Assignment 1

Question 1:

The volume of a ball in P dimensional space is

$$V_p(R) = \frac{\pi^{P/2}}{\Gamma(\frac{P}{2}+1)} R^P$$

$\frac{\pi^{\frac{P}{2}}}{\Gamma(\frac{P}{2}+1)}$ part is constant given ball is P dimensional, radius R changes.

We have N datapoints which are uniformly distributed in a unit ball. (R=1). The probability that a point falls inside a ball of radius r can be found through the cdf:

$$P(X \in V_p(r)) = \frac{V_p(r)}{V_p(1)}$$

$$= \frac{\frac{\pi^{\frac{P}{2}}}{\Gamma(\frac{P}{2}+1)} r^P}{\frac{\pi^{\frac{P}{2}}}{\Gamma(\frac{P}{2}+1)} 1^P}$$

$$= r^P$$

Both the ball of radius r and unit ball has the same origin. So, if a point is inside a ball its distance from the origin is smaller than the radius of this ball.

Euclidean norm in p dimension space of vector $x = ||x||$

So, $P(x \in V_p(r)) = P(||x|| \leq r) = r^P$

Now let us consider a random variable

$$R = \min_i \{\|x_i\|\}$$

which denotes smallest value of all samples.

$$P(R \leq r) = P(\min_i \{\|x_i\|\} \leq r)$$

$$= 1 - P(\min_i \{\|x_i\|\} > r)$$

$$= 1 - \prod_{i=1}^{N} P(\|x_i\| > r)$$

$$= 1 - \prod_{i=1}^{N} (1 - P(\|x_i\| < r))$$

$$= 1 - \prod_{i=1}^{N} (1 - r^P)$$

$$= 1 - (1 - r^P)^N$$

This cdf of R denotes the probability that the shortest distance of a point to origin in all random variable is less than or equal to cdf's parameter r.

Now median is 0.5 quantile of its cdf. So,

$$1 - (1 - r^P)^N = 0.5$$

$$\Rightarrow (1 - r^P)^N = \frac{1}{2}$$

$$\Rightarrow 1 - r^P = \frac{1}{2}^{\frac{1}{N}}$$

$$\Rightarrow r^P = 1 - \frac{1}{2}^{\frac{1}{N}}$$

$$\Rightarrow r = \left(1 - \frac{1}{2}^{\frac{1}{N}}\right)^{\frac{1}{P}}$$

Thus, The median distance from origin to the closest data point is,

$$\boxed{d(p, N) = \left(1 - \frac{1}{2}^{\frac{1}{N}}\right)^{\frac{1}{p}}}$$

(proved)

4 For $N = 10000$, $p = 1000$

$$d(1000, 10000) = \left(1 - 0.5^{\frac{1}{10000}}\right)^{\frac{1}{1000}} = \boxed{0.9905} \text{ (ans.)}$$

Question 02:

Given, $f(x) = (x_1 + x_2)(x_1 x_2 + x_1 x_2^2)$

$$= x_1^2 x_2 + x_1^2 x_2^2 + x_1 x_2^2 + x_1 x_2^3$$

Now, the gradient, $\nabla f(x) = \begin{bmatrix} \frac{\partial}{\partial x_1} f(x) \\ \frac{\partial}{\partial x_2} f(x) \end{bmatrix}$

$$\therefore \nabla f(x) = \begin{bmatrix} 2x_1 x_2 + 2x_1 x_2^2 + x_2^2 + x_2^3 \\ x_1^2 + 2x_1^2 x_2 + 2x_1 x_2 + 3x_1 x_2^2 \end{bmatrix}$$

Finding 3 stationary points, $\nabla f(x) = 0$

$$\therefore 2x_1 x_2 + 2x_1 x_2^2 + x_2^2 + x_2^3 = 0$$

$$\Rightarrow 2x_1(x_2 + x_2^2) + x_2(x_2 + x_2^2) = 0$$

$$\Rightarrow (x_2 + x_2^2)(2x_1 + x_2) = 0$$

$2x_1 + x_2 = 0$	$x_2 + x_2^2 = 0$
$x_2 = 2x_1$	$x_2(1 + x_2) = 0$
	$x_2 = 0 \quad x_2 = -1$

$$x_1^2 + 2x_1^2 x_2 + 2x_1 x_2 + 3x_1 x_2^2 = 0$$

$$x_1 (x_1 + 2x_1 x_2 + 2x_2 + 3x_2^2) = 0$$

$x_1 = 0$ or $x_1 + 2x_1 x_2 + 2x_2 + 3x_2^2 = 0$ ——— (1)

From previous page,

let, $x_2 = 0$ so from (1) $x_1 = 0$

let $x_2 = -1$ so from (1) $x_1 - 2x_1 - 2 + 3 = 0$

$\Rightarrow -x_1 + 1 = 0$

$\therefore x_1 = 1$

Let $x_2 = -2x_1$ so from (1)

$$x_1 - 4x_1^2 - 4x_1 + 12x_1^2 = 0$$

$$\Rightarrow -3x_1 + 8x_1^2 = 0$$

$$\Rightarrow x_1(-3 + 8x_1) = 0$$

$x_1 = 0$ or $-3 + 8x_1 = 0$

$x_1 = \frac{3}{8}$

$x_2 = -2 \cdot x_1 = -2 \cdot \frac{3}{8} = -\frac{6}{8}$

So, stationary points, $\boxed{(x_1, x_2) = (0,0), (1,-1), (\frac{3}{8}, -\frac{6}{8})}$

Now, the Hessian matrix:

$$\nabla^2 f(x) = \begin{bmatrix} 2x_2 + 2x_2^2 & 2x_1 + 4x_1x_2 + 2x_2 + 3x_2^2 \\ 2x_1 + 4x_1x_2 + 2x_2 + 3x_2^2 & 2x_1^2 + 2x_1 + 6x_1x_2 \end{bmatrix}$$

at point (0, 0)

$$\nabla^2 f(x) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

If a Hessian matrix is negative definite then ~~the~~ its corresponding critical points are local maxima.

We can check if a matrix is negative definite through its eigenvalues.

The eigen value of the hessian matrix at 0, 0 is $[0, 0]$ so it's not negative definite.

at point (1, −1)

$$\nabla^2 f(x) = \begin{bmatrix} -2+2 & 2+4\cdot1(-1)-2+3 \\ 2-4-2+3 & 2+2-6 \end{bmatrix}$$

$$= \boxed{\begin{bmatrix} 0 & -1 \\ -1 & -2 \end{bmatrix}}$$

It's eigenvalue $[0.414 \quad -2.414]$

so its also not a local maxima. Rather a saddle point.

Now, at point $\left(\frac{3}{8}, \frac{-6}{8}\right)$

$$\nabla^2 f(x) = \begin{bmatrix} -\frac{2\times 6}{8} + 2\cdot\left(\frac{6}{8}\right)^2 & 2\cdot\frac{3}{8} + 4\cdot\frac{3}{8}\left(-\frac{6}{8}\right) + 2\left(\frac{-6}{8}\right) + 3\left(-\frac{6}{8}\right) \\ 2\cdot\frac{3}{8} + 4\cdot\frac{3}{8}\left(-\frac{6}{8}\right) + 2\left(-\frac{6}{8}\right) + 3\left(-\frac{6}{8}\right) & -\frac{21}{32} \end{bmatrix}$$

$$= \boxed{\begin{bmatrix} -\frac{3}{8} & -\frac{3}{16} \\ -\frac{3}{16} & -\frac{21}{32} \end{bmatrix}}$$

Its corresponding eigenvalue is: $[-.28125 \quad -0.75]$

So the Hessian is negative definite.

So the point $\left(\frac{3}{8}, \frac{-6}{8}\right)$ is local maximum.

and it is the only local maximum of the given function.

Question 03:

$f(x) = 8x_1 + 12x_2 + x_1^2 - 2x_2^2$

$$\nabla f(x) = \begin{bmatrix} 8 + 0 + 2x_1 + 0 \\ 0 + 12 + 0 - 4x_2 \end{bmatrix} = \begin{bmatrix} 2x_1 + 8 \\ -4x_2 + 12 \end{bmatrix}$$

$$\nabla^2 f(x) = \begin{bmatrix} 2 & 0 \\ 0 & -4 \end{bmatrix}$$

Now there is only constants in the Hessian which implies there is only one solution to the given quadratic equation.

The stationary point

$$\nabla f(x) = 0 \quad \begin{matrix} 2x_1 + 8 = 0 \Rightarrow x_1 = -4 \\ -4x_2 + 12 = 0 \Rightarrow x_2 = 3 \end{matrix}$$

The stationary point doesn't have any effect on the hessian.

The eigenvalue of the hessian is $[2 \ \ -4]$

There is a ~~mi~~ positive eigenvalue and also a negative eigenvalue So, the hessian is neither positive or negative definite.

So, the staionary point is a saddle point.

(Proved)

Question 04:

Let A be a $n \times n$ matrix

B be a $m \times m$ matrix.

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & \cdots & \cdots & A_{2n} \\ \vdots & & & \\ A_{n1} & \cdots & \cdots & A_{nn} \end{bmatrix} \qquad B = \begin{bmatrix} B_{11} & \cdots & B_{1m} \\ B_{21} & \cdots & B_{2m} \\ \vdots & & \vdots \\ B_{m1} & \cdots & B_{mm} \end{bmatrix}$$

as A is positive definite we can say

$x'^T A x' > 0$ where x' is its corresponding eigenvect

Similarly B is also ~~semid~~ positive definite

$x''^T B x'' > 0$ where x'' is B's eigenvector

$$x' = \begin{bmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_n \end{bmatrix} \qquad x'' = \begin{bmatrix} x''_1 \\ x''_2 \\ \vdots \\ x''_m \end{bmatrix}$$

We can represent $x'^T A x' > 0$ in terms of matrix multiplied form as the following

$$\sum_{j=1}^{n} \sum_{i=1}^{n} x'_i A_{ij} x'_j > 0 \quad \text{——— (1)}$$

similarly $x''^T B x'' > 0$ as

$$\sum_{j=1}^{m} \sum_{i=1}^{m} x''_i \, B_{ij} \, x''_j > 0 \qquad (11)$$

Now let's consider $C = \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$ which can be expressed as the following:

$$C = \begin{array}{c} n\left\{\begin{array}{c} \\ \\ \\ \\ \end{array}\right. \\ m\left\{\begin{array}{c} \\ \\ \\ \\ \end{array}\right. \end{array} \left[\begin{array}{cccc|cccc} A_{11} & \cdots & \cdots & A_{1n} & 0 & \cdots & \cdots & 0 \\ A_{21} & \cdots & \cdots & A_{2n} & 0 & \text{All} & \cdots & 0 \\ \vdots & & & \vdots & \vdots & 0\text{'s} & & \vdots \\ A_{n1} & \cdots & \cdots & A_{nn} & 0 & \cdots & \cdots & 0 \\ \hline 0 & \cdots & \cdots & 0 & B_{11} & \cdots & \cdots & B_{1m} \\ 0 & \text{All} & & 0 & B_{21} & \cdots & \cdots & B_{2m} \\ \vdots & 0\text{'s} & & \vdots & \vdots & & & \vdots \\ 0 & \cdots & \cdots & 0 & B_{m1} & \cdots & \cdots & B_{mm} \end{array}\right]$$

$$\underbrace{\qquad\qquad}_{n} \quad \underbrace{\qquad\qquad}_{m}$$

dimension $(n+m) \times (n+m)$

We want to prove C is also positive definite. So, C also has an eigenvector x for which $x^T C x > 0$.

where,

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \\ x_{n+1} \\ \vdots \\ x_{n+m} \end{bmatrix} \quad \begin{array}{l} \left.\begin{array}{c} \\ \\ \\ \end{array}\right\} = \begin{bmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_n \end{bmatrix} = x' \\ \left.\begin{array}{c} \\ \\ \end{array}\right\} = \begin{bmatrix} x''_1 \\ x''_2 \\ \vdots \\ x''_m \end{bmatrix} = x'' \end{array}$$

So, x is the join of both eigenvector of A and B.

If $x^T C x > 0$ then we would be able to say C is positive definite.

$$x^T C x$$

$$\begin{bmatrix} x_1 & x_2 & \cdots & x_n & x_{n+1} & \cdots & x_{m+n} \end{bmatrix} \overset{C}{\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \\ x_{n+1} \\ \vdots \\ x_{n+m} \end{bmatrix}$$

doing matrix multiplication we can get the following result:

$$\sum_{j=1}^{n} \left[\sum_{i=1}^{n} x_i C_{ij} x_j + \sum_{i=n+1}^{n+m} x_i \cdot 0 \cdot x_j \right]$$

$$+ \sum_{j=n+1}^{n+m} \left[\sum_{i=1}^{n} x_i \cdot 0 \cdot x_j + \sum_{i=n+1}^{n+m} x_i \cdot C_{ij} \cdot x_j \right]$$

$$\Rightarrow \sum_{j=1}^{n} \sum_{i=1}^{n} x_i C_{ij} x_j + \sum_{j=n+1}^{n+m} \sum_{i=n+1}^{n+m} x_i C_{ij} x_j$$

This part of C corresponds to A and is equal to (I)

This part of C corresponds to B and is equal to (II)

as (I) and (II) are >0 so addition of (I) + (II) >0.

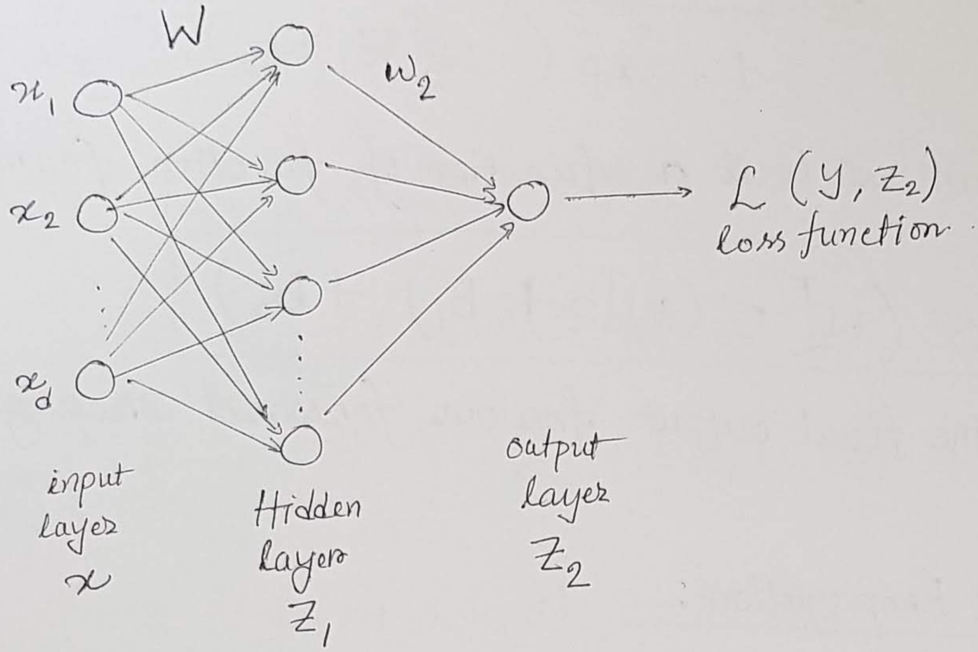
So, $x^T C x > 0$

$\therefore$ C or $\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}$ is a positive definite matrix

(proved)

Question 05

Forward Propagation:



The input is a vector x.

ith node of $Z_1^{[i]}$ is calculated the following way:

[we are using sigmoid function for activation and b_1 is bias term in 1^{st} layer along with parameter W where $W^{[i]}$ is parameter corresponding to i-th node in $Z_1^{[i]}$]

$$Z_1^{[i]} = \sigma\left(W^{[i]T}x + b_1\right) = \frac{1}{1 + \exp\left(-W^{[i]T}x + b_1\right)}$$

Vectorizing for the entire layer we get:

$$Z_1 = \sigma\left(W^T x + b_1\right) = \frac{1}{1 + \exp\left(-W^T x + b_1\right)}$$

Now for the output layer Z_1 is used as input and W_2, b_2 is our parameter. Similarly using sigmoid

We get $Z_2 = \sigma(\omega_2^T Z_1 + b_2)$

$$= \frac{1}{1 + \exp(-\omega_2^T Z_1 + b_2)}$$

OR we can write it as function of function form in below:

$$\boxed{Z_2 = \sigma(W_2^T \sigma(W^T x + b_1) + b_2)}$$

Which is the final output for our forward propagation.

Backward Propagation:

Our loss function is a 2-class cross-entropy loss function

$$\mathcal{L}(Z_2, y) = -(y \log Z_2 + (1-y)\log(1-Z_2)).$$

Now we start backprop from our last layer, so

$$\frac{\partial \mathcal{L}}{\partial W_2} = \frac{\partial \mathcal{L}}{\partial Z_2} \cdot \frac{\partial Z_2}{\partial W_2}$$

using chain rule.

$$\frac{\partial \mathcal{L}}{\partial Z_2} = -\frac{\partial}{\partial Z_2}(y \log Z_2 + (1-y)\log(1-Z_2))$$

$$= -\frac{y}{Z_2} + \frac{1-y}{1-Z_2}.$$

$$\frac{\partial \mathcal{L}}{\partial W_2} = \frac{\partial \mathcal{L}}{\partial Z_2} \frac{\partial Z_2}{\partial W_2}$$

$$= \left(-\frac{y}{Z_2} + \frac{1-y}{1-Z_2}\right) \frac{\partial}{\partial W_2} \sigma(W_2^T Z_1 + b_2)$$

$$= \left(-\frac{y}{Z_2} + \frac{1-y}{1-Z_2}\right) \sigma(W_2^T Z_1 + b_2)\left(1 - \sigma(W_2^T Z_1 + b_2)\right)$$

$$\cdot \frac{\partial}{\partial W_2}(W_2^T Z_1 + b_2)$$

$$= \left(-\frac{y}{Z_2} + \frac{1-y}{1-Z_2}\right) \cdot Z_2 (1-Z_2) \cdot Z_1$$

$$= \left(-y(1-Z_2) + (1-y) \cdot Z_2\right) Z_1$$

$$= \left(-y + y \cdot Z_2 + Z_2 - y \cdot Z_2\right) \cdot Z_1$$

$$= (Z_2 - y) \cdot Z_1$$

To match dimension of $\frac{\partial \mathcal{L}}{\partial W_2}$ with W_2 we rearrange and get,

$$\boxed{\frac{\partial \mathcal{L}}{\partial W_2} = Z_1(Z_2 - y)}$$

For parameter b_2

$$\frac{\partial \mathcal{L}}{\partial b_2} = \frac{\partial \mathcal{L}}{\partial Z_2} \frac{\partial Z_2}{\partial b_2}$$

$$\frac{\partial Z_2}{\partial b_2} = \frac{\partial}{\partial b_2} \sigma(W_2^T Z_1 + b_2) = Z_2(1 - Z_2)$$

$$\boxed{\frac{\partial \mathcal{L}}{\partial b_2} = (Z_2 - y)}$$

Now for the parameters of the 1^{st} hidden layer:

$$\frac{\partial \mathcal{L}}{\partial W_1} = \frac{\partial \mathcal{L}}{\partial Z_2} \cdot \frac{\partial Z_2}{\partial Z_1} \cdot \frac{\partial Z_1}{\partial W_1}$$

$$\frac{\partial}{\partial Z_1} Z_2 = \frac{\partial}{\partial Z_1} \sigma(W_2^T Z_1 + b_2)$$

$$= \sigma(W_2^T Z_1 + b_2)\,(1 - \sigma(W_2^T Z_1 + b_2))\,\frac{\partial}{\partial Z_1}(W_2^T Z_1 +$$

$$= Z_2\,(1 - Z_2)\cdot W_2$$

$$\frac{\partial}{\partial W_1} Z_1 = \frac{\partial}{\partial W} \sigma(W^T X + b_1)$$

$$= Z_1\,(1 - Z_1)\cdot \frac{\partial}{\partial W}(W^T x + b_1)$$

$$= Z_1\,(1 - Z_1)\cdot X$$

$$\frac{\partial \mathcal{L}}{\partial W_1} = \frac{\partial \mathcal{L}}{\partial Z_2}\,\frac{\partial Z_2}{\partial Z_1}\,\frac{\partial Z_1}{\partial W}$$

$$= \left(-\frac{y}{Z_2} + \frac{1-y}{1-Z_2}\right) Z_2(1-Z_2)\,W_2\;Z_1\,\cancel{Z}(1-Z_1)\,X$$

$$\boxed{\frac{\partial \mathcal{L}}{\partial W} = Z_1(Z_2 - y)\cdot W_2 \cdot Z_1\,(1 - Z_1)\cdot X}$$

To match the dimension of $\frac{\partial \mathcal{L}}{\partial W}$ with W we rearrange and get the following.

$$\frac{\partial \mathcal{L}}{\partial W} = X\,[Z_1\,(Z_2 - y)]^T\,[Z_1 \odot (1 - Z_1)]\,W_2^T$$

For parameter b_1

$$\frac{\partial L}{\partial b_1} = \frac{\partial L}{\partial z_2} \frac{\partial z_2}{\partial z_1} \frac{\partial z_1}{\partial b_1}$$

$$\frac{\partial}{\partial b_1} z_1 = \frac{\partial}{\partial b_1} \sigma(W^T x + b_1)$$

$$= z_1 (1 - z_1) \frac{\partial}{\partial b_1} (W^T x + b_1)$$

$$= z_1 (1 - z_1)$$

$$\therefore \boxed{\frac{\partial L}{\partial b_1} = z_1 (z_2 - y) \cdot W_2 \cdot z_1 (1 - z_1).}$$

All gradients based on all parameter W, b_1, W_2, b_2 has been shown.