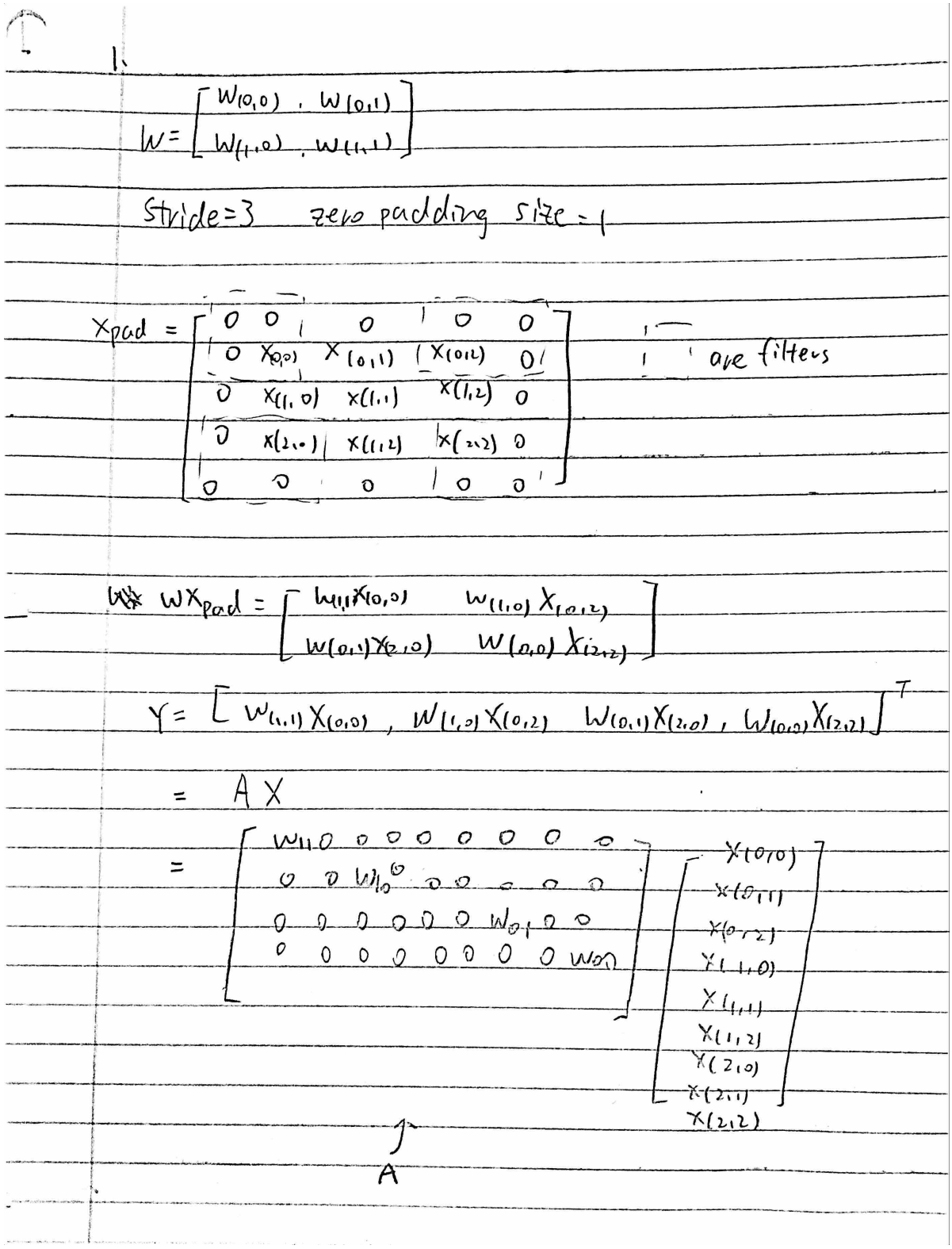
CS7643: Deep Learning
Fall 2020
HW3 Solutions

Tianxue Hu

October 7, 2020

Problem 1


1.
$W = \begin{bmatrix} W_{(0,0)} & W_{(0,1)} \\ W_{(1,0)} & W_{(1,1)} \end{bmatrix}$
Stride=3 zero padding size=1
$X_{pad} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & X_{(0,0)} & X_{(0,1)} & X_{(0,2)} & 0 \\ 0 & X_{(1,0)} & X_{(1,1)} & X_{(1,2)} & 0 \\ 0 & X_{(2,0)} & X_{(2,1)} & X_{(2,2)} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$
are filters
$WX_{pad} = \begin{bmatrix} W_{(1,1)}X_{(0,0)} & W_{(1,0)}X_{(0,2)} \\ W_{(0,1)}X_{(2,0)} & W_{(0,0)}X_{(2,2)} \end{bmatrix}$
$Y = [W_{(1,1)}X_{(0,0)}, W_{(1,0)}X_{(0,2)}, W_{(0,1)}X_{(2,0)}, W_{(0,0)}X_{(2,2)}]^T$
$= AX$
$= \begin{bmatrix} W_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & W_{10} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & W_{01} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & W_{00} \end{bmatrix} \begin{bmatrix} X_{(0,0)} \\ X_{(0,1)} \\ X_{(0,2)} \\ X_{(1,0)} \\ X_{(1,1)} \\ X_{(1,2)} \\ X_{(2,0)} \\ X_{(2,1)} \\ X_{(2,2)} \end{bmatrix}$
A

Problem 2

2.

$$W = \begin{bmatrix} w_{(0,0)} & w_{(0,1)} \\ w_{(1,0)} & w_{(1,1)} \end{bmatrix}$$

Stride=2 no padding

Affine transformation

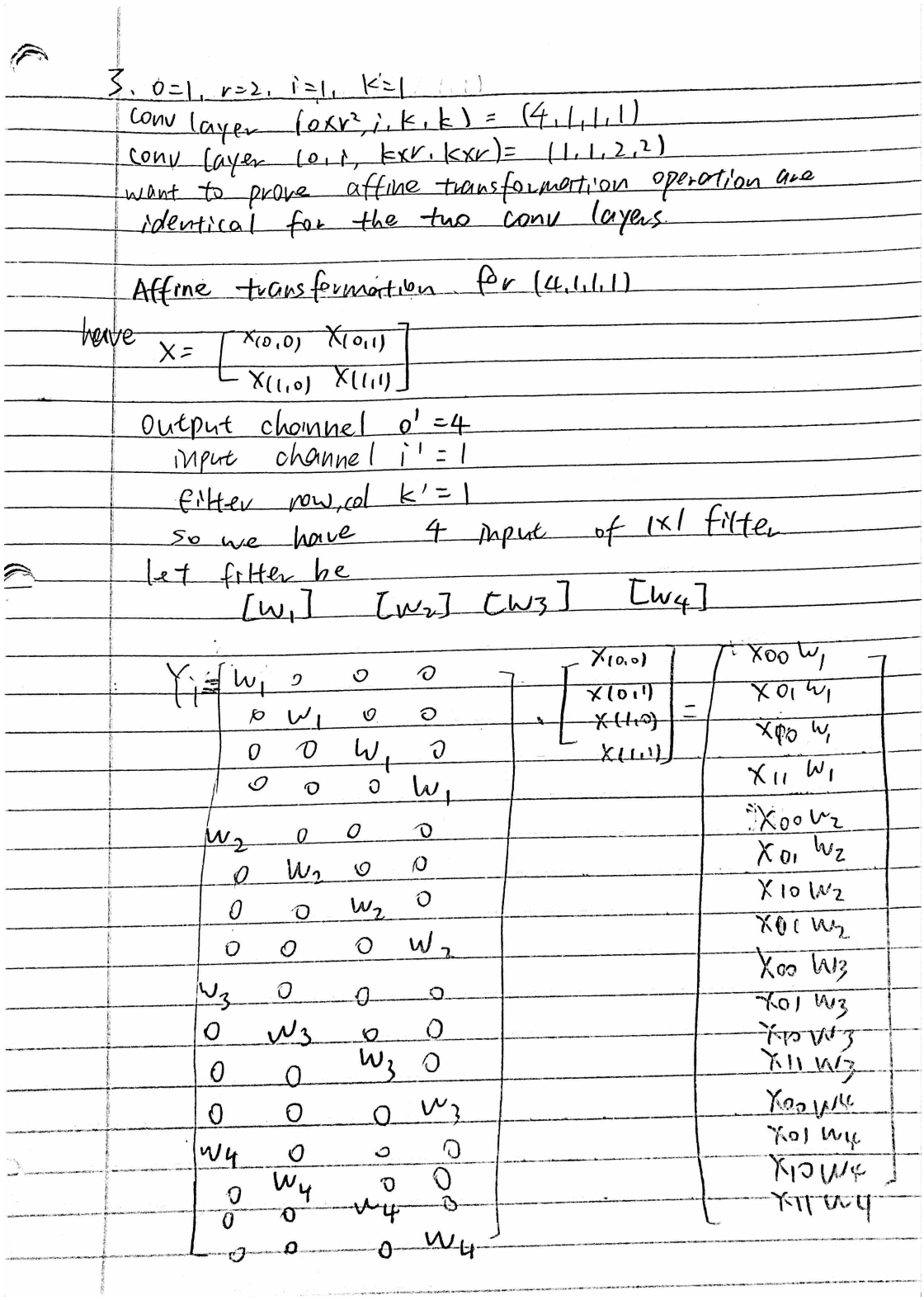
$$\begin{matrix} x_{(0,0)} \\ x_{(1,0)} \end{matrix} \begin{bmatrix} w_{(0,0)} & w_{(0,1)} & w_{(0,0)} & w_{(0,1)} \\ w_{(1,0)} & w_{(1,1)} & w_{(1,0)} & w_{(1,1)} \\ w_{(0,0)} & w_{(0,1)} & w_{(0,0)} & w_{(0,1)} \\ w_{(1,0)} & w_{(1,1)} & w_{(1,0)} & w_{(1,1)} \end{bmatrix} \begin{matrix} x_{(0,1)} \\ x_{(1,1)} \end{matrix}$$

$$Y = [x_{(0,0)} w_{(0,0)},\ x_{(0,1)} w_{(0,1)},\ x_{(0,1)} w_{(0,0)},\ x_{(0,1)} w_{(0,1)},$$
$$x_{(0,0)} w_{(1,0)},\ x_{(0,0)} w_{(1,1)},\ x_{(0,1)} w_{(1,0)},\ x_{(0,1)} w_{(1,1)},$$
$$x_{(1,0)} w_{(0,0)},\ x_{(1,0)} w_{(0,1)},\ x_{(1,1)} w_{(0,0)},\ x_{(1,1)} w_{(0,1)},$$
$$x_{(1,0)} w_{(1,0)},\ x_{(1,0)} w_{(1,1)},\ x_{(1,1)} w_{(1,0)},\ x_{(1,1)} w_{(1,1)}]^T$$

$$= A \cdot X$$

$$= \begin{bmatrix} w_{(0,0)} & 0 & 0 & 0 \\ w_{(0,1)} & 0 & 0 & 0 \\ 0 & w_{(0,0)} & 0 & 0 \\ 0 & w_{(0,1)} & 0 & 0 \\ w_{(1,0)} & 0 & 0 & 0 \\ w_{(1,1)} & 0 & 0 & 0 \\ 0 & w_{(1,0)} & 0 & 0 \\ 0 & w_{(1,1)} & 0 & 0 \\ 0 & 0 & w_{(0,0)} & 0 \\ 0 & 0 & w_{(0,1)} & 0 \\ 0 & 0 & 0 & w_{(0,0)} \\ 0 & 0 & 0 & w_{(0,1)} \\ 0 & 0 & w_{(1,0)} & 0 \\ 0 & 0 & w_{(1,1)} & 0 \\ 0 & 0 & 0 & w_{(1,0)} \\ 0 & 0 & 0 & w_{(1,1)} \end{bmatrix} \cdot \begin{bmatrix} x_{(0,0)} \\ x_{(0,1)} \\ x_{(1,0)} \\ x_{(1,1)} \end{bmatrix}$$

Problem 3


3. o=1, r=2, i=1, k=1
conv layer (o×r², i, k, k) = (4, 1, 1, 1)
conv layer (o, i, k×r, k×r) = (1, 1, 2, 2)
want to prove affine transformation operation are identical for the two conv layers
Affine transformation for (4,1,1,1)
have X = [X(0,0) X(0,1); X(1,0) X(1,1)]
Output channel o' = 4
input channel i' = 1
filter row, col k' = 1
so we have 4 input of 1×1 filter
let filter be
[W1] [W2] [W3] [W4]

Affine transformation for (1, 1, 2, 2)

output channel $o' = 1$

input channel $i' = 1$

filter row, col $k' = 2$

which gives 2x2 filter matrix

$$\begin{bmatrix} w_1 & w_2 \\ w_3 & w_4 \end{bmatrix}$$

which is the same as problem 2. , given X

so $Y_2 =$

$$\begin{bmatrix}
w_1 & 0 & 0 & 0 \\
w_2 & 0 & 0 & 0 \\
0 & w_1 & 0 & 0 \\
0 & w_2 & 0 & 0 \\
w_3 & 0 & 0 & 0 \\
w_4 & 0 & 0 & 0 \\
0 & w_3 & 0 & 0 \\
0 & w_4 & 0 & 0 \\
0 & 0 & w_1 & 0 \\
0 & 0 & w_2 & 0 \\
0 & 0 & 0 & w_1 \\
0 & 0 & 0 & w_2 \\
0 & 0 & w_3 & 0 \\
0 & 0 & w_4 & 0 \\
0 & 0 & 0 & w_3 \\
0 & 0 & 0 & w_4
\end{bmatrix}
\cdot
\begin{bmatrix} x_{00} \\ x_{01} \\ x_{10} \\ x_{11} \end{bmatrix}
=
\begin{bmatrix}
x_{00} w_1 \\
x_{00} w_2 \\
x_{01} w_1 \\
x_{01} w_2 \\
x_{00} w_3 \\
x_{00} w_4 \\
x_{01} w_3 \\
x_{01} w_4 \\
x_{10} w_1 \\
x_{10} w_2 \\
x_{11} w_1 \\
x_{11} w_2 \\
x_{10} w_3 \\
x_{10} w_4 \\
x_{11} w_3 \\
x_{11} w_4
\end{bmatrix}$$

So $Y_1 = Y_2$. proved that the affine transformation operation are identical

Problem 4

4.

$$f(x)=\begin{cases}1 & \text{if } w^Tx+b\geq 0\\ 0 & \text{if } w^Tx+b<0\end{cases}$$

AND:

x_1	x_2	$f_{AND}(x)$
0	0	0
0	1	0
1	0	0
1	1	1

one possible solution:

$$w_{AND}=\begin{bmatrix}0.5\\0.5\end{bmatrix} \qquad b_{AND}=-1.6$$

prove table

$f(x)=w^Tx+b = [0.5\ \ 0.5]\begin{bmatrix}0\\0\end{bmatrix}-1 = -1<0 \Rightarrow 0$

$= [0.5\ \ 0.5]\begin{bmatrix}0\\1\end{bmatrix}-1 = -0.5<0 \Rightarrow 0$

$= [0.5\ \ 0.5]\begin{bmatrix}1\\0\end{bmatrix}-1 = -0.5<0 \Rightarrow 0$

$= [0.5\ \ 0.5]\begin{bmatrix}1\\1\end{bmatrix}-1 = 0=0 \Rightarrow 1$

OR : x_1 x_2 $f_{OR}(x)$

0	0	0
0	1	1
1	0	1
1	1	1

one possible soln:

$W_{OR} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$ $b = -0.1$

Prove table

$f(x) = W^T x + b = [0.5 \quad 0.5] \begin{bmatrix} 0 \\ 0 \end{bmatrix} - 0.1 = -0.1 < 0 \Rightarrow 0$

$[0.5 \quad 0.5] \begin{bmatrix} 0 \\ 1 \end{bmatrix} - 0.1 = 0.4 > 0 \Rightarrow 1$

$[0.5 \quad 0.5] \begin{bmatrix} 1 \\ 0 \end{bmatrix} - 0.1 = 0.4 > 0 \Rightarrow 1$

$[0.5 \quad 0.5] \begin{bmatrix} 1 \\ 1 \end{bmatrix} - 0.1 = 0.9 > 0 \Rightarrow 1.$

Problem 5

5.

x_1	x_2	$f_{XOR}(x)$
0	0	0
0	1	1
1	0	1
1	1	0

From table

$f_{XOR}\left(\begin{bmatrix}0\\1\end{bmatrix}\right) > f_{XOR}\left(\begin{bmatrix}0\\0\end{bmatrix}\right)$

$w^T\begin{bmatrix}0\\1\end{bmatrix}+b > w^T\begin{bmatrix}0\\0\end{bmatrix}+b$

$w_2+b > b$

$w_2 > 0$ ①

$f_{XOR}\left(\begin{bmatrix}0\\1\end{bmatrix}\right) > f_{XOR}\left(\begin{bmatrix}1\\1\end{bmatrix}\right)$

$w^T\begin{bmatrix}0\\1\end{bmatrix}+b > w^T\begin{bmatrix}1\\1\end{bmatrix}+b$

$w_2+b > w_1+w_2+b$

$w_1 < 0$ ②

$f_{XOR}\left(\begin{bmatrix}1\\0\end{bmatrix}\right) > f_{XOR}\left(\begin{bmatrix}1\\1\end{bmatrix}\right)$

$w^T\begin{bmatrix}1\\0\end{bmatrix}+b > w^T\begin{bmatrix}1\\1\end{bmatrix}+b$

$w_1+b > w_1+w_2+b$

$w_2 < 0$ ③

Then ① and ③ are conflict.
Thus XOR cannot be represented using linear model

②

Problem 6

6. $h(x) = W^{(3)} \max\{0, W^{(2)} \max\{0, W^{(1)}x + b^{(1)}\} + b^{(2)}\} + b^{(3)}$

$W^{(1)} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$ $\quad b^{(1)} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

$W^{(2)} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ $\quad b^{(2)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$W^{(3)} = \begin{bmatrix} 1 & 1 \end{bmatrix}$ $\quad b^{(3)} = [1]$

$x = 1$

$h(x) = [1\ 1] \max\{0, \begin{bmatrix} 1 \\ 1 \end{bmatrix} \max\{0, \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}[1] + \begin{bmatrix} 0 \\ 1 \end{bmatrix}\} + \begin{bmatrix} 0 \\ 0 \end{bmatrix}\} + [1]$

$\downarrow$

$\begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix}$

$\downarrow$

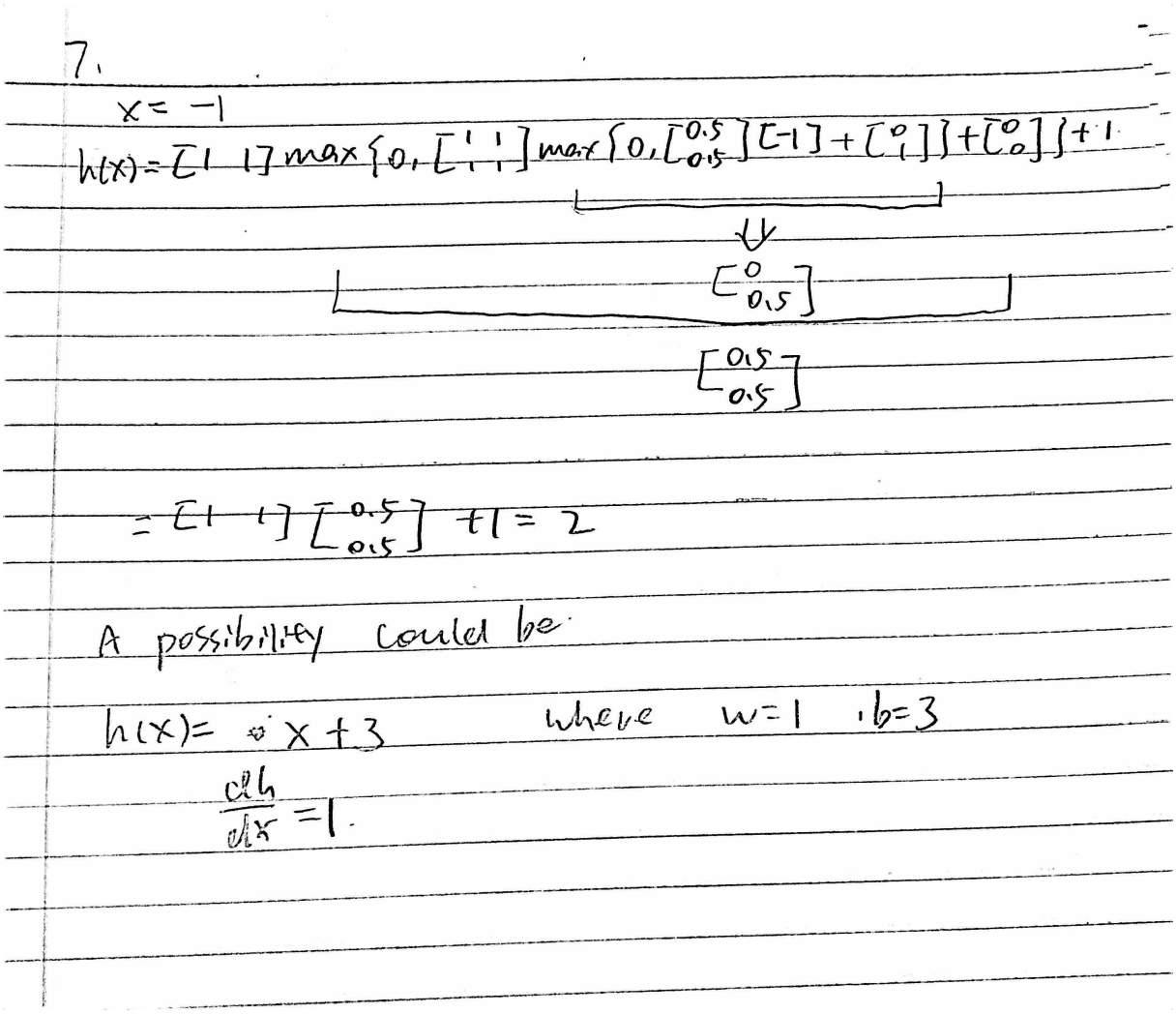
$\begin{bmatrix} 2 \\ 2 \end{bmatrix}$

$= [1\ 1]\begin{bmatrix} 2 \\ 2 \end{bmatrix} + 1 = 5$

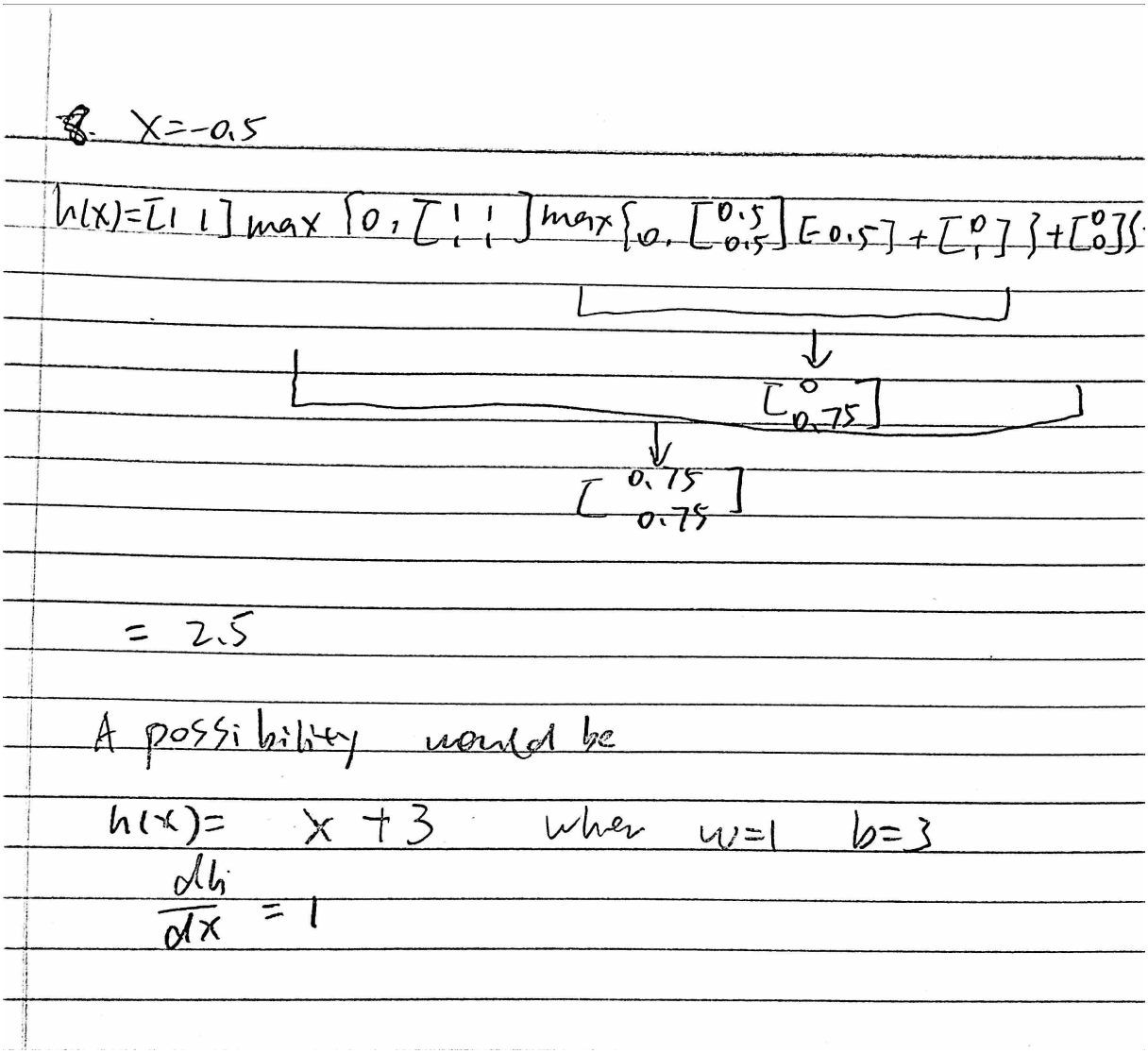
$h(x) = Wx + b = 3x + 2$ where $w = 3$ $b = 2$

$\frac{dh}{dx} = 3$ $\quad h(x) = 3x + 2$.

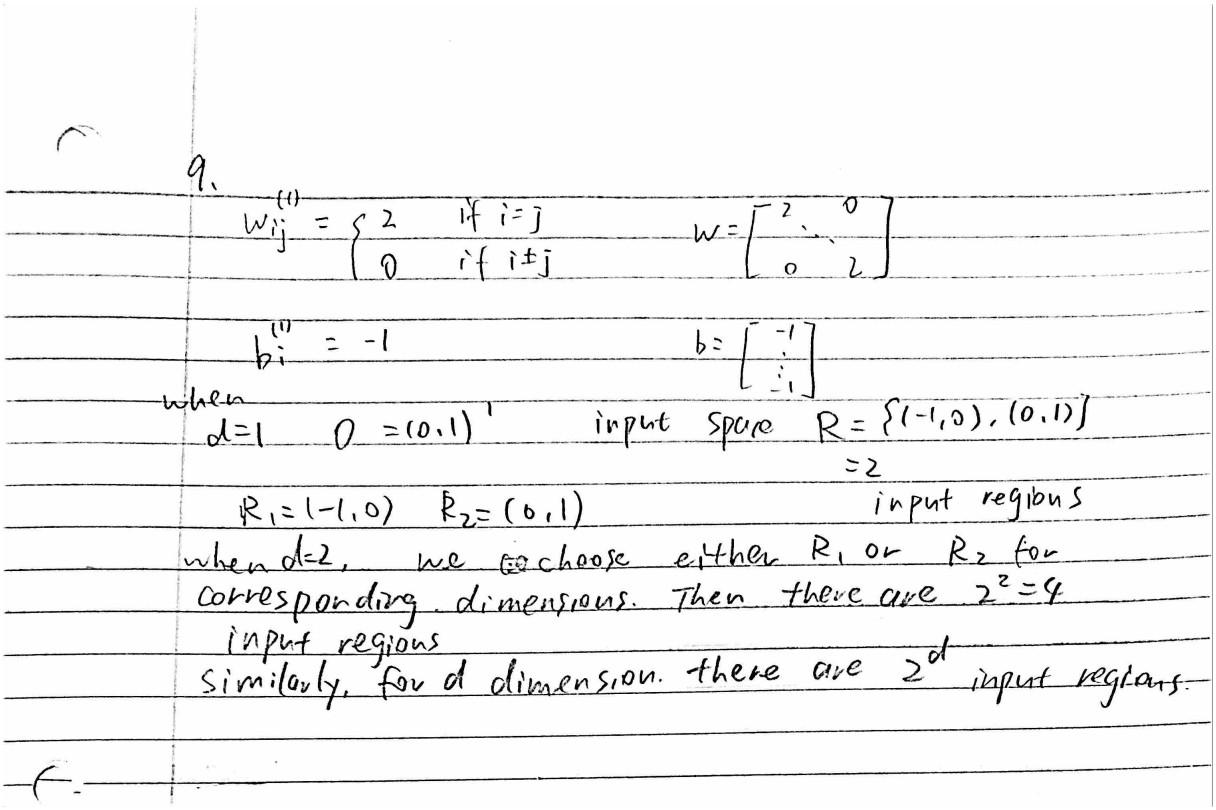
Problem 7


7.
x = -1
h(x) = [1 1] max{0, [1 1] max{0, [0.5 0.5] [-1] + [0 1]} + [0 0]} + 1
[0 0.5]
[0.5 0.5]
= [1 1] [0.5 0.5] + 1 = 2
A possibility could be
h(x) = 1 x + 3 where w = 1 , b = 3
dh/dx = 1.

Problem 8


8. X=-0.5
h(x)=[1 1] max [0, [1 1] max [0, [0.5 0.5] [-0.5] + [0 1]] + [0 0]]
[0 0.75]
[0.75 0.75]
= 2.5
A possibility would be
h(x)= x + 3 when w=1 b=3
dh/dx = 1

Problem 9



9.

$$W_{ij}^{(1)} = \begin{cases} 2 & \text{if } i=j \\ 0 & \text{if } i \neq j \end{cases} \qquad W = \begin{bmatrix} 2 & & 0 \\ & \ddots & \\ 0 & & 2 \end{bmatrix}$$

$$b_i^{(1)} = -1 \qquad b = \begin{bmatrix} -1 \\ \vdots \\ -1 \end{bmatrix}$$

when $d=1$ $\quad 0 = (0,1)^T$ input space $R = \{(-1,0), (0,1)\}$

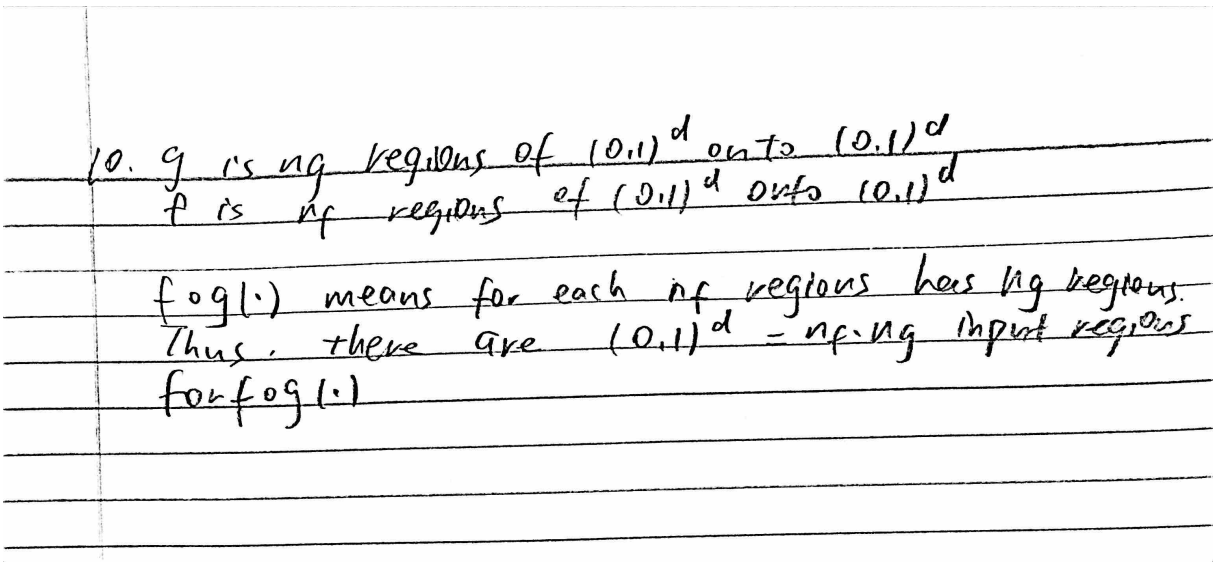
$= 2$ input regions

$R_1 = (-1,0)$ $\quad R_2 = (0,1)$

when $d=2$, we choose either R_1 or R_2 for corresponding dimensions. Then there are $2^2 = 4$ input regions

similarly, for d dimension. there are 2^d input regions.

Problem 10


10. g is n_g regions of $(0,1)^d$ onto $(0,1)^d$
f is n_f regions of $(0,1)^d$ onto $(0,1)^d$
f∘g(·) means for each n_f regions has n_g regions.
Thus, there are $(0,1)^d$ - $n_f \cdot n_g$ input regions
for f∘g(·)

Problem 11

11.

$h_1(x) = |W_1 x + b_1|$

$h_2(x) = |W_2 h_1 + b_1| = |W_2(W_1 x + b_1) + b_2|$

$h_3(x) = |W_3(W_2(W_1 x + b_1) + b_2) + b_3|$

$\vdots$

$f(x) = h_L(x) = |W_L h_{L-1} + b_L|$

where $x \in (0,1)^d$

we proved in 9 that each layer has 2^d input regions for d dimensions.

we prove by induction

Base case: for h_1 there are 2^d regions $2^d = 2^{1 \times d} = 2^{2d}$ when $L=1$.

Inductive case: assume for h_k there're $2^{k \cdot d}$ input regions, we want to prove for h_{k+1} there're $2^{(k+1) \cdot d}$ input regions.

for L_{k+1}: $h_{k+1} = h_{k+1}(h_k(x))$

which has $2^d \cdot 2^{k \cdot d}$ input regions

$= 2^{d + kd}$

$= 2^{(k+1)d}$

inductive case proved.

proved.

Problem 12

12.

The optimization problem by least square method is

$$w = \underset{w \in \mathbb{R}^d}{\arg\min} \sum_{i=1}^{n} (y_i - w^T x_i)^2$$

to get optimized soln.

we take derivative w.r.t. to w and set to 0

$$\frac{d}{dw} \Sigma (y_i - w x_i)^2$$

$$= 2 \sum_i - x_i (y_i - w x_i) = 0$$

$$\Rightarrow 2 \sum_i x_i (y_i - w x_i) = 0$$

$$2 \sum_i x_i y_i - 2 \sum w x_i x_i = 0$$

$$\sum_i x_i y_i = \sum_i w x_i^2$$

$$w_{gd} = \frac{\sum_i x_i y_i}{\sum_i x_i^2}$$

as $Xw = y$

$$w_{gd} = \frac{\sum_i \frac{y_i}{w_i} y_i}{\sum x_i^2} = \frac{1}{w_i} \frac{\sum y_i^2}{\sum x_i^2}$$

$$= \frac{1}{w_i} \sum_i w_i^2$$

$$= \underset{w \in \mathbb{R}^d}{\arg\min} \|w\|_2^2$$

Problem 13

As $||x||_2^2 = \sum_i x_i$, (25) is calculating the optimized gradient descent by finding at which w, the sum of w_i has the least value.

Problem 14

(23) tries to find the optimiazed w that have the least sum of differences between actual y and prediction $y_hat = w^T x$. The solution is not necessarily unique since $f(w)$ may falls into an local minimum and the result depends on w_0.

Problem 15

Approximation error is the minimum generalization error achievable by a predictor in the hypothesis class Estimation error is minimize training error being the only estimate of predictor minimizing the generalization error.

Problem 16

Briefly summarize the key contributions, strengths and weaknesses of this paper.
Key contributions: This paper shows that double descent is a robust phenomenon that occurs in a variety of tasks, architectures, and optimization methods. They defined the effective model complexity (EMC) of a training procedure as the maximum number of samples on which it can achieve close to zero training error. And increasing training time will increase the EMC. The phenomenon indicates that training a deep network on a larger train set performs worse.
Strength: The paper investigate the double decent effect in different senarios: with/without data augmentation, different optimization methods, different number of epochs, different noise settings, and trained on different kinds of networks. Their experiments making the conclusion reputable and universal.
Weakness: The paper proposed several EMC regimes: under-paremeterized regime, over-parameterized regime, and critically parameterized regime. They indicates that in the critically parameterized regime the double descent phenomenon causes poor performance when there are more training data. However, they failed to find out how to define the critical interval if others want to apply it. The paper says "The width of the critical interval depends on both the distribution and the training procedure in ways we do not yet completely understand".

Problem 17

What is your personal takeaway from this paper? This could be expressed either in terms of relating the approaches adopted in this paper to your traditional understanding of learning parameterized models, or potential future directions of research in the area which the authors havenât addressed, or anything else that struck you as being noteworthy.

Despite the conventional wisdom of over-fitting and under-fitting usually happens in statistical modeling, deep learning researchers believe the larger the model the better the performance. This paper discovers the double descent phenomenon that the performance of deep learning models also has a "negative" relationship to the number of training data and model size (when EMC is in a critical interval). As mentioned in the weakness part, the researchers still cannot understand how to find the critical interval, I hope it will be defined in the future for effective model training.