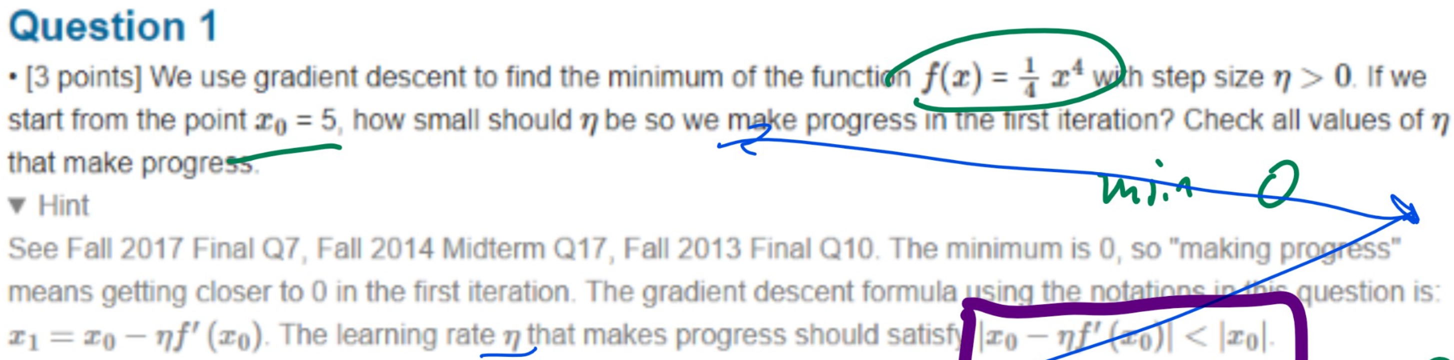
M2Q1, M2Q2



Question 1

- [3 points] We use gradient descent to find the minimum of the function $f(x) = \frac{1}{4}\,x^4$ with step size $\eta > 0$. If we start from the point $x_0 = 5$, how small should η be so we make progress in the first iteration? Check all values of η that make progress.

▼ Hint

See Fall 2017 Final Q7, Fall 2014 Midterm Q17, Fall 2013 Final Q10. The minimum is 0, so "making progress" means getting closer to 0 in the first iteration. The gradient descent formula using the notations in this question is: $x_1 = x_0 - \eta f'(x_0)$. The learning rate η that makes progress should satisfy $|x_0 - \eta f'(x_0)| < |x_0|$.

- Choices:

☐ 0.1544

☐ 0.0488

☐ 0.0563

☐ 0.0438

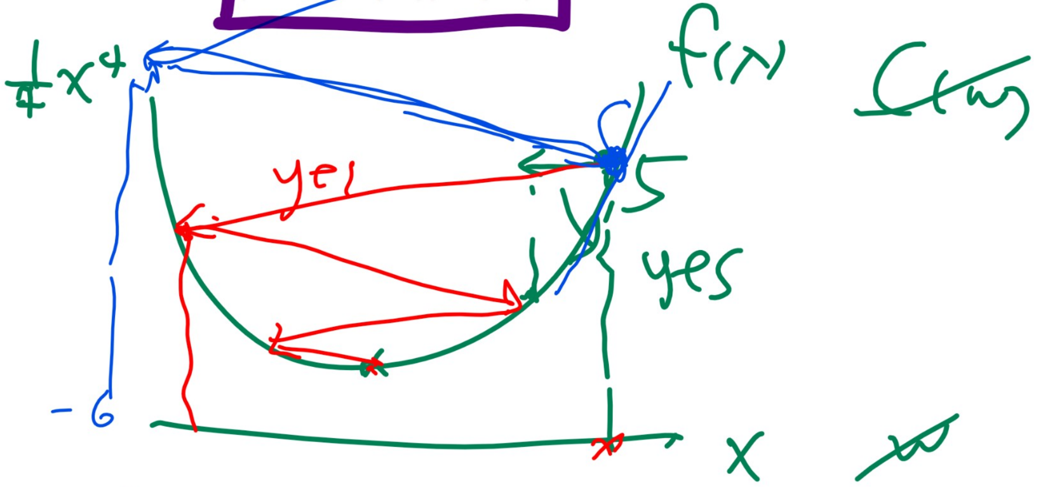
☐ 0.0742

☐ None of the above

- Calculator: ☐ Calculate

min 0

not making progress

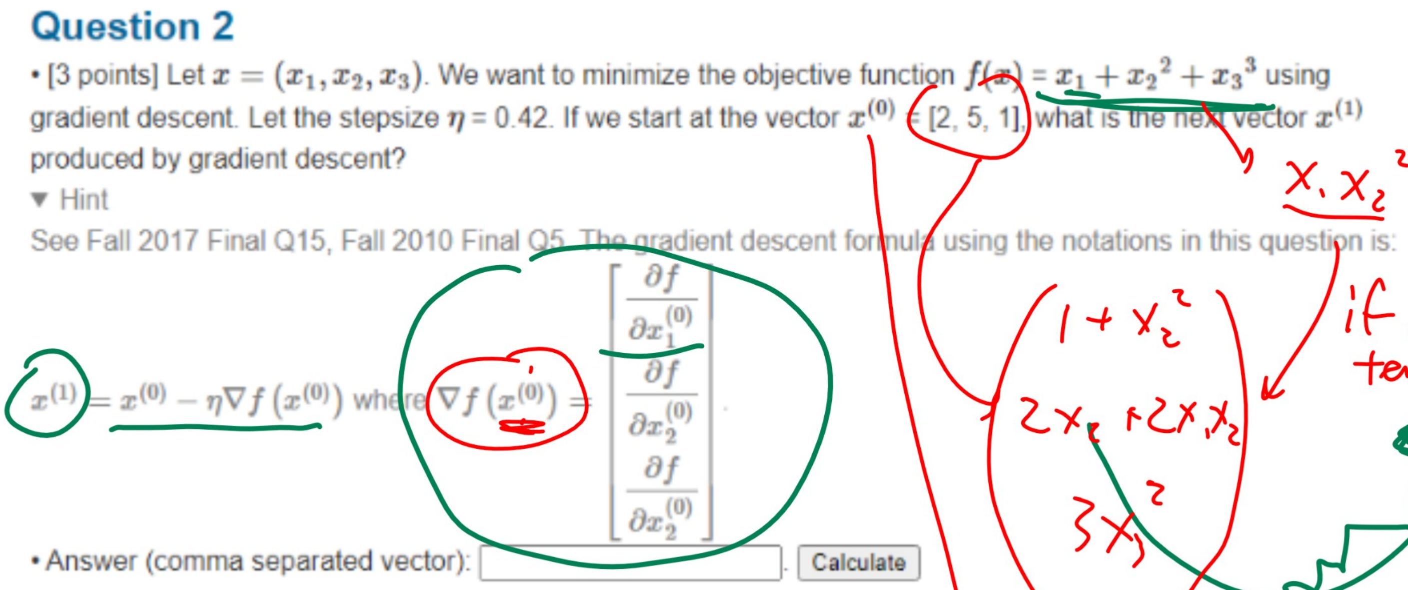


$$|5 - \eta\, \overset{5^3}{\cancel{x^3}}\,| < |5|$$

$$\begin{cases} 5 - \eta 5^3 < 5 \\ \eta 5^3 - 5 < 5 \end{cases} \Rightarrow \quad \eta < \underline{\qquad}$$

Converge getting closer to local min. as $t \to \infty$

M2Q2



Question 2

• [3 points] Let $x = (x_1, x_2, x_3)$. We want to minimize the objective function $f(x) = x_1 + {x_2}^2 + {x_3}^3$ using gradient descent. Let the stepsize η = 0.42. If we start at the vector $x^{(0)}$ = [2, 5, 1], what is the next vector $x^{(1)}$ produced by gradient descent?

▼ Hint

See Fall 2017 Final Q15, Fall 2010 Final Q5. The gradient descent formula using the notations in this question is:

$$x^{(1)} = x^{(0)} - \eta \nabla f\left(x^{(0)}\right) \text{ where } \nabla f\left(x^{(0)}\right) = \begin{bmatrix} \dfrac{\partial f}{\partial x_1^{(0)}} \\ \dfrac{\partial f}{\partial x_2^{(0)}} \\ \dfrac{\partial f}{\partial x_2^{(0)}} \end{bmatrix}.$$

• Answer (comma separated vector): [ ] Calculate

$x_1 x_2^2$

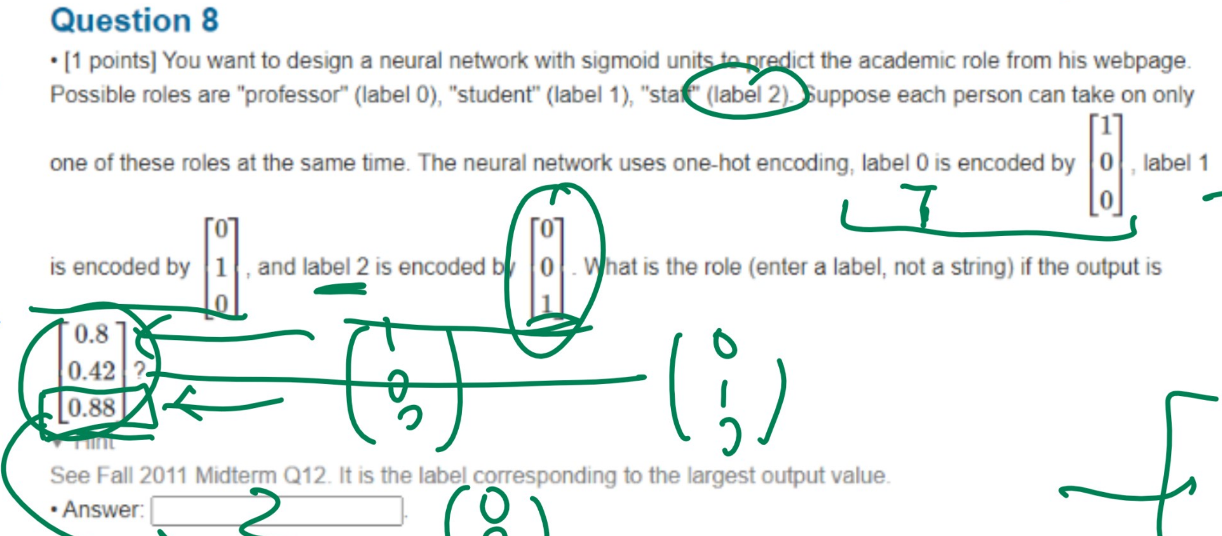
$$\begin{pmatrix} 1 + x_2^2 \\ 2x_2 + 2x_1x_2 \\ 3x_3^2 \end{pmatrix}$$

if add term

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{(0)} = \begin{pmatrix} 2 \\ 5 \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} 2 \\ 5 \\ 1 \end{pmatrix} - 0.42 \begin{pmatrix} 1 \\ 2\cancel{x_2}\ 5 \\ 3\cancel{x_3}^2\ 1 \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 10 \\ 3 \end{pmatrix}$$

M3 Q8



Question 8

• [1 points] You want to design a neural network with sigmoid units to predict the academic role from his webpage. Possible roles are "professor" (label 0), "student" (label 1), "staff" (label 2). Suppose each person can take on only one of these roles at the same time. The neural network uses one-hot encoding, label 0 is encoded by $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$, label 1 is encoded by $\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$, and label 2 is encoded by $\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$. What is the role (enter a label, not a string) if the output is $\begin{bmatrix} 0.8 \\ 0.42 \\ 0.88 \end{bmatrix}$?

▼ Hint

See Fall 2011 Midterm Q12. It is the label corresponding to the largest output value.

• Answer: [ ] 2

$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ $\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$ $\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$

$a_i = [0, 1]$

$a_i \geq 0.5 \Rightarrow \hat{y} = 1$

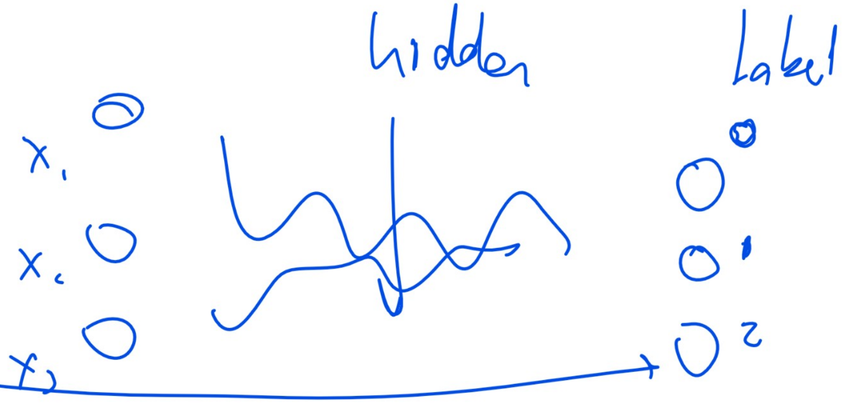
$a_i < 0.5 \Rightarrow \hat{y} = 0$

$\underbrace{a_i}_{\Pr\ 1}, \quad \underbrace{1 - a_i}_{\Pr\ 0}$

Softmax → $$\underbrace{\frac{e^{-z_1}}{e^{-z_1}+e^{-z_2}+e^{-z_3}} \quad \frac{e^{-z_2}}{\sim\sim} \quad \frac{e^{-z_3}}{\sim\sim}}_{\text{Sum up to } \underline{1}}$$

(X_1, y_1) (X_2, y_2)

$\begin{pmatrix} 5 \\ 2 \\ 1 \end{pmatrix}, 2$ $\begin{pmatrix} 3 \\ 3 \\ 3 \end{pmatrix}, 0$



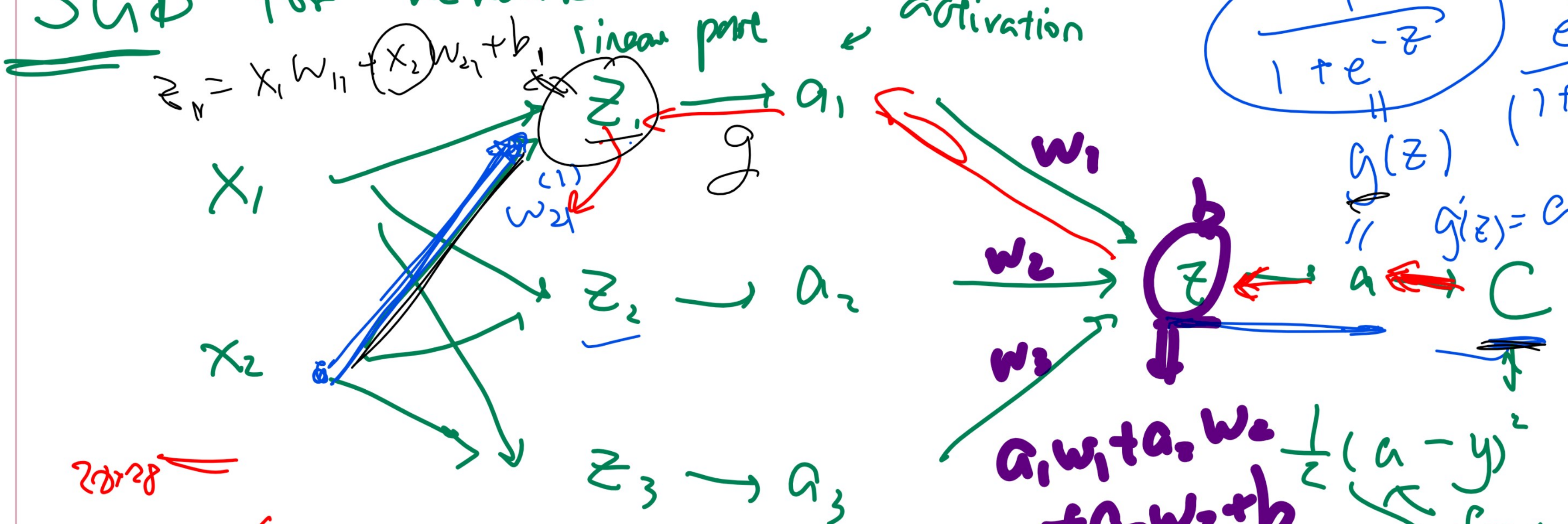
$\left[\begin{pmatrix} 5 \\ 2 \\ 1 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 \\ 3 \\ 3 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \right]$

$C = (y - a)^2$

$C = \sum_{j=1}^{3} \left(\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} - \begin{pmatrix} 0.5 \\ 0.2 \\ 0.3 \end{pmatrix} \right)^2$

SGD for network

$z_1 = x_1 w_{11} + x_2 w_{21} + b_1$



$$\frac{1}{1+e^{-z}} = g(z) = a$$

$$\frac{e^{-z}}{(1+e^{-z})^2} \Downarrow g'(z) = a(1-a)$$

$$\underset{\text{weights in L1}}{6} + \underset{\text{biases}}{3} + \underset{\text{ws in L}}{3} + \underset{\text{bias}}{1} \Longrightarrow 13 \text{ parameters.}$$

(0,1) (0.2) (-1, 1)

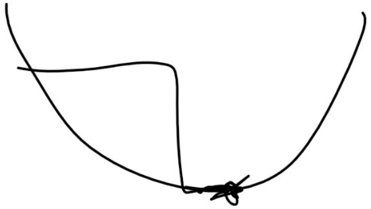
start

$w \in [-1, 1]$ random $w = [\text{Math.random}() * 2 - 1]$

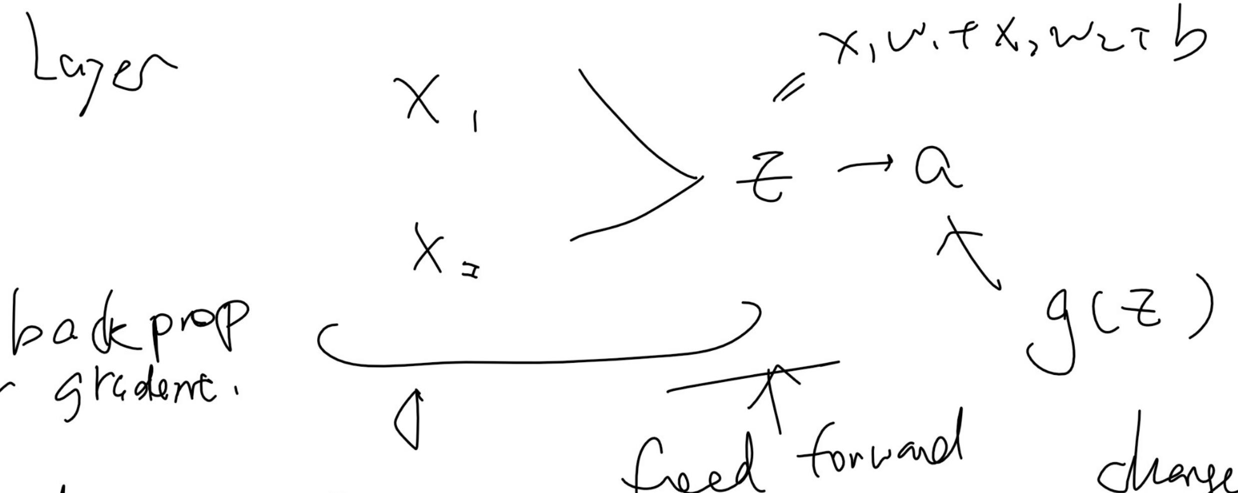
back prop

$$w_{21} = w_{21} - \alpha \boxed{\frac{\partial C}{\partial w_{21}}}$$

$$\underbrace{\frac{\partial C}{\partial a}}_{(a-y)} \underbrace{\frac{\partial a}{\partial z}}_{a(1-a)} \underbrace{\frac{\partial z}{\partial a_1}}_{w_1} \cdot \underbrace{\frac{\partial a_1}{\partial z_1}}_{a_1(1-a_1)} \underbrace{\frac{\partial z_1}{\partial w_{21}}}_{x_2}$$



Single Layer



update

w_{11}

w_{21}

w_{12}

w_{22}

w_1

w_2

w_{11} from $t-1$ $\Longrightarrow$ all a, a_1, ... change each iteration / epoch.

update at the same time.

x_i y_i stay same