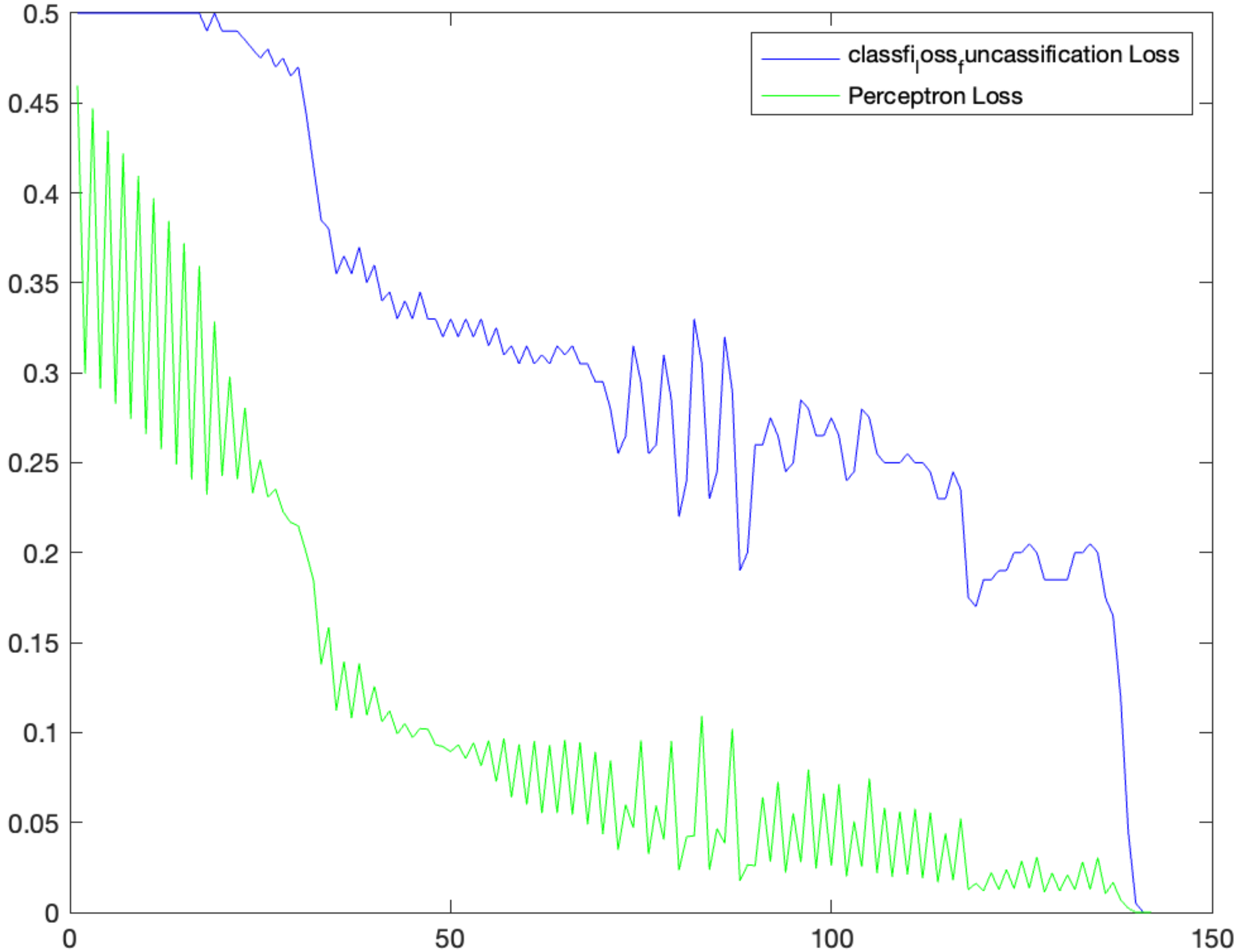
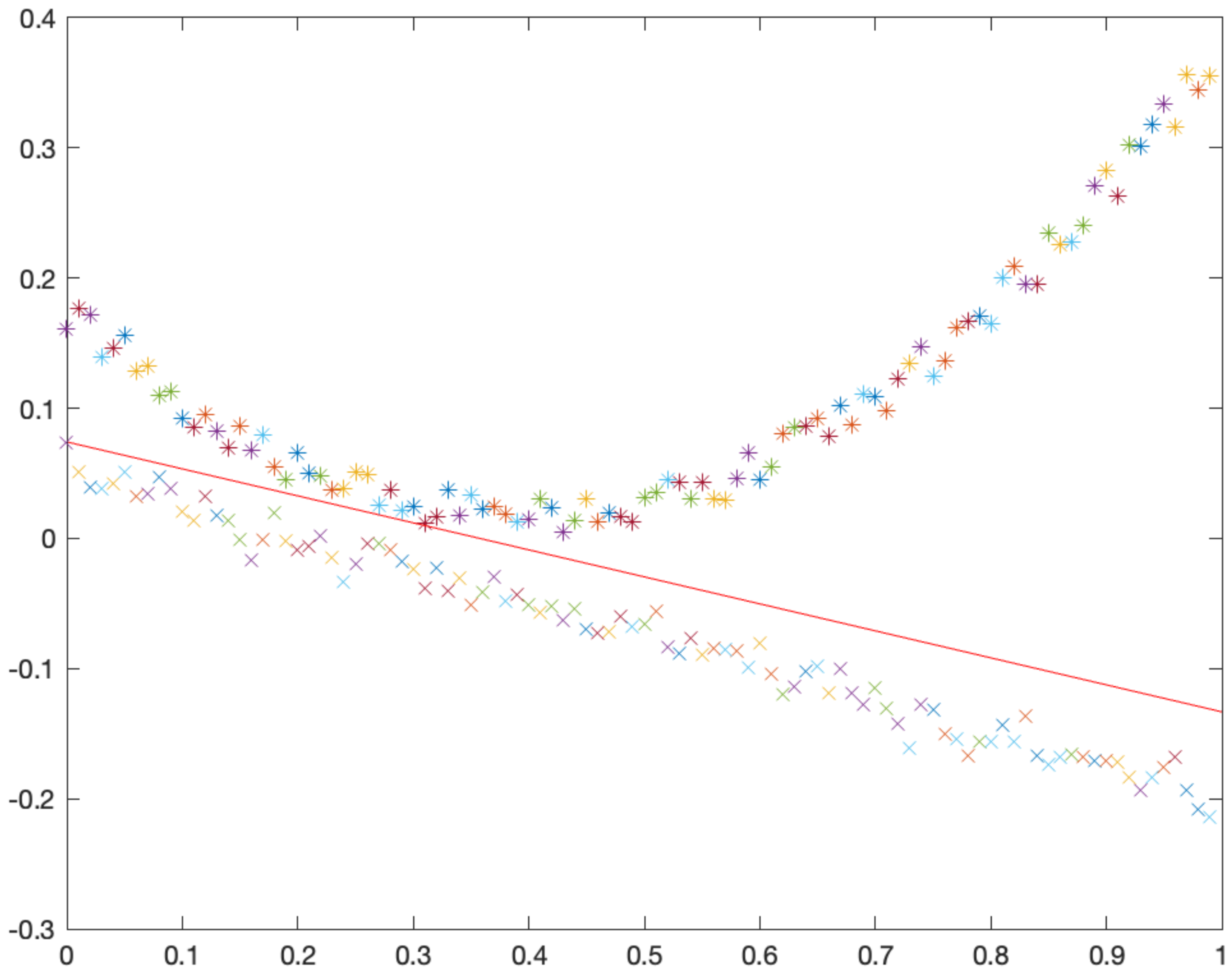
Jincheng Tian machine learning hw2

Problem1:

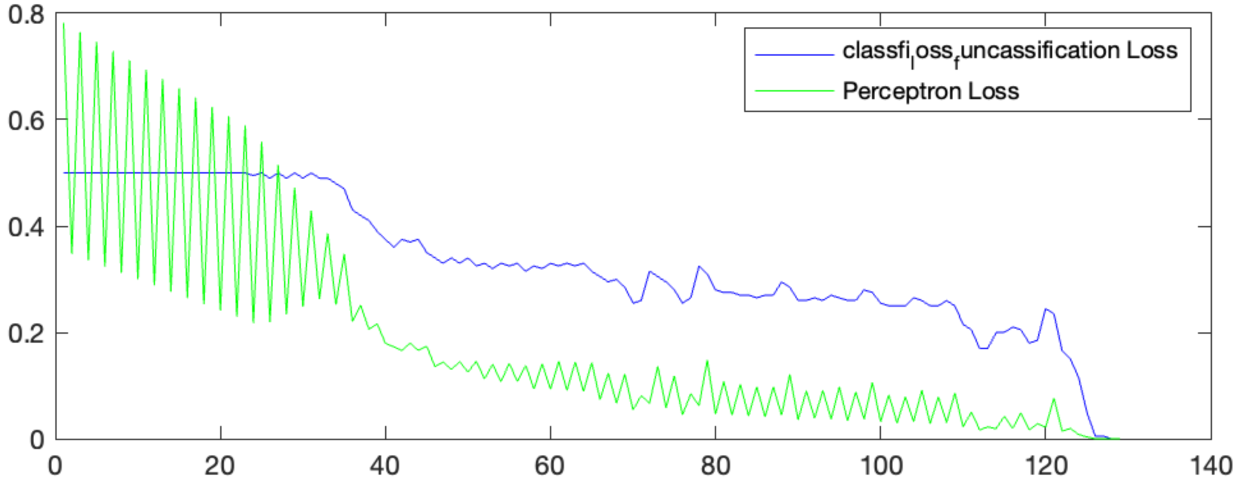
Step size = 2.5

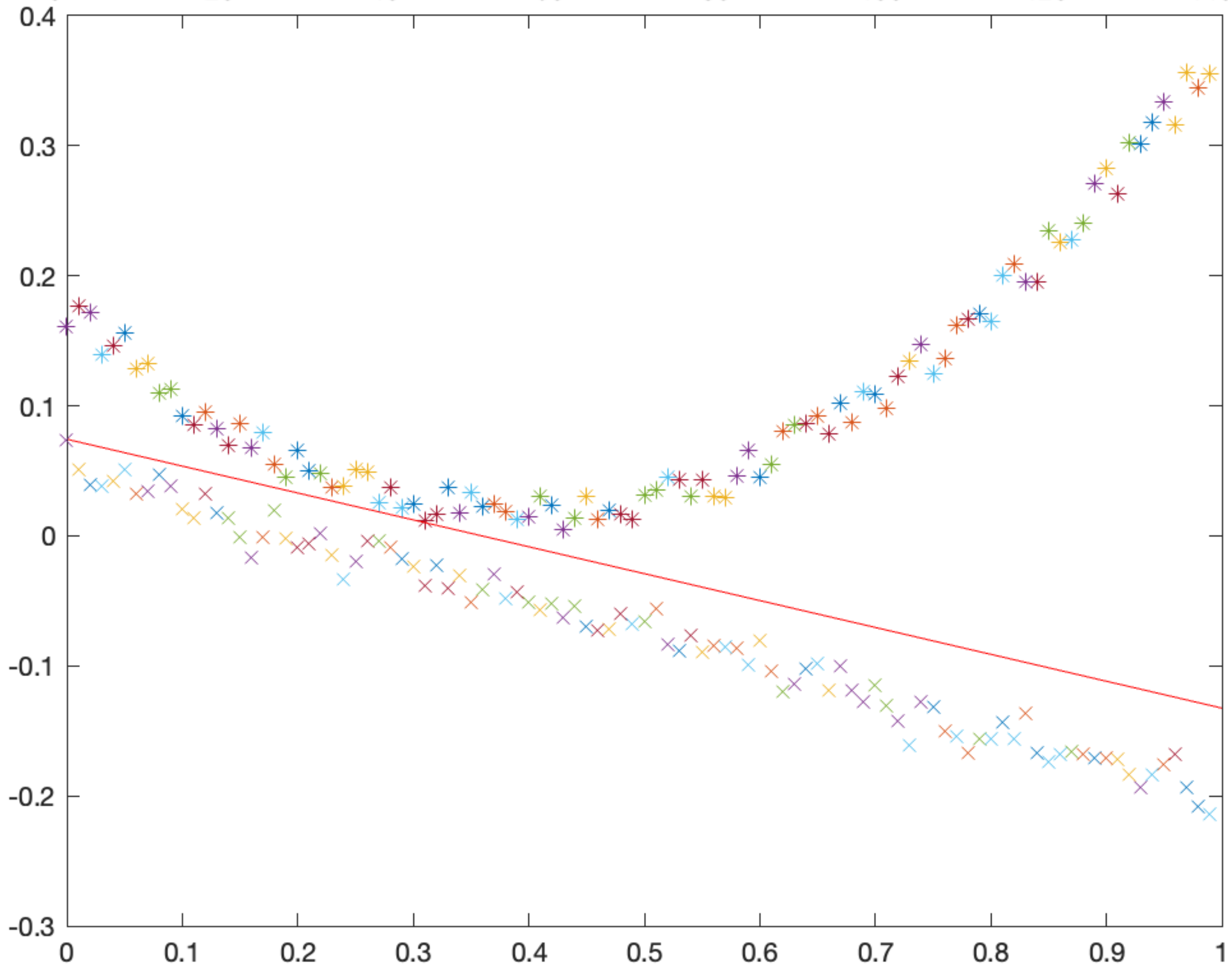


0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
50
100
150
classfi_loss_funcassification Loss
Perceptron Loss

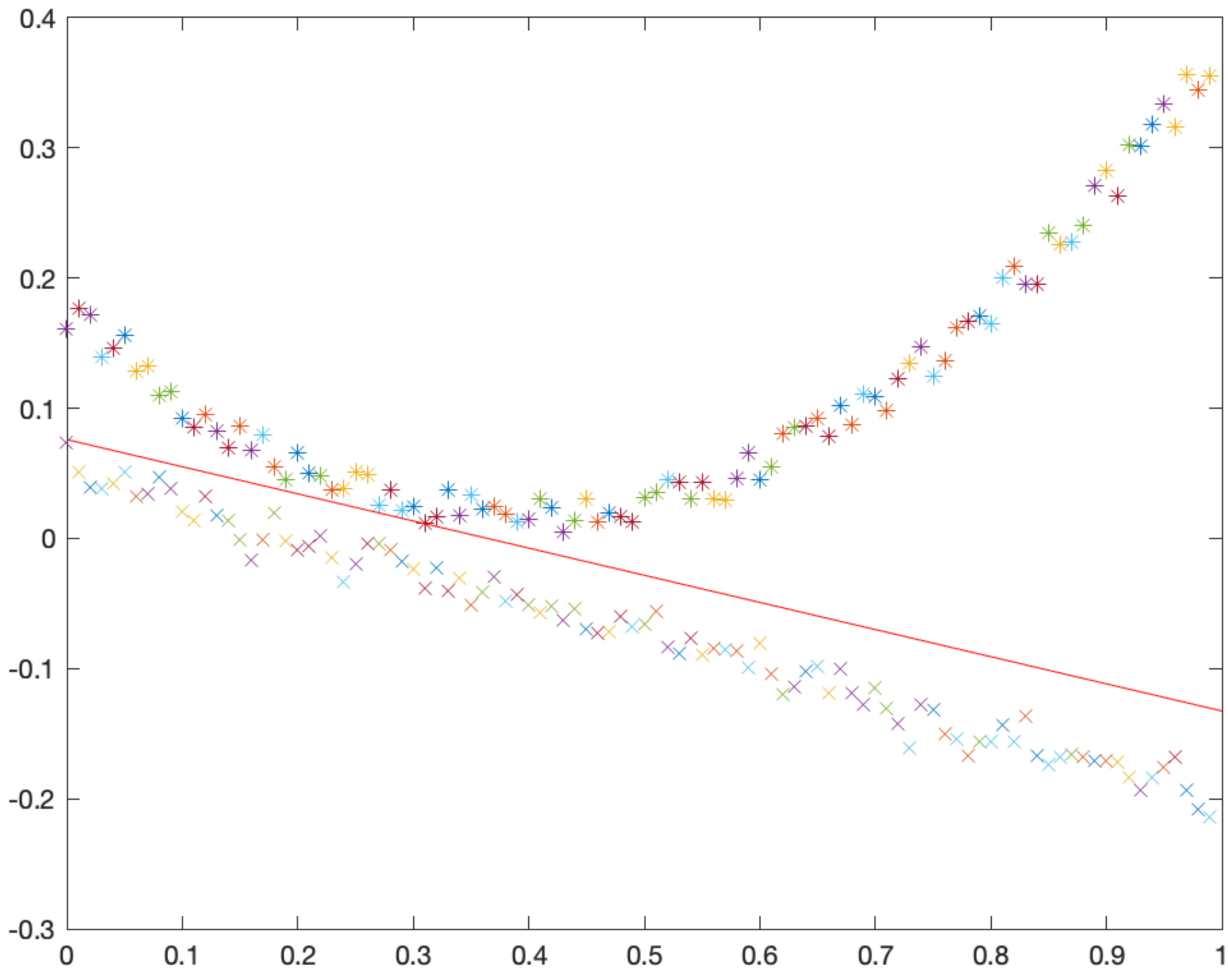


Step size = 3.5

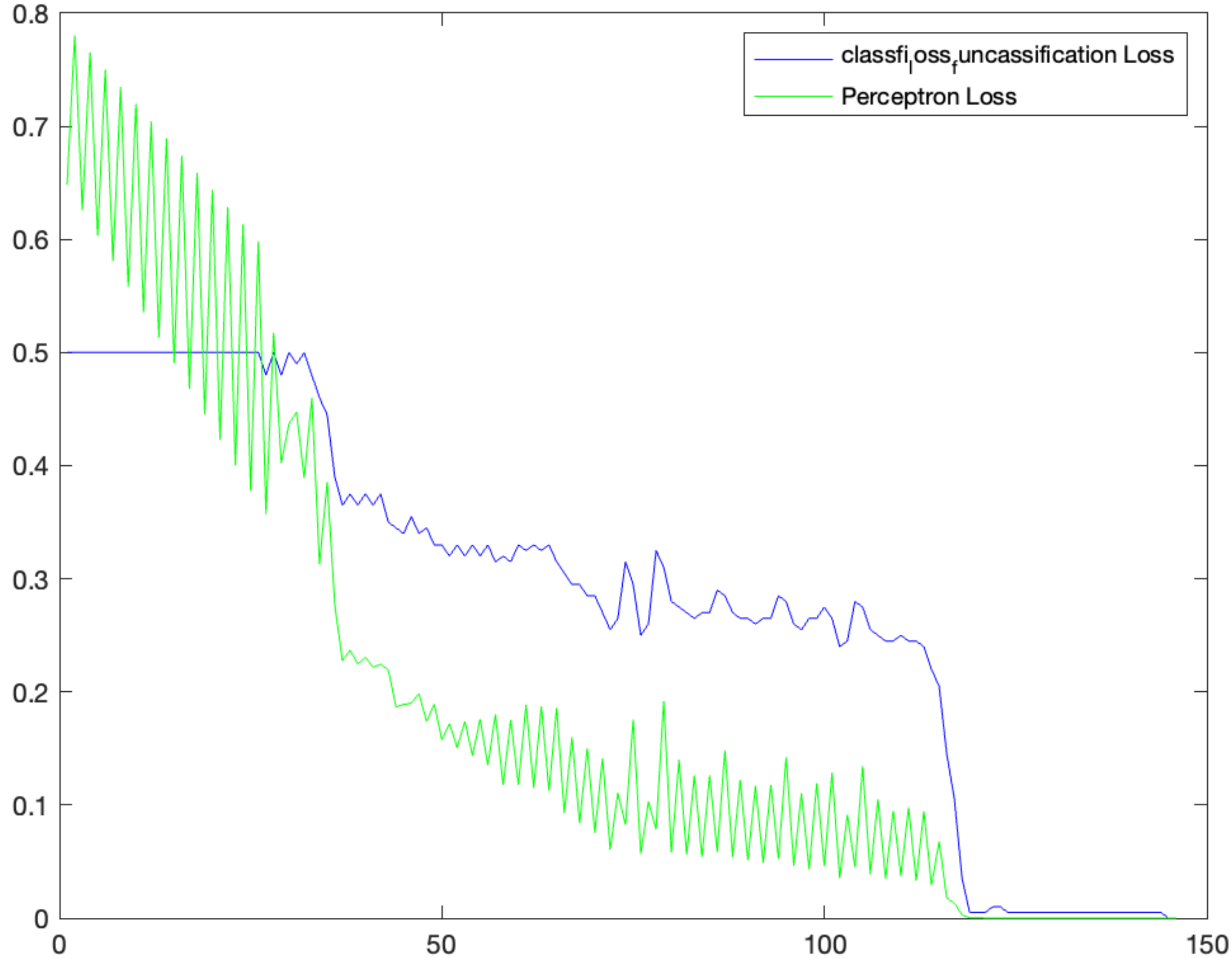




Step size. = 4.5



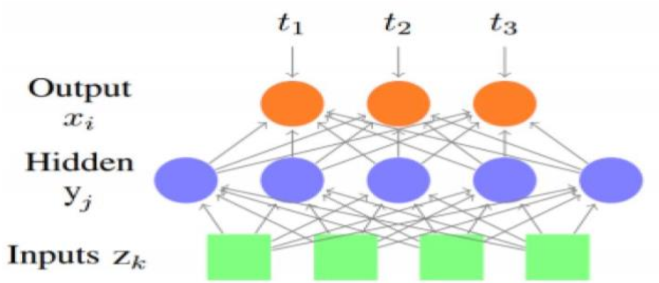
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1



Therefore, as shown above, we have the graph of step size 2.5, 3.5, 4.5, as the step size increase, the loss would be fluctuated more.

Problem 2 (15 points)

Consider the following network, where x denotes output units, y denotes hidden units, and z denotes input units.



problem 2.

a). ~~E~~ first. we get the derivation of $\frac{\partial E}{\partial w_{ji}}$, hidden layer y_j output layer x_i backpropogation on w_{ji}

$$\frac{\partial E}{\partial w_{ji}} = \frac{\partial E}{\partial x_i} \cdot \frac{\partial x_i}{\partial s_i} \cdot \frac{\partial s_i}{\partial w_{ji}}$$

$$= \frac{\partial}{\partial x_i}\left(-\sum_i (t_i \log(x_i) + (1-t_i)\log(1-x_i))\right) \frac{\partial x_i}{\partial s_i} \cdot \frac{\partial s_i}{\partial w_{ji}}$$

$$= \left(\frac{1-t_i}{1-x_i} - \frac{t_i}{x_i}\right) \frac{\partial x_i}{\partial s_i} \cdot \frac{\partial s_i}{\partial w_{ji}}$$

$$= \left(\frac{1-t_i}{1-x_i} - \frac{t_i}{x_i}\right) \frac{\partial x_i}{\partial s_i}\, \sigma(s_i) \frac{\partial s_i}{\partial w_{ji}}$$

$$\cdots$$

$$= (x_i - t_i) \frac{\partial}{\partial w_{ji}} \sum_j y_j w_{ji}$$

$$= (x_i - t_i)\, y_j$$

Next we use same way to derive $\frac{\partial E}{\partial w_{kj}}$, denote $x_i - t_i$ as δ

$$\frac{\partial E}{\partial w_{kj}} = \frac{\partial E}{\partial s_j} \cdot \frac{\partial s_j}{\partial w_{kj}}$$

$$= \sum_i \frac{\partial E}{\partial s_i} \cdot \frac{\partial s_i}{\partial s_j} \cdot \frac{\partial s_j}{\partial w_{kj}}$$

$$= \sum_i \delta_i \frac{\partial s_i}{\partial s_j} \cdot \frac{\partial s_j}{\partial w_{kj}}$$

$$= \sum_i \delta_i \frac{\partial}{\partial s_j}\left(\sum \sigma(s_j) \cdot w_{ji}\right) \cdot \frac{\partial s_j}{\partial w_{kj}}$$

$$= \sum_i \delta_i w_{ji} \sigma'(s_j) \frac{\partial s_j}{\partial w_{kj}}$$

$$\cdots$$

$$= \sum_i (x_i - t_i) \cdot w_{ji} \cdot y_j (1 - y_j) \cdot z_k$$

b). backpropogation on w_{ji}, denote $-t_i(1-x_i)$ as δ

first, we derive $\frac{\partial E}{\partial w_{ji}} = \frac{\partial}{\partial x_i}(-\sum_i t_i \log(x_i)) \frac{\partial x_i}{\partial s_i} \cdot \frac{\partial s_i}{\partial w_{ji}}$

$$= -\frac{t_i}{x_i} \frac{\partial x_i}{\partial s_i} \frac{\partial s_i}{\partial w_{ji}}$$

$$= -\frac{t_i}{x_i} \frac{\partial}{\partial s_i} \sigma_2 s_i \frac{\partial s_i}{\partial w_{ji}}$$

$$= -\frac{t_i}{x_i} \cdot \frac{(e^{s_i})'(\sum_i e^{s_i}) - e^{s_i}(\sum_i e^{s_i})'}{(\sum_i e^{s_i})^2} \cdot \frac{\partial s_i}{\partial w_{ji}}$$

$$= -\frac{t_i}{x_i}(x_i - x_i^2) \cdot \frac{\partial s_i}{\partial w_{ji}}$$

$$= -t_i(1-x_i) \frac{\partial}{\partial w_{ji}} \cdot \sum_j y_i w_{ji}$$

$$= y_i(x_i - t_i)$$

for the same way, we derive $\frac{\partial E}{\partial w_{kj}}$

$$= \sum_i (x_i - t_i) w_{ji} y_i (1-y_i) z_k$$

update: $w_{ji}^{T+1} = w_{ji}^{T} - \eta \frac{\partial E}{\partial w_{ji}}$

$$w_{kj}^{T+1} = w_{kj}^{T} - \eta \frac{\partial E}{\partial w_{kj}}$$

Problem 3 (10 points)

Consider the discrete distribution $\{p_k | k = 1, 2, \ldots, N\}$. The entropy of this distribution is given as $H = -\sum_{k=1}^{N} p_k \log p_k$. What is the distribution that maximizes this entropy? Show formal derivations using the method of Lagrange multipliers.

problem 3:

we have

discrete distribution $\{p_k \mid k=1,2,\ldots N\}$

entropy of this distribution is $H = -\sum_{k=1}^{N} p_k \cdot \log p_k$

We want the distribution to maximize entropy.

$$H = -\sum_{k=1}^{N} \cdot p_k \cdot \log p_k$$

to max, is to minimize the value of H'

which is $\min_p H' = \min_p \sum_{k=1}^{N} p_k \log(p_k)$

$$\Leftrightarrow \quad \min_x H'(x) = \min_x x^T \log(x)$$

using Lagrange multiplier λ, using I as N-d vector with each entry set to 1

we have

$$\min_x \max_\lambda f(x) = \min_x \max_\lambda x^T \log(x) - \lambda(I^T x - 1)$$

derivation would be

$$\frac{\partial}{\partial p} f(p) = (I + \log(x)) - \lambda I$$

when derivative to 0

we have $\hat{x} = e^{\lambda - 1} I$

$\Rightarrow \because I^T \cdot x = 1$

$\Rightarrow \lambda = 1 - \log N$

put λ back to $\hat{\lambda}$,

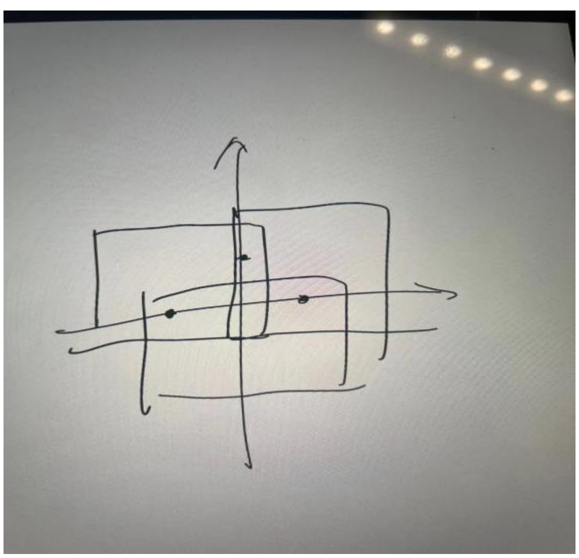
we have $\hat{\lambda} = \frac{1}{N} I$

$\Rightarrow$ the max distribution would be

$\{ p_K = \frac{1}{N} \mid k = 1, 2, 3, \ldots, N \}$

Problem 4 (10 points)
What is the VC dimension of axis-aligned squares? Justify your answer.

3 is the VC dimension of axis-aligned squares. For example, we could have (1, 0), (0, 1), and (−1, 0) in one axis that are shattered by axis-aligned squres. To label two of these points, put two points at corner, then we have at least 3 as the vc dimension.



If we have four points, it is the same situation. Therefore, the VC-demisino in the plane would be 3