CS7643: Deep Learning
Fall 2020
HW3 Solutions

Ruize Yang

October 7, 2020

Problem 1

$$Y = AX = \begin{bmatrix} w_{1,1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & w_{1,0} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & w_{0,1} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & w_{0,0} \end{bmatrix} \begin{bmatrix} x_{0,0} \\ x_{0,1} \\ \vdots \\ x_{2,2} \end{bmatrix} = \begin{bmatrix} w_{1,1}x_{0,0} \\ w_{1,0}x_{0,2} \\ w_{0,1}x_{2,0} \\ w_{0,0}x_{2,2} \end{bmatrix}$$

Problem 2

$$Y = AX = \begin{bmatrix} w_{0,0} & 0 & 0 & 0 \\ w_{0,1} & 0 & 0 & 0 \\ 0 & w_{0,0} & 0 & 0 \\ 0 & w_{0,1} & 0 & 0 \\ w_{1,0} & 0 & 0 & 0 \\ w_{1,1} & 0 & 0 & 0 \\ 0 & w_{1,0} & 0 & 0 \\ 0 & w_{1,1} & 0 & 0 \\ 0 & 0 & w_{0,0} & 0 \\ 0 & 0 & w_{0,1} & 0 \\ 0 & 0 & 0 & w_{0,0} \\ 0 & 0 & 0 & w_{0,1} \\ 0 & 0 & w_{1,0} & 0 \\ 0 & 0 & w_{1,1} & 0 \\ 0 & 0 & 0 & w_{1,0} \\ 0 & 0 & 0 & w_{1,1} \end{bmatrix} \begin{bmatrix} x_{0,0} \\ x_{0,1} \\ x_{1,0} \\ x_{1,1} \end{bmatrix} = \begin{bmatrix} x_{00}w_{00} \\ x_{00}w_{01} \\ x_{01}w_{00} \\ x_{01}w_{01} \\ x_{00}w_{10} \\ x_{01}w_{11} \\ x_{01}w_{10} \\ x_{01}w_{11} \\ x_{10}w_{00} \\ x_{10}w_{01} \\ x_{11}w_{00} \\ x_{11}w_{01} \\ x_{10}w_{10} \\ x_{10}w_{11} \\ x_{11}w_{10} \\ x_{11}w_{11} \end{bmatrix}$$

Problem 3

$(4,1,1,1)$ kernel:

The filters in four channels are: $\begin{bmatrix} w_1 \end{bmatrix}, \begin{bmatrix} w_2 \end{bmatrix}, \begin{bmatrix} w_3 \end{bmatrix}, \begin{bmatrix} w_4 \end{bmatrix}$, $X = \begin{bmatrix} x_{0,0} & x_{0,1} \\ x_{1,0} & x_{1,1} \end{bmatrix}$, $Y_i = w_i * X$, with stride 1 padding 0, the outputs in four channels are: $\begin{bmatrix} w_1x_{00} & w_1x_{01} \\ w_1x_{10} & w_1x_{11} \end{bmatrix}, \begin{bmatrix} w_2x_{00} & w_2x_{01} \\ w_2x_{10} & w_2x_{11} \end{bmatrix}, \begin{bmatrix} w_3x_{00} & w_3x_{01} \\ w_3x_{10} & w_3x_{11} \end{bmatrix}, \begin{bmatrix} w_4x_{00} & w_4x_{01} \\ w_4x_{10} & w_4x_{11} \end{bmatrix}$.

$$Y = Ax = \begin{bmatrix} w_1 & 0 & 0 & 0 \\ 0 & w_1 & 0 & 0 \\ 0 & 0 & w_1 & 0 \\ 0 & 0 & 0 & w_1 \\ w_2 & 0 & 0 & 0 \\ 0 & w_2 & 0 & 0 \\ 0 & 0 & w_2 & 0 \\ 0 & 0 & 0 & w_2 \\ w_3 & 0 & 0 & 0 \\ 0 & w_3 & 0 & 0 \\ 0 & 0 & w_3 & 0 \\ 0 & 0 & 0 & w_3 \\ w_4 & 0 & 0 & 0 \\ 0 & w_4 & 0 & 0 \\ 0 & 0 & w_4 & 0 \\ 0 & 0 & 0 & w_4 \end{bmatrix} \begin{bmatrix} x_{00} \\ x_{01} \\ x_{10} \\ x_{11} \end{bmatrix} = \begin{bmatrix} w_1x_{00} \\ w_1x_{01} \\ w_1x_{10} \\ w_1x_{11} \\ w_2x_{00} \\ w_2x_{01} \\ w_2x_{10} \\ w_2x_{11} \\ w_3x_{00} \\ w_3x_{01} \\ w_3x_{10} \\ w_3x_{11} \\ w_4x_{00} \\ w_4x_{01} \\ w_4x_{10} \\ w_4x_{11} \end{bmatrix}$$

$(1,1,2,2)$ kernel:

$W = \begin{bmatrix} w_1 & w_2 \\ w_3 & w_4 \end{bmatrix}, X = \begin{bmatrix} x_{0,0} & x_{0,1} \\ x_{1,0} & x_{1,1} \end{bmatrix}$

With stride 2 padding 0, $Y = AX =$

$$\begin{bmatrix} w_1 & 0 & 0 & 0 \\ w_2 & 0 & 0 & 0 \\ 0 & w_1 & 0 & 0 \\ 0 & w_2 & 0 & 0 \\ w_3 & 0 & 0 & 0 \\ w_4 & 0 & 0 & 0 \\ 0 & w_3 & 0 & 0 \\ 0 & w_4 & 0 & 0 \\ 0 & 0 & w_1 & 0 \\ 0 & 0 & w_2 & 0 \\ 0 & 0 & 0 & w_1 \\ 0 & 0 & 0 & w_2 \\ 0 & 0 & w_3 & 0 \\ 0 & 0 & w_4 & 0 \\ 0 & 0 & 0 & w_3 \\ 0 & 0 & 0 & w_4 \end{bmatrix} \begin{bmatrix} x_{00} \\ x_{01} \\ x_{10} \\ x_{11} \end{bmatrix} = \begin{bmatrix} w_1x_{00} \\ w_2x_{00} \\ w_1x_{01} \\ w_2x_{01} \\ w_3x_{00} \\ w_4x_{00} \\ w_3x_{01} \\ w_4x_{01} \\ w_1x_{10} \\ w_2x_{10} \\ w_1x_{11} \\ w_2x_{11} \\ w_3x_{10} \\ w_4x_{10} \\ w_3x_{11} \\ w_4x_{11} \end{bmatrix}$$

The elements in two Y are the same, only their ordering are different.

Problem 4

AND:

$$\begin{cases} w_1 + w_2 + b \geq 0 \\ w_1 + b < 0 \\ w_2 + b < 0 \\ b < 0 \end{cases}$$

, let $w_{AND} = (2, 2), b_{AND} = -3$, then

x_1	x_2	$w^T x + b$	f(x)
0	0	-3	0
0	1	-1	0
1	0	-1	0
1	1	1	1

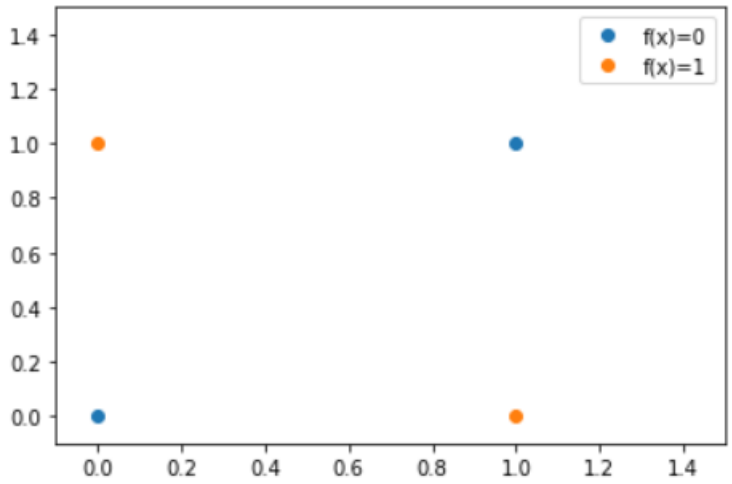
OR:

$$\begin{cases} w_1 + w_2 + b \geq 0 \\ w_1 + b \geq 0 \\ w_2 + b \geq 0 \\ b < 0 \end{cases}$$

, let $w_{OR} = (2, 2), b_{OR} = -1$, then

x_1	x_2	$w^T x + b$	f(x)
0	0	-1	0
0	1	1	1
1	0	1	1
1	1	3	1

Problem 5



There is not a line that can separate f(x)=1 and f(x)=0 points.

For XOR:

$$\begin{cases} w_1 + w_2 + b < 0 \\ w_1 + b \geq 0 \\ w_2 + b \geq 0 \\ b < 0 \end{cases}$$

From 1∼3 equation $w_1 < 0, w_2 < 0$; from 2∼4 equation, $w_1 > 0, w_2 > 0$, so there is not a solution.

Problem 6

$x = 1$:
First layer: $\max\{0, W^{(1)}x + b^{(1)}\} = \begin{bmatrix} 0.5 \\ 1.5 \end{bmatrix}$.
Second layer: $\max\{0, W^{(2)}h_1 + b^{(2)}\} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$.
Third layer: $h(x) = W^{(3)}h_2 + b^{(3)} = 5$.

$W = 4, b = 1 \Rightarrow Wx + b = 5$. $\dfrac{dh}{dx} = 4$

Problem 7

$x = -1$:

First layer: $\max\{0, W^{(1)}x + b^{(1)}\} = \begin{bmatrix} 0 \\ 0.5 \end{bmatrix}$.

Second layer: $\max\{0, W^{(2)}h_1 + b^{(2)}\} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$.

Third layer: $h(x) = W^{(3)}h_2 + b^{(3)} = 2$.

$W = 1, b = 3 \Rightarrow Wx + b = 2$. $\dfrac{dh}{dx} = 1$

Problem 8

$x = -0.5$:

First layer: $\max\{0, W^{(1)}x + b^{(1)}\} = \begin{bmatrix} 0 \\ 0.75 \end{bmatrix}$.

Second layer: $\max\{0, W^{(2)}h_1 + b^{(2)}\} = \begin{bmatrix} 0.75 \\ 0.75 \end{bmatrix}$.

Third layer: $h(x) = W^{(3)}h_2 + b^{(3)} = 2.5$.

$W = 1, b = 3 \Rightarrow Wx + b = 2.5$. $\dfrac{dh}{dx} = 1$

Problem 9

$$f_1(x) = |W^{(1)}x + b| = | \begin{bmatrix} 2 & 0 & \cdots & 0 \\ 0 & 2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} + \begin{bmatrix} -1 \\ -1 \\ \vdots \\ -1 \end{bmatrix} |, f_1(x)_i = |2x_i - 1|.$$

For $O = (0, 1)$ in each dimension (for each x_i), there are two identified regions $x_i \in (0, 1/2)$ and $x_i \in (1/2, 1)$.

This is a bijection from either of the two regions onto $(0, 1)$, because for each $f_1(x)_i$, there is a unique $x_i = (y_i + 1)/2$ for $x_i \in (1/2, 1)$, or $x_i = (1 - y_i)/2$ for $x_i \in (0, 1/2)$.

Assume each x_i is independent, then each x_i can choose in either of the two regions, there are 2^d regions identified onto $O = (0, 1)^d$.

Problem 10

As g and f are bijections, $f(g(x))$ is also a bijection between each region of input of g and $(0,1)^d$. Because for output y and any region in n_f, there is a unique $f^{-1}(y)$; then for any region in n_g there is unique $g^{-1}(f^{-1}(y))$.
For each of the n_g regions that g identifies, its image is $(0,1)^d$ and could be partitioned into n_f regions for f. So for n_g regions, there are $n_g \times n_f$ regions that identified onto $(0,1)^d$ by $f(g(x))$.

Problem 11

From the above two questions, for each of the layer, it identifies 2^d regions of input $\in (0,1)^d$ to output $\in (0,1)^d$, and is a bijection.
Assume the function for layer h_{L-1} is $f_{h_{L-1}} \circ \cdots \circ f_{h_1}$, it identifies $2^{(L-1)d}$ regions in input, and is bijection. Then $h_L = f_{h_L} \circ (f_{h_{L-1}} \circ \cdots \circ f_{h_1})$, f_{h_L} identifies 2^d regions, and $f_{h_L} \circ f_{h_{L-1}}$ is a bijection. So for all L layers, f_{h_L} identify 2^{Ld} regions of input $(0,1)^d$.

Problem 12

For gradient descent:
$f(w) = \sum_{i=1}^{n}(y_i - w^T x_i)^2 = (y - wX)^T(y - Xw) = w^T X^T Xw - 2y^T Xw + y^T y$,
$\frac{\partial f}{\partial w} = 2X^T Xw - 2X^T y$.
$w^t = w^{t-1} - \eta \nabla f(w^{t-1}) = w^{t-2} - \eta \nabla f(w^{t-2}) - \eta \nabla f(w^{t-1}) = w^0 - \eta \sum_{i=0}^{t-1} \eta \nabla f(w^i)$.
$w^0 = 0, w^t = -2\eta \sum_{i=0}^{t-1}(X^T Xw^i - X^T y) = -2\eta \sum_{i=0}^{t-1} X^T(Xw^i - y)$.
$Xw^i - y$ is the error for each i, let $-2\eta(Xw^i - y) = a_i$ be a fixed vector for each i, $A = \sum_i a_i$.
Then $w^t = X^T \sum_{i=1}^{t-1} a_i = X^T A$ for some vector A, $y = XX^T A$.

For minimizing norm:
$w^* = \arg\min ||w||_2^2 \ s.t. Xw^* = y$, $L(w^*, \lambda) = ||w||_2^2 + \lambda^T(y - Xw)$.
$\frac{\partial L}{\partial w^*} = 2w^* - X^T\lambda$, let $\frac{\partial L}{\partial w^*} = 0$, then $2w^* = X^T\lambda, w^* = \frac{1}{2}X^T\lambda$.
Multiply X at both sides of $2w^* = X^T\lambda \Rightarrow 2Xw^* = XX^T\lambda$.
$XW^* = y \Rightarrow 2y = XX^T\lambda$.
Assume XX^T is invertible, multiply $(XX^T)^{-1}$ at both sides $\Rightarrow 2(XX^T)^{-1}y = (XX^T)^{-1}(XX^T)\lambda \Rightarrow$
$2(XX^T)^{-1}y = \lambda$.
$w^* = \frac{1}{2}X^T\lambda \Rightarrow w^* = X^T(XX^T)^{-1}y$.
$y = XX^T A \Rightarrow w^* = X^T(XX^T)^{-1}XX^T A = X^T A$.

The results of gradient descent and minimizing norm are the same.

Problem 13

$Xw = y$ means this optimization looks for a w that gives zero training error.
$w^{gd} = \arg\min ||w||_2^2$ means for all possible w that gives zero error, looks for the one with smallest L2 norm (L2 regularization).
So this optimization looks for a w with smallest training error and norm.

Problem 14

As $d > n$, there are infinite solutions to $Xw = y$. Gradient descent finds the global minimum with smallest norm, which is an approximation to L2 regularization, even if there is no explicit regularizer.

Problem 15

Optimization error cannot be clearly separated from estimation error, because the result of optimization depends on the whole dataset X and also assumptions on X.

Problem 16

The author shows that there is double descent phenomenon occurs in some deep learning model, where the testing error first decrease, then increase, then decrease again. They show that this could happen with the increasing of number of parameters in the model and training epochs. Sometimes this phenomenon leads to worse performance for large dataset. They show that neither the statistical view of "too large models are worse" nor the machine learning view of "bigger models are better" are correct, and the performance is related to the effective model complexity.

But for very complex model that could fit large dataset, it may not cross the interpolation threshold in practice. And it seems this phenomenon also depends on label noise, so maybe using better data can do the same thing as regularization.

Problem 17

The phenomenon of keep increasing parameter number even after overfitting would actually gives better performance is new to me. Another thing that I'm curious about is whether changing the input and output, e.g. dimension, output type, would have similar effects as number of samples.