CS795/895: Fundamentals of Deep Learning (Spring 2024)
Homework Assignment 2
Name: Sushmitha Halli Sudhakara

Problem 1 (20 points)

Let P be the vector space of all polynomial functions on $\mathbb{R}$ with real coefficients. Define linear transformation $T, D : P \to P$ by

$$(Dp)(x) = p'(x)$$

and

$$(Tp)(x) = xp(x)$$

for all $x \in \mathbb{R}$.

(a) Let $p(x) = x^3 - 7x^2 + 5x + 6$ for all $x \in \mathbb{R}$. Then $((D+T)(p))(x) = x^4 - ax^3 + bx^2 - bx + c$ where $a = \underline{7}$, $b = \underline{8}$, and $c = \underline{5}$.

(a) Given $p(x) = x^3 - 7x^2 + 5x + 6$

$(Dp)(x) = p'(x) = \frac{d}{dx} p(x)$

$= \frac{d}{dx}(x^3 - 7x^2 + 5x + 6)$

$\boxed{(Dp)(x) = 3x^2 - 14x + 5}$ —— ①

$(Tp)(x) = x\,p(x)$

$= x(x^3 - 7x^2 + 5x + 6)$

$\boxed{(Tp)(x) = x^4 - 7x^3 + 5x^2 + 6x}$ —— ②

$((D+T)p)(x) = (Dp)(x) + (Tp)(x)$

$= (3x^2 - 14x + 5) + (x^4 - 7x^3 + 5x^2 + 6x)$

$\boxed{((D+T)p)(x) = x^4 - 7x^3 + 8x^2 - 8x + 5}$ —— ③

Given

$\boxed{((D+T)p)(x) = x^4 - ax^3 + bx^2 - bx + c}$ —— ④

Comparing ③ & ④ we get

$a = 7$, $b = 8$ and $c = 5$

(b) Let p be as in (a). Then $(DTp)(x) = ax^3 - bx^2 + cx + 6$ where
$a = \underline{4}$,
$b = \underline{21}$,
$c = \underline{10}$.

(b) Given $p(x) = x^3 - 7x^2 + 5x + 6$

$$\boxed{(Tp)(x) = x^4 - 7x^3 + 5x^2 + 6x} \longrightarrow \text{from (2)}$$

$$(DTp)(x) = \frac{d}{dx}(Tp)(x)$$

$$= \frac{d}{dx}(x^4 - 7x^3 + 5x^2 + 6x)$$

$$\boxed{(DTp)(x) = 4x^3 - 21x^2 + 10x + 6} \quad \text{— (5)}$$

Given,

$$\boxed{(DTp)(x) = ax^3 - bx^2 + cx + 6} \quad \text{— (6)}$$

Comparing (5) & (6) we get

$\underline{\underline{a = 4, \ b = 21 \text{ and } c = 10}}$

(c) Let p be as in (a). Then $(TDp)(x) = ax^3 - bx^2 + cx$ where
$a = \underline{3}$,
$b = \underline{14}$,
$c = \underline{5}$.

(c) Given $p(x) = x^3 - 7x^2 + 5x + 6$

$\boxed{(Dp)(x) = 3x^2 - 14x + 5}$ $\longrightarrow$ from (1)

$(TDp)(x) = x\,(Dp)(x)$

$= x\,(3x^2 - 14x + 5)$

$\boxed{(TDp)(x) = 3x^3 - 14x^2 + 5x}$ —— (7)

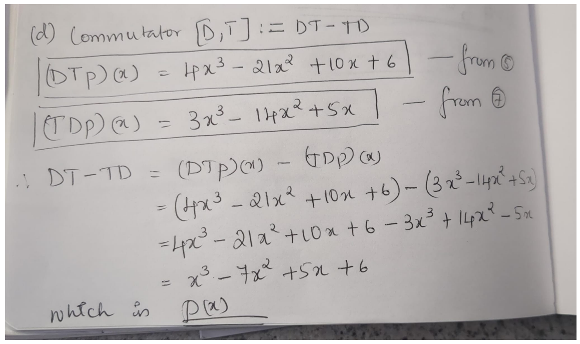
Given

$\boxed{(TDp)(x) = ax^3 - bx^2 + cx}$ —— (8)

Comparing (7) & (8) we get

$\underline{a = 3,\ b = 14 \text{ and } c = 5}$

(d) Evaluate (and simplify) the commutator $[D, T] := DT - TD$.
Answer: $[D, T] = \underline{p(x) = x^3 - 7x^2 + 5x + 6}$.



(d) Commutator $[D,T] := DT - TD$

$\boxed{(DTp)(x) = 4x^3 - 21x^2 + 10x + 6}$ — from 6

$\boxed{(TDp)(x) = 3x^3 - 14x^2 + 5x}$ — from 7

$$\begin{aligned} \therefore DT - TD &= (DTp)(x) - (TDp)(x) \\ &= (4x^3 - 21x^2 + 10x + 6) - (3x^3 - 14x^2 + 5x) \\ &= 4x^3 - 21x^2 + 10x + 6 - 3x^3 + 14x^2 - 5x \\ &= x^3 - 7x^2 + 5x + 6 \end{aligned}$$

which is $\underline{p(x)}$

(e) Find a number p such that $(TD)^p = T^p D^p + TD$.
Answer: $p =$ a real number.

Problem 2 (15 points)

Let $T : \mathbb{R}^3 \to \mathbb{R}^4$ be defined by

$$Tx = (x_1 - 3x_3, x_1 + x_2 - 6x_3, x_2 - 3x_3, x_1 - 3x_3)$$

for every $x = (x_1, x_2, x_3) \in \mathbb{R}^3$. (The map T is linear, but you need not prove this.) Then

(a)

$$[T] = \begin{bmatrix} \quad & \quad & \quad \\ \quad & \quad & \quad \\ \quad & \quad & \quad \\ \quad & \quad & \quad \end{bmatrix}$$

The standard basis vectors in $\mathbb{R}^3$ are:
$E_1 = (1, 0, 0)$
$E_2 = (0, 1, 0)$
$E_3 = (0, 0, 1)$
Applying $T\mathbf{x} = (x_1 - 3x_3, x_1 + x_2 - 6x_3, x_2 - 3x_3, x_1 - 3x_3)$, we get:
Apply T to E_1:
$T(E_1) = T(1, 0, 0) = (1 - 3 \cdot 0, 1 + 0 - 6 \cdot 0, 0 - 3 \cdot 0, 1 - 3 \cdot 0) = (1, 1, 0, 1)$
Apply T to E_2:
$T(E_2) = T(0, 1, 0) = (0 - 3 \cdot 0, 0 + 1 - 6 \cdot 0, 1 - 3 \cdot 0, 0 - 3 \cdot 0) = (0, 1, 1, 0)$
Apply T to E_3:
$T(E_3) = T(0, 0, 1) = (0 - 3 \cdot 1, 0 + 0 - 6 \cdot 1, 0 - 3 \cdot 1, 0 - 3 \cdot 1) = (-3, -6, -3, -3)$
The columns of the matrix $[T]$ are the transformed basis vectors:

$$[T] = \begin{bmatrix} 1 & 0 & -3 \\ 1 & 1 & -6 \\ 0 & 1 & -3 \\ 1 & 0 & -3 \end{bmatrix}$$

(b) $T(3, -2, 4) =$________________.
Given $T(\mathbf{x}) = (x_1 - 3x_3, x_1 + x_2 - 6x_3, x_2 - 3x_3, x_1 - 3x_3)$
Substituting $x_1 = 3, x_2 = -2$, and $x_3 = 4$ into the transformation we get,

$$T(3, -2, 4) = (3 - 3 \cdot 4, 3 - 2 - 6 \cdot 4, -2 - 3 \cdot 4, 3 - 3 \cdot 4)$$

$$T(3, -2, 4) = (-9, -23, -14, -9).$$

Problem 3 (15 points)

The in-sample error of a linear regression problem can be expressed as

$$E_{\text{in}} = \frac{1}{N}\sum_{n=1}^{N}(\hat{y}_n - y_n)^2 = \frac{1}{N}||\hat{\mathbf{y}} - \mathbf{y}||_2^2$$

in which

$$\hat{\mathbf{y}} = [\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_N]^T = [\mathbf{w}^T\mathbf{x}_1, \mathbf{w}^T\mathbf{x}_2, \cdots, \mathbf{w}^T\mathbf{x}_N]^T = X\mathbf{w}$$

is the predicted labels, and

$$\mathbf{y} = [y_1, y_2, \cdots, y_N]^T$$

are the true labels.

$X = [\mathbf{x}_1, \mathbf{x}_2, \cdots .\mathbf{x}_N]^T$ is the input.

Prove that

$$E_{\text{in}} = \frac{1}{N}\left(\mathbf{w}^T X^T X\mathbf{w} - 2\mathbf{w}^T X^T\mathbf{y} + \mathbf{y}^T\mathbf{y}\right)$$

Solution 3

Solution 3

In-sample error of a linear regression problem

$$E_{in} = \frac{1}{N}\sum_{n=1}^{N}(\hat{y}_n - y_n)^2 = \frac{1}{N}\|\hat{y} - y\|_2^2$$

Given $\hat{y} = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N]^T = [w^T x_1, w^T x_2, \ldots, w^T x_N]^T$

$\hat{y} = Xw \Rightarrow$ predicted labels

$y = [y_1, y_2, \ldots, y_N]^T \Rightarrow$ true labels

$X = [x_1, x_2, \ldots; x_N]^T \Rightarrow$ input

Proof

$$E_{in} = \frac{1}{N}\sum_{n=1}^{N}(\hat{y}_n - y_n)^2$$

$$= \frac{1}{N}\|\hat{y} - y\|_2^2$$

Given $\hat{y} = Xw$

$$\Rightarrow E_{in} = \frac{1}{N}\|Xw - y\|_2^2$$

The squared norm can be expanded as

$$\|Xw - y\|_2^2 = (Xw - y)^T(Xw - y)$$

$$= (Xw)^T Xw - 2(Xw)^T y + y^T y$$

We know that $(AB)^T = B^T A^T$

$$\therefore (Xw)^T(Xw) - 2(Xw)^T y + y^T y$$

$$= w^T X^T Xw - 2w^T X^T y + y^T y$$

Substituting the above in in-sample error of linear regression problem equation

$$E_{in} = \frac{1}{N}\|Xw - y\|_2^2$$

$$\boxed{E_{in} = \frac{1}{N}\left(w^T X^T Xw - 2w^T X^T y + y^T y\right)}$$

Hence proved

Problem 4 (25 points)

Proof the shift-invariance property of the softmax function. The softmax function in multi-class logistic regression has an invariance property when shifting the parameters. Given the weights $\mathbf{w} = (\mathbf{w}_1, \cdots, \mathbf{w}_K)$, suppose we subtract the same vector $\mathbf{u}$ from each of the K weight vectors, the outputs of softmax function will remain the same. You may denote $\mathbf{w}' = \{\mathbf{w}'_i\}_{i=1}^K$, where $\mathbf{w}'_i = \mathbf{w}_i - \mathbf{u}$. Prove that

$$P(y = k|\mathbf{x}; \mathbf{w}') = P(y = k|\mathbf{x}; \mathbf{w})$$

Solution 4

Solution 4

Given weights $w = (w_1, \cdots, w_K)$

$\Rightarrow$ Suppose we subtract the vector u from each of the K weight vectors, the output of softmax function will remain the same.

$$w' = \{w'_i\}_{i=1}^{K}, \text{ where } w'_i = w_i - u$$

To prove, $P(y=k \mid x; w') = P(y=k \mid x; w)$

$$\text{Softmax}(z_k) = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}} \quad \text{— Softmax equation}$$

Considering a multi-class logistic regression problem with K classes. The probability of the class k given the input x and the weights w is:

$$P(y=k \mid x; w) = \frac{e^{w_k^T x}}{\sum_{j=1}^{K} e^{w_j^T x}}$$

Now shifted weight are given by

$$w'_i = w_i - u \text{ for all } i.$$

The probability with the shifted weights becomes

$$P(y=k \mid x; w') = \frac{e^{(w_K - u)^T x}}{\sum_{j=1}^{K} e^{(w_j - u)^T x}}$$

Simplifying the numerator & denominator, we have;

$$P(y=k|x;w') = \frac{e^{w_k^T x - u^T x}}{\sum_{j=1}^{K} e^{w_j^T x - u^T x}}$$

$-u^T x$ is a constant for all classes in the softmax function, therefore we can factor it out from both numerator & denominator.

$$P(y=k|x;w') = \frac{e^{w_k^T x} \cdot \cancel{e^{-u^T x}}}{\sum_{j=1}^{K} e^{w_j^T x} \cdot \cancel{e^{-u^T x}}}$$

$$P(y=k|x;w') = \frac{e^{w_k^T x}}{\sum_{j=1}^{K} e^{w_j^T x}}$$

which is equal to $P(y=k|x;w)$

Hence proved,

$$P(y=k|x;w') = P(y=k|x;w)$$

Problem 5 (25 points)

Prove that the softmax-based multiclass logistic regression is equivalent to the sigmoid-based binary logistic regression.

Solution 5

Proof that softmax-based multiclass logistic regression is equivalent to the sigmoid-based binary logistic regression.

Sigmoid-based Binary Logistic Regression:
In binary logistic regression, we model the probability that an instance x belongs to a class (labelled as 1) as follows:

$$p(y=1|x) = \sigma(\theta^T x)$$
$$p(y=0|x) = 1 - \sigma(\theta^T x)$$

where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function, θ is the parameter vector, and x is the feature vector.

Softmax-based Multiclass Logistic Regression:
In multiclass logistic regression, we generalize this concept to multiple classes. Given K classes, the probability that an instance x belongs to class k is modeled as:

$$p(y=k|x) = \frac{e^{\theta_k^T x}}{\sum_{j=1}^{K} e^{\theta_j^T x}}$$

This is known as the softmax function, where θ_k is the parameter vector for class k.

Equivalence of Softmax in Binary Case to Sigmoid:
To show the equivalence, let's consider the softmax function in the case of $K = 2$ classes. We will denote θ_1 and θ_2 as the parameter vectors for class 1 and 2, respectively.

$$p(y=1|x) = \frac{e^{\theta_1^T x}}{e^{\theta_1^T x} + e^{\theta_2^T x}}$$
$$p(y=2|x) = \frac{e^{\theta_2^T x}}{e^{\theta_1^T x} + e^{\theta_2^T x}}$$

Notice that if we set $\theta = \theta_1 - \theta_2$ and rewrite $p(y = 1|x)$, we get:

$$p(y=1|x) = \frac{1}{1+e^{-\theta^T x}}$$

This is exactly the form of the sigmoid function used in binary logistic regression. Therefore, when $K = 2$, the softmax function simplifies to the sigmoid function, demonstrating that the softmax-based multiclass logistic regression is equivalent to the sigmoid-based binary logistic regression for the binary case.