Name+Email ID Brian Bertness - bert0078@umn.edu 1478201

CSCI 5521: Introduction to Machine Learning (Spring 2024)[1]

Final Exam

Due on Gradescope by 1pm, May 6

Instructions:

- The final exam has 5+1 questions, 100+2 points, on 8 pages, including one extra credit problem worth 2 points.
- Please **write your name & ID on this cover page**.
- **For full credit, show how you arrive at your answers**.

1. **(30 points)** In I-III, select the correct option(s) (it is not necessary to explain).

(I)	(II)	(III)

I. Select all the option(s) that are **activation functions**:

[✓ (a)] Softmax [✓ (b)] Sigmoid ~~(c)~~ Entropy ~~(d)~~ Linear Discrimination [✓ (e)] ReLU

II. Select all the option(s) that are true about **model selection**:

[✓ (a)] Nonparametric methods do not assume distributions of data to start with.
~~(b)~~ Support vector machines and decision trees are both methods for only classification, but not for regression.
~~(c)~~ Random forests are used over decision trees in cases where interpretability is needed.
~~(d)~~ Kernel methods are designed to reduce overfitting.
[✓ (e)] We would like to always select simpler models.

III. Select all the option(s) that are true:

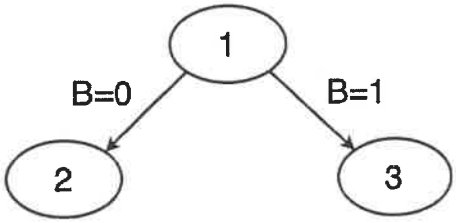
[✓ (a)] Activation functions can be used in hidden layers and output layers.
[✓ (b)] Multilayer Perceptron models can go with cross entropy or other loss functions depending on the task and data.
~~(c)~~ We can augment either training or test data to reduce overfitting.
~~(d)~~ Simple and intuitive manipulation of data (e.g., scaling or rotating an image) do not help with data augmentation.
~~(e)~~ Graphical models encode relationships therefore they have to be fully connected.

[1] Instructor: Catherine Qi Zhao. TA: Xianyu Chen, Amarachi Nzeukwu, James Yang, Yifeng Zhang. Email: csci5521.s2024@gmail.com

Name+Email ID Brian Bertness - bert0078@umn.edu 1478201

2. **(12 points)** Given the decision tree in the figure below, the node 1 was split using feature B. Now suppose we wish to split node 2. What is the feature that you will be using to split? Show your work.

So we are only going to split on node 2!



↓ A	B	↓ C	Class
0	0*	0	0 m
1	0*	0	0 m
1	1	0	1
0	0*	0	0 m
0	1	1	0
0	1	1	0
1	1	1	1
1	1	0	1
1	0*	1	1 m
1	1	1	1

For node 2 we have B=0. There are 4 rows in the table with B=0. We calculate entropy on features A and C:

A:

$A_0 = -\frac{2}{4}(\frac{2}{4}\log_2(\frac{2}{4}) + 0\log_2(0)) = 0.0$

$A_1 = -\frac{2}{4}(\frac{1}{2}\log_2(\frac{1}{2}) + \frac{1}{2}\log_2(\frac{1}{2})) = 0.5$

$A_{tot} = A_0 + A_1 = 0.0 + 0.5 = 0.5$

C:

$C_0 = -\frac{3}{4}(\frac{3}{4}\log_2(\frac{3}{4}) + 0\log_2(0)) = 0.0$

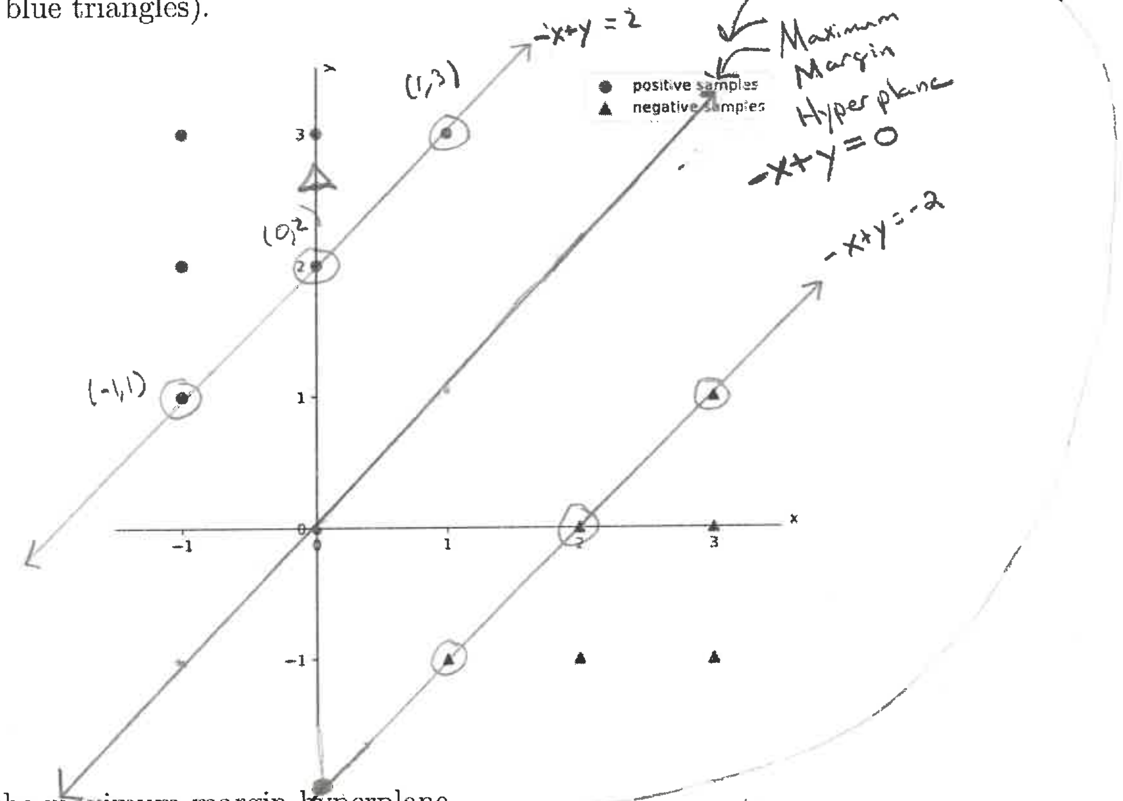
$C_1 = -\frac{1}{4}(0\log_2(0) + \frac{1}{4}\log_2(\frac{1}{4})) = 0.0$

$C_{tot} = C_0 + C_1 = 0.0 + 0.0 = 0.0$

Since $C_{tot} \leq A_{tot}$ We split node 2 on feature C.

Name+Email ID Brian Bertness - bert0078@umn.edu 1478201

3. **(15 points)** Suppose we are training a linear SVM on a tiny dataset of 12 points shown in the figure below. Samples with positive labels are (-1, 1), (-1, 2), (-1, 3), (0, 2), (0, 3), (1, 3) (denoted as red dots) and samples with negative labels are (1, -1), (2, -1), (2, 0), (3, -1), (3, 0), (3, 1) (denoted as blue triangles).



(a) Draw the maximum-margin hyperplane.

(b) Circle the support vectors.

(c) Pick one positive and one negative sample, and calculate their distances to the hyperplane.

Given $d^t = \frac{w^T x^t + w_0}{\|w\|}$, we get with $g(x) = w^T x^t + w_0$ and hyperplane equation $-x+y=0$.

Thus, $w = [-1,1]$ and $w_0 = 0$

Positive: $d(0,2) = \frac{|g(x)|}{\|w\|} = \frac{|0+2|}{\sqrt{(-1)^2+(1)^2}} = \frac{2}{\sqrt{2}}$

Negative: $d(3,0) = \frac{|g(x)|}{\|w\|} = \frac{|-3+0|}{\sqrt{(-1)^2+(1)^2}} = \frac{3}{\sqrt{2}}$

(d) If a new sample (0, 2.5) comes as a negative sample on top of the original 12 points, answer the following questions (it is not necessary to explain):

i. Will the decision boundary remain the same?

A. Yes

[✓ B. No]

ii. Which SVM algorithm is the best option here? (Select one)

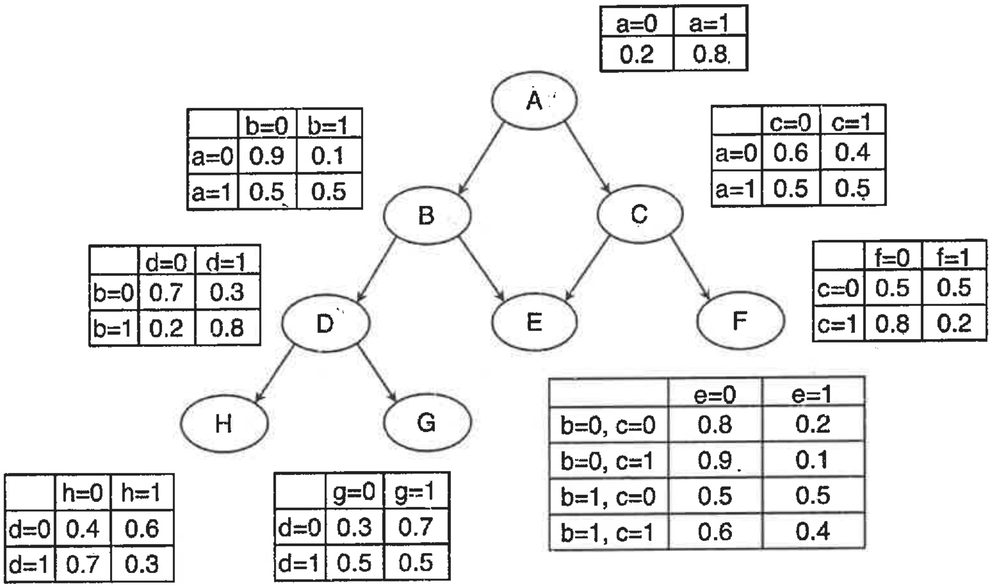
A. Hard-margin linear SVM

[✓ B. Soft-margin linear SVM]

C. Kernel SVM

Name+Email ID Brian Bertness - bert∅∅78 @ umn.edu 147820

4. **(25 points)** Consider the Bayesian Network below:



a=0	a=1
0.2	0.8

	b=0	b=1
a=0	0.9	0.1
a=1	0.5	0.5

	c=0	c=1
a=0	0.6	0.4
a=1	0.5	0.5

	d=0	d=1
b=0	0.7	0.3
b=1	0.2	0.8

	f=0	f=1
c=0	0.5	0.5
c=1	0.8	0.2

	e=0	e=1
b=0, c=0	0.8	0.2
b=0, c=1	0.9	0.1
b=1, c=0	0.5	0.5
b=1, c=1	0.6	0.4

	h=0	h=1
d=0	0.4	0.6
d=1	0.7	0.3

	g=0	g=1
d=0	0.3	0.7
d=1	0.5	0.5

Note: The numerical values of the probabilities are for part (e). You do not need to use them for (a)-(d).

(a) Find the joint probability $P(A, B, C, D, E, F, G, H)$ as the product of conditional probabilities, according to the graphical model given above.

$P(A,B,C,D,E,F,G,H) = P(A)P(B|A)P(C|A)P(D|B)P(E|B,C)P(F|C)P(G|D)P(H|D)$

(b) List all conditional independence given the graph.

$A \perp\!\!\!\perp D \mid B$

$A \perp\!\!\!\perp F \mid C$

$B \perp\!\!\!\perp H \mid D$

$B \perp\!\!\!\perp G \mid D$

$B \perp\!\!\!\perp G,H \mid D$

$B \perp\!\!\!\perp C \mid A$

$D \perp\!\!\!\perp E \mid B$

$E \perp\!\!\!\perp F \mid C$

$G \perp\!\!\!\perp H \mid D$

$A \perp\!\!\!\perp H \mid B, D$

$A \perp\!\!\!\perp G \mid B, D$

$A \perp\!\!\!\perp GH \mid B, D$

$D \perp\!\!\!\perp A, E \mid B$

$F \perp\!\!\!\perp A, E \mid C$

$H \perp\!\!\!\perp B, G \mid D$

(c) Show how to find the conditional probability $P(A|B)$.

Bayes Thm

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\sim A)P(\sim A)}$$

(d) Show how to find the probability $P(A, H)$.

Textbook 440'

$$P(A,H) = \sum_{B,D} P(A,B,D,H)$$

$$\begin{aligned}
&= P(A,B,D,H) + \quad \overset{①}{=} P(A)P(B|A)P(D|B)P(H|D) + \\
&\quad P(A,\sim B,D,H) + \quad P(A)P(\sim B|A)P(D|\sim B)P(H|D) + \\
&\quad P(A,B,\sim D,H) + \quad P(A)P(B|A)P(\sim D|B)P(H|\sim D) + \\
&\quad P(A,\sim B,\sim D,H) \quad P(A)P(\sim B|A)P(\sim D|\sim B)P(H|\sim D)
\end{aligned}$$

(e) Using the conditional probability distribution (CPD) tables in the figure, find :

i. $P(a = 1|b = 0)$

bayes thm

$$P(a=1|b=0) = \frac{P(b=0|a=1)\,P(a=1)}{P(b=0|a=1)P(a=1) + P(b=0|a=0)P(a=0)}$$

$$= \frac{(0.5)(0.8)}{(0.5)(0.8) + (0.9)(0.2)} \approx 0.69$$

ii. $P(a = 1, h = 0)$ Plugging into ① above we get

$$\begin{aligned}
P(a=1,h=0) = {} & P(a=1)P(b=1|a=1)P(d=1|b=1)P(h=0|d=1) + \\
& P(a=1)P(b=0|a=1)P(d=1|b=0)P(h=0|d=1) + \\
& P(a=1)P(b=1|a=1)P(d=0|b=1)P(h=0|d=0) + \\
& P(a=1)P(b=0|a=1)P(d=0|b=0)P(h=0|d=0)
\end{aligned}$$

$$\begin{aligned}
= {} & (0.8)(0.5)(0.8)(0.7) + \\
& (0.8)(0.5)(0.3)(0.7) + \\
& (0.8)(0.5)(0.2)(0.4) + \\
& (0.8)(0.5)(0.7)(0.4) \\
= {} & (0.224) + (0.084) + (0.032) + (0.112) \\
= {} & 0.452
\end{aligned}$$

iii. $P(a = 1, b = 0, c = 1, d = 0, e = 0, f = 0, g = 0, h = 0)$

Sample Find'

$$= P(a=1)P(b=0|a=1)P(c=1|a=1)P(d=0|b=0)P(e=0|b=0,c=1)P(f=0|c=1)P(g=0|d=0)P(h=0|d=0)$$

$$= (0.8)(0.5)(0.5)(0.7)(0.9)(0.8)(0.3)(0.4) \approx 0.012$$

iv. $P(b = 0, c = 1, d = 0, e = 0, f = 0, g = 0|a = 1, h = 0)$ =

Since $P(B,C,D,E,F,G|A,H) = \dfrac{P(A,B,C,D,E,F,G,H)}{P(A,H)}$

$$\doteq \frac{P(A)P(B|A)P(C|A)P(D|B)P(E|B,C)P(F|C)P(G|D)P(H|D)}{P(A,H)}$$

$$= \frac{\overset{1}{P(a=1)}\,\overset{2}{P(b=0|a=1)}\,\overset{3}{P(c=1|a=1)}\,\overset{4}{P(d=0|b=0)}\,\overset{5}{P(e=0|b=0,c=1)}\,\overset{6}{P(f=0|c=1)}\,\overset{7}{P(g=0|d=0)}\,\overset{8}{P(h=0|d=0)}}{P(a=1,h=0)}$$

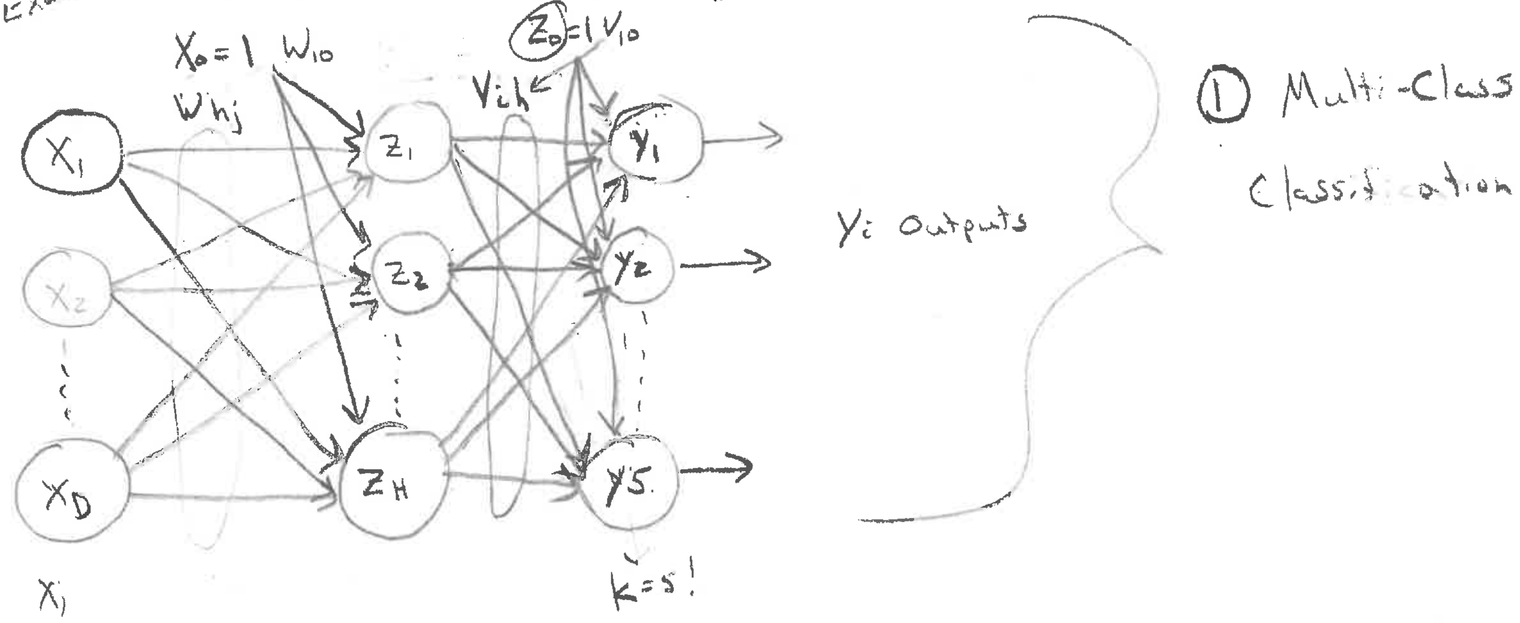
$$= \frac{(0.8)(0.5)(0.5)(0.7)(0.9)(0.8)(0.3)(0.4)}{0.452} = 2.68\,E{-2}$$

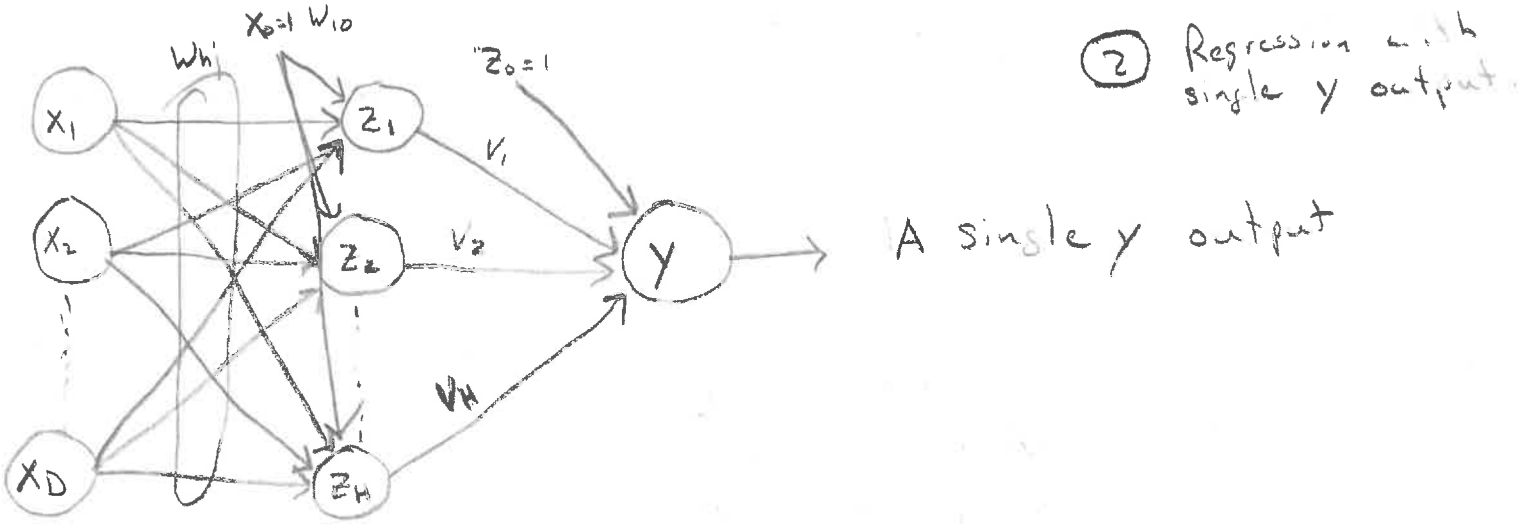
5. **(18 points)** Consider a Multi-layer Perceptron (MLP) for the following two general tasks: (1) multi-class classification of K=5 categories with 5 output units; and (2) regression with a single output unit, where each hidden unit in both tasks uses a hyperbolic tangent function such that $z_h^t = \tanh(\sum_{j=1}^{D} w_{hj}x_j^t + w_{h0})$. The output unit in classification uses a softmax activation function such that $y_i^t = \frac{\exp(\sum_h v_{ih}z_h^t + v_{i0})}{\sum_j \exp(\sum_h v_{jh}z_h^t + v_{j0})}$. The error functions for tasks (1) and (2) are given below respectively:

- Multi-class classification: $E(W, V|X) = -\sum_{t=1}^{N}\sum_{i=1}^{K} r_i^t \log y_i^t + \frac{\lambda}{2}\sum_{h=1}^{H} ||w_h||_2^2 + \frac{\sigma}{2}\sum_{k=1}^{K} ||v_k||_2^2$
- Regression: $E(W, v|X) = \frac{1}{2}\sum_{t=1}^{N}(r^t - y^t)^2 + \frac{\lambda}{2}\sum_{h=1}^{H} ||w_h||_2^2 + \frac{\sigma}{2}||v||_2^2$.

Pg or 5 of Practice Exam

(a) Draw two Multi-layer Perceptrons, each for one of the above tasks, showing: input values $x_0 ... x_D$, output of the hidden units $z_0 ... z_H$, weights W and V (or v), and the output(s) (*i.e.*, y_i of output unit i for multi-class classification, and y for regression). Note the difference in the structure between the two tasks (you may write or draw).





(b) Derive the Forward Step equations for y for both tasks.
Hint: Think about whether or not you should apply an activation function at output unit.

$\boxed{1}$

$$Z_h^t = \tanh\left(\sum_{j=1}^{D} W_{hj} X_j^t + W_{ho}\right)$$

$$Y_i^t = \frac{\exp\left(\Sigma_h V_{ih} Z_h^t + V_{io}\right)}{\Sigma_j \exp\left(\Sigma_h V_{jh} Z_h^t + V_{jo}\right)}$$

$\boxed{2}$

$$Z_h^t = \tanh\left(\sum_{j=1}^{D} W_{hj} X_j^t + W_{ho}\right)$$

$$y^t = \sum_{h=1}^{H} V_h Z_h^t + V_0$$

(c) Pick one of the two tasks, derive the Backward Step equation for w_{hj}.
Hint:

- $\tanh'(x) = 1 - \tanh^2(x)$
- Given the softmax function $f(\alpha_i) = \frac{exp(\alpha_i)}{\sum_j exp(\alpha_j)}$, then $\frac{\partial f(\alpha_i)}{\partial \alpha_j} = f(\alpha_i)(\delta_{ij} - f(\alpha_j))$, in which δ_{ij} is an indicator function, such that $\delta_{ij} = 1$ if $i = j$, and 0 otherwise.

pg 289 book

$$E(w,v|x) = \frac{1}{2}\underbrace{\sum_{t=1}^{N}(r^t - y^t)^2}_{E_1} + \underbrace{\frac{\lambda}{2}\sum_{h=1}^{H}\|W_h\|_2^2}_{E_2} + \underbrace{\frac{\lambda}{2}\|V\|_2^2}_{E_3}$$

$$\Delta W_{hj} = -\eta \frac{\partial E}{\partial W_{hj}}$$

$$= -\eta \sum_t \frac{\partial E}{\partial y^t}\frac{\partial y^t}{\partial \alpha^t}\frac{\partial \alpha^t}{\partial z_h^t}\frac{\partial z_h^t}{\partial W_{hj}}$$

$$= \eta \sum_t (r^t - y^t)(1 - \tanh^2(V_h Z_h^t + V_0))(V_h) Z_h^t(1 - Z_h^t)(X_j^t)$$

$$+ \frac{\lambda}{2}\Sigma 2W_h$$

with the complete equation being

$W_{hj} = W_{hj} + \Delta W_{hj}$ where ΔW_{hj} is as defined above.

Now, $E = E_1 + E_2 + E_3$

For E_1 $\frac{\partial E}{\partial y^t} = -(r^t - y^t)$ ✓

$\frac{\partial y^t}{\partial \alpha^t} = 1 - \tanh^2(V_h Z_h^t + V_0)$ ✓

$\frac{\partial \alpha^t}{\partial Z_h^t} = V_h$ ✓, $\frac{\partial Z_h^t}{\partial W_{hj}} = Z_h^t(1 - Z_h^t)X_j^t$

For E_2 $\frac{\partial E}{\partial W_{hj}} = \frac{\lambda}{2}\Sigma 2W_h$

For E_3 $\frac{\partial E}{\partial W_{hj}} = 0$

Name+Email ID Brian Bertness - bert0078@umn.edu 1478201

6. **(2 points, extra credit)** Mary is an intern working in a biomedical research team, which is tasked with developing an AI-driven decision support system for diagnosing and treating complex diseases. The system leverages a large language model (LLM), which is a form of deep neural network, trained on vast amounts of data on **common diseases**, such as diabetes or cardiovascular diseases. The model accesses a large database of well-labeled, comprehensive data, such as medical literature, patient histories, and clinical trial data.

   (a) Would the current team's LLM be a good choice for diagnosing and treating common diseases? Explain why. If not a good choice, please suggest a machine learning method/strategy they could use, and explain your suggestion.

   (b) For rare genetic disorders, the available data is much smaller and more limited. Would the current team's LLM be a good choice for diagnosing and treating rare genetic disorders? Explain why. If not a good choice, please suggest a machine learning method/strategy they could use, and explain your suggestion.