(1)

$f \equiv (x_1-4)^2 + (x_2-3)^2 \to \min$

$g \equiv x_1^2 + 2x_2^2 = 1$

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

$(x_1-4)^2 + (x_2-3)^2 + \lambda(x_1^2 + 2x_2^2 - 1) \to \min$

$(x_1-4)^2 + (x_2-3)^2 + \lambda x_1^2 + 2x_2^2\lambda - \lambda$

~~$2(x_1-4) + \lambda(x_1^2 + 2x_2^2 + 0)(2x_1)$~~

~~$2x_1 - 8 + x_1^2\lambda + 2x_2^2\lambda$~~

~~$2(x_1-4)$~~

$\frac{\partial}{\partial x_1}: \quad 2(x_1-4) + 2\lambda x_1 = 0$

$2x_1 - 8 + 2\lambda x_1 = 0$

$2x_1 + 2\lambda x_1 = 8$

$x_1(2 + 2\lambda) = 8$

$x_1 = \frac{8}{2+2\lambda}$

$\frac{\partial}{\partial x_2}: \quad 2(x_2-3) + 4x_2\lambda = 0$

$2x_2 - 6 + 4x_2\lambda = 0$

$2x_2 + 4x_2\lambda = 6$

$x_2(2 + 4\lambda) = 6$

$x_2 = \frac{6}{2+4\lambda}$

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

$x_1^2 + 2x_2^2 = 1$

~~$\frac{64}{(2+2\lambda)^2} - \frac{72}{(2+4\lambda)^2}$~~

$\left(\frac{8}{2+2\lambda}\right)^2 + 2\left(\frac{6}{2+4\lambda}\right)^2 = 1$

$\frac{64}{4(2+2\lambda)^2} + 2\left(\frac{36}{(2+4\lambda)^2}\right) = 1$

$\frac{64}{(2+2\lambda)^2} + \frac{72}{(2+4\lambda)^2} = 1$

~~$\frac{(2+4\lambda)^2}{72} \cdot \frac{64}{(2+2\lambda)^2} = \frac{(2+4\lambda)^2}{72}$~~

~~$64 \cdot (2+4\lambda)^2 + 72 \cdot (2+2\lambda)^2 = (2+2\lambda)^2(2+4\lambda)^2$~~

$$\frac{64}{24+8\lambda+4\lambda^2} + \frac{12}{4+16\lambda+16\lambda^2} = 1$$

$$16(4+16\lambda+16\lambda^2) + 72(4+8\lambda+4\lambda^2) = (4+16\lambda+16\lambda^2)(4+8\lambda+4\lambda^2)$$

$$64 + 256\lambda + 256\lambda^2 + 288 + 576\lambda + 288\lambda^2$$

$$\frac{304 + 632\lambda + 544\lambda^2}{4 + 8\lambda + 4\lambda^2} = \frac{16+32\lambda+16\lambda^2+64\lambda+128\lambda^2}{4+16\lambda+16\lambda^2}$$

$$76 + 79 + 136 = 4 + 16\lambda + 16\lambda^2$$

$$291 = 4 + 16\lambda + 16\lambda^2$$

$$287 = 16\lambda + 16\lambda^2$$

$$287 = 16\lambda(1+\lambda^2) \quad 16(\lambda+\lambda^2)$$

$$18 \approx \lambda + \lambda^2$$

$$291 = 4(1+4\lambda+4\lambda^2)$$

$$72.75 = 4\lambda^2 + 4\lambda + 1$$

$$287 = 16\lambda + 16\lambda^2$$

$$17.94 = \lambda + \lambda^2$$

$$17.94 = \lambda(1+1^2)$$

$$\frac{17.94}{\lambda} = 1 + 1^2$$

$$\frac{17.94}{\lambda} = 2$$

$$17.94 = 2\lambda$$

$$\lambda = 8.97$$

$$x_1 = \frac{8}{2+2\lambda} = \frac{8}{2+2(8.97)}$$

$$= \frac{8}{19.94} = .4$$

$$x_2 = \frac{6}{2+4\lambda} = \frac{6}{37.88} \approx .158$$

(2) $\alpha = .1 \quad x_1^{(0)} = 0 \quad x_2^{(0)} = 0$

$F = 2(x_1+2)^2 + 3(x_2-2)^2$

$$x_1 = x_1^{(0)} - \alpha \cdot \frac{\partial f}{\partial x_1} = x_1^{(0)} - .1\left(4(x_1^{(0)}+2)\right) = 0 - .1(8) = 0 - .8 = -.8$$

$$x_2 = x_2^{(0)} - \alpha \cdot \frac{\partial f}{\partial x_2} = x_2^{(0)} - .1\left(6(x_2^{(0)}-2)\right) = 0 - .1(-12) = 0 + 1.2 = 1.2$$

$$2(-.8+2)^2 + 3(1.2-2)^2$$

$$= 2.88 + 1.92$$

$$= 4.8$$

$x_1^{(0)} = -.8 \quad x_2^{(0)} = 1.2$

$$x_1^{(0)} - .1\left(4(x_1^{(0)}+2)\right) = -.8 - .1\left(4(-.8+2)\right) = -.8 - .48$$

$$= -1.28$$

$$x_2^{(0)} - .1\left(6(x_2^{(0)}-2)\right) = 1.2 - .1\left(6(1.2-2)\right) = 1.2 + .48 = 1.68$$

$$2(-1.28+2)^2 + 3(1.68-2)^2$$

$$= 1.0368 + .3072$$

$$= 1.344$$

(3.3) What is the main objective of the traditional NN?

Speed -> get all computations of f as fast as possible

(3.6) Why use L-NL-L networks not NL-L-NL ones?

L is faster than NL. If L-NL-L has 2 Ls which NL-L-NL has 1 L, then having two Ls is faster than having one L. Therefore we use L-NL-L rather than NL-L-NL.

3.9 What do we want to minimize when we train a neural network?

mean ~~square~~ square approximation error

3.12 Explain how gradient descent leads to backpropagation formulas for $w_0 \dots w_k$.

given: $x_1, \dots, x_n, \tilde{y}$ we want to find w_{ki}, w_k

we use gradient descent $w \to w - \alpha \cdot \frac{\partial J}{\partial w}$

$$\frac{\partial J}{\partial w} = 2(y - \tilde{y}) \cdot \frac{\partial y}{\partial w} = 2e \cdot \frac{\partial y}{\partial w} \to w \to w - 2\alpha e \cdot \frac{\partial y}{\partial w}$$

$$\frac{\partial y}{\partial w_0} = -1 \qquad \Delta w = -2\alpha \cdot e \frac{\partial y}{\partial w}$$

$$\Delta w_0 = -2\alpha e \cdot (-1) = 2\alpha e$$

$$\frac{\partial y}{\partial w_k} \qquad \Delta w_k = -2\alpha e \cdot y_k = -\Delta w_0 \cdot y_k$$

$$\frac{\partial y}{\partial w_{ki}} = w_k \cdot s'(w_k, x_1 + \dots) \cdot (-1)$$

$$s'(z) = \frac{0 - 1 \cdot (0 + e^{-z}(-1))}{(1+e^{-z})^2} = \frac{e^{-z}}{(1+e^{-z})^2} = \frac{e^{-z}}{1+e^{-z}} = \frac{1}{1+e^{-z}}$$

$$1 - y_k = 1 - \frac{1}{1+e^{-z}} = \frac{e^{-z}}{1+e^{-z}}$$

$$s'(z) = s \cdot (1 - s)$$

$$\frac{\partial s}{\partial z} = s - s^2$$

$$\frac{\partial y}{\partial w_{k0}} = w_k \cdot s'(w_{k1} x_1 + \dots) \cdot (-1)$$

$$-w_k \cdot y_k (1 - y_k)$$

$$\Delta w_{k0} = -2\alpha e (-w_k) \cdot y_k (1 - y_k)$$

$$= 2\alpha e w_k y_k (1 - y_k)$$

$$\Delta w_{ki} = -2\alpha e \cdot (-w_k) \cdot y_k (1 - y_k) x_i$$

$$\Delta w_{k0} = 2\alpha \cdot e \cdot w_k \cdot y_k (1 - y_k)$$

$\Delta w_{ki} = -\Delta w_{k0} \cdot x_i$

$\Delta w_{k0} = -\Delta w_k \cdot w_k \cdot (1 - y$

$\Delta w_k = -\Delta w_0 \cdot y_k$

$\Delta w_0 = 2\alpha \cdot e$

(3.2) What is ML?

we know: $x^{(k)} = (x_1^{(k)}, \ldots, x_n^{(k)})$ and $y^{(k)}$

$$1 \le k \le K$$

we want: a function $f(x_1, \ldots, x_n)$

such that for all k,

$$f(x_1^{(k)}, \ldots, x_n^{(k)}) \approx y^{(k)}$$

(3.4) Why we only use linear & 1-input nonlinear neurons?

linear is faster but we need nonlinear since not all dependencies are linear. The more inputs we have the longer it takes to ~~process~~ process.

(3.5) Why cannot we ~~use~~ have 1- or 2-layer neural networks?

With 1 layer we only get linear functions. Why Not 2??

With 1 nonlinear layer we only get functions of one variable

(3.8) Which activation function is used in traditional NN? Why?

We use sigmoid function because biological neurons use $s_0(z) = \dfrac{1}{1 + \exp(-z)}$ as well.

(3.10) What method do we use to minimize the corresponding objective function? gradient descent.

(3.1) Proof smoothest line connecting two pts is a straight line

$$(y_1-y_0)^2 + (y_2-y_1)^2 + \dots + (y_i-y_{i-1})^2 + (y_{i+1}-y_i)^2 + \dots + (y_n-y_{n-1})^2$$

$$2(y_1-y_0) + 2(y_2-y_1)\cdot(-1) = 0$$

$$y_1 - y_0 = y_2 - y_1$$

$$2(y_i-y_{i-1}) + 2(y_{i+1}-y_i)\cdot(-1) = 0$$

$$y_i - y_{i-1} = y_{i+1} - y_i$$

$i = \dfrac{x_1 - x_0}{\Delta x}$

therefore $y_1 - y_0 = y_2 - y_1 = y_3 - y_2 = \dots$

$$y_i = y_0 + \frac{x_1 - x_0}{\Delta x}\cdot\Delta y = x_i\cdot\frac{\Delta y}{\Delta x} + \left(y_0 - x_0\cdot\frac{\Delta y}{\Delta x}\right)$$

i.e. $y = a\cdot x + b$

$$a = \frac{\Delta y}{\Delta x} \quad ; \quad y_0 - x_0\cdot\frac{\Delta y}{\Delta x}$$

Extra Credit: Why Sigmoid?

$$s(x+x_0) = \frac{a(x_0)\cdot s(x) + b(x_0)}{c(x_0)\cdot s(x) + d(x_0)}$$

$x^k \in L \quad \dfrac{(x+h)^k - x^k}{h} = k\cdot x^{k-1}$

then $x^{k-1} \in L,\ x^{k-2} \in L, \dots$

$x^2 \in L$

$\{x^k, x^2\} = (x^k)'\cdot x^2 - x^k\cdot(x^2)'$

$= k\cdot x^{k-1}\cdot x^2 - x^k\cdot 2x = k\cdot x^{k+1} - 2x^{k+1}$

$= (k-2)\cdot x^{k+1}$

if for some $k>2$, $x^k \in L$, then $x^{k+1} \in L$,

$x^{k+2} \in L \dots$

So we get $a_k x^k + a_{k+1} x^{k+1} + \dots$

$a_k x^k + a_{k+1} x^{k+1} + \dots \in L$

then $a_k x^k \in L$

$a_{k+1} x^{k+1} \in L$