ECE C147/C247, Winter 2025 **Homework #3**
Neural Networks & Deep Learning Prof. J.C. Kao
UCLA ECE TAs: B. Qu, K. Pang, S. Dong, S. Rajesh, T. Monsoor, S. Yan

Due Friday, 7 Feb 2025, by 11:59pm to Gradescope.
100 points total.

1. (15 points) **Backpropagation for autoencoders.** In an autoencoder, we seek to reconstruct the original data after some operation that reduces the data's dimensionality. We may be interested in reducing the data's dimensionality to gain a more compact representation of the data.

   For example, consider $\mathbf{x} \in \mathbb{R}^n$. Further, consider $\mathbf{W} \in \mathbb{R}^{m \times n}$ where $m < n$. Then $\mathbf{Wx}$ is of lower dimensionality than $\mathbf{x}$. One way to design $\mathbf{W}$ so that $\mathbf{Wx}$ still contains key features of $\mathbf{x}$ is to minimize the following expression

   $$\mathcal{L} = \frac{1}{2} \left\| \mathbf{W}^T \mathbf{W} \mathbf{x} - \mathbf{x} \right\|^2$$

   with respect to $\mathbf{W}$. (To be complete, autoencoders also have a nonlinearity in each layer, i.e., the loss is $\frac{1}{2} \left\| f(\mathbf{W}^T f(\mathbf{Wx})) - x \right\|^2$. However, we'll work with the linear example.)

   (a) (3 points) In words, describe why this minimization finds a $\mathbf{W}$ that ought to preserve information about $\mathbf{x}$.

   (b) (3 points) Draw the computational graph for $\mathcal{L}$. **Hint:** You can set up the computational graph to this problem in a way that will allow you to solve for part (d) without taking 4D tensor derivative.

   (c) (3 points) In the computational graph, there should be two paths to $\mathbf{W}$. How do we account for these two paths when calculating $\nabla_{\mathbf{W}} \mathcal{L}$? Your answer should include a mathematical argument.

   (d) (6 points) Calculate the gradient: $\nabla_{\mathbf{W}} \mathcal{L}$.

2. (20 points) **Backpropagation for Gaussian-process latent variable model. (Optional for students in C147: Please write 'I am a C147 student' in the solution and you will get full credit for this problem).** An important component of unsupervised learning is visualizing high-dimensional data in low-dimensional spaces. One such nonlinear algorithm to do so is from Lawrence, NIPS 2004, called GP-LVM. GP-LVM optimizes the maximum-likelihood of a probabilistic model. We won't get into the details here, but rather to the bottom line: in this paper, a log-likelihood has to be differentiated with respect to a matrix to derive the optimal parameters.

   To do so, we will apply the chain rule for multivariate derivatives via backpropagation. The log-likelihood is:

   $$\mathcal{L} = -c - \frac{D}{2} \log |\mathbf{K}| - \frac{1}{2} \mathrm{tr}(\mathbf{K}^{-1} \mathbf{Y} \mathbf{Y}^T)$$

where $\mathbf{K} = \alpha\mathbf{X}\mathbf{X}^T + \beta^{-1}\mathbf{I}$ and c is a constant. The $|.|$ symbol in this context refers to the determinant of a matrix. To solve this, we'll take the derivatives with respect to the two terms with dependencies on $\mathbf{X}$:

$$\begin{aligned}\mathcal{L}_1 &= -\frac{D}{2}\log|\alpha\mathbf{X}\mathbf{X}^T + \beta^{-1}\mathbf{I}| \\ \mathcal{L}_2 &= -\frac{1}{2}\text{tr}\left((\alpha\mathbf{X}\mathbf{X}^T + \beta^{-1}\mathbf{I})^{-1}\mathbf{Y}\mathbf{Y}^T\right)\end{aligned}$$

Hint: To receive full credit, you will be required to show all work. You may use the following matrix derivative without proof:

$$\frac{\partial\mathcal{L}}{\partial\mathbf{K}} = -\mathbf{K}^{-T}\frac{\partial\mathcal{L}}{\partial\mathbf{K}^{-1}}\mathbf{K}^{-T}.$$

Also, consider the matrix operation, $\mathbf{Z} = \mathbf{X}\mathbf{Y}$. If we have an upstream derivative, $\partial\mathcal{L}/\partial\mathbf{Z}$, then backpropagate the derivatives in the following way:

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\mathbf{X}} &= \frac{\partial\mathcal{L}}{\partial\mathbf{Z}}\mathbf{Y}^T \\ \frac{\partial\mathcal{L}}{\partial\mathbf{Y}} &= \mathbf{X}^T\frac{\partial\mathcal{L}}{\partial\mathbf{Z}}\end{aligned}$$

(a) (3 points) Draw a computational graph for $\mathcal{L}_1$.

(b) (6 points) Compute $\frac{\partial\mathcal{L}_1}{\partial\mathbf{X}}$.

(c) (3 points) Draw a computational graph for $\mathcal{L}_2$.

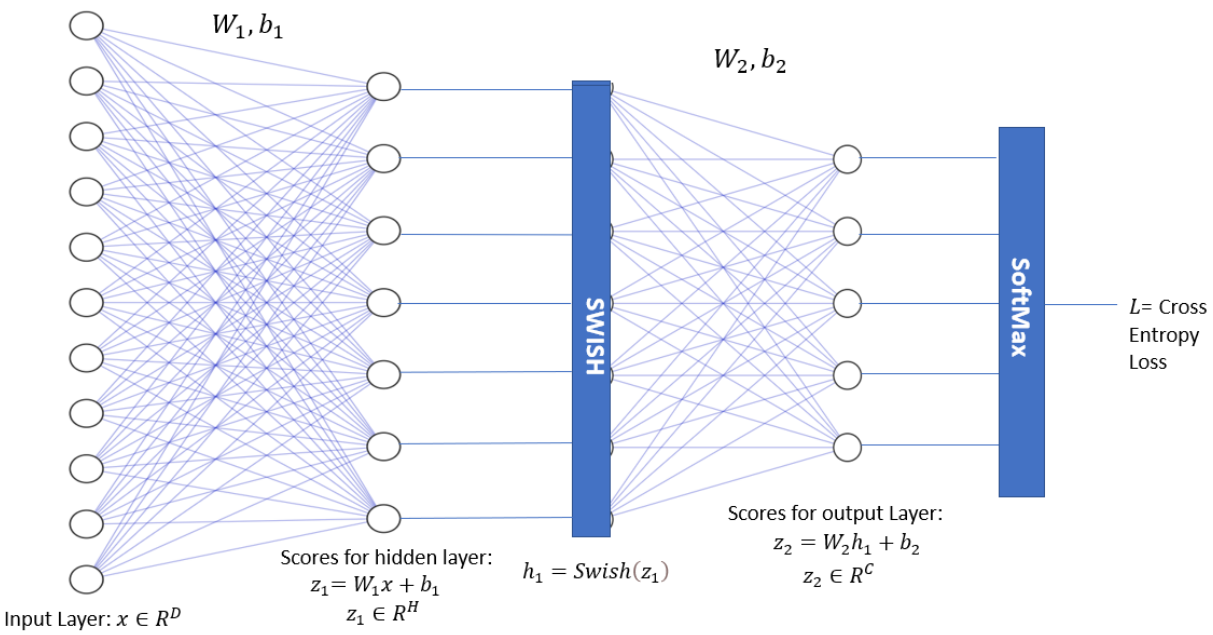
(d) (6 points) Compute $\frac{\partial\mathcal{L}_2}{\partial\mathbf{X}}$.

(e) (2 points) Compute $\frac{\partial\mathcal{L}}{\partial\mathbf{X}}$.

3. (15 points) **NNDL to the rescue!!**

It looks like a calm Monday morning and you are almost done with NNDL HW for the week (sigh)! but then suddenly (tring tring ...)your phone starts buzzing, you pick the call and the person from the other end sounds tense, the person exclaims ...There is a national Emergency !!: *7 different Pandora creature species (from Avatar) have been spotted in 1000's of numbers across various places in the country . They are having a hard time to adjust to earth's climate and are causing chaos. As a result there has been a Power outage in many cities. Luckily LA is an exception. UCLA's engineering division is helping out with this emergency, and you have been summoned to contribute to the same.* You quickly take a bird to the secret facility and meet with director in charge of this operation. The director gives you a dataset consisting of images of these creatures along with their species type and instructs you to design a machine learning model to classify the images into species type. The only design constraint that the director has imposed is that the model should not have a very large number of parameters because some of UCLA's compute facilities are overloaded due to the power outages.

You have just learned about Fully connected neural networks (FC net) in class and decide to use it for accomplishing the task. To satisfy the design constraint, you have decided to build a 2-layer FC net and train it using the provided dataset. The trained model will not only



enable you to classify the images into species type but the hidden representations (outputs of intermediate layers) can be used to analyze the various properties of the species. A pictorial representation of the 2-layer FC net is shown above:

In the architecture shown, D represents the number of neurons in input layer, H represents the number of neurons in hidden layer , C represents the number of neurons in the output layer (in our design $C = 7$). The output is then passed through a softmax classifier. Although we learned about the ReLu activation in class, but we decided to use the Swish activation function (introduced by google brain) for the hidden layer. Swish activation function for any scalar input k is defined as,

$$\text{swish}(k) = \frac{k}{1 + e^{-k}} = k\sigma(k),$$

where, $\sigma(k)$, is the sigmoid activation function you have seen in lectures

You will train the 2-layer FC net using gradient descent and for that you will need to compute the gradients. For the gradient computations, you are allowed to keep your final answer in terms of $\frac{\partial L}{\partial z_2}$.

(a) (3 points) Draw the computational graph for the 2-layer FC net.

(b) (5 points) Compute $\nabla_{W_2} L$, $\nabla_{b_2} L$.

(c) (7 points) Compute $\nabla_{W_1} L$, $\nabla_{b_1} L$.

4. (30 points) **2-layer neural network.**

Complete the two-layer neural network Jupyter notebook. Print out the entire notebook and relevant code and submit it as a pdf to gradescope. Download the CIFAR-10 dataset, as you did in HW #2.

5. (20 points) **General FC neural network.**

   Complete the FC Net Jupyter notebook. Print out the entire notebook and relevant code and submit it as a pdf to gradescope.

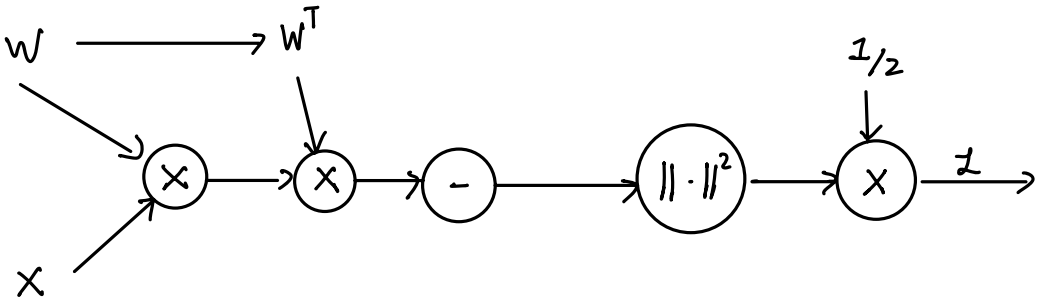
ECE 247
Sujit Silas Armstrong
HW3

1. a) $x \in \mathbb{R}^n$
$W \in \mathbb{R}^{m \times n}$ $(m < n)$

$$L = \frac{1}{2} \| W^T W x - x \|^2$$

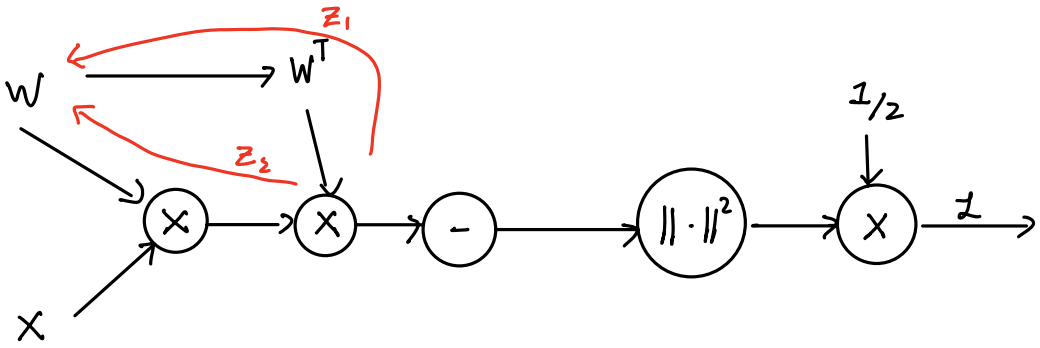
This minimization preserves the information about x by projecting important features into a lower dimensional space constrained by L2-distance. This preserves information about x because W is optimized to retain the most relevant features, minimizing the loss of critical information in the process.

b) Computational graph



$$l = Wx \; ; \quad m = W^T W x - x \; ; \quad k = \|m\|^2 \; ; \quad g = \frac{1}{2} k$$

c)



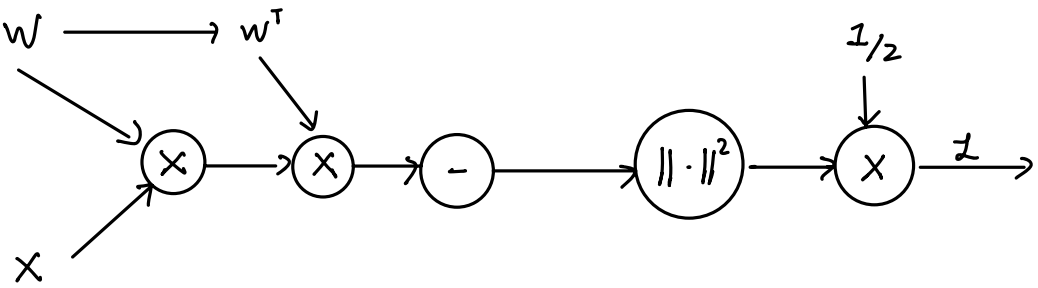
→ we have 1 path going through W^T and the other one going through the multiplication gate.
→ we can take the gradient for each path individually & add them to find $\nabla_W L$

By chain rule & the law of total derivatives we have

$$\nabla_W \mathcal{L} = \frac{\partial z_1}{\partial w}\frac{\partial \mathcal{L}}{\partial z_1} + \frac{\partial z_2}{\partial w}\frac{\partial \mathcal{L}}{\partial z_2}$$

$$z_1 = w \; ; \; z_2 = w, \quad \frac{\partial z_1}{\partial w} = \frac{\partial z_2}{\partial w} = 1$$

d)



$$l = Wx \; ; \quad m = W^T W x - x \; ; \quad k = \|m\|^2 \; ; \quad g = \frac{1}{2}k$$

$l \in \mathbb{R}^m$ $\quad m \in \mathbb{R}^n$

$$\frac{\partial \mathcal{L}}{\partial g} = \frac{1}{2} \qquad \mathcal{L} = \frac{1}{2} m^T m$$

$$\frac{\partial \mathcal{L}}{\partial k} = \frac{\partial g}{\partial k}\frac{\partial \mathcal{L}}{\partial g} = \frac{1}{2} \cdot 2$$

$$\frac{\partial \mathcal{L}}{\partial m} = m = W^T W x - x$$

$$\frac{\partial \mathcal{L}}{\partial l} = \frac{\partial l}{\partial m}\frac{\partial \mathcal{L}}{\partial l} = Wm \qquad \frac{\partial \mathcal{L}}{\partial W^T} = \frac{\partial m}{\partial W^T}\frac{\partial \mathcal{L}}{\partial m} = m(Wx)^T$$

$$\frac{\partial \mathcal{L}}{\partial W} = \frac{\partial m}{\partial W}\frac{\partial \mathcal{L}}{\partial m} = Wmx^T$$

$$\therefore \frac{\partial \mathcal{L}}{\partial W} = W((W^T W x - x)x^T) + Wx(W^T W x - x)^T$$

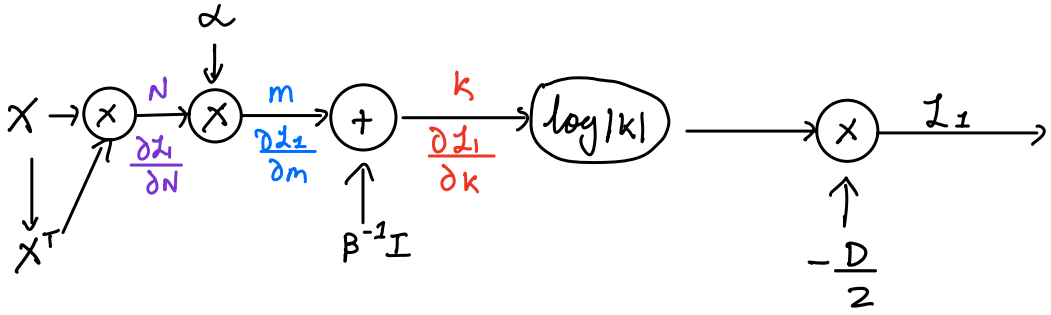
2. a) & b) $\mathcal{L} = -C - \frac{D}{2}\log|K| - \frac{1}{2}\mathrm{tr}(K^{-1}YY^T)$

$$K = \alpha XX^T + \beta^{-1}I$$

$$\mathcal{L}_1 = -\frac{D}{2}\log|\alpha XX^T + \beta^{-1}I|$$

$$\mathcal{L}_2 = -\frac{1}{2}\mathrm{tr}((\alpha XX^T + \beta^{-1}I)^{-1}YY^T)$$

Note: $\frac{\partial \mathcal{L}}{\partial K} = -K^{-T}\frac{\partial \mathcal{L}}{\partial K^{-1}}K^{-T}$



$$\mathcal{L}_1 = -\frac{D}{2}\log|K| = \frac{-D}{2}\log(\det(K))$$

$$\frac{\partial \mathcal{L}_1}{\partial K} = \frac{-D}{2}(K^{-1})^T = \frac{-D}{2}(K^T)^{-1}$$

$\mathcal{L}_1 = \frac{D}{2}\log(K)$ ⊕ gradient flows $\frac{\partial \mathcal{L}_1}{\partial m} = \frac{-D}{2}(K^T)^{-1}$

$K = m\alpha$ ⊗ Switches gradients $\frac{\partial \mathcal{L}_1}{\partial N} = -\frac{\alpha D}{2}(K^T)^{-1}$

$N = XX^T$; $\frac{\partial \mathcal{L}_1}{\partial X} = \frac{\partial \mathcal{L}_1}{\partial N}(X^T)^T = -\frac{\alpha D}{2}(K^T)^{-1}\cdot X$

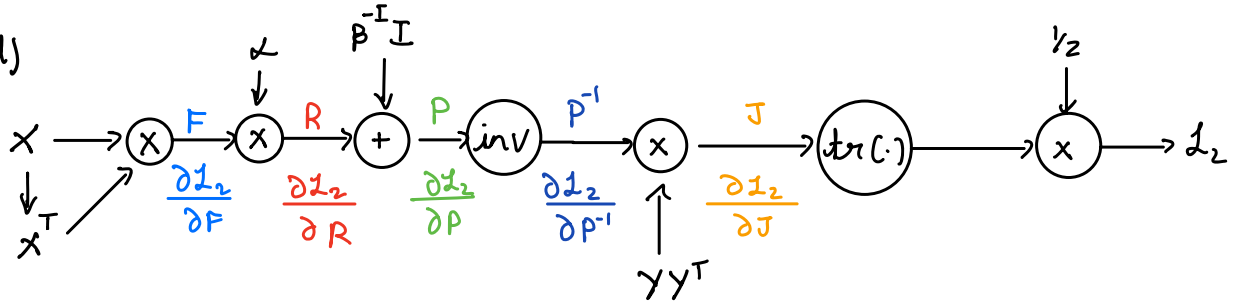
$$\frac{\partial \mathcal{L}_1}{\partial X^T} = X^T\frac{\partial \mathcal{L}_1}{\partial N} = X^T\left[-\frac{\alpha D}{2}(K^T)^{-1}\right] = -\frac{\alpha D}{2}(K^{-1}X)^T$$

Adding the gradient path through X & X^T we get

$$\frac{\partial \mathcal{L}_1}{\partial X} = \frac{-\alpha D}{2}(K^T)^{-1}X - \left[\frac{\alpha D}{2}(K^{-1}X)^T\right]^T$$

K is symmetric $\therefore$ $\boxed{\frac{\partial \mathcal{L}_1}{\partial X} = -\alpha D K^{-1}X}$

c) & d)



$$P = \alpha X X^T + \beta^{-1} I$$

$$\mathcal{L}_2 = -\frac{1}{2} tr(J)$$

$$\frac{\partial \mathcal{L}_2}{\partial J} = -\frac{1}{2} I$$

$$J = P^{-1}(YY^T)$$

$$\frac{\partial \mathcal{L}_2}{\partial P^{-1}} = -\frac{1}{2}(YY^T)$$

$$\frac{\partial \mathcal{L}_2}{\partial P} = -P^{-1} \frac{\partial \mathcal{L}_2}{\partial P^{-1}} P^{-1}$$

$$\frac{\partial \mathcal{L}_2}{\partial R} = -P^{-1} \frac{\partial \mathcal{L}_2}{\partial P^{-1}} P^{-1}$$

$$\frac{\partial \mathcal{L}_2}{\partial F} = -\alpha P^{-1} \frac{\partial \mathcal{L}_2}{\partial P^{-1}} P^{-1}$$

$$F = YY^T$$

$$\frac{\partial \mathcal{L}_2}{\partial X} = -\alpha P^{-1} \frac{\partial \mathcal{L}_2}{\partial P^{-1}} P^{-1} X$$

$$\frac{\partial \mathcal{L}_2}{\partial X} = \frac{\alpha}{2} P^{-1} YY^T P^{-1} X$$

$$\frac{\partial \mathcal{L}_2}{\partial X^T} = -\frac{\alpha}{2} X^T P^{-1} YY^T P^{-1}$$

$$= \frac{\partial \mathcal{L}_2}{\partial X} + \left[\frac{\partial \mathcal{L}_2}{\partial X^T}\right]^T$$

$$= \frac{\alpha}{2} P^{-1} YY^T P^{-1} X + \left[\frac{\alpha}{2} X^T P^{-1} YY^T P^{-1}\right]^T$$

$$= \frac{\alpha}{2} P^{-1} YY^T P^{-1} X + \frac{\alpha}{2} (P^{-1})^T YY^T (P^T)^{-1} X$$

P is symmetric so :

$$= \frac{\alpha}{2} P^{-1} YY^T P^{-1} X + \frac{\alpha}{2} P^{-1} YY^T P^{-1} X$$

$$\boxed{\frac{\partial \mathcal{L}_2}{\partial X} = \alpha P^{-1} YY^T P^{-1} X}$$

e) By the law of total derivation, we have :

$$\frac{\partial \mathcal{L}}{\partial X} = \frac{\partial \mathcal{L}_1}{\partial X} + \frac{\partial \mathcal{L}_2}{\partial X}$$

$$= -\alpha D (\alpha X X^T + \beta^{-1} I)^{-1} X + \alpha (\alpha X X^T + \beta^{-1} I)^{-1} YY^T (\alpha X X^T + \beta^{-1} I)^{-1} X$$

$$= \alpha \left[-DI + (\alpha X X^T + \beta^{-1} I)^{-1} YY^T\right] (\alpha X X^T + \beta^{-1} I)^{-1} X$$

3. a) 2-layer FC

H → hidden layers

C → 7 (neurons in the output layer)

$\text{Swish}(k) = \frac{k}{1+e^{-k}} = k\,\sigma(k)$, where $\sigma(k)$ is the sigmoid activation function.

$= k\,(1+e^{-k})^{-1}$

$z_1 = W_1 x + b_1$, $z_1 \in \mathbb{R}^H$ (affine function) , $x \in \mathbb{R}^D$

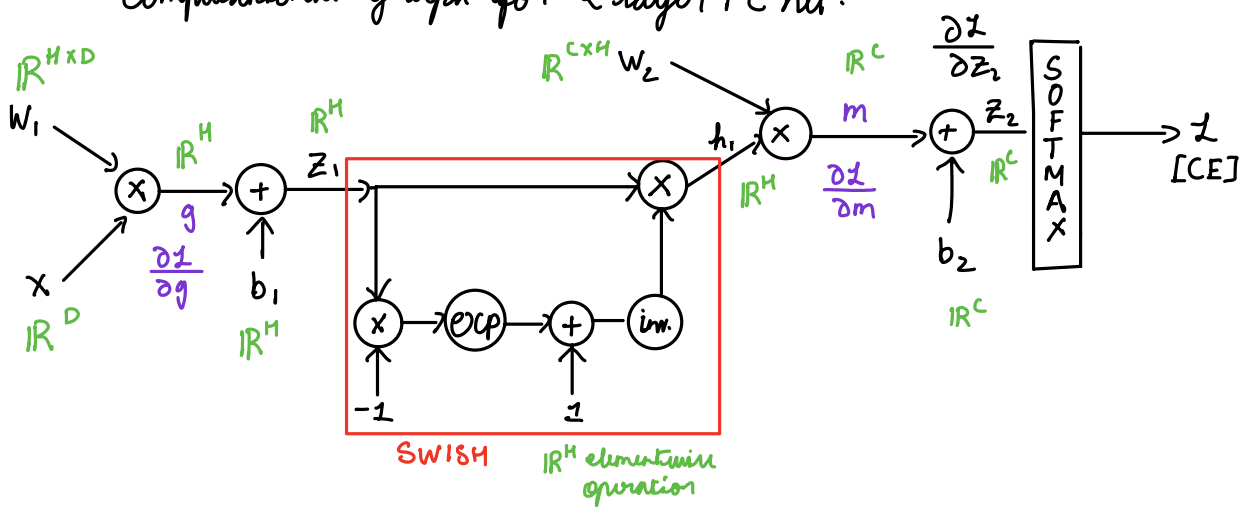
$h_1 = \text{Swish}(z_1) = k\,\sigma(k)$ $\qquad h_1' = z_1 + \sigma(z_1)(1-\sigma(z_1))$

$z_2 = W_2 h_1 + b_2$, $z_2 \in \mathbb{R}^C$

$\hat{y} = \text{Softmax}(z_2)$

$\mathcal{L}$ = cross entropy $\Rightarrow$ $\mathcal{L} = -\sum_i y_i \log(\hat{y}_i)$

Computational graph for 2 layer FC net:



b)&c) $\nabla_{W_2}\mathcal{L}$, $\nabla_{b_2}\mathcal{L}$, $\nabla_{W_1}\mathcal{L}$, $\nabla_{b_1}\mathcal{L}$

$\frac{\partial \mathcal{L}}{\partial m} = \frac{\partial \mathcal{L}}{\partial z_2}$ & $\frac{\partial \mathcal{L}}{\partial b_2} = \frac{\partial \mathcal{L}}{\partial z_2}$ $\qquad \therefore z_2 = m + b_2$

$m = W_2 h_1$ $\qquad \frac{\partial m}{\partial h_1} = W_2^T \in \mathbb{R}^{H\times C}$ $\qquad \frac{\partial m}{\partial W_2} = h_1^T \in \mathbb{R}^{C\times H\times C}$

$\frac{\partial \mathcal{L}}{\partial h_1} = \frac{\partial m}{\partial h_1}\frac{\partial \mathcal{L}}{\partial m} = W_2^T \frac{\partial \mathcal{L}}{\partial m}$

$\frac{\partial \mathcal{L}}{\partial W_2} = \frac{\partial m}{\partial W_2}\frac{\partial \mathcal{L}}{\partial m} = \frac{\partial \mathcal{L}}{\partial m} h_1^T$

$z_1 = \text{Swish}(h_1)$

$$z_1 = \begin{bmatrix} z_{1,1} \\ z_{1,2} \\ \vdots \\ z_{1,H} \end{bmatrix} \in \mathbb{R}^H \qquad h_1 = \text{Swish}(z_1) = \begin{bmatrix} \text{Swish}(z_{1,1}) \\ \text{Swish}(z_{1,2}) \\ \vdots \\ \text{Swish}(z_{1,H}) \end{bmatrix} = \begin{bmatrix} h_{1,1} \\ h_{1,2} \\ \vdots \\ h_{1,H} \end{bmatrix}$$

JACOBIAN

$$\mathbb{R}^H \to \mathbb{R}^H \quad \frac{\partial h_1}{\partial z_1} = \begin{bmatrix} \frac{\partial h_1}{\partial z_{1,1}} \\ \frac{\partial h_1}{\partial z_{1,2}} \\ \vdots \\ \frac{\partial h_1}{\partial z_{1,H}} \end{bmatrix} = \begin{bmatrix} \frac{\partial h_{1,1}}{\partial z_{1,1}} & \frac{\partial h_{1,2}}{\partial z_{1,1}} \to 0 & \cdots & \frac{\partial h_{1,H}}{\partial z_{1,1}} \to 0 \\ \frac{\partial h_{1,1}}{\partial z_{1,2}} \to 0 & \frac{\partial h_{1,2}}{\partial z_{1,2}} & \cdots & \frac{\partial h_{1,H}}{\partial z_{1,2}} \to 0 \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial h_{1,1}}{\partial z_{1,H}} \to 0 & \frac{\partial h_{1,2}}{\partial z_{1,H}} \to 0 & \cdots & \frac{\partial h_{1,H}}{\partial z_{1,H}} \end{bmatrix}$$

Only diagonal elements remain non-zero

$$\frac{\partial h_{1,i}}{\partial z_{1,j}} = \begin{cases} \frac{\partial}{\partial z_{1,j}} \left[\frac{z_{1,j}}{1+\exp(-z_{1,j})} \right], & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}$$

by Quotient rule we have

$$\frac{\partial}{\partial k} = \left[\frac{k}{1+\exp(-k)} \right] = \frac{1 \times (1+\exp(-k)) - [-\exp(-k)]k}{(1+\exp(-k))^2}$$

$$= \frac{1 + (k+1)e^{-k}}{(1+e^{-k})}$$

$$\frac{\partial h_1}{\partial z_1} = \text{diag}\left(\frac{1 + (z_{1,1}+1)e^{-z_{1,1}}}{(1+e^{-z_{1,1}})^2}, \frac{1+(z_{1,2}+1)e^{-z_{1,2}}}{(1+e^{-z_{1,2}})^2}, \ldots \right)$$

$$\frac{\partial L}{\partial z_1} = \frac{\partial h_1}{\partial z_1} \frac{\partial L}{\partial h_1} = S_1 \odot W_2^T \frac{\partial L}{\partial z_2}$$

where $S_1 = \begin{bmatrix} \frac{1 + (z_{1,1}+1)e^{-z_{1,1}}}{(1+e^{-z_{1,1}})^2} \\ \frac{1 + (z_{1,2}+1)e^{-z_{1,2}}}{(1+e^{-z_{1,2}})^2} \\ \vdots \\ \frac{1 + (z_{1,H}+1)e^{-z_{1,H}}}{(1+e^{-z_{1,H}})^2} \end{bmatrix}$

$$\frac{\partial L}{\partial W_1} = \frac{\partial z_1}{\partial W_1}\frac{\partial L}{\partial z_1} = \frac{\partial L}{\partial z_1}x^T = \left[S_1 \odot W_2^T \frac{\partial L}{\partial z_2}\right]x^T$$

$$\frac{\partial L}{\partial x} = \frac{\partial x}{\partial W_1}\frac{\partial L}{\partial x} = W_1^T\frac{\partial L}{\partial x} = W_1^T\left[S_1 \odot W_2^T\frac{\partial L}{\partial z_2}\right]$$

$$\nabla_{W_2} L = \frac{\partial L}{\partial W_2} = \frac{\partial L}{\partial z_2}h_1^T$$

$$\nabla_{b_2} L = \frac{\partial L}{\partial b_2} = \frac{\partial L}{\partial z_2}$$

$$\nabla_{W_1} L = \frac{\partial L}{\partial W_1} = \frac{\partial z_1}{\partial W_1}\frac{\partial L}{\partial z_1} = \frac{\partial L}{\partial z_1}x^T = \left[S_1 \odot W_2^T\frac{\partial L}{\partial z_2}\right]x^T$$

$$\nabla_{b_1} L = \frac{\partial L}{\partial z_1} = \frac{\partial h_1}{\partial z_1}\frac{\partial L}{\partial h_1} = S_1 \odot W_2^T\frac{\partial L}{\partial z_2}$$