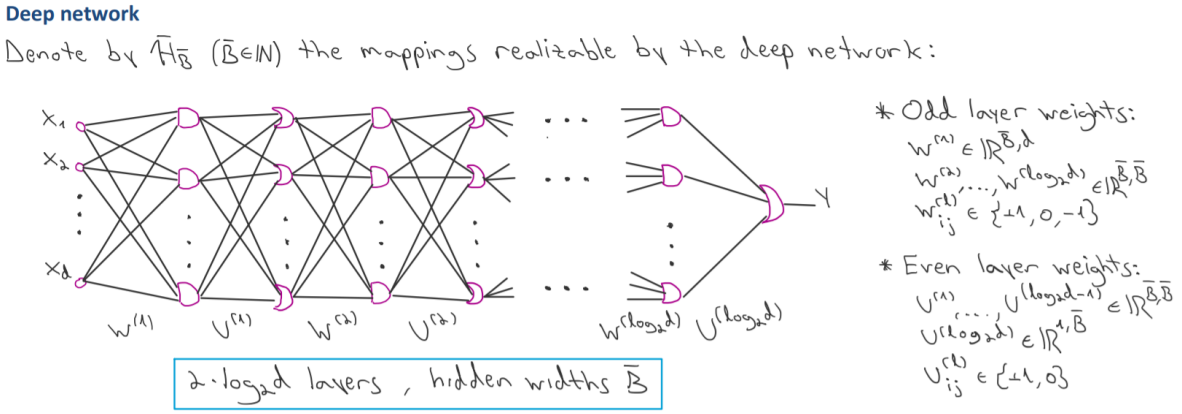
Solution:

Denote by $\bar{H}_{\bar{B}}$ the class of functions realizable by deep AND-OR networks with width $\bar{B}$ (as defined in class). We will prove that $\bar{H}_{\bar{B}} \neq y^X$ by combinatorial reasoning for $\bar{B} \in \mathcal{O}(poly(d))$. And that for $\bar{H}_{\bar{B}} = y^X$ we need $\bar{B}$ to be exponential in d.

- $|y^X| = |y|^{|X|} = 2^{2^d}$
- $|\bar{H}_{\bar{B}}| = ?$

$$|\bar{H}_{\bar{B}}| \leq \#\{Parameter\ combinations\ in\ AND-OR\ deep\ nets\ of\ width\ \bar{B}\}$$

Let's count the number of parameter combinations. Reminder:



- $W^{(1)}$ has $\bar{B} \times d$ parameters from $\{-1, 0, +1\}$, yielding: $3^{\bar{B} \times d}$ combinations
- $W^{(j)}$ has $\bar{B}^2$ parameters from $\{-1, 0, +1\}$, for $j \in \{2, \ldots, \log_2(d)\}$, yielding: $3^{(\log_2(d)-1)\times \bar{B}^2}$ combinations
- $U^{(\log_2(d))}$ has $\bar{B}$ parameters from $\{0, +1\}$, yielding: $2^{\bar{B}}$ combinations
- $U^{(j)}$ has $\bar{B}^2$ parameters from $\{0, +1\}$, yielding: $2^{(\log_2(d)-1)\times \bar{B}^2}$ combinations.

Bringing it all together:

$$|\bar{H}_{\bar{B}}| \leq 3^{\bar{B}\times d} \times 3^{(\log_2(d)-1)\times \bar{B}^2} \times 2^{\bar{B}} \times 2^{(\log_2(d)-1)\times \bar{B}^2}$$

Let $\bar{B} \in \mathcal{O}(poly(d))$.

$$|\bar{H}_{\bar{B}}| \leq 3^{\bar{B}\times d} \times 3^{(\log_2(d)-1)\times \bar{B}^2} \times 2^{\bar{B}} \times 2^{(\log_2(d)-1)\times \bar{B}^2} = 3^{d\bar{B}+(\log_2(d)-1)\bar{B}^2} 2^{\bar{B}+(\log_2(d)-1)\times \bar{B}^2} \leq$$

$$\leq 4^{d\bar{B}+(\log_2(d)-1)\bar{B}^2} 4^{\bar{B}+(\log_2(d)-1)\times \bar{B}^2} = 4^{d\bar{B}+(\log_2(d)-1)\bar{B}^2+\bar{B}+(\log_2(d)-1)\times \bar{B}^2} =$$

$$4^{(d+1)\bar{B}+2(\log_2(d)-1)\bar{B}^2} \overset{(*)}{\leq} 4^{(d+1)\bar{B}^2+2(d+1)\bar{B}^2} = 4^{3(d+1)\bar{B}^2} = 2^{6(d+1)\bar{B}^2}$$

(Where $(*)$ holds because $\bar{B} \leq \bar{B}^2$, $\log_2(d) - 1 \leq d + 1$, and monotonicity of $x \mapsto 4^x$)

For any $\bar{B} \in \mathcal{O}(poly(d))$:
$2^{6(d+1)\bar{B}^2} = 2^{\mathcal{O}(poly(d))}$ and in particular $2^{6(d+1)\bar{B}^2} < 2^{2^d}$ so $|\bar{H}_{\bar{B}}| < |y^X|$.
So,
$\bar{H}_{\bar{B}} \neq y^X$ for $B \in \mathcal{O}(poly(d))$.

We will now derive an exponential lower bound for $\bar{B}$ for us to achieve $\overline{H}_{\bar{B}} = y^X$.

We saw

$$|\overline{H}_{\bar{B}}| \leq 2^{6(d+1)\bar{B}^2}$$

For $\overline{H}_{\bar{B}} = y^X$ we must have

$$2^{6(d+1)\bar{B}^2} \geq 2^{2^d}$$

$\Longrightarrow$

$6(d+1)\bar{B}^2 \geq 2^d \Longrightarrow \bar{B}^2 \geq \frac{2^d}{6(d+1)}$.

So for $\overline{H}_{\bar{B}} = y^X$, we must have $\bar{B}^2 \geq \frac{2^d}{6(d+1)}$ And in particular $\bar{B} \geq \frac{2^{\frac{d}{2}}}{\sqrt{6(d+1)}}$.

An exponential in d lower bound on $\bar{B}$, as desired.

■

Part 2

Question 1

Proposition: A shallow network of width $B \geq 2$ can realize any piecewise linear mapping with $\leq B$ pieces.
Conversely, any mapping realizable by such network is piecewise linear with $\leq B+1$ pieces.

Proof

First, I will explain why the condition $B \geq 2$ is necessary.

Explaination:

By definition of a piecewise linear function, we only have to explain why the proposition doesn't hold for the case $B = 1$.

In that case, it means that the function is linear in $(-\infty, \infty)$ which means, each function f which is piecewise linear with 1 piece is

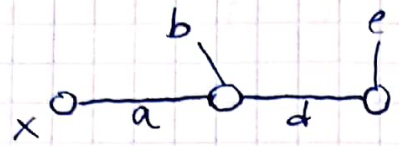
$f(x) = \alpha_f x + \beta_f \quad \forall x \in (-\infty, \infty)$

Let f_1 be a linear function $f_1(x) = \alpha_1 x + \beta_1$ s.t $\alpha_1 \neq 0$

Assume, in contradiction, that the proposition holds for the case $B = 1$.

Then, By the proposition, there exists a shallow network with width $B=1$ that can realize f_1.
Lets sign this network as N_{f_1} and the weighes as:

N_{f_1} :



which means there exist, $a,b,d,e \in \mathbb{R}$ s.t

$$N_{f_1}(x) = d \cdot \text{RELU}(ax+b) + e$$

Of course $a \neq 0$, otherwise N_{f_1} would output the same output for each input, in contradiction that it realizes f_1 which outputs other values for each two inputs ($f_1(x) = \alpha_1 \cdot x + \beta_1$, and $\alpha_1 \neq 0$).

It holds that $ax+b=0$ for $x = \frac{-b}{a}$.
If $a<0$ then $ax+b \leq 0$ for $-\frac{b}{a}$, $-\frac{b}{a}+1$.
If $a>0$ then $ax+b \leq 0$ for $-\frac{b}{a}$, $-\frac{b}{a}-1$.
In both cases we found 2 different points which $ax+b \leq 0$, and thus $\text{RELU}(ax+b) = \max(0, ax+b) = 0$, thus $N_{f_1}(x)$ outputs $(d \cdot 0 + e = e)$ for both points in contradiction that f_1 never outputs the same value for 2 different points, so N_{f_1} doesnt realize f_1, in contradiction.

$\Rightarrow$ For our proposition, it must holds $B \geq 2$.

□

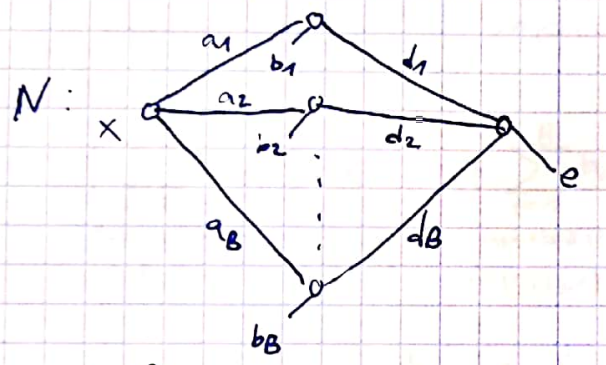
Part 2 - question 1

Now, we will prove the following statement/proposition:
Any mapping realizable by such network is piecewise linear with $\leq B+1$ pieces.

proof

Let sign such network as N and its weights as

$$\{a_i, b_i, d_i\}_{i=1}^{B}, e$$



$$N(x) = \sum_{i=1}^{B} RELU(a_i x + b_i) \cdot d_i + e$$

For each $1 \leq i \leq B$ s.t $a_i \neq 0$ it holds that at $c_i := \frac{-b_i}{a_i}$: $a_i c_i + b_i = 0$. We will sign these points as $\{c_1, \dots c_j\}$ for $j = |\{a_i \mid a_i \neq 0\}| \leq B$ and assume w.l.o.g that $c_1 < c_2 < \dots < c_j$ (otherwise we would rename them) and $c_0 = -\infty$, $c_{j+1} = \infty$.
Then, by definition of these points, we conclude that for each interval (c_i, c_{i+1}) $[0 \leq i \leq j]$ it holds for each $1 \leq i \leq B$ $a_i x + b_i \geq 0$ or $a_i x + b_i \leq 0$, for all $x \in (c_i, c_{i+1})$.

We have got that each interval (c_i, c_{i+1}) holds that in it $a_i x + b_i \geq 0$ or $a_i x + b_i \leq 0$ for each $1 \leq i \leq B$.

Thus: For a specific interval (c_i, c_{i+1}):

$$N(x) = \sum_{i=1}^{B} \text{RELU}(a_i x + b_i) d_i + e$$

$$= \sum_{\substack{i=1 \\ a_i x + b_i \geq 0 \text{ in } (c_i, c_{i+1})}}^{B} (a_i x + b_i) d_i + e = \sum_{\substack{i=1 \\ a_i x + b_i \geq 0 \text{ in } (c_i, c_{i+1})}}^{B} a_i d_i x + b_i d_i + e$$

(when $a_i x + b_i \leq 0$ then RELU() $= 0$ ✓)

$$= \left(\sum_{\substack{i=1 \\ a_i x + b_i \geq 0 \text{ in } (c_i, c_{i+1})}}^{B} a_i d_i \right) x + \left(\sum_{\substack{i=1 \\ a_i x + b_i \geq 0 \text{ in } (c_i, c_{i+1})}}^{B} b_i d_i + e \right)$$

Which means that N is linear in (c_i, c_{i+1}), but that's true for each $0 \leq i \leq j$ and thus we conclude that N realizes a piecewise linear function with $j+1 \leq B+1$ pieces as we wished.

□

Now, we will prove the last part of the question:

A shallow network of width $B \geq 2$ can realize any piecewise linear mapping with $\leq B$ pieces.

Proof

We will prove the proposition for linear mappings with B pieces and if there would be less than B pieces, we would use the same solution ~~for the worst~~, but put in the last neurons of the network zeros on their weights.

Let f be a linear mapping with B pieces.

By definition, there exise:

$$-\infty = c_0 < c_1 < c_2 < \dots < c_{B-1} < c_B = \infty$$

and $\{\alpha_i, \beta_i\}_{i=1}^{B}$

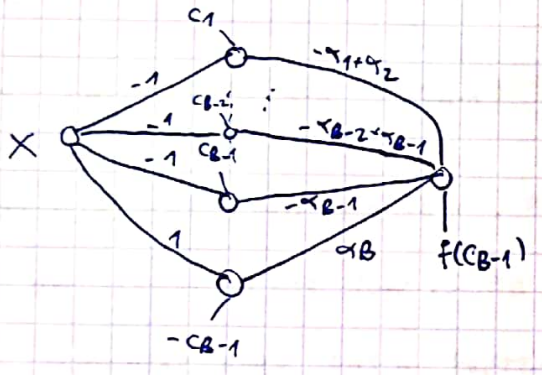
such that:

$$\forall\, 0 \le i \le B-1: \quad \forall x \in (c_i, c_{i+1}) \quad f(x) = \alpha_{i+1} x + \beta_{i+1}.$$

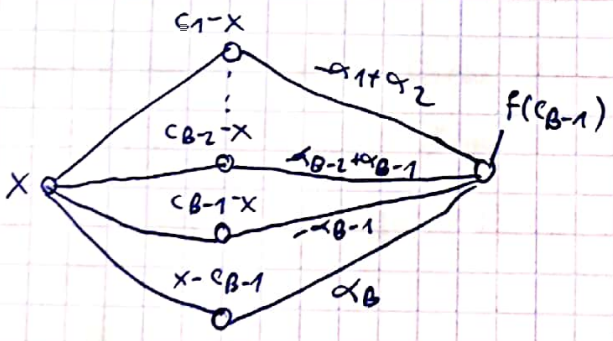
We will build a network with width B that realizes f:

We will define the network N as the following:

N:



more clearly:



We'll prove that $N(x) = f(x)$ for every x and we will finish.

① we will prove that $\forall x \ge c_{B-1}$ $N(x) = f(x)$ and then

② we will prove that $N(x) = f(x)$ on each (c_i, c_{i+1})

$0 \le i \le B-2$

① Let $x \geq c_{B-1}$, then $\forall 1 \leq i \leq c_{B-1} \quad c_i - x \leq 0$,

thus $RELU(c_i - x) = 0$

and we conclude that

$$N(x) = \sum_{i=1}^{B-1} RELU(c_i - x) \cdot d_i + RELU(x - c_{B-1}) \cdot \alpha_B + f(c_{B-1})$$

(the weights as defined)

$$= RELU(x - c_{B-1}) \cdot \alpha_B + f(c_{B-1})$$

$x - c_{B-1} \geq 0$

$$= (x - c_{B-1}) \alpha_B + f(c_{B-1})$$

[Note that at $x = c_{B-1}$ $N(x) = f(c_{B-1})$]

$$= \alpha_B \cdot x - \alpha_B c_{B-1} + f(c_{B-1})$$

Thus, $N(x)$ is linear on $x \geq c_{B-1}$, and has the same slope as f has on $x \geq c_{B-1}$ (the slope is α_B) and also $N(c_{B-1}) = f(c_{B-1})$

Then, we conclude that they represent the same line on $x \geq c_{B-1}$ $\Rightarrow N(x) = f(x) \quad \forall x \geq c_{B-1}$. □

② Let $0 \leq i \leq B-2$. We'll show that $N(x) = f(x)$ on (c_i, c_{i+1}). It holds for every $x \in (c_i, c_{i+1})$:

$$N(x) = \sum_{j=1}^{B-1} RELU(c_j - x) d_j + RELU(x - c_{B-1}) \alpha_B + f(c_{B-1})$$

$$= \sum_{j=i+1}^{B-1} (c_j - x) d_j + f(c_{B-1})$$

$$= -\left(\sum_{j=i+1}^{B-1} d_j\right) x + \sum_{j=i+1}^{B-1} c_j d_j + f(c_{B-1})$$

Thus N is linear on (c_i, c_{i+1}) and the slope there is

$$-\left(\sum_{j=i+1}^{B-1} d_j\right) = -(d_{i+1} + d_{i+2} + \dots + d_{B-1})$$

$$= -(-\alpha_{i+1} + \alpha_{i+2} - \alpha_{i+2} + \alpha_{i+3} - \dots + \alpha_{B-1} - \alpha_{B-1})$$

$$= -(-\alpha_{i+1}) = \alpha_{i+1}$$

Until now, we have got that for each $0 \leq i \leq B-2$ N acts on (c_i, c_{i+1}) as a linear function with the same slope which f acts on that interval. that means:

$$f(x) = \alpha_{i+1} x + \beta_{i+1} \quad \text{(by definition)}$$

$$N(x) = \alpha_{i+1} x + \gamma_{i+1}$$

where $\gamma_{i+1} = \sum_{j=i+1}^{B-1} c_j d_j + f(c_{B-1})$

So it only remains for us to explain $\gamma_{i+1} = \beta_{i+1}$

So we conclude that N realizes the function f.

We have already proved that $N(c_{B-1}) = f(c_{B-1})$

and since f and N are continous (N is composition of continous functions, and f by definition is continous) then it means that N acts on $[c_{B-2}, c_{B-1}]$ exactly like f does (*) → which means $\gamma_{B-1} = \beta_{B-1}$.

Same slope and same value at one point

The last also means that $N(c_{B-2}) = f(c_{B-2})$, and we can yield that $\gamma_{B-2} = \beta_{B-2}$.

Iteratively, we can yield that $\gamma_{i+1} = \beta_{i+1}$ $\forall\, 0 \leq i \leq B-2$

Thus, N realizes the function f as desired.

□

Question 2

In the context of the expressive efficiency with inapproximability analysis delivered in class - derive a lower bound on the number of intervals in $S_>$ and $S_<$ that a piecewise linear mapping with $\leq B+1$ pieces must miss.

Solution

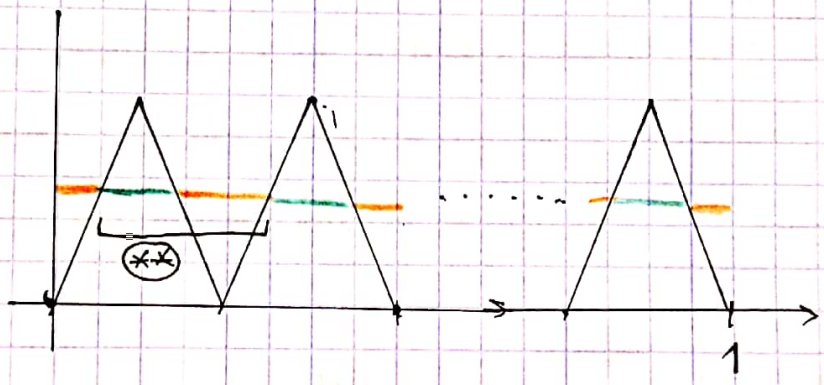
We'll remind the definitions for $S_>$, $S_<$:

$$S_> := \{ x \in [0,1] : g^{\circ L-1}(x) > \frac{1}{2} \}$$

$$S_< := \{ x \in [0,1] : g^{\circ L-1}(x) < \frac{1}{2} \}$$

— : $S_>$

— : $S_<$



A function misses interval in $S_>$ if $f \leq \frac{1}{2}$ on that interval.

" " " " " $S_<$ " $f \geq \frac{1}{2}$ " " " .

We proved that the number of intervals is $2^{L-1}+1$ - there are $2^{L-2}+1$ intervals in $S_<$ and 2^{L-2} intervals in $S_>$.

We'll look at the first interval of $S_<$, we'll sign it $I_1^<$.

Now we'll define 2^{L-2} tuples such the i'th tuple $1 \leq i \leq 2^{L-2}$ consists of the i'th interval of $S_>$ and the following interval in $S_<$. For example, (XX) will be the first tuple.

Lemma

Let f be a piecewise linear with $\leq B+1$ pieces.
We'll prove that f can avoid missing at most
$\lceil \frac{1}{2}(2^{L-1}+1) + \frac{1}{2}(B+1) \rceil$ intervals.

Proof of lemma

Since f is piecewise linear with $\leq B+1$ pieces, it has at most B "breaking points" where the slope is change there.
We'll look at the j breaking points which inside one of our tuple $[\{x \mid x \text{ is breaking point of } f \text{ and } x \in [\frac{1}{4}\cdot 2^{-L+2}, 1]\}]$.

Of course $j \leq B$ (there are at most B breaking points on all $\mathbb{R}$).

We'll count how many of the pairs of intervals which are the tuples we've difined - f can avoid missing both intervals in the tuple (the one in $S_>$, and the one in $S_<$).

Note that between each two breaking points f is linear and thus if there exises a tuple of intervals $(S_>^i, S_<^i)$ that f didn't miss, then it means that after this tuple (or inside him, depends where the next breaking point is) $f < \frac{1}{2}$ and thus there cannot be another turle which f avoids missing both intervals there.
Thus, beceen each two breaking points (or the last breaking point and the point 1), f can avoid missing at most 1 tuple (which means f can avoid missind only once a tuple, two intervals)

Therefore, from all the tuples we have, f can avoid missing at most j tuples.

Now, there are two cases:

① For any tuple until the first breaking point, f misses one of the intervals in that tuple:

In that case we've got that at most j tuples f avoided missing the intervals in them (from the 2^{L-2} tuples), and of course that it missed a tuple (which means that f wouldn't able to avoid missing both intervals in the tuple) than f avoided missing one of them and missed the second. In total we look at $2^{L-2} \cdot 2 = 2^{L-1}$ intervals, and we know that f avoided missing at most j tuples, thus f avoided missing at most $\lceil \frac{1}{2} \cdot 2^{L-1} + \frac{1}{2} \cdot j \rceil$ intervals from the 2^{L-1} intervals. In addition we have the first interval $I_1^<$ which maybe f avoided missing it too, thus in total f avoided missing at most

$$\lceil \tfrac{1}{2} \cdot 2^{L-1} + \tfrac{1}{2} \cdot j \rceil + 1 = \lceil \tfrac{1}{2} \cdot (2^{L-1}+1) + \tfrac{1}{2}(j+1) \rceil$$

② There exists a tuple which f avoided missing both its intervals [the tuple is until the first breaking point]:

In that case we get that f avoided $j+1$ tuples at most but in that case f must miss the interval $I_1^<$ because we assume f is linear until the first breaking point and because it avoided a tuple it means that $f > \frac{1}{2}$ in the beginning of the tuple, and f has negative slope, thus on $I_1^<$ $f > \frac{1}{2}$ necessarily.

Thus, from this explanation + the explanation at ①:

f avoided missing at most $\lceil \frac{1}{2} \cdot 2^{L-1} + \frac{1}{2}(j+1) \rceil$

We've got that f avoided missing at most

$$\left\lceil \frac{1}{2}\cdot(2^{L-1}+1)+\frac{1}{2}(j+1) \right\rceil \leq \left\lceil \frac{1}{2}\cdot(2^{L-1}+1)+\frac{1}{2}(B+1) \right\rceil$$

$j \leq B$ intervals.

Thus, because the number of total intervals is $2^{L-1}+1$ it means that f must miss at least:

$$\left\lfloor \frac{1}{2}\cdot(2^{L-1}+1)-\frac{1}{2}(B+1) \right\rfloor$$

□

Part 2 - Question 3

We have to modify the universality and expressive efficiency analysis given in class so that they apply to leaky RELU activation $\sigma(z) = \max\{az, z\}$ for $a \in (0,1)$ instead of RELU.

Proof

Notice that in the regular proof of universality (with RELU) we only used the proposition ⊛ from question 1:

> A shallow network of width $B \geq 2$ can realize any piecewise linear mapping with $\leq B$ pieces.

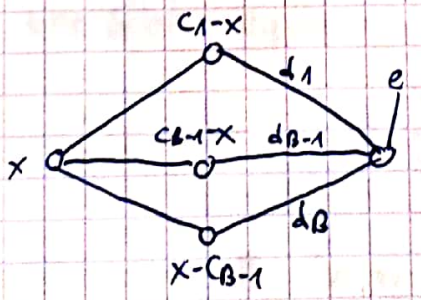
Thus, it's enough for us to prove it now with shallow network that uses leaky RELU and we'll conclude universality also for networks that use leaky RELU.

Proof of proposition ⊛

Let f be a piecewise linear with $\leq B$ pieces, defined by $-\infty = c_0 < c_1 < c_2 < \dots < c_{B-1} < c_B = \infty$

$$f(x) = \alpha_{i+1} x + \beta_{i+1} \qquad 0 \leq i \leq B-1$$

We'll build a network N which relizes f:



where we haven't defined $d_1, \dots, d_B, e$ yet.

We've already proved that on each interval (c_i, c_{i+1}) N is linear (in question 1), and we would like to see $d_1 \dots d_B, e$ s.t $N(x) = f(x)$ on each interval. [The proof that N is linear on each interval was with RELU but it is exacely the same for leaky RELU - instead looking at the line $y=0$ for $x \le 0$, the line will be $y = ax$]

$$N(x) = \sum_{i=1}^{B-1} LRELU(c_i - x) \cdot d_i + LRELU(x - c_{B-1}) d_B + e$$

1. $x \ge c_{B-1}$: on that part, we will get $LRELU(x - c_{B-1}) = x - c_{B-1}$ and $LRELU(c_i - x) = a(c_i - x)$ $\forall 1 \le i \le B-1$

thus

$$N(x) = \sum_{i=1}^{B-1} a(c_i - x) d_i + d_B(x - c_{B-1}) + e$$

$$= \left(\sum_{i=1}^{B-1} -a d_i + d_B\right) x + \sum_{i=1}^{B-1} a c_i d_i - c_{B-1} + e$$

$\Rightarrow$ we would want $\sum_{i=1}^{B-1} -a d_i + d_B = \alpha_B$ (slope)

2. $c_{B-2} \le x \le c_{B-1}$:

With the same process we'll get:

$$\sum_{i=1}^{B-2} -a d_i - d_{B-1} + a d_B = \alpha_{B-1}$$

$\vdots$

We'll get the B'th equation

$$\sum_{i=1}^{B-1} -d_i + a d_B = \alpha_1$$

Thus, we've got B linear equation.

We'll write our B equations as matrix:

$$\begin{bmatrix} -a & -a & \cdots & -a & 1 \\ & & & -1 & a \\ & -a & & & \vdots \\ & & -1 & & \\ -a & & & & \\ -1 & & \cdots & -1 & a \end{bmatrix} \begin{pmatrix} d_1 \\ \vdots \\ d_B \end{pmatrix} = \begin{pmatrix} \alpha_B \\ \alpha_{B-1} \\ \vdots \\ \alpha_1 \end{pmatrix}$$

Let's sign the matrix as A.

Then if A is invertible, then there exists a solution for $\begin{pmatrix} d_1 \\ \vdots \\ d_B \end{pmatrix}$ and all the equations we want to exist will exist. In that case, we will assign the left variable e to be such $N(C_{B-1}) = f(C_{B-1})$ and then by the explanation in question 1 we will get that $N(x) = f(x)$ [the explanation will be the same because LRELU is continous function and thus N is still continous because it is composition of continous functions]

Now, it's easy to see that A is invertible because by B-1 actions on the first B-1 rows ($r_i \leftarrow r_i - a r_B$) we'll get the matrix:

$$\begin{bmatrix} -a & -a & \cdots & -a & 1 \\ & & & -1 & a \\ & -a & & & \vdots \\ & & -1 & & \\ -a & & & & \\ -1 & & \cdots & -1 & a \end{bmatrix} \longrightarrow \begin{bmatrix} \bigcirc & & & 1-a^2 \\ & & & -1+a \\ & & & * \\ & & -1+a & \\ & -1 & & \end{bmatrix} \quad i=B$$

where B is invertible because $\det(B) = (-1)^{B+1}(1-a^2)\cdot(-1+a)^{B-2}\cdot(-1) \neq 0$ ($a \in (0,1)$)

(the only permutation which doesn't have the value 0)

$\Rightarrow$ A is invertible [$\det(A) \neq 0$]

Now, lets focus on the proof of expressive eficiency.

In that part, we have used the proposition from question 1 (not only the first part of it, we have proved in the universality with leaky RELU)

Note the second part of the proposition: (**)

> any mapping realitable by shallow network with width $B \geq 2$ is piecewise linear with $\leq B+1$ pieces.

Then the proof was given to this part in question 1 with regular RELU holds exacely for networks with leaky RELU [instead $y=0$ when $a_i x+b_i \leq 0$ we'll get $a(a_i x+b_i)$]

Moreover

~~Except~~, when we proved that $\exists \bar{h} \in \bar{H}_B$ for $B \in O(1)$ s.t $\bar{h} \notin H_B$ unless $B \in \exp(L)$

we built $\bar{h} = g^{\circ k}$ when we proof that

$$g(x) = RELU[2x] - RELU[4x-2] + RELU[2x-2]$$

Now, from the recitation, we proved that:

$$RELU[wx+b] = \alpha LRELU[wx+b] + \beta LRELU[-wx-b]$$

for fixed $\alpha, \beta \in \mathbb{R}$

Thus, we can write:

$\alpha = \frac{1}{1-q^2}$ $\beta = \frac{q}{1-q^2}$

$$g(x) = (\alpha RELU[2x] + \beta LRELU[-2x])$$
$$- (\alpha LRELU[4x-2] + \beta LRELU[-4x+2])$$
$$+ (\alpha LRELU[2x-2] + \beta LRELU[-2x+2])$$

And therefore we can write g as sum of $\bar{B} = 6 \in O(1)$ leaky RELU, thus by composing on it k times g^k can still be implemented in the network with $\bar{B}=6$ and (eaky RELU.

The other parts of the proof remain the <u>same</u>.

Question 4:

Prove that the shallow and deep networks are $\mathcal{F}$-universal in the sense of $d(\cdot,\cdot)$, where $\mathcal{F} \subseteq \mathbb{R}^{\mathbb{R}}$ is the set of Riemann integrable functions and $d: \mathcal{F}\times\mathcal{F} \to \mathbb{R}_{\geq 0}$ is defined by:

$$d(f_1, f_2) = \int_0^1 |f_1(x) - f_2(x)|\, dx$$

Proof

We would like to prove the following proposition: ⊛

$$\forall \varepsilon > 0 \quad \forall f \in \mathcal{F} \quad \exists_B \exists N \in H_B : \quad d(f, N) \leq \varepsilon$$

That will prove that the shallow network is $\mathcal{F}$-universal and since we've already proved that $H_B \subseteq \overline{H}_{\bar{B}}$ where $\bar{B} = B$ that will prove that the deep network is $\mathcal{F}$-universal too.

lemma: Each $f \in \mathcal{F}$ is bounded: $\exists M \in \mathbb{R}: |f(x)| \leq M \quad \forall x \in [0,1]$

proof:

Recall that a function is supposed to be Riemann integrable if for all $\varepsilon > 0$ there exists a partition s.t $U - L < \varepsilon$

[when $U = \sum_{i=0}^{n} \sup_{t \in [x_i, x_{i+1}]} f(t) \cdot (x_{i+1} - x_i)$

$L = \sum_{i=0}^{n} \inf_{t \in [x_i, x_{i+1}]} f(t) \cdot (x_{i+1} - x_i)$]

Suppose f becomes unbounded (say, unbounded above) near the point x_0. Then the definition of U isn't defined well because in the interval where x_0 belongs to, the supremum is not defined well, contradiction. [the same for L]

Thus, there exists M s.t $|f(x)| \leq M \quad \forall x \in [0,1]$

Proof of proposition ⊛:

Let $\varepsilon>0$, $f\in F$. By the lamma, there exists $M\in\mathbb{R}$ s.t $|f(x)|\le M$.

In addition, $f\in F$ and therefore there exises a seq of intervals $[x_0,x_1],[x_1,x_2],\dots,[x_n,x_{n+1}]$ s.t: $x_0=0$, $x_{n+1}=1$,

, $U=\sum_{i=0}^{n}\sup_{t\in[x_i,x_{i+1}]} f(t)\cdot(x_{i+1}-x_i)$, $L=\sum_{i=0}^{n}\inf_{t\in[x_i,x_{i+1}]} f(t)\ (x_{i+1}-x_i)$

and $U-L<\frac{\varepsilon}{2}$.

Lets define $U:[0,1]\to\mathbb{R}$ $\quad U(x)=\sup_{y\in[x_i,x_{i+1}]} f(y)$ i s.t $x\in[x_i,x_{i+1}]$

$L:[0,1]\to\mathbb{R}$ $\quad L(x)=\inf_{y\in[x_i,x_{i+1}]} f(y)$ i s.t $x\in[x_i,x_{i+1}]$

Then, by definition $L(x)\le f(x)\le U(x)$ $\forall x\in[0,1]$

and thus:

$$\int_0^1|U(x)-f(x)|\le\int_0^1 U(x)-f(x)\le\int_0^1 U(x)-L(x)$$

$$=\int_0^1 U(x)\,dx-\int_0^1 L(x)=U-L\le\frac{\varepsilon}{2}$$

Thus, we've got $\int_0^1|U(x)-f(x)|\le\frac{\varepsilon}{2}$

Now we'll define $N(x)\in H_B$ for $B=2n-1$ (which means a function that we can represent with shallow network with $B=2n-1$ width), we will prove that

$\int_0^1|N(x)-U(x)|\le\frac{\varepsilon}{2}$.

Say it is true, then we'll get:

$$d(N,f)=\int_0^1|N(x)-f(x)|=\int_0^1|N(x)-U(x)+U(x)-f(x)|\overset{\Delta}{\le}\int_0^1|N(x)-U(x)|\,dx+\int_0^1|U(x)-f(x)|\,dx$$

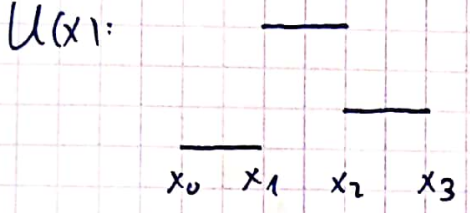
$$\le\frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon$$

as we wished.

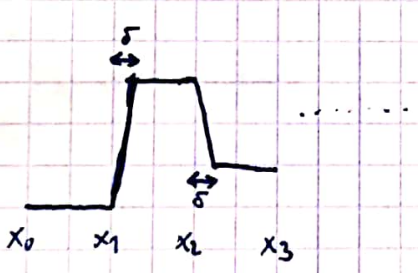
Defining the function N(x):

We'll define a function which is piecewise linear with $B=2n-1$ pieces, and from question 1 we already know that a shallow network of width B can realize it.

We'll draw the function U(x) which was defined earlier:



U(x) is composed from set of intervals s.t on each interval U is const. We'll define N(x) by the following [N is almost like U]:



where $\delta = \frac{\varepsilon}{2n \cdot M}$

[From each interval $[x_i, x_{i+1}]$ $i \geq 1$, we take δ from the beginning and make a line to the end of the previous interval]

Of course N(x) is defined well and has $n+(n-1)=2n-1$ pieces which compose the function as we wished

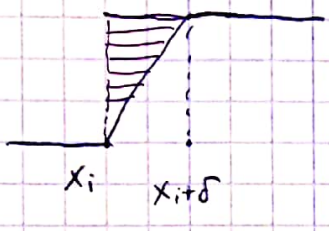
So a network with width $B=2n-1$ can implement it.

Now, we'll prove that $\int_0^1 |N(x)-U(x)| \leq \frac{\varepsilon}{2}$

Notice that N and u are only differ on each beginning of size δ for each interval and thus:

$$\int_0^1 |N(x)-u(x)|dx = \sum_{i=1}^{n} \int_{x_i}^{x_i+\delta} |N(x)-u(x)|dx$$

Now for each $i \geq 1$, $\int_{x_i}^{x_i+\delta} |N(x)-u(x)|dx$ is actually the area of the following triangle:



As we know, $|f(x)| \leq M$, and therefore for every two different points x, y $|f(x)-f(y)| \leq 2M$.

Now we've got:

$$\int_{x_i}^{x_i+\delta} |N(x)-u(x)|dx = \text{Area of triangle} = \frac{\delta \cdot \left|f\left(\frac{x_i+x_{i+1}}{2}\right)-f\left(\frac{x_i+x_{i-1}}{2}\right)\right|}{2}$$

$$\leq \frac{\delta \cdot 2M}{2} = \frac{\frac{\varepsilon}{2Mn} \cdot 2M}{2} = \frac{\varepsilon}{2n}$$

we chose $\delta = \frac{\varepsilon}{2Mn}$

Therefore,

$$\int_0^1 |N(x)-u(x)|dx = \sum_{i=1}^{n} \int_{x_i}^{x_i+\delta} |N(x)-u(x)|dx \leq \sum_{i=1}^{n} \frac{\varepsilon}{2n} = n \cdot \frac{\varepsilon}{2n} = \frac{\varepsilon}{2}$$

as we wished.

□

Foundations of Deep Learning – Homework Assignment #2

Adi Album & Tomer Epshtein

Part 3: (1)

Prove (Lemma): ("matricization of outer product = Kronecker product of matricizations")

Let T and $\bar{T}$ be tensors of order n and $\bar{n}$ respectively. Let $I \subset [n+\bar{n}]$ and denote by $I-n$ the set obtained by subtracting n from each element in I. Then:

$$[|T \otimes \bar{T}|]_I = [|T|]_{I \cap [n]} \odot [|\bar{T}|]_{I-n \cap [\bar{n}]}$$

Proof:

Let $T \in \mathbb{R}^{m_1 \times \ldots \times m_n}$ be a tensor of order n and $\bar{T} \in \mathbb{R}^{\overline{m_1} \times \ldots \times \overline{m_{\bar{n}}}}$ be a tensor of order $\bar{n}$.

Denote $I = \{i_1, \ldots, i_{|I|}\} \subset [n+\bar{n}]$ where $1 \leq i_1 < \cdots < i_{|I|} \leq n+\bar{n}$.

Are the two sides of the target equation equal?

1. Sanity check – they have equal dimensions… Both sides of the equation are matrices of dimensions:
$$\left(\prod_{i \in I \cap [n]} m_i\right) \cdot \left(\prod_{i \in I-n \cap [\bar{n}]} \overline{m_i}\right) \times \left(\prod_{i \in I^C \cap [n]} m_i\right) \cdot \left(\prod_{i \in I^C-n \cap [\bar{n}]} \overline{m_i}\right)$$

2. We'll use linearity to make our proof less complicated.
   a. Linearity of tensor outer product:
   $$(\alpha A + \beta B) \otimes C = \alpha(A \otimes B) + \beta(B \otimes C)$$
   $$C \otimes (\alpha A + \beta B) = \alpha(C \otimes A) + \beta(C \otimes B)$$
   For any scalars $\alpha, \beta \in \mathbb{R}$ and any tensors A, B, C such that A and B have equal dimensions.

   b. Linearity of Matricization:
   $$[|\alpha A + \beta B|]_I = \alpha[|A|]_I + \beta[|B|]_I$$
   For any scalars $\alpha, \beta \in \mathbb{R}$ and any tensors A, B such that A and B have equal dimensions.

   c. Linearity of Kronecker Product:
   $$(\alpha A + \beta B) \odot C = \alpha(A \odot C) + \beta(B \odot C)$$
   $$C \odot (\alpha A + \beta B) = \alpha(C \odot A) + \beta(C \odot B)$$
   For any scalars $\alpha, \beta \in \mathbb{R}$ and any tensors A, B, C such that A and B have equal dimensions.

   Any tensor $T \in \mathbb{R}^{m_1 \times \ldots \times m_n}$ can be represented as a linear combinations of tensors $E_i \in \mathbb{R}^{m_1 \times \ldots \times m_n}$ with scalars $\alpha_i \in \mathbb{R}\ i \in [m_1 \cdot \ldots \cdot m_n]$, such that E_i is a tensor with '1' in a single entry and '0' in all other entries.
   $$T = \sum_{i=1}^{m_1 \cdot \ldots \cdot m_n} \alpha_i E_i$$

3. Claim:
   Let $E_i \in \mathbb{R}^{m_1 \times \ldots \times m_n}$ be some tensor with '1' in a single entry and '0' in all other entries,
   Let $\bar{E_i} \in \mathbb{R}^{\overline{m_1} \times \ldots \times \overline{m_{\bar{n}}}}$ be some tensor with '1' in a single entry and '0' in all other entries.
   Let $I \subset [n+\bar{n}]$ and denote by $I-n$ the set obtained by subtracting n from each element in I.
   Then:
   $$[|E_i \otimes \bar{E_i}|]_I = [|E_i|]_{I \cap [n]} \odot [|\bar{E_i}|]_{I-n \cap [\bar{n}]}$$

4. Assume we proved Claim (3).
   We'll first show this completes our proof:

Let $T \in \mathbb{R}^{m_1 \times \ldots \times m_n}$ be a tensor of order n and $\bar{T} \in \mathbb{R}^{\overline{m_1} \times \ldots \times \overline{m_{\bar{n}}}}$ be a tensor of order $\bar{n}$.

By (2) we have $T = \sum_{i=1}^{m_1 \cdot \ldots \cdot m_n} \alpha_i E_i$ and $\bar{T} = \sum_{\bar{\imath}=1}^{\overline{m_1} \cdot \ldots \cdot \overline{m_{\bar{n}}}} \overline{\alpha_\imath} \overline{E_\imath}$

$[|T \otimes \bar{T}|]_I = \left[\left|\sum_{i=1}^{m_1 \cdot \ldots \cdot m_n} \alpha_i E_i \otimes \sum_{\bar{\imath}=1}^{\overline{m_1} \cdot \ldots \cdot \overline{m_{\bar{n}}}} \overline{\alpha_\imath} \overline{E_\imath}\right|\right]_I =$
{Linearity of tensor outer product}
$= \left[\left|\sum_{i=1}^{m_1 \cdot \ldots m_n} \sum_{\bar{\imath}=1}^{\overline{m_1} \cdot \ldots \cdot \overline{m_{\bar{n}}}} \alpha_i \overline{\alpha_\imath} \cdot (E_i \otimes \overline{E_\imath})\right|\right]_I =$
{Linearity of matricization}
$= \sum_{i=1}^{m_1 \cdot \ldots m_n} \sum_{\bar{\imath}=1}^{\overline{m_1} \cdot \ldots \cdot \overline{m_{\bar{n}}}} \alpha_i \overline{\alpha_\imath} \cdot [|E_i \otimes \overline{E_\imath}|]_I =$
{(3)}
$= \sum_{i=1}^{m_1 \cdot \ldots m_n} \sum_{\bar{\imath}=1}^{\overline{m_1} \cdot \ldots \cdot \overline{m_{\bar{n}}}} \alpha_i \overline{\alpha_\imath} \cdot [|E_i|]_{I \cap [n]} \odot [|\overline{E_\imath}|]_{I-n \cap [\bar{n}]} =$
{Linearity of keronecker product}
$= \sum_{i=1}^{m_1 \cdot \ldots \cdot m_n} \alpha_i [|E_i|]_{I \cap [n]} \odot \sum_{\bar{\imath}=1}^{\overline{m_1} \cdot \ldots \cdot \overline{m_{\bar{n}}}} \overline{\alpha_\imath} [|\overline{E_\imath}|]_{I-n \cap [\bar{n}]} =$
{Linearity of matricization}

$$= \left[\left|\sum_{i=1}^{m_1 \cdot \ldots \cdot m_n} \alpha_i E_i\right|\right]_{I \cap [n]} \odot \left[\left|\sum_{\bar{\imath}=1}^{\overline{m_1} \cdot \ldots \cdot \overline{m_{\bar{n}}}} \overline{\alpha_\imath} \overline{E_\imath}\right|\right]_{I-n \cap [\bar{n}]} = [|T|]_{I \cap [n]} \odot [|\bar{T}|]_{I-n \cap [\bar{n}]}$$

Which gives us:

$$[|T \otimes \bar{T}|]_I = [|T|]_{I \cap [n]} \odot [|\bar{T}|]_{I-n \cap [\bar{n}]}$$

As desired.

It remains to prove claim 3:

Claim 3:
Let $E_i \in \mathbb{R}^{m_1 \times \ldots \times m_n}$ be some tensor with '1' in a single entry and '0' in all other entries,
Let $\overline{E_\imath} \in \mathbb{R}^{\overline{m_1} \times \ldots \times \overline{m_{\bar{n}}}}$ be some tensor with '1' in a single entry and '0' in all other entries.
Let $I \subset [n + \bar{n}]$ and denote by $I - n$ the set obtained by subtracting n from each element in I.
Then:

$$[|E_i \otimes \overline{E_\imath}|]_I = [|E_i|]_{I \cap [n]} \odot [|\overline{E_\imath}|]_{I-n \cap [\bar{n}]}$$

Proof of claim 3:

Let $E_i \in \mathbb{R}^{m_1 \times \ldots \times m_n}$ be some tensor with '1' in a single entry and '0' in all other entries,
I.e. $(E_i)_{d_1, \ldots, d_n} = 1$ for some $d_k \in [m_k], k \in [n]$. And E_i is 0 for all other entries.

Let $\overline{E_\imath} \in \mathbb{R}^{\overline{m_1} \times \ldots \times \overline{m_{\bar{n}}}}$ be some tensor with '1' in a single entry and '0' in all other entries.
I.e. $(\overline{E_\imath})_{\overline{d_1}, \ldots, \overline{d_{\bar{n}}}} = 1$ for some $\overline{d_k} \in [\overline{m_k}], k \in [\bar{n}]$. And $\overline{E_\imath}$ is 0 for all other entries.

Start by viewing LHS: $[|E_i \otimes \overline{E_\imath}|]_I$

- **Tensor outer product:**

  $E_i \otimes \overline{E_\imath} \in \mathbb{R}^{m_1 \times \ldots \times m_n \times \overline{m_1} \times \ldots \times \overline{m_{\bar{n}}}}$ is a tensor defined by:

  $$(E_i \otimes \overline{E_\imath})_{l_1, \ldots, l_n, \overline{l_1}, \ldots, \overline{l_{\bar{n}}}} = E_{i_{l_1, \ldots, l_n}} \cdot \overline{E_{\imath}}_{\overline{l_1}, \ldots, \overline{l_{\bar{n}}}}$$

  So this produces a tensor with '1' in a single entry. The entry with index:

$$(d_1, \dots, d_n, \overline{d_1}, \dots, \overline{d_{\bar{n}}})$$

And '0' in every other entry.

- **Matricization:**

$[|E_i \otimes \overline{E_{\bar{\imath}}}|]_I \in \mathbb{R}^{(\prod_{i \in I \cap [n]} m_i)\cdot(\prod_{i \in I - n \cap [\bar{n}]} \overline{m_i}) \times (\prod_{i \in I^C \cap [n]} m_i)\cdot(\prod_{i \in I^C - n \cap [\bar{n}]} \overline{m_i})}$ is a matrix

Defined such that:

$$([|E_i \otimes \overline{E_{\bar{\imath}}}|]_I)_{u,v} = (E_i \otimes \overline{E_{\bar{\imath}}})_{d_1,\dots,d_n,\overline{d_1},\dots,\overline{d_{\bar{n}}}}$$

for every $d_k \in [m_k], k \in [n]$ and $\overline{d_k} \in [\overline{m_k}], \bar{k} \in [\bar{n}]$

Where:

$$u = 1 + \sum_{t=1}^{|I \cap [n]|} (d_{i_t} - 1) \prod_{t'=t+1}^{|I \cap [n]|} m_{i_{t'}} + \sum_{t=1}^{|I - n \cap [\bar{n}]|} (\overline{d_{\imath_t}} - 1) \prod_{t'=t+1}^{|I - n \cap [\bar{n}]|} \overline{m_{\imath_{t'}}} \prod_{l=1}^{|I \cap n|} m_{i_l}$$

$$v = 1 + \sum_{t=1}^{|I^C \cap [n]|} (d_{i_t} - 1) \prod_{t'=t+1}^{|I^C \cap [n]|} m_{i_{t'}} + \sum_{t=1}^{|I^C - n \cap [\bar{n}]|} (\overline{d_{\imath_t}} - 1) \prod_{t'=t+1}^{|I^C - n \cap [\bar{n}]|} \overline{m_{\imath_{t'}}} \prod_{l=1}^{|I^C \cap n|} m_{i_l}$$

I.e. We have a matrix with a single '1' entry in index (u, v) corresponding to index $(d_1, \dots, d_n, \overline{d_1}, \dots, \overline{d_{\bar{n}}})$ in $E_i \otimes \overline{E_{\bar{\imath}}}$.

Simply putting it, if we denote the matrix in LHS as

$$A \in \mathbb{R}^{(\prod_{i \in I \cap [n]} m_i)\cdot(\prod_{i \in I - n \cap [\bar{n}]} \overline{m_i}) \times (\prod_{i \in I^C \cap [n]} m_i)\cdot(\prod_{i \in I^C - n \cap [\bar{n}]} \overline{m_i})}$$

We have a matrix with a single '1' entry at (u_A, v_A) and all other entries are '0' where

$$u_A = 1 + \sum_{t=1}^{|I \cap [n]|} (d_{i_t} - 1) \prod_{t'=t+1}^{|I \cap [n]|} m_{i_{t'}} + \sum_{t=1}^{|I - n \cap [\bar{n}]|} (\overline{d_{\imath_t}} - 1) \prod_{t'=t+1}^{|I - n \cap [\bar{n}]|} \overline{m_{\imath_{t'}}} \prod_{l=1}^{|I \cap n|} m_{i_l}$$

$$v_A = 1 + \sum_{t=1}^{|I^C \cap [n]|} (d_{i_t} - 1) \prod_{t'=t+1}^{|I^C \cap [n]|} m_{i_{t'}} + \sum_{t=1}^{|I^C - n \cap [\bar{n}]|} (\overline{d_{\imath_t}} - 1) \prod_{t'=t+1}^{|I^C - n \cap [\bar{n}]|} \overline{m_{\imath_{t'}}} \prod_{l=1}^{|I^C \cap n|} m_{i_l}$$

Let's move on to viewing RHS: $[|E_i|]_{I \cap [n]} \odot [|\overline{E_{\bar{\imath}}}|]_{I - n \cap [\bar{n}]}$

- **Matricizations:**
  - We have an order n tensor $E_i \in \mathbb{R}^{m_1 \times \dots \times m_n}$, so its matricization w.r.t $I \cap [n]$ is:

    $$[|E_i|]_{I \cap [n]} \in \mathbb{R}^{(\prod_{i \in I \cap [n]} m_i) \times (\prod_{i \in I^C \cap [n]} m_i)}$$

    Is defined such that:

    $$([|E_i|]_{I \cap [n]})_{u,v} = E_{i_{d_1,\dots,d_n}}$$

    for every $d_k \in [m_k], k \in [n]$.
    Where:

    $$u = 1 + \sum_{t=1}^{|I \cap [n]|} (d_{i_t} - 1) \prod_{t'=t+1}^{|I \cap [n]|} m_{i_{t'}}$$

    $$v = 1 + \sum_{t=1}^{|I^C \cap [n]|} (d_{i_t} - 1) \prod_{t'=t+1}^{|I^C \cap [n]|} m_{i_{t'}}$$

    Simply putting it, this yields a matrix $B \in \mathbb{R}^{(\prod_{i \in I \cap [n]} m_i) \times (\prod_{i \in I^C \cap [n]} m_I)}$ with a single '1' entry at (u_B, v_B) and all other entries are '0'. Where:

$$u_B = 1 + \sum_{t=1}^{|I\cap[n]|} (d_{i_t} - 1) \prod_{t'=t+1}^{|I\cap[n]|} m_{i_{t'}}$$

$$v_B = 1 + \sum_{t=1}^{|I^C\cap[n]|} (d_{i_t} - 1) \prod_{t'=t+1}^{|I^C\cap[n]|} m_{i_{t'}}$$

- And an order $\bar{n}$ tensor $\overline{E_{\bar{\imath}}} \in \mathbb{R}^{\overline{m_1}\times\ldots\times\overline{m_{\bar{n}}}}$, so its matricization w.r.t $I - n \cap [\bar{n}]$ is:

$$[|\overline{E_{\bar{\imath}}}|]_{I-n\cap[\bar{n}]} \in \mathbb{R}^{\left(\prod_{i\in I-n\cap[\bar{n}]}\overline{m_i}\right)\times\left(\prod_{i\in I^C-n\cap[\bar{n}]}\overline{m_i}\right)}$$

Is defined such that:

$$\left([|\overline{E_{\bar{\imath}}}|]_{I-n\cap[\bar{n}]}\right)_{u,v} = \overline{E_{\bar{\imath}_{\overline{d_1}},\ldots,\overline{d_{\bar{n}}}}}$$

for every $d_k \in [\overline{m_k}]$, $k \in [\bar{n}]$.
Where:

$$u = 1 + \sum_{t=1}^{|I-n\cap[\bar{n}]|} (\overline{d_{\iota_t}} - 1) \prod_{t'=t+1}^{|I-n\cap[\bar{n}]|} \overline{m_{\iota_{t'}}}$$

$$v = 1 + \sum_{t=1}^{|I^C-n\cap[\bar{n}]|} (\overline{d_{\iota_t}} - 1) \prod_{t'=t+1}^{|I^C-n\cap[\bar{n}]|} \overline{m_{\iota_{t'}}}$$

Simply putting it, this yields a matrix $C \in \mathbb{R}^{\left(\prod_{i\in I-n\cap[\bar{n}]}\overline{m_i}\right)\times\left(\prod_{i\in I^C-n\cap[\bar{n}]}\overline{m_i}\right)}$ with a single '1' entry at (u_C, v_C) and all other entries are '0'. Where:

$$u_C = 1 + \sum_{t=1}^{|I-n\cap[\bar{n}]|} (\overline{d_{\iota_t}} - 1) \prod_{t'=t+1}^{|I-n\cap[\bar{n}]|} \overline{m_{\iota_{t'}}}$$

$$v_C = 1 + \sum_{t=1}^{|I^C-n\cap[\bar{n}]|} (\overline{d_{\iota_t}} - 1) \prod_{t'=t+1}^{|I^C-n\cap[\bar{n}]|} \overline{m_{\iota_{t'}}}$$

Kronecker Product:

We now have two matrices:
$B \in \mathbb{R}^{\left(\prod_{i\in I\cap[n]} m_i\right)\times\left(\prod_{i\in I^C\cap[n]} m_I\right)}$, $C \in \mathbb{R}^{\left(\prod_{i\in I-n\cap[\bar{n}]}\overline{m_i}\right)\times\left(\prod_{i\in I^C-n\cap[\bar{n}]}\overline{m_i}\right)}$
Each matrix with a single '1' entry and all other entries being '0'.
B's '1' entry: (u_B, v_B), $C's$ '1' entry: (u_C, v_C).

B & C's Kronecker product yields a matrix $D \in \mathbb{R}^{\left(\prod_{i\in I\cap[n]} m_i\right)\cdot\left(\prod_{i\in I-n\cap[\bar{n}]}\overline{m_i}\right)\times\left(\prod_{i\in I^C\cap[n]} m_i\right)\cdot\left(\prod_{i\in I^C-n\cap[\bar{n}]}\overline{m_i}\right)}$ with a single '1' entry at index (u_D, v_D) where:

$$u_D = \left(\prod_{i\in I-n\cap[\bar{n}]} \overline{m_i}\right)\cdot(u_B - 1) + u_C$$

$$v_D = \left(\prod_{i\in I^C-n\cap[\bar{n}]} \overline{m_i}\right)\cdot(v_B - 1) + v_C$$

Bringing it **ALL** together…

- On the LHS we got a matrix A:

$$A \in \mathbb{R}^{\left(\prod_{i\in I\cap[n]} m_i\right)\cdot\left(\prod_{i\in I-n\cap[\bar{n}]}\overline{m_i}\right)\times\left(\prod_{i\in I^C\cap[n]} m_i\right)\cdot\left(\prod_{i\in I^C-n\cap[\bar{n}]}\overline{m_i}\right)}$$

with a single '1' entry at (u_A, v_A) and all other entries are '0' where

$$u_A = 1 + \sum_{t=1}^{|I\cap[n]|} (d_{i_t} - 1) \prod_{t'=t+1}^{|I\cap[n]|} m_{i_{t'}} + \sum_{t=1}^{|I-n\cap[\bar{n}]|} (\overline{d_{\iota_t}} - 1) \prod_{t'=t+1}^{|I-n\cap[\bar{n}]|} \overline{m_{\iota_{t'}}} \prod_{l=1}^{|I\cap n|} m_{i_l}$$

$$v_A = 1 + \sum_{t=1}^{|I^C \cap [n]|} (d_{i_t} - 1) \prod_{t'=t+1}^{|I^C \cap [n]|} m_{i_{t'}} + \sum_{t=1}^{|I^C - n \cap [\bar{n}]|} (\overline{d_{i_t}} - 1) \prod_{t'=t+1}^{|I^C - n \cap [\bar{n}]|} \overline{m_{i_{t'}}} \prod_{l=1}^{|I^C \cap n|} m_{i_l}$$

- On the RHS we got a matrix D:

$$D \in \mathbb{R}^{(\prod_{i \in I \cap [n]} m_i) \cdot (\prod_{i \in I - n \cap [\bar{n}]} \overline{m_i}) \times \left(\prod_{i \in I^C \cap [n]} m_i\right) \cdot \left(\prod_{i \in I^C - n \cap [\bar{n}]} \overline{m_i}\right)}$$

with a single '1' entry at (u_D, v_D) and all other entries are '0' where

$$u_D = \left(\prod_{i \in I - n \cap [\bar{n}]} \overline{m_i} \right) \cdot (u_B - 1) + u_C$$

$$v_D = \left(\prod_{i \in I^C - n \cap [\bar{n}]} \overline{m_i} \right) \cdot (v_B - 1) + v_C$$

And:

$$u_B = 1 + \sum_{t=1}^{|I \cap [n]|} \left(d_{i_t} - 1\right) \prod_{t'=t+1}^{|I \cap [n]|} m_{i_{t'}}$$

$$v_B = 1 + \sum_{t=1}^{|I^C \cap [n]|} \left(d_{i_t} - 1\right) \prod_{t'=t+1}^{|I^C \cap [n]|} m_{i_{t'}}$$

$$u_C = 1 + \sum_{t=1}^{|I - n \cap [\bar{n}]|} \left(\overline{d_{i_t}} - 1\right) \prod_{t'=t+1}^{|I - n \cap [\bar{n}]|} \overline{m_{i_{t'}}}$$

$$v_C = 1 + \sum_{t=1}^{|I^C - n \cap [\bar{n}]|} \left(\overline{d_{i_t}} - 1\right) \prod_{t'=t+1}^{|I^C - n \cap [\bar{n}]|} \overline{m_{i_{t'}}}$$

Plotting u_B, v_B, u_C, and v_C into u_D and v_D yields equality:

$$u_A = u_D \text{ and } v_A = v_D$$

I.e matrix A (=LHS) is equal to matrix D (=RHS)

■

Foundations of Deep Learning – Homework Assignment #2

Adi Album & Tomer Epshtein

Part 3: (2)

Problem:

Modify the expressive efficiency analysis given in class to the case in which the deep network has $\log_4(N)$ hidden layers (Assume N is a power of 4), with pooling windows of size 4 (In class we treated $\log_2(N)$ hidden layers with pooling windows of size 2).

Solution:

Assume N is a power of 4.
Denote by $\overline{H}_{\bar{B}}$, $\bar{B} \in \mathbb{N}$ the hypothesis space corresponding to a deep network with pooling windows of size 4 and with depth $\log_4(N)$.

Similarly to notation in class, we'll denote its learnable weights by:

$$\left\{\left(a^{l,1,\gamma}, a^{l,2,\gamma}, \dots, a^{l,\frac{N}{4^l},\gamma}\right)\right\}_{\gamma=1}^{r_l} \subset (\mathbb{R}^{r_l})^{N/4^l}, l = 0,1,\dots,L$$

Where $L = \log_4(N)$.

Let's now state the expressive efficiency statement for this hypothesis class w.r.t H_B, the hypothesis space corresponding to shallow nets:

1. $\forall B$, $\exists \bar{B} \in \mathcal{O}(B)$ such that $H_B \subseteq \overline{H}_{\bar{B}}$
2. $\exists \bar{h} \in \overline{H}_{\bar{B}}$ for $\bar{B} \in \mathcal{O}(M)$ such that $\bar{h} \notin H_B$ unless $B \in \mathcal{O}(\exp(N))$.

Proof 1:
Nearly identical to the proof in class for deep nets with windows of size 2 and depth $\log_2(N)$.
Let $B \in \mathbb{N}$. Choose $\bar{B} = B$. Let $h \in H_B$. I.e. h is a function realizable by a shallow network of width B.
We will prove it is realizable by a deep network with window size 4, depth $\log_4(N)$ and width $\bar{B} = B$.

Denote by $\{(a_S^{z,1}, a_S^{z,2}, \dots, a_S^{z,N})\}_{z=1}^{B} \subset (\mathbb{R}^M)^N$ and $a_S \in \mathbb{R}^B$ the learnable weights of the shallow net realizing h. (S – Denoting shallow net weights)

We will now define the weights for the deep net so that it realizes h:
With $r_0 = r_1 = \cdots = r_{L-1} = B$:
For $l = 1, \dots, L$: We define the weights for the conv layers $\left\{\left(a^{l,1,\gamma}, a^{l,2,\gamma}, \dots, a^{l,\frac{N}{4^l},\gamma}\right)\right\}_{\gamma=1}^{r_l}$ by:

$$a^{l,1,\gamma} = e^{\gamma}$$

Where $e^{\gamma} \in \mathbb{R}^{r_l}$ is a vector with a single entry "1" at position γ and "0" in all other entries.
Under this definition, the 1x1 conv layers act as "passthru" layers, and the $4 \to 1$ pooling layers *together* constitute a global pooling (multiplying 4 entries at a time).

Altering the weights in hidden layer 0: $\{(a^{0,1,\gamma}, a^{0,2,\gamma}, \dots, a^{0,N,\gamma})\}_{\gamma=1}^{r_0} \subset (\mathbb{R}^{r_0})^N$ and output weights $a^L \in \mathbb{R}^B$ we can realize h using a deep network.

Define:

$$a^{0,k,\gamma} \coloneqq a_S^{\gamma,k}$$

$$a^L \coloneqq a_S$$

Using above weights of layer 0 and output layer L, together with "passthru" conv layers $1, \ldots, L-1$ out deep-net realizes h.

Proof 2:

For this part of the proof we will need to use a decomposition of the network's behavior, similar to the HT decomposition we saw in class for deep nets with windows of size 2 and depth $\log_2(N)$.

In class we saw the equivalence of shallow nets to tensors defined by the CP decomposition, and the equivalence of deep nets (with windows of size 2 and depth $\log_2(N)$) to tensors defined by the HT decomposition:

$$shallow - nets \Leftrightarrow CP\ decomp, \quad deep - nets \Leftrightarrow HT\ decomp$$

In our case, deep nets with windows of size 4 and depth $\log_4(N)$ we obtain the following decomposition which we will call the STAR decomposition denoted by $(*)$:

$$\Phi^{1,j,\gamma} = \sum_{\alpha=1}^{r_0} a_\alpha^{1,j,\gamma} \cdot (a^{0,4j-3,\alpha} \otimes a^{0,4j-2,\alpha} \otimes a^{0,4j-1,\alpha} \otimes a^{0,4j,\alpha})$$

…

$$\Phi^{l,j,\gamma} = \sum_{\alpha=1}^{r_{l-1}} a_\alpha^{l,j,\gamma} \cdot (\Phi^{l-1,4j-3,\alpha} \otimes \Phi^{l-1,4j-2,\alpha} \otimes \Phi^{l-1,4j-1,\alpha} \otimes \Phi^{l-1,4j,\alpha})$$

…

$$\Phi^{L-1,j,\gamma} = \sum_{\alpha=1}^{r_{L-2}} a_\alpha^{L-1,j,\gamma} \cdot (\Phi^{L-2,4j-3,\alpha} \otimes \Phi^{L-2,4j-2,\alpha} \otimes \Phi^{L-2,4j-1,\alpha} \otimes \Phi^{L-2,4j,\alpha})$$

$$A^{(*)} = \sum_{\alpha=1}^{r_{L-1}} a_\alpha^L \cdot (\Phi^{L-1,1,\alpha} \otimes \Phi^{L-1,2,\alpha} \otimes \Phi^{L-1,3,\alpha} \otimes \Phi^{L-1,4,\alpha})$$

As we saw in class, by viewing the matricization of the CP decomposition (↔shallow-nets) we saw that for any choice of learnable parameters, any shallow net is equivalent to a tensor A such that $rank[|A^{CP}|] \leq Z$ as a sum of Z rank 1 matrices.

We'll consider a similar analysis for STAR decomposition (Denoted by $(*)$) corresponding to deep nets with windows of size 4 and depth $\log_4(N)$.

We'll define the weights of the deep net as follows:

- <u>Layer 0:</u>
  Let $r_0 \geq M$ be the width of hidden layer 0. Denote "Layer 0" weights by:
  $$\{(a^{0,1,\gamma}, a^{0,2,\gamma}, \ldots, a^{0,N,\gamma})\}_{\gamma=1}^{r_0} \subset (\mathbb{R}^{r_0})^N$$
  And define:
  $$a^{0,j,\gamma} \coloneqq e^\gamma \in \mathbb{R}^{r_0}$$
  for all $\gamma \in [M]$ where e^γ is "1" at entry γ and "0" elsewhere.
  And:
  $$a^{0,j,\gamma} \coloneqq \bar{0} \in \mathbb{R}^{r_0}$$
  for all $\gamma > M$, where $\bar{0}$ is the all zeros vector.

- Layer 1:
  Let r_1 be the width of hidden layer 1. Denote "Layer 1" weights by:
  $$\left\{\left(a^{1,1,\gamma}, a^{1,2,\gamma}, \dots, a^{1,N/4,\gamma}\right)\right\}_{\gamma=1}^{r_1} \subset (\mathbb{R}^{r_1})^{N/4}$$
  $$a^{1,j,1} = \bar{1} \in \mathbb{R}^{r_1}, a^{1,j,\gamma} = \bar{0} \in \mathbb{R}^{r_1}$$
  for all $\gamma > 1$ where $\bar{1}$ is the all ones vector and $\bar{0}$ is the all zeros vector.

  Here we achieve:
  $[|\Phi^{1,j,\gamma}|]$ is the identity matrix for $\gamma = 1$ and the zero matrix for $\gamma > 1$.

- Layers 2, …, (L-1):
  Let $r_2, \dots, r_{L-1}$ be the widths of layers $2, \dots, (L-1)$. Denote their weights by:
  $$\left\{\left(a^{l,1,\gamma}, a^{l,2,\gamma}, \dots, a^{0,N/4^l,\gamma}\right)\right\}_{\gamma=1}^{r_l} \subset (\mathbb{R}^{r_l})^{\frac{N}{4^l}}, l = 2, \dots, (L-1)$$

  And define:
  $$a^{l,j,\gamma} = \begin{cases} e^1, & \gamma = 1 \\ \bar{0}, & \gamma > 1 \end{cases}$$
  For $l = 2, \dots, (L-1), j = 1,2, \dots, N/4^l, \gamma = 1,2, \dots, r_l$.

- Output weights:
  Define $a^L = e^1$.

Under our above assignments we achieve with the canonical decomposition:

$$\left[|A^{(*)}|\right] = \left[\left|\sum_{\alpha=1}^{r_{L-1}} a_\alpha^L \cdot (\Phi^{L-1,1,\alpha} \otimes \Phi^{L-1,2,\alpha} \otimes \Phi^{L-1,3,\alpha} \otimes \Phi^{L-1,4,\alpha})\right|\right] =$$

{Linearity of matricization}

$$= \sum_{\alpha=1}^{r_{L-1}} a_\alpha^L \cdot [|(\Phi^{L-1,1,\alpha} \otimes \Phi^{L-1,2,\alpha} \otimes \Phi^{L-1,3,\alpha} \otimes \Phi^{L-1,4,\alpha})|] =$$

{"matricization of outer product equals Kronecker product of matricizations" property}

$$= \sum_{\alpha=1}^{r_{L-1}} a_\alpha^L \cdot [|(\Phi^{L-1,1,\alpha}|] \odot [|\Phi^{L-1,2,\alpha}|] \odot [|\Phi^{L-1,3,\alpha}|] \odot [|\Phi^{L-1,4,\alpha})|] =$$

… Keep going down the network with (*) decomposition … Calculations identical to those held in class …

$$\dots = Id_{M^2} \odot Id_{M^2} \odot \dots \odot Id_{M^2} = Id_{(M^2)^{\frac{N}{4}}}$$

Id_{M^2} appears $N/4$ times, where Id_n is a $n \times n$ identity matrix.

In particular we showed a realization of a tensor by $(*)$ decomposition which corresponds to a deep net with windows of size 4 and depth $\log_4(N)$. The matricization via the canonical matricization produces a matrix with $rank\left(\left[|A^{(*)}|\right]\right) = (M^2)^{\frac{N}{4}} = M^{\frac{N}{2}}$.

We'll summarize: We saw in class that any realization of a shallow net produces a tensor denoted by A such that it holds the property: $rank([|A|]) \leq Z$ where Z is shallow net's width.

We now saw an implementation of a deep net with windows of size 4 and depth $\log_4(N)$ such that if we denote it's corresponding tensor by $A^{(*)}$ we have $rank\left(\left[\left|A^{(*)}\right|\right]\right) = M^{\frac{N}{2}}$.

For this tensor to be realized by a shallow net we must have $Z \geq M^{\frac{N}{2}}$ and in particular we must have a network with exponential width (exponential in N).

■

Foundations of Deep Learning – Homework Assignment #2

Adi Album & Tomer Epshtein

Part 3: (3)

Problem:

Consider the following "quadrant" partition of the N input elements:

$$I_{quad} = \left\{1,2,\dots,\frac{1}{4}N,\frac{1}{2}N+1,\frac{1}{2}N+2,\dots\frac{3}{4}N\right\}$$

$$I_{quad}^{C} = \left\{\frac{1}{4}N+1,\frac{1}{4}N+2,\dots,\frac{1}{2}N,\frac{3}{4}N+1,\frac{3}{4}N+2,\dots,N\right\}$$

Prove that under this partition, the separation rank of a function realized by a deep network (with $L = \log_2(N)$ hidden layers) is no greater than $r_{L-1} \cdot r_{L-2}^2$ where r_l stands for the width of layer l.

Solution:

First, a visualization of the partition:

1		$\frac{1}{4}N+1$		$\frac{1}{2}N+1$		$\frac{3}{4}N+1$	
2		$\frac{1}{4}N+2$		$\frac{1}{2}N+2$		$\frac{3}{4}N+2$	
3		⋮		⋮		⋮	
⋮							
	$\frac{1}{4}N$				$\frac{3}{4}N$		N

Let $\bar{h} \in \overline{H_{\bar{B}}}$ be a deep net. Denote by A^{HT} its corresponding tensor given by the HT decomposition:

$$\Phi^{1,j,\gamma} = \sum_{\alpha=1}^{r_0} a_\alpha^{1,j,\gamma} \cdot (a^{0,2j-1,\alpha} \otimes a^{0,2j,\alpha})$$

…

$$\Phi^{l,j,\gamma} = \sum_{\alpha=1}^{r_{l-1}} a_\alpha^{l,j,\gamma} \cdot (\Phi^{l-1,2j-1,\alpha} \otimes \Phi^{l-1,2j,\alpha})$$

…

$$\Phi^{L-1,j,\gamma} = \sum_{\alpha=1}^{r_{L-2}} a_\alpha^{L-1,j,\gamma} \cdot (\Phi^{L-2,2j-1,\alpha} \otimes \Phi^{L-2,2j,\alpha})$$

$$A^{HT} = \sum_{\alpha=1}^{r_{L-1}} a_\alpha^{L} \cdot (\Phi^{L-1,1,\alpha} \otimes \Phi^{L-1,2,\alpha})$$

Let's view A^{HT}'s matricization w.r.t $I \coloneqq I_{quad}$:

$$[|A^{HT}|]_I = \left[\left|\sum_{\alpha=1}^{r_{L-1}} a_\alpha^L \cdot (\Phi^{L-1,1,\alpha} \otimes \Phi^{L-1,2,\alpha})\right|\right]_I =$$

{Linearity of matricization and "matricization of outer product equals Kronecker product of matricizations" property}

$$= \sum_{\alpha=1}^{r_{L-1}} a_\alpha^L \cdot \left([|\Phi^{L-1,1,\alpha}|]_{I \cap [N/2]} \odot [|\Phi^{L-1,2,\alpha}|]_{I-\frac{N}{2} \cap [N/2]}\right)$$

Let's take another step back through the HT decomposition:
$\forall I' \subset [N/2], j = 1,2, \gamma \in [r_{L-1}]$:

$$\left[\left|\Phi^{L-1,j,\gamma}\right|\right]_{I'} = \left[\left|\sum_{\beta=1}^{r_{L-2}} a_\beta^{L-1,j,\gamma} \cdot \left(\Phi^{L-2,2j-1,\beta} \otimes \Phi^{L-2,2j,\beta}\right)\right|\right]_{I'} =$$

{Linearity of matricization and "matricization of outer product equals Kronecker product of matricizations" property}

$$\sum_{\beta=1}^{r_{L-2}} a_\beta^{L-1,j,\gamma} \cdot \left(\left[\left|\Phi^{L-2,2j-1,\beta}\right|\right]_{I' \cap [N/4]} \odot \left[\left|\Phi^{L-2,2j,\beta}\right|\right]_{I'-\frac{N}{4} \cap [N/4]}\right)$$

Putting it all together:

$$[|A^{HT}|]_{I_{quad}} = \sum_{\alpha=1}^{r_{L-1}} a_\alpha^L \cdot \left([|\Phi^{L-1,1,\alpha}|]_{I_{quad} \cap [N/2]} \odot [|\Phi^{L-1,2,\alpha}|]_{I_{quad}-\frac{N}{2} \cap [N/2]}\right) =$$

$$\sum_{\alpha=1}^{r_{L-1}} a_\alpha^L \cdot \left(\begin{array}{l} \sum_{\beta=1}^{r_{L-2}} a_\beta^{L-1,1,\alpha} \left(\left[\left|\Phi^{L-2,1,\beta}\right|\right]_{I_{quad} \cap [N/2] \cap [N/4]} \odot \left[\left|\Phi^{L-2,2,\beta}\right|\right]_{((I_{quad} \cap [N/2])-\frac{N}{4}) \cap [N/4]}\right) \odot \\ \left(\sum_{\beta=1}^{r_{L-2}} a_\beta^{L-1,2,\alpha} \left(\left[\left|\Phi^{L-2,3,\beta}\right|\right]_{\left(I_{quad}-\frac{N}{2} \cap [N/2]\right) \cap [N/4]} \odot \left[\left|\Phi^{L-2,4,\beta}\right|\right]_{\left(\left(I_{quad}-\frac{N}{2} \cap [N/2]\right)-\frac{N}{4}\right) \cap [N/4]}\right) \end{array} \right)$$

Let's understand that matricizations' indices:

1. $I_{quad} \cap [N/2] \cap [N/4] = [N/4]$
2. $\left(\left(I_{quad} \cap [N/2]\right) - \frac{N}{4}\right) \cap [N/4] = \emptyset$
3. $\left(I_{quad} - \frac{N}{2} \cap [N/2]\right) \cap [N/4] = [N/4]$
4. $\left(\left(I_{quad} - \frac{N}{2} \cap [N/2]\right) - \frac{N}{4}\right) \cap [N/4] = \emptyset$

So

$$= \sum_{\alpha=1}^{r_{L-1}} a_\alpha^L \cdot \left(\begin{array}{l} \left(\sum_{\beta=1}^{r_{L-2}} a_\beta^{L-1,1,\alpha} \left(\left[\left|\Phi^{L-2,1,\beta}\right|\right]_{\left[\frac{N}{4}\right]} \odot \left[\left|\Phi^{L-2,2,\beta}\right|\right]_{\emptyset}\right)\right) \odot \\ \left(\sum_{\beta=1}^{r_{L-2}} a_\beta^{L-1,2,\alpha} \left(\left[\left|\Phi^{L-2,3,\beta}\right|\right]_{\left[\frac{N}{4}\right]} \odot \left[\left|\Phi^{L-2,4,\beta}\right|\right]_{\emptyset}\right)\right) \end{array} \right) =$$

$$\sum_{\alpha=1}^{r_{L-1}} a_\alpha^L \cdot \begin{pmatrix} \sum_{\beta=1}^{r_{L-2}} a_\beta^{L-1,1,\alpha} (vec(\Phi^{L-2,1,\beta}) \cdot vec\left(\Phi^{L-2,2,\beta}\right)^T) \odot \\ (\sum_{\beta=1}^{r_{L-2}} a_\beta^{L-1,2,\alpha} (vec(\Phi^{L-2,3,\beta}) \cdot vec\left(\Phi^{L-2,4,\beta}\right)^T) \end{pmatrix}$$

- $vec(u) \cdot vec(v)^T$ produces a rank 1 matrix
- A sum of r_{L-2} rank 1 matrix produces a $\leq r_{L-2}$ ranked matrix

Denote:
$\forall \alpha \in [r_{L-1}]$:

$$U_\alpha \coloneqq \sum_{\beta=1}^{r_{L-2}} a_\beta^{L-1,1,\alpha} (vec(\Phi^{L-2,1,\beta}) \cdot vec\left(\Phi^{L-2,2,\beta}\right)^T)$$

$$V_\alpha \coloneqq (\sum_{\beta=1}^{r_{L-2}} a_\beta^{L-1,2,\alpha} (vec(\Phi^{L-2,3,\beta}) \cdot vec\left(\Phi^{L-2,4,\beta}\right)^T)$$

So we have $rank(U_\alpha) \leq r_{L-2}$, $rank(V_\alpha) \leq r_{L-2}$ for all $\alpha \in [r_{L-1}]$.

We have:

$$[|A^{HT}|]_{I_{quad}} = \sum_{\alpha=1}^{r_{L-1}} a_\alpha^L \cdot (U_\alpha \odot V_\alpha)$$

We saw in class that $rank(A \odot B) = rank(A) \cdot rank(B)$ for any two matrices A & B.
So $\forall \alpha \in [r_{L-1}]$: $rank(U_\alpha \odot V_\alpha) = rank(U_\alpha) \cdot rank(V_\alpha) \leq r_{L-2}^2$

Therefore:

$$rank\left([|A^{HT}|]_{I_{quad}}\right) \leq r_{L-1} \cdot r_{L-2}^2$$

As the sum of r_{L-1} matrices with rank $\leq r_{L-2}^2$.

■