ECE/CS 559 - Fall 2017 - Midterm Answers.

Full Name: **ID Number:**

Q1 (25 pts). Let $f(x) = x^2$.

(a) **(3 pts)** Let $x^\star$ be the global minimizer of f, i.e. $x^\star = \min_{x\in\mathbb{R}} f(x)$, where $\mathbb{R}$ is the set of real numbers. Find $x^\star$.

Solution: $x^\star = 0$.

(b) **(5 pts)** Recall that the gradient descent equations are given by $x_{n+1} = x_n - \eta f'(x_n)$, $n \in \{0, 1, 2, \ldots\}$, where $\eta > 0$ is the learning parameter, and $f'(x)$ is the derivative of $f(x)$ with respect to x. Find x_n for any $n \geq 1$ given $x_0 = 2$ and $\eta = \frac{1}{4}$.

Solution: $f'(x) = 2x$, so that the update equation is $x_{n+1} = x_n - 2\eta x_n = x_n(1 - 2\eta)$. In particular, for $\eta = \frac{1}{4}$, we obtain $x_{n+1} = \frac{x_n}{2}$. For $x_0 = 2$, this yields $x_1 = 1$, $x_2 = \frac{1}{2}$, or, in general, $x_n = 2^{-n+1}$.

(c) **(11 pts)** Describe the $n \to \infty$ asymptotic behavior of x_n and $f(x_n)$ for every possible initial condition $x_0 \in \mathbb{R}$ and learning parameter $\eta > 0$. For example, your answer should be able to describe where x_n and $f(x_n)$ go as $n \to \infty$ given initial conditions $x_0 = -2444$ and $\eta = 120$ (or any other x_0 and η that will be given to you).

Solution: From (b), we have $x_{n+1} = (1 - 2\eta)^{n+1} x_0$. First, if $x_0 = 0$, we have $x_n = 0$, $\forall n$ regardless of how η is chosen. Otherwise, we can observe that if $|1 - 2\eta| < 1$ (or, equivalently, if $0 < \eta < 1$), x_n converges to 0 regardless of what x_0 is. If $1 - 2\eta = 1$ (or, equivalently, if $\eta = 0$), we have $x_n = x_0$, $\forall n$. If $1 - 2\eta = -1$ (or, equivalently, if $\eta = 1$), x_n will oscillate between x_0 and $-x_0$. If $1 - 2\eta < -1$ (or, equivalently, if $\eta > 1$), x_n will oscillate between $+\infty$ and $-\infty$ as $n \to \infty$. If $1 - 2\eta > 1$ (or, equivalently, if $\eta < 0$), x_n will diverge to either $+\infty$ or $-\infty$, depending on the sign of x_0. This is a complete solution for any $\eta \in \mathbb{R}$; it is OK if you only consider $\eta > 0$.

Not Covered

(d) **(6 pts)** Let $k(y, z) = y^2 + z^4$. Let $g(y, z) \triangleq \begin{bmatrix} \frac{\partial k}{\partial y} \\ \frac{\partial k}{\partial z} \end{bmatrix}$ and $H(y, z) \triangleq \begin{bmatrix} \frac{\partial^2 k}{\partial y^2} & \frac{\partial^2 k}{\partial y \partial z} \\ \frac{\partial^2 k}{\partial z \partial y} & \frac{\partial^2 k}{\partial z^2} \end{bmatrix}$ denote the gradient and the Hessian of k, respectively. Recall that the update equations for Newton's method are given by $\begin{bmatrix} y_{n+1} \\ z_{n+1} \end{bmatrix} = \begin{bmatrix} y_n \\ z_n \end{bmatrix} - \eta \left(H(y_n, z_n)\right)^{-1} g(y_n, z_n)$, $n \in \{0, 1, 2, \ldots\}$. Given $y_0 = z_0 = \eta = 1$, calculate y_n, z_n for every $n \geq 1$.

Solution: By the definitions, we obtain $g(y, z) = \begin{bmatrix} 2y \\ 4z^3 \end{bmatrix}$ and $H(y, z) = \begin{bmatrix} 2 & 0 \\ 0 & 12z^2 \end{bmatrix}$ so that $(H(y, z))^{-1} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{12z^2} \end{bmatrix}$. For $\eta = 1$, this gives us the update equations $\begin{bmatrix} y_{n+1} \\ z_{n+1} \end{bmatrix} = \begin{bmatrix} y_n \\ z_n \end{bmatrix} - \begin{bmatrix} y_n \\ \frac{z_n}{3} \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{2}{3} z_n \end{bmatrix}$, $n \in \{0, 1, 2, \ldots\}$. Therefore, $y_n = 0$, and $z_n = (\frac{2}{3})^n$.

Q2 (25 pts). Consider a neuron with $n \geq 1$ inputs $x_1, \ldots, x_n$, and output $y = \theta(w_0 + w_1 x_1 + \cdots + w_n x_n)$, where $w_0, w_1, \ldots, w_n$ are the neuron bias and weights, and the activation function is given by $\theta(x) = 1$ if $x \in [0, 1]$, and $\theta(x) = 0$ if $x \notin [0, 1]$. Note that the activation function is different than the functions that we have encountered throughout the lectures.

(a) **(9 pts):** Let $n = 1$. Does there exist w_0, w_1 such that $y = 1 - x_1$ for $x_1 \in \{0, 1\}$? In other words, can a single neuron with activation function θ implement the `NOT` gate? If your answer is "Yes," find specific w_0, w_1 such that the neuron implements the `NOT` gate. If your answer is "No," prove that no choice for w_0, w_1 can result in a neuron that implements the `NOT` gate.

Solution: Yes. Geometrically, in the 2D plane, the activation function provides an output of 1 on a strip whose width and orientation you can adjust by adjusting the weights and the bias. With this observation, it is immediate that any gate of 2 inputs is, in fact, implementable. In particular, $w_0 = \frac{1}{2}$ and $w_1 = -1$ can implement the NOT gate. Note that these are the same weights that we have chosen in class for the step activation function.

(b) **(8 pts):** Let $n = 2$. Does there exist w_0, w_1, w_2 such that $y = x_1 x_2$ for $x_1, x_2 \in \{0, 1\}$? In other words, can a single neuron with activation function θ implement the `AND` gate? Justify your answer as in (a).
Solution: Yes. The choices $w_0 = -\frac{3}{2}$ and $w_1 = w_2 = 1$ can implement the AND gate. Note that these are the same weights that we have chosen in class for the step activation function.

(c) **(8 pts):** Let $n = 2$. Does there exist w_0, w_1, w_2 such that $y = ((x_1 + x_2) \mod 2)$ for $x_1, x_2 \in \{0, 1\}$? In other words, can a single neuron with activation function θ implement the `XOR` gate? Justify your answer as in (a).

Solution: Yes. One solution is $w_0 = -\frac{1}{2}$, $w_1 = w_2 = 1$.

Q3 (25 pts). Let u be the step activation function with $u(x) = 1$ if $x \geq 0$, and $u(x) = 0$, otherwise. Consider the perceptron $y = u(w_0 + w_1 x_1 + w_2 x_2)$, where w_1 and w_2 are the weights for inputs x_1 and x_2, respectively, w_0 is the perceptron bias, and y is the perceptron output. Let $\mathcal{C}_0 = \left\{ \begin{bmatrix} 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \end{bmatrix} \right\}$, and $\mathcal{C}_1 = \left\{ \begin{bmatrix} 1 & 1 \end{bmatrix} \right\}$. The desired output for class $\mathcal{C}_0$ is 0, and the desired output for class $\mathcal{C}_1$ is 1. Correspondingly, let $d(\mathbf{x}) = 0$ if $\mathbf{x} \in \mathcal{C}_0$, and otherwise, let $d(\mathbf{x}) = 1$ if $\mathbf{x} \in \mathcal{C}_1$.

(a) **(8 pts)** If possible, find w_0, w_1, w_2 that can separate $\mathcal{C}_0$ and $\mathcal{C}_1$ (i.e., provide the desired output for all 4 possible input vectors). Otherwise, prove that no choice of weights can separate the two classes.

Solution: This is equivalent to designing the AND gate. So, $w_0 = -\frac{3}{2}$, $w_1 = w_2 = 1$ will work.

(b) **(10 pts)** Recall that the perceptron training algorithm relies on the update $\mathbf{w} \leftarrow \mathbf{w} + \eta(d(\mathbf{x}) - y) \begin{bmatrix} 1 & \mathbf{x} \end{bmatrix}$, where $\mathbf{w} = \begin{bmatrix} w_0 & w_1 & w_2 \end{bmatrix}$ is the weight vector. Let $\eta = 1$ and the initial weight vector be given by $\mathbf{w} = \begin{bmatrix} -0.5 & 1 & 0 \end{bmatrix}$. Calculate the updated weights after two epochs of training.

Solution: Straightforward calculation. The final weights you get should be $\begin{bmatrix} -0.5 & 2 & 1 \end{bmatrix}$.

(c) **(7 pts)** Will the weights provided by the algorithm (as setup in (b)) eventually converge after a sufficiently larger number of epochs? Justify your answer.

Solution: Yes. The patterns are linearly separable so the PTA will converge.

Q4 (15 pts). Let

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \phi \left(\begin{bmatrix} w_{10} & w_{11} & w_{12} \\ w_{20} & w_{21} & w_{22} \end{bmatrix} \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix} \right), \text{ and } \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \phi \left(\begin{bmatrix} u_{10} & u_{11} & u_{12} \\ u_{20} & u_{21} & u_{22} \end{bmatrix} \begin{bmatrix} 1 \\ y_1 \\ y_2 \end{bmatrix} \right) \tag{1}$$

with the understanding that the activation function ϕ is applied component-wise. These equations define a two-layer neural network with 2 input nodes, 2 neurons in the hidden layer, and 2 output nodes.

(a) **(5 pts)** Draw the block diagram of the neural network with all inputs, outputs, weights labeled.

Solution: Left as exercise.

(b) **(10 pts)** Let $E = (d_1 - z_1)^2 + (d_2 - z_2)^4 + u_{22}^2$. Write down the expressions for $\frac{\partial E}{\partial w_{10}}$ and $\frac{\partial E}{\partial u_{22}}$. You may use the backpropagation algorithm. Your expressions may contain intermediate variables that you shall clearly define on the feedforward/feedback graphs.

Solution: Left as exercise.

Q5 (10 pts). Consider the activation function $\phi(v) = \frac{v}{1+|v|}$ defined for all real numbers.

(a) **(5 pts)** Find $\phi'(v) = \frac{\partial \phi}{\partial v}$.

Solution: For $v \geq 0$, we have $\phi(v) = \frac{v}{1+v}$ so that $\phi'(v) = \frac{1}{(1+v)^2}$. For $v \leq 0$, we have $\phi(v) = \frac{v}{1-v}$ so that $\phi'(v) = \frac{1}{(1-v)^2}$. Hence, for any v, we have $\phi'(v) = \frac{1}{(1+|v|)^2}$.

(b) **(5 pts)** Express $\phi'(v)$ in terms of $\phi(v)$ only.

Solution (by Aria Ameri, my solution was much less neater): We have $|\phi(v)| = \frac{|v|}{1+|v|}$. Negating both sides and adding 1, we get $1 - |\phi(v)| = \frac{1}{1+|v|}$. Therefore, $\phi'(v) = \frac{1}{(1+|v|)^2} = (1 - |\phi(v)|)^2$!

Ironic that this turned out to be toughest question of the exam with the worst overall performance. Only one person got this right. Some of you found essentially the same result, but you had conditioned on the negativity of positivity of v (Hence your expressions did not depend on $\phi(v)$ only).