Q1.

$f = d = (x_1 - 4)^2 + (x_2 - 3)^2$ function to maximized

$g = x_1^2 + 2x_2^2 - 1 = 0$ Constraint

we use lagrange multiplier method

$$(x_1 - 4)^2 + (x_2 - 3)^2 + \lambda (x_1^2 + 2x_2^2 - 1) \longrightarrow \text{maximize}$$

differentiating w.r.to x_1 & x_2 & equating to 0

$$2(x_1 - 4) + 2\lambda x_1 = 0 \Rightarrow x_1 - 4 + \lambda x_1 = 0 \Rightarrow x_1 = \frac{4}{1+\lambda}$$

$$2(x_2 - 3) + 4\lambda x_2 = 0 \Rightarrow x_2 - 3 + 2\lambda x_2 = 0 \Rightarrow x_2 = \frac{3}{1+2\lambda}$$

Substituting in g

$$\left(\frac{4}{1+\lambda}\right)^2 + 2 \cdot \left(\frac{3}{1+2\lambda}\right)^2 - 1 = 0$$

$$16(1+2\lambda)^2 + 18(1+2\lambda)^2 - (1+\lambda)^2(1+2\lambda)^2 = 0$$

Solving above equation we get the value of λ

we plug this λ in values of x_1 & x_2 and end up with two linear equation's solving these two linear equations will give us the point (x_1, x_2)

Q2.

$x_1^0 = 0 \quad x_2^0 = 0 \quad \alpha = 0.1$

$x_1^1 = x_1^0 - \alpha \cdot \frac{\partial f}{\partial x_1} \qquad \frac{\partial f}{\partial x_1} = 4(x_1 + 2)$

$= 0 - 0.1 * 4(0+2)$

$= 0 - 0.1 * 8$

$= -0.8$

$x_2^1 = x_2^0 - \alpha \cdot \frac{\partial f}{\partial x_2} \qquad \frac{\partial f}{\partial x_2} = 6(x_2 - 2)$

$= 0 - 0.1 * 6(0-2)$

$= 0 - 0.1 * (-12)$

$= 1.2$

The next point for iteration is $(-0.8, 1.2)$

Q 3.1. If we know the dependence $y = f(x)$ is smooth then the values of y are close whenever the values of x are close

i.e $y_1 = y_0 \quad y_2 = y_1 \quad \cdots \quad y_n = y_{n-1}$

In other words we want all the differences $e_i = y_i - y_{i-1}$ to be close to 0

The square of distance between points is $\sum_{i=1}^{n} e_i^2$

i.e $(y_1 - y_0)^2 + (y_2 - y_1)^2 + \cdots (y_i - y_{i-1})^2 + \cdots (y_n - y_{n-1})^2 \quad - (1)$

We have to minimize equation (1)

Differentiating w.r.to y_i & equating to 0

$\frac{\partial}{\partial y_1} \Rightarrow 2(y_1 - y_0) + 2(y_2 - y_1)(-1) = 0 \Rightarrow y_1 - y_0 = y_2 - y_1$

$\frac{\partial}{\partial y_i} \Rightarrow 2(y_i - y_{i-1}) + 2(y_{i+1} - y_i)(-1) = 0 \Rightarrow y_i - y_{i-1} = y_{i+1} - y_i$

Q 3.1 continued

Thus we have

$y_1 - y_0 = y_2 - y_1 = y_3 - y_2 \cdots y_n - y_{n-1}$ i.e the difference between the two neighbouring points is same. let us denote this distance by Δy

Then we have

$$y_1 = y_0 + \Delta y$$

$$y_2 = y_1 + \Delta y = y_0 + 2\Delta y$$

$$\vdots$$

$$y_i = y_{i-1} + \Delta y = y_0 + i\,\Delta y \quad -(11)$$

from

$$x_i = x_0 + i.\Delta x$$

we have

$$i = \frac{x_i - x_0}{\Delta x}$$

Substituting in (11)

$$y_i = y_0 + \frac{x_i - x_0}{\Delta x} \cdot \Delta y$$

$$y_i = y_0 + \frac{\Delta y}{\Delta x} \cdot x_i - \frac{x_0}{\Delta x} \cdot \Delta y$$

$$\Rightarrow y_i = \frac{\Delta y}{\Delta x} \cdot x_i + \left(y_0 - \frac{\Delta y}{\Delta x} \cdot x_0\right)$$

This is eq form $y = ax + b$ which is a straight line

Q3.4 The main objective of traditional neural network is to make data processing faster thus we need many fast processing neurons working in parallel. In general linear neurons are faster compared to non-linear neurons thus we use the linear neuron's. However not all the processes in real world are linear thus we need to have a non-linear neuron as well and for a non-linear neuron's the more input's we have the longer it takes to process thus we use non-linear neuron's with single input.

3.7 Newton showed that white light (any light) can be represented as a combination of monochromatic signals. In math terms monochromatic signal is a sinusoid which means that every function $f(x)$ can be represented as a linear combination of sinusoids.

i.e $f(x) = \Sigma A_i \sin(w_i x) + B_i \cos(w_i x)$

further for function's of two variables

$f(x,y) = \Sigma A_i(y) \sin(w_i x) + B_i(y) \cos(w_i x)$

Again the function's $A_i(y)$ & $B_i(y)$ can also be written as

$$A_i(y) = \sum_k C_{ik} \sin(w_k y) + D_{ik} \cos(w_k y) \quad \text{and similarly for } B_i(y)$$

Thus the original function becomes

$$f(x,y) = \sum_i \left(\sum_k C_{ik} \sin(w_k y) + D_{ik} \cos(w_k y) \cdot \sin(w_i x) + \dots \right.$$

$$= \Sigma C_{ik} \sin(w_k y) \cdot \sin(w_i x) + D_{ik} \cos(w_k y) \cdot \sin(w_i x) + \dots$$

Using the formulas from trigonometry for $\sin a \sin b$ & $\sin a \cos b$ we have

$$f = \Sigma \left(C_{ik} \sin(w_k y + w_i x) + D_{ik} \sin(w_k y - w_i x) \right.$$

This is the same as function for 3-layer Neural Network i.e

$$f = \sum_k W_k S_k \left(\sum_i \omega_{ki} x_i - \omega_{k0} \right) - W_0$$

Thus we can say that 3-layer Neural Networks are universal approximators.

Q 3.10. We use gradient descent method to minimize the corresponding Objective function.

we start with $(x_1, \ldots x_n)$ and we want to find a point $x_i' = x_i + \Delta x_i$ in some small δ visinity of this point for which

$$f(x_1 + \Delta x_1, \ldots x_n + \Delta x_n) \simeq f(x_1, \ldots x_n) + \frac{\partial f}{\partial x_1} \Delta x_1 + \ldots \frac{\partial f}{\partial x_n} \Delta x_n$$

is smallest as possible

Thus we minimize above equation under constraint

$$(\Delta x_1)^2 + \ldots (\Delta x_n)^2 - \delta^2 = 0$$

Using lagrange multiplier method.

$$f(x_1, \ldots x_n) + \frac{\partial f}{\partial x_1} \Delta x_1 + \ldots \frac{\partial f}{\partial x_n} \Delta x_n + \lambda (\Delta x_1^2 + \ldots \Delta x_n^2 - \delta^2) \longrightarrow \min.$$

differentiating we get

$$\frac{\partial f}{\partial x_i} + 2\lambda \Delta x_i = 0$$

$$\Rightarrow \Delta x_i = -\alpha \frac{\partial f}{\partial x_i} \quad \text{where } \alpha = \frac{1}{2\lambda}$$ Thus from original point x_i we move to new point

$$x_i' = x_i - \alpha \frac{\partial f}{\partial x_i}$$ This is gradient descent.

84. We want the class of transformation to be Natural and linear transformations are Natural. $x \to \lambda x$, $x \to x + x_0$ $\Rightarrow$ $x \to ax + b$.

If f and g are Natural reasonable then their composition is also reasonable Natural.

We also want the family to be described by finitely many parameters

We use the theorem: Every reasonable transformation is fractional linear i.e $f(x) = \frac{ax+b}{cx+d}$

So we want a activation function that will not change if we change the starting point

$y = s(x)$ $\quad s(x+x_0) = f(s(x))$ for some reasonable transformation $f(z)$

$$s(x+x_0) = \frac{A(x_0) * s(x) + B(x_0)}{c(x_0) \cdot s(x) + d(x_0)}$$ dividing the numerator & denominator by $d(x_0)$

$$s(x+x_0) = \frac{a(x_0) \cdot s(x) + b(x_0)}{c(x_0) \cdot s(x) + 1}$$ This is a functional equation

Differentiating w.r. to x_0

$$s'(x+x_0) = \frac{\left(a'(x_0) \cdot s(x) + b'(x_0)\right)\{c(x_0) \cdot s(x) + 1\} - (a(x_0) \cdot s(x) + b(x_0) \cdot \{c'(x_0) \cdot s(x) + 0\}}{\{c(x_0) \cdot s(x) + 1\}^2}$$

Plug $x_0 = 0$

$$s'(x) = \frac{\left(a'(0) \cdot s(x) + b'(0)\right)\{c(0) \cdot s(x) + 1\} - (a(0) \cdot s(x) + b(0)\{c'(0) \cdot s(x)\}}{\{c(0) \cdot s(x) + 1\}^2}$$

Using $a(0)=1$, $b(0)=0$ & $c(0)=0$ we have

$$s'(x) = a'(0)\cdot s(x) + b'(0) - c'(0)\cdot s(x)^2$$

i.e of form

$$s'(x) = A\,s(x) - B\,s(x)^2 + C$$

$$\frac{d\,s(x)}{A\,s(x) - B\,s(x)^2 + C} = dx$$

Integrating we get the sigmoid function thus sigmoid function is used as activation function

Above integration is similar to

$$\int \frac{dy}{y^2 - a^2} = \int da$$

$$\frac{1}{2a}\left[\ln(y-a) - \ln(y+a)\right] = a$$

$$\ln\left(\frac{y+a}{y-a}\right) = -z$$

$$\frac{y+a}{y-a} = e^{-z}$$

$$\Rightarrow y = \frac{\text{constant}}{1+e^{-z}}$$ sigmoid.