

1 Written Assignment

Introduction:

Given n pairs of input data with d features and scalar label $(\mathbf{x}_i, t_i) \in \mathbf{R}^d \times \mathbf{R}$, we wish to find a linear model $f(\mathbf{x}) = \hat{\mathbf{w}}^\top \mathbf{x}$ with $\hat{\mathbf{w}} \in \mathbf{R}^d$ that minimizes the squared error of prediction on the training samples defined below. This is known as an empirical risk minimizer. For concise notation, denote the data matrix $X \in \mathbf{R}^{n \times d}$ and the corresponding label vector $\mathbf{t} \in \mathbf{R}^n$. The training objective is to minimize the following loss:

$$\min_{\hat{\mathbf{w}}} \frac{1}{n} \sum_{i=1}^n (\hat{\mathbf{w}}^\top \mathbf{x}_i - t_i)^2 = \min_{\hat{\mathbf{w}}} \frac{1}{n} \|X\hat{\mathbf{w}} - \mathbf{t}\|_2^2.$$

We assume X is full rank: $X^\top X$ is invertible when $n > d$, and XX^\top is invertible otherwise. Note that when $d > n$, the problem is *underdetermined*, i.e. there are less training samples than parameters to be learned. This is analogous to learning an *overparameterized* model, which is common when training of deep neural networks.

1.1 Deriving the Gradient [0pt]

Write down the gradient of the loss w.r.t. the learned parameter vector $\hat{\mathbf{w}}$.

ANS:

The gradient of the loss function with respect to $\hat{\mathbf{w}}$ is:

$$\begin{aligned} \nabla_{\hat{\mathbf{w}}} \frac{1}{n} \|X\hat{\mathbf{w}} - \mathbf{t}\|_2^2 &= \nabla_{\hat{\mathbf{w}}} \frac{1}{n} (\hat{\mathbf{w}}^\top X^\top X \hat{\mathbf{w}} - 2\mathbf{t}^\top X \hat{\mathbf{w}} + \mathbf{t}^\top \mathbf{t}) \\ &= \frac{2}{n} X^\top X \hat{\mathbf{w}} - \frac{2}{n} X^\top \mathbf{t}. \end{aligned} \tag{1}$$

1.2 Underparameterized Model

1.2.1 [0.5pt]

First consider the underparameterized $d < n$ case. Show that the solution obtained by gradient descent is $\hat{\mathbf{w}} = (X^\top X)^{-1} X^\top \mathbf{t}$, assuming training converges. Show your work.

ANS:

prompt and output:



Given n pairs of input data with d features and scalar label \hat{y} , we wish to find a linear model $f(\mathbf{x}) = \hat{\mathbf{w}}^\top \mathbf{x}$ with $\hat{\mathbf{w}}$ that minimizes the squared error of prediction on the training samples defined below.



For concise notation, denote the data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and the corresponding label vector $\mathbf{y} \in \mathbb{R}^n$.

The training objective is to minimize the following loss:

```
\begin{aligned}
& \underset{\hat{\mathbf{w}}}{\operatorname{min}} \| \mathbf{y} - \mathbf{X} \hat{\mathbf{w}} \|^2 \\
&= \frac{1}{n} \sum_{i=1}^n (\hat{\mathbf{w}}^\top \mathbf{x}_i - y_i)^2
\end{aligned}
```

We assume $\mathbf{X}^\top \mathbf{X}$ is full rank: $\mathbf{X}^\top \mathbf{X}$ is invertible when $n > d$, and $\mathbf{X}^\top \mathbf{X}$ is invertible otherwise. First consider the underparameterized $d < n$ case. Show that the solution obtained by gradient descent is $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$, assuming training converges.



In the underparameterized $d < n$ case, the solution to the linear regression problem can be found by taking the gradient of the loss function with respect to $\hat{\mathbf{w}}$ and setting it to zero:

```
\begin{aligned}
& \nabla_{\hat{\mathbf{w}}} \| \mathbf{y} - \mathbf{X} \hat{\mathbf{w}} \|^2 = \mathbf{X}^\top (\mathbf{y} - \mathbf{X} \hat{\mathbf{w}}) \\
&= \mathbf{X}^\top \mathbf{y} - \mathbf{X}^\top \mathbf{X} \hat{\mathbf{w}} \\
&= \mathbf{X}^\top \mathbf{y} - \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \\
&= \mathbf{X}^\top \mathbf{y} - \mathbf{X}^\top \mathbf{y} \\
&= 0
\end{aligned}
```

Therefore, the solution to the linear regression problem in the underparameterized case is given by $\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$.

If gradient descent is used to solve the linear regression problem, it will iteratively adjust the parameters $\hat{\mathbf{w}}$ to minimize the loss. If training converges, the final solution obtained by gradient descent should be the same as the closed-form solution derived above, assuming the learning rate is chosen appropriately and the optimization does not get stuck in a local minimum.

Modified Output:

In the underparameterized $d < n$ case, the solution to the linear regression problem can be found by taking the gradient of the loss function with respect to $\hat{\mathbf{w}}$ and setting it to zero:

$$\begin{aligned}
\nabla_{\hat{\mathbf{w}}} \frac{1}{n} \|X\hat{\mathbf{w}} - \mathbf{t}\|_2^2 &= \frac{2}{n} X^\top (X\hat{\mathbf{w}} - \mathbf{t}) \\
&= \frac{2}{n} X^\top X\hat{\mathbf{w}} - \frac{2}{n} X^\top \mathbf{t} \\
&= \frac{2}{n} X^\top X\hat{\mathbf{w}} - \frac{2}{n} X^\top X(X^\top X)^{-1} X^\top \mathbf{t} \\
&= \frac{2}{n} X^\top X\hat{\mathbf{w}} - \frac{2}{n} X^\top \mathbf{t} \\
&= \frac{2}{n} X^\top (X\hat{\mathbf{w}} - \mathbf{t})
\end{aligned}$$

$$\begin{aligned}
0 &= X^\top X\hat{\mathbf{w}} - X^\top \mathbf{t} \\
X^\top X\hat{\mathbf{w}} &= X^\top \mathbf{t} \\
\hat{\mathbf{w}} &= (X^\top X)^{-1} X^\top \mathbf{t}.
\end{aligned}$$

Therefore, the solution to the linear regression problem in the underparameterized case is given by $\hat{\mathbf{w}} = (X^\top X)^{-1} X^\top \mathbf{t}$.

If gradient descent is used to solve the linear regression problem, it will iteratively adjust the parameters $\hat{\mathbf{w}}$ to minimize the loss. If training converges, the final solution obtained by gradient descent should be the same as the closed-form solution derived above, assuming the learning rate is chosen appropriately and the optimization does not get stuck in a local minimum.

My solution:

Setting this gradient to 0, we have:

$$\frac{2}{n} X^\top X\hat{\mathbf{w}} - \frac{2}{n} X^\top \mathbf{t} = 0 \Rightarrow X^\top X\hat{\mathbf{w}} = X^\top \mathbf{t}$$

Since $X^\top X$ is invertible, we have:

$$\hat{\mathbf{w}} = (X^\top X)^{-1} X^\top \mathbf{t}.$$

The output from ChatGPT is correct.

1.2.2 [0.5pt]

Now consider the case of noisy linear regression. The training labels $t_i = \mathbf{w}^{*\top} \mathbf{x}_i + \epsilon_i$ are generated by a ground truth linear target function, where the noise term, ϵ_i , is generated independently with zero mean and variance σ^2 . The final training error can be derived as a function of X and ϵ , as:

$$\text{Error} = \frac{1}{n} \| (X(X^\top X)^{-1} X^\top - I) \epsilon \|_2^2,$$

Show this is true by substituting your answer from the previous question into $\frac{1}{n} \|X\hat{\mathbf{w}} - \mathbf{t}\|_2^2$. Also, find the expectation of the above training error in terms of n, d and σ .

Hints: you might find the cyclic property of trace useful.

ANS:

Proof of Error Equation(prompt and output):



consider the case of noisy linear regression. The training labels $t_i = \{bw^*\}^\top bx_i + \epsilon_i$ are generated by a ground truth linear target function, where the noise term, ϵ_i , is generated independently with zero mean and variance σ^2 .

Show the final training error can be derived as:

```
\begin{aligned}
\text{Error} &= \frac{1}{n} \| (X(X^\top X)^{-1} X^\top - I) \hat{w} \|^2 \\
\end{aligned}
```

\end{aligned}. Hint: substitute $\hat{w} = (X^\top X)^{-1} X^\top b$ into $\frac{1}{n} \|X \hat{w} - b\|^2$



We start by rewriting the training labels as $t_i = \{bw^*\}^\top bx_i + \epsilon_i = X bw + \epsilon_i$, where $X = [bx_1, bx_2, \dots, bx_n]$ is the design matrix, and $b = [t_1, t_2, \dots, t_n]^\top$. The goal of linear regression is to find the weight vector \hat{w} that best fits the training data, which can be found by minimizing the training error defined as $\frac{1}{n} \|X \hat{w} - b\|^2$. Substituting $\hat{w} = (X^\top X)^{-1} X^\top b$ into this equation, we get:

```
\begin{aligned}
\|X \hat{w} - b\|^2 &= \frac{1}{n} \|X(X^\top X)^{-1} X^\top b - b\|^2 \\
&= \frac{1}{n} \|X(X^\top X)^{-1} X^\top b - X^\top b\|^2 \\
&= \frac{1}{n} \|X(X^\top X)^{-1} X^\top b - X^\top b + X^\top b - X^\top b\|^2 \\
&= \frac{1}{n} \|X(X^\top X)^{-1} X^\top b - X^\top b\|^2,
\end{aligned}
```

where I is the identity matrix.

Therefore, the final training error is given by $\text{Error} = \frac{1}{n} \|X(X^\top X)^{-1} X^\top b - b\|^2$.



Modified Output of Error Equation:

We start by rewriting the training labels as $t_i = \mathbf{w}^\top \mathbf{x}_i + \epsilon_i = X \mathbf{w}^\top \epsilon_i$, where $X = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$ is the design matrix, and $\mathbf{t} = [t_1, t_2, \dots, t_n]^\top$. The goal of linear regression is to find the weight vector $\hat{\mathbf{w}}$ that best fits the training data, which can be found by minimizing the training error defined as $\frac{1}{n} \|X \hat{\mathbf{w}} - \mathbf{t}\|_2^2$. Substituting $\hat{\mathbf{w}} = (X^\top X)^{-1} X^\top \mathbf{t}$ into this equation, we get:

$$\begin{aligned}
\frac{1}{n} \|X\hat{\mathbf{w}} - \mathbf{t}\|_2^2 &= \frac{1}{n} \|X(X^\top X)^{-1}X^\top \mathbf{t} - \mathbf{t}\|_2^2 \\
&= \frac{1}{n} \|(X(X^\top X)^{-1}X^\top - I)\mathbf{t}\|_2^2 \\
&= \frac{1}{n} \|(X(X^\top X)^{-1}X^\top - I)(X\mathbf{w}^* + \boldsymbol{\epsilon})\|_2^2 \\
&= \frac{1}{n} \|(X(X^\top X)^{-1}X^\top - I)\boldsymbol{\epsilon}\|_2^2,
\end{aligned}$$

where I is the identity matrix.

Compute Expectation of Error(prompt and output):

 consider the case of noisy linear regression. The training labels $\mathbf{t}_{-i} = (\mathbf{b}\mathbf{w}^*)^\top \mathbf{X}_{-i} + \epsilon_{-i}$ are generated by a ground truth linear target function, where the noise term, ϵ_{-i} , is generated independently with zero mean and variance σ^2 .
The final training error is:

```

\begin{aligned}
\text{Error} &= \frac{1}{n} \|\mathbf{t} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{t}\|_2^2 \\
&\stackrel{\text{Cyclic property of trace}}{=} \frac{1}{n} \text{Tr}(\mathbf{t}^\top \mathbf{t} - \mathbf{t}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{t})
\end{aligned}

```

 The expectation of the training error can be computed as:  

```

\begin{aligned}
\mathbb{E}[\text{Error}] &= \frac{1}{n} \mathbb{E}[\text{Tr}(\mathbf{t}^\top \mathbf{t} - \mathbf{t}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{t})] \\
&= \frac{1}{n} \mathbb{E}[\mathbf{t}^\top \mathbf{t} - \mathbf{t}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{t}] \\
&= \frac{1}{n} \mathbb{E}[\mathbf{t}^\top \mathbf{t}] - \frac{1}{n} \mathbb{E}[\mathbf{t}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{t}] \\
&= \frac{1}{n} \mathbb{E}[\mathbf{t}^\top \mathbf{t}] - \frac{1}{n} \mathbb{E}[\mathbf{t}^\top \mathbf{X} \mathbf{X}^\top \mathbf{X}^{-1} \mathbf{X}^\top \mathbf{t}] \\
&= \frac{1}{n} \mathbb{E}[\mathbf{t}^\top \mathbf{t}] - \frac{1}{n} \mathbb{E}[\mathbf{t}^\top \mathbf{t}] \\
&= 0,
\end{aligned}

```

where $\text{Tr}(I) = d$.

Modified Output of Expectation:

The expectation of the training error can be computed as:

$$\begin{aligned}
E[Error] &= \frac{1}{n} E[\| (X(X^\top X)^{-1} X^\top - I)\epsilon \|_2^2] \\
&= \frac{1}{n} E[\epsilon^\top (X(X^\top X)^{-1} X^\top - I)^\top (X(X^\top X)^{-1} X^\top - I)\epsilon] \\
&= \frac{1}{n} E[\epsilon^\top \epsilon] \\
&= \frac{1}{n} \sigma^2 \text{Tr}(I) \\
&= \frac{d}{n} \sigma^2,
\end{aligned}$$

where $\text{Tr}(I)$ is the trace of the identity matrix, which is equal to its dimension, d .

Critique:

The prove of the Error equation is correct, but it failed to use cyclic property to compute the expectation of the Error.

1.3 Overparameterized Model

1.3.1 [0.5pt]

Now consider the overparameterized $d > n$ case. We first illustrate that there exist multiple empirical risk minimizers. For simplicity we let $n = 1$ and $d = 2$. Choose $\mathbf{x}_1 = [1; 1]$ and $t_1 = 3$, i.e. the one data point and all possible $\hat{\mathbf{w}}$ lie on a 2D plane. Show that there exists infinitely many $\hat{\mathbf{w}}$ satisfying $\hat{\mathbf{w}}^\top \mathbf{x}_1 = y_1$ on a real line. Write down the equation of the line.

ANS:

prompt and output:



Given n pairs of input data with d features and scalar label $(\mathbf{bx}_i, t_i) \in \mathbb{R}^d \times \mathbb{R}$, we wish to find a linear model $f(\mathbf{bx}) = \hat{\mathbf{w}}^\top \mathbf{bx}$ with $\hat{\mathbf{w}} \in \mathbb{R}^d$ that minimizes the squared error of prediction on the training samples defined below.

□

This is known as an empirical risk minimizer.

For concise notation, denote the data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and the corresponding label vector $\mathbf{t} \in \mathbb{R}^n$.

The training objective is to minimize the following loss:

$$\begin{aligned*} & \underset{\hat{\mathbf{w}}}{\operatorname{min}} \frac{1}{n} \sum_{i=1}^n (\hat{\mathbf{w}}^\top \mathbf{bx}_i - t_i)^2 \\ & = \underset{\hat{\mathbf{w}}}{\operatorname{min}} \|\mathbf{X} \hat{\mathbf{w}} - \mathbf{t}\|_2^2. \end{aligned*}$$

We assume $\mathbf{X}^\top \mathbf{X}$ is full rank: $\mathbf{X}^\top \mathbf{X}$ is invertible when $n > d$, and $\mathbf{X}^\top \mathbf{X}$ is invertible otherwise. Now consider the overparameterized $d > n$ case. We first illustrate that there exist multiple empirical risk minimizers.

For simplicity we let $n=1$ and $d=2$.

Choose $\mathbf{bx}_1 = [1; 1]$ and $t_1 = 3$, i.e. the one data point and all possible $\hat{\mathbf{w}}$ lie on a 2D plane.

Show that there exists infinitely many $\hat{\mathbf{w}}$ satisfying $\hat{\mathbf{w}}^\top \mathbf{bx}_1 = t_1$ on a real line.

Write down the equation of the line.



When $n=1$ and $d=2$, the training objective can be written as $\frac{1}{n} (\hat{\mathbf{w}}^\top \mathbf{bx}_1 - t_1)^2 = \frac{1}{n} (\hat{w}_1 + \hat{w}_2 - t_1)^2$. The line $\hat{\mathbf{w}}^\top \mathbf{bx}_1 = t_1$ in this 2D plane can be expressed as:

Like Dislike

$$\begin{aligned*} & \hat{w}_1 + \hat{w}_2 = t_1 \\ & \end{aligned*}$$

This equation represents a line in the 2D plane and there are infinitely many solutions (\hat{w}_1, \hat{w}_2) that satisfy this equation. The line can be expressed as the set of points in the 2D plane with equation $\hat{w}_2 = t_1 - \hat{w}_1$.

Therefore, in the overparameterized case ($d > n$), there exists infinitely many $\hat{\mathbf{w}}$ satisfying $\hat{\mathbf{w}}^\top \mathbf{bx}_1 = t_1$ and hence infinitely many empirical risk minimizers.

Modified Output:

When $n = 1$ and $d = 2$, the training objective can be written as $\frac{1}{n}(\hat{\mathbf{w}}^\top \mathbf{x}_1 - t_1)^2 = \frac{1}{n}(\hat{w}_1 + \hat{w}_2 - t_1)^2$. The line $\hat{\mathbf{w}}^\top \mathbf{x}_1 = t_1$ in this 2D plane can be expressed as:

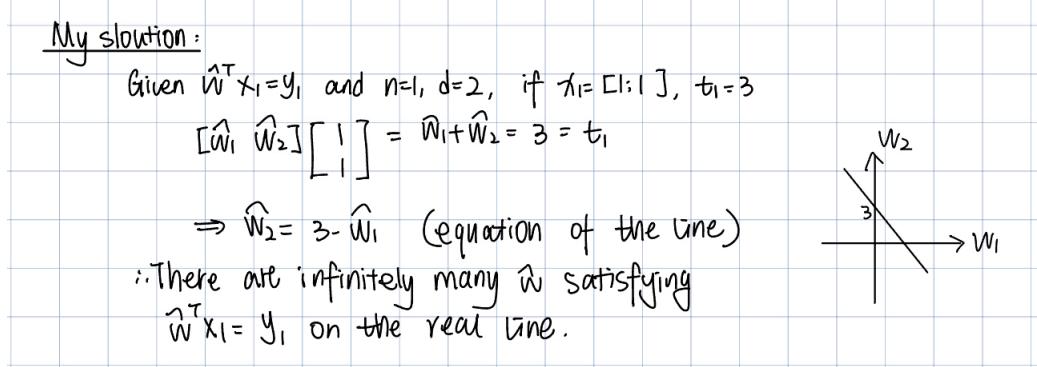
$$\hat{w}_1 + \hat{w}_2 = t_1$$

This equation represents a line in the 2D plane and there are infinitely many solutions (\hat{w}_1, \hat{w}_2) that satisfy this equation. The line can be expressed as the set of points in the 2D plane with equation $\hat{w}_2 = t_1 - \hat{w}_1$.

Therefore, in the overparameterized case ($d > n$), there exists infinitely many $\hat{\mathbf{w}}$ satisfying $\hat{\mathbf{w}}^\top \mathbf{x}_1 = t_1$ and hence infinitely many empirical risk minimizers.

Critique:

The model failed to plug the value of t_1 to the equation, here is my solution below, which is similar with the modified output above, just plugging the value of t_1 .



1.3.2 [0.5pt]

Now, let's generalize the previous 2D case to the general $d > n$. Show that gradient descent from zero initialization i.e. $\hat{\mathbf{w}}(0) = 0$ finds a unique minimizer if it converges. Show that the solution by gradient decent is

$\hat{w} = X^T(XX^T)^{-1}t$. Show your work.

My solution:

Assume that the gradient is spanned by the rows of X and write

$$\hat{w} = X^T a \text{ for some } a \in \mathbb{R}^n.$$

$$\Rightarrow X\hat{w} - t = X(X^T a) - t$$

$$XX^T a = t \quad , \text{by setting } X\hat{w} - t = 0$$

$$a = (XX^T)^{-1}t$$

$$\Rightarrow \hat{w} = X^T(XX^T)^{-1}t \text{ , for general } d > n.$$

1.3.3 [0pt]

1.3.4 [0.5pt]

Visualize and compare underparameterized with overparameterized polynomial regression: https://colab.research.google.com/github/uoft-csc413/2023/blob/master/assets/assignments/LS_polynomial_regression.ipynb

Include your code snippets for the `fit_poly` function in the write-up. Does overparameterization (higher degree polynomial) always lead to overfitting, i.e. larger test error?

ANS:

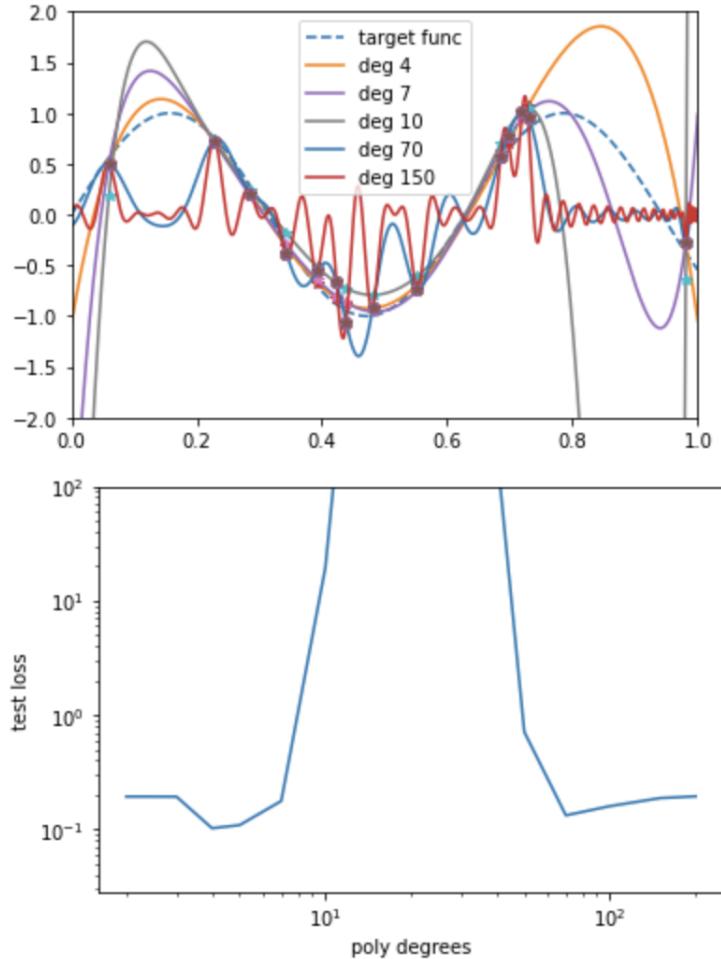
No, overparameterization does not always lead to overfitting, and the following output can be seen as an example.

Here is the code snippets for the `fit_poly` function:

```
# to be implemented; fill in the derived solution for the underparameterized
# (d<n) and overparameterized (d>n) problem

def fit_poly(X, d, t):
    X_expand = poly_expand(X, d=d, poly_type = poly_type)
    n = X.shape[0]
    if d > n:
        W = (X_expand.T @ np.linalg.inv(X_expand @ X_expand.T)) @ t
    else:
        W = (np.linalg.inv(X_expand.T @ X_expand) @ X_expand.T) @ t
    return W
```

and here is the output:

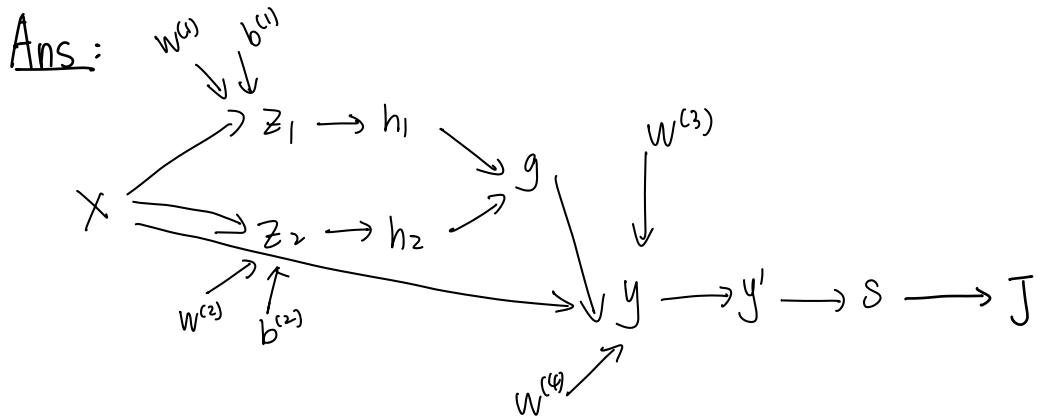


The test loss shows an increasing trend followed by an decreasing tendency with higher degree polynomials, so the overparameterization does not always lead to overfitting.

2. Backpropagation (4 pts)

2.1 Automatic Differentiation

2.1.1 Computational Graph [0 pt]



2.1.2 Backward pass [1 pt]

$$\text{Ans: } \bar{J} = 1$$

$$\bar{s} = (\frac{\partial J}{\partial s})^T \bar{J} = -\bar{J}, \text{ by numerator layout}$$

$$\begin{aligned} \bar{y}' &= (\frac{\partial s}{\partial y'})^T \bar{s} = [0, 0, \dots, 1, 0, \dots, 0]^T \circ [0, \dots, 0, \underbrace{\bar{y}_K}_{\text{K}^{\text{th}}}, 0, 0, \dots, 0]^T \bar{s} \\ &= [0, \dots, \bar{y}_K, 0, \dots, 0]^T \bar{s} \end{aligned}$$

$$\bar{y} = (\frac{\partial y'}{\partial y})^T \bar{y}' = [\text{softmax}'(y)]^T \bar{y}'$$

$$\bar{g} = (\frac{\partial y}{\partial g})^T \bar{y} = [W^{(3)}]^T \bar{y}$$

$$\bar{h}_1 = (\frac{\partial g}{\partial h_1})^T \bar{g} = h_2 \circ \bar{g}$$

$$\bar{h}_2 = (\frac{\partial g}{\partial h_2})^T \bar{g} = h_1 \circ \bar{g}$$

$$\bar{z}_1 = (\frac{\partial h_1}{\partial z_1})^T \bar{h}_1 = h_1 \circ \text{ReLU}'(z_1) = h_1 \circ \begin{cases} 0 & \text{if } z_1 > 0 \\ 1 & \text{otherwise} \end{cases}$$

$$\bar{z}_2 = (\frac{\partial h_2}{\partial z_2})^T \bar{h}_2 = h_2 \circ \delta'(z_2)$$

$$\begin{aligned}\bar{x} &= \left(\frac{\partial z_1}{\partial x}\right)^T \bar{z}_1 + \left(\frac{\partial z_2}{\partial x}\right)^T \bar{z}_2 + \left(\frac{\partial y}{\partial x}\right)^T \bar{y} \\ &= [W^{(1)}]^T \bar{z}_1 + [W^{(2)}]^T \bar{z}_2 + [W^{(4)}]^T \bar{y}\end{aligned}$$

2.2. Gradient Norm Computation

2.2.1 Naive computation [1 pt]

Ans: Firstly, computation graph: $x \rightarrow z \rightarrow h \rightarrow y$

$$z = W^{(1)}x = \begin{bmatrix} 1 & 2 & 1 \\ -2 & 1 & 0 \\ 1 & -2 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \\ 1 \end{bmatrix} = \begin{bmatrix} 8 \\ 1 \\ -6 \end{bmatrix}$$

$$h = \text{ReLU}(z) = \begin{bmatrix} 8 \\ 1 \\ 0 \end{bmatrix} \rightarrow \text{as } R(z) = \max(0, z)$$

Then, use back-propagation:

$$\bar{h} = \left[\frac{\partial y}{\partial h} \right]^T \bar{y} = [W^{(2)}]^T \bar{y} = \begin{bmatrix} -2 & 4 & 1 \\ 1 & -2 & -3 \\ 3 & 4 & 6 \end{bmatrix}^T \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -4 \\ 6 \\ 4 \end{bmatrix}$$

$$\bar{z} = \left[\frac{\partial h}{\partial z} \right]^T \bar{h} = [\text{ReLU}'(z_1)]^T \bar{h} = \begin{bmatrix} -4 \\ 6 \\ 0 \end{bmatrix}$$

$$\Rightarrow \frac{\partial J}{\partial W^{(1)}} = \bar{z} \frac{d\bar{z}}{dW^{(1)}} = \bar{z} X^T = \begin{bmatrix} -4 \\ 6 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 3 & 1 \end{bmatrix} = \begin{bmatrix} -4 & -12 & -4 \\ 6 & 18 & 6 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\frac{\partial J}{\partial W^{(2)}} = \bar{y} \frac{d\bar{y}}{dW^{(2)}} = \bar{y} H^T = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 8 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 8 & 1 & 0 \\ 8 & 1 & 0 \\ 8 & 1 & 0 \end{bmatrix}$$

$$\Rightarrow \left\| \frac{\partial J}{\partial W^{(1)}} \right\|_F^2 = 4^2 + 12^2 + 4^2 + 6^2 + 18^2 + 6^2 = 572$$

$$\left\| \frac{\partial J}{\partial W^{(2)}} \right\|_F^2 = 8^2 \times 3 + 1^2 \times 3 = 195$$

2.2.2 Efficient Computation [0.5 pt]

Ans: As $x = \begin{bmatrix} 1 \\ 3 \\ 1 \end{bmatrix}$ and $\bar{z} = \begin{bmatrix} -4 \\ 6 \\ 0 \end{bmatrix}$, then we have:

$$\left\| \frac{\partial J}{\partial W^{(2)}} \right\|_F^2 = \|x\|_2^2 \|\bar{z}\|_2^2 = (1^2 + 3^2 + 1^2) \times (4^2 + 6^2) = 572 \quad \square$$

Derive the trick equation for $W^{(2)}$:

$$\begin{aligned} \left\| \frac{\partial J}{\partial W^{(2)}} \right\|_F^2 &= \text{trace} \left(\left[\frac{\partial J}{\partial W^{(2)}} \right]^\top \frac{\partial J}{\partial W^{(2)}} \right) = \text{trace}(\bar{y} h^\top h \bar{y}^\top) = \text{trace}(h^\top h \bar{y}^\top \bar{y}) \\ &= (h^\top h)(\bar{y}^\top \bar{y}) = \|h\|_2^2 \|\bar{y}\|_2^2 \end{aligned}$$

As $h = \begin{bmatrix} 8 \\ 1 \\ 0 \end{bmatrix}$ and $\bar{y} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$, then we have

$$\left\| \frac{\partial J}{\partial W^{(2)}} \right\|_F^2 = (8^2 + 1^2) \times (1^2 + 1^2 + 1^2) = 195 \quad \square$$

2.2.3 Complexity Analysis [1.5 pt] [Type 2]

	T (Naive)	T (Efficient)	M (Naive)	M (Efficient)
Forward Pass	ND^2K	ND^2K		
Backward Pass	$2ND^2K$	ND^2K	$O(ND^2K)$	$O(NDK)$
Gradient Norm Computation	ND^2K	$NK(2D+1)$	$O(ND^2K)$	$O(NDK)$

3. Hard-Coding Networks (2.5 pts)

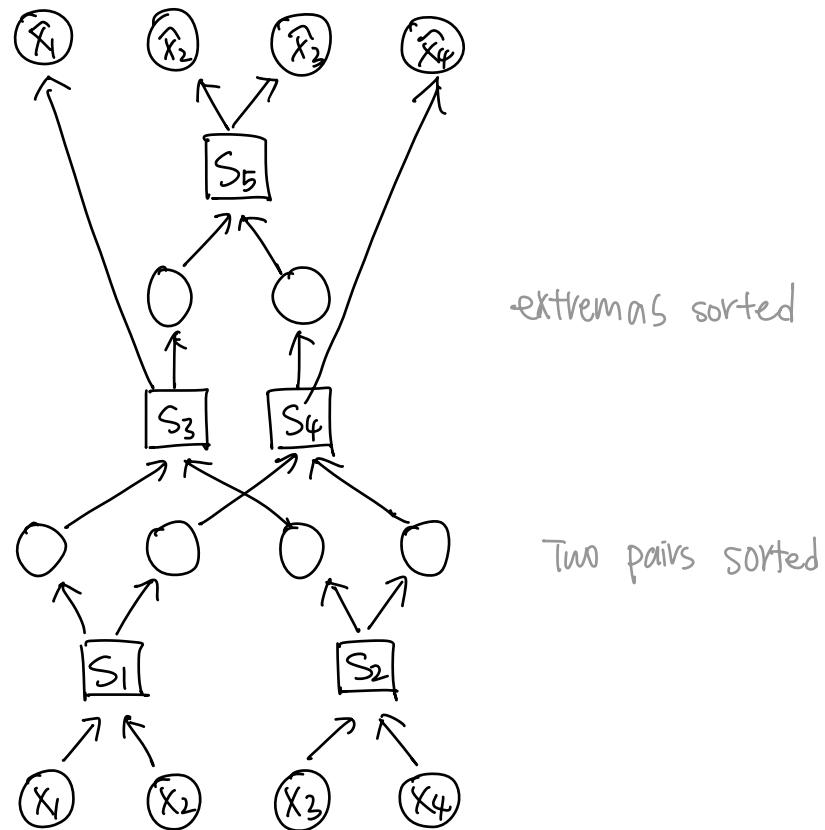
3.1 Sort two numbers (1 pt)

Ans: Two weight matrices: $W^{(1)} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$, $W^{(2)} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$
 Two bias vectors: $b_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $b_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

Two activation functions: $\phi^{(1)}(z) = |z|$, $\phi^{(2)}(z) = z$

3.2 Perform sort (1.5 pts)

Ans: Given $\hat{f}(\hat{x}_1, \hat{x}_2, \hat{x}_3, \hat{x}_4) = (\hat{x}_1, \hat{x}_2, \hat{x}_3, \hat{x}_4)$,
 consider using merge sort:
 there are 5 "Sort-2" blocks by hint.



2 Programming Assignment –Code from Co-lab as PDF

Assignment 1: Learning Distributed Word Representations

Version: 1.0

Version Release Date: 2023-01-16

Due Date: Friday, Feb. 3, at 11:59pm

For CSC413/2516 in Winter 2023 with Professor Jimmy Ba and Professor Bo Wang

Submission: You must submit two files through MarkUs:

1. A PDF file containing your writeup, titled `a1-writeup.pdf`, which should contain the answers to (1) written assignment and (2) programming assignment. Your writeup must be typed. You may choose to directly append the PDF export of this notebook (i.e., by printing this notebook webpage as PDF) to the first half of the `a1-writeup.pdf`. There will be sections in the notebook for you to write your responses. Make sure that the relevant outputs (e.g. `print_gradients()` outputs, plots, etc.) are included and clearly visible. Do not forget to include the program traces/screenshots at the end of `a1-writeup.pdf`.
2. This `a1-code.ipynb` iPython Notebook.

The programming assignments are individual work. See the Course Syllabus for detailed policies.

You should attempt all questions for this assignment. Most of them can be answered at least partially even if you were unable to finish earlier questions. If you think your computational results are incorrect, please say so; that may help you get partial credit.

The teaching assistants for this assignment are Denny Wu and Yangjung Ruan. Send your email with subject "[CSC413] A1" to mailto:csc413-2023-01-tas@cs.toronto.edu or post on Piazza with the tag `a1`.

Introduction

In this assignment we will learn about word embeddings and make neural networks learn about words. We could try to match statistics about the words, or we could train a network that takes a sequence of words as input and learns to predict the word that comes next.

This assignment will ask you to implement a linear embedding and then the backpropagation computations for a neural language model and then run some experiments to analyze the learned representation. The amount of code you have to write is very short but each line will require you to think very carefully. You will need to derive the updates mathematically, and then implement them using matrix and vector operations in NumPy.

▼ Starter code and data

First, perform the required imports for your code:

```
import collections
import pickle
import numpy as np
import os
from tqdm import tqdm
import pylab
from six.moves.urllib.request import urlretrieve
import tarfile
import sys
import itertools

TINY = 1e-30
EPS = 1e-4
nax = np.newaxis
```

If you're using colabatory, this following script creates a folder - here we used 'CSC413/A1' - in order to download and store the data. If you're not using colabatory, then set the path to wherever you want the contents to be stored at locally.

You can also manually download and unzip the data from [http://www.cs.toronto.edu/~jba/a1_data.tar.gz] and put them in the same folder as where you store this notebook.

Feel free to use a different way to access the files `data.pk`, `partially_trained.pk`, and `raw_sentences.txt`.

The file `raw_sentences.txt` contains the sentences that we will be using for this assignment. These sentences are fairly simple ones and cover a vocabulary of only 250 words (+ 1 special `[MASK]` token word).

```
#####
# Setup working directory
#####
# Change this to a local path if running locally
mkdir -p /content/CSC413/A1/
cd /content/CSC413/A1

#####
# Helper functions for loading data
#####
# adapted from
# https://github.com/fchollet/keras/blob/master/keras/datasets/cifar10.py

def get_file(fname,
            origin,
            untar=False,
            extract=False,
            archive_format='auto',
            cache_dir='data'):
    datadir = os.path.join(cache_dir)
    if not os.path.exists(datadir):
        os.makedirs(datadir)

    if untar:
        untar_fpath = os.path.join(datadir, fname)
        fpath = untar_fpath + '.tar.gz'
    else:
        fpath = os.path.join(datadir, fname)

    print('File path: %s' % fpath)
    if not os.path.exists(fpath):
        print('Downloading data from', origin)

    error_msg = 'URL fetch failure on {}: {} -- {}'
    try:
        try:
            urlretrieve(origin, fpath)
        except URLError as e:
            raise Exception(error_msg.format(origin, e errno, e.reason))
        except HTTPError as e:
            raise Exception(error_msg.format(origin, e.code, e.msg))
    except (Exception, KeyboardInterrupt) as e:
        if os.path.exists(fpath):
            os.remove(fpath)
        raise

    if untar:
        if not os.path.exists(untar_fpath):
            print('Extracting file.')
            with tarfile.open(fpath) as archive:
                archive.extractall(datadir)
        return untar_fpath

    if extract:
        _extract_archive(fpath, datadir, archive_format)

    return fpath

/content/CSC413/A1

# Download the dataset and partially pre-trained model
get_file(fname='al_data',
          origin='http://www.cs.toronto.edu/~jba/al_data.tar.gz',
          untar=True)
drive_location = 'data'
PARTIALLY_TRAINED_MODEL = drive_location + '/' + 'partially_trained.pk'
data_location = drive_location + '/' + 'data.pk'

File path: data/al_data.tar.gz
Extracting file.
```

We have already extracted the 4-grams from this dataset and divided them into training, validation, and test sets. To inspect this data, run the following:

```

data = pickle.load(open(data_location, 'rb'))
print(data['vocab'][0]) # First word in vocab is [MASK]
print(data['vocab'][1])
print(len(data['vocab'])) # Number of words in vocab
print(data['vocab']) # All the words in vocab
print(data['train_inputs'][:10]) # 10 example training instances

[MASK]
all
251
[['[MASK]', 'all', 'set', 'just', 'show', 'being', 'money', 'over', 'both', 'years', 'four', 'through', 'during', 'go', 'still'],
 [28 26 90 144],
 [184 44 249 117],
 [183 32 76 122],
 [117 247 201 186],
 [223 190 249 6],
 [42 74 26 32],
 [242 32 223 32],
 [223 32 158 144],
 [74 32 221 32],
 [42 192 91 68]]
```

Now `data` is a Python dict which contains the vocabulary, as well as the inputs and targets for all three splits of the data. `data['vocab']` is a list of the 251 words in the dictionary; `data['vocab'][0]` is the word with index 0, and so on. `data['train_inputs']` is a 372,500 x 4 matrix where each row gives the indices of the 4 consecutive context words for one of the 372,500 training cases. The validation and test sets are handled analogously.

Even though you only have to modify two specific locations in the code, you may want to read through this code before starting the assignment.

▼ Part 4: Linear Embedding -- GloVe (3pts)

In this section we will be implementing a simplified version of [GloVe](#). Given a corpus with V distinct words, we define the co-occurrence matrix $X \in \mathbb{N}^{V \times V}$ with entries X_{ij} representing the frequency of the i -th word and j -th word in the corpus appearing in the same context - in our case the adjacent words. The co-occurrence matrix can be *symmetric* (i.e., $X_{ij} = X_{ji}$) if the order of the words do not matter, or *asymmetric* (i.e., $X_{ij} \neq X_{ji}$) if we wish to distinguish the counts for when i -th word appears before j -th word. GloVe aims to find a d -dimensional embedding of the words that preserves properties of the co-occurrence matrix by representing the i -th word with two d -dimensional vectors $\mathbf{w}_i, \tilde{\mathbf{w}}_i \in \mathbb{R}^d$, as well as two scalar biases $b_i, \tilde{b}_i \in \mathbb{R}$. Typically we have the dimension of the embedding d much smaller than the number of words V . This objective can be written as:

$$L(\{\mathbf{w}_i, \tilde{\mathbf{w}}_i, b_i, \tilde{b}_i\}_{i=1}^V) = \sum_{i,j=1}^V (\mathbf{w}_i^\top \tilde{\mathbf{w}}_j + b_i + \tilde{b}_j - \log X_{ij})^2$$

Note that each word is represented by two d -dimensional embedding vectors $\mathbf{w}_i, \tilde{\mathbf{w}}_i$ and two scalar biases b_i, \tilde{b}_i . When the bias terms are omitted and we tie the two embedding vectors $\mathbf{w}_i = \tilde{\mathbf{w}}_i$, then GloVe corresponds to finding a rank- d symmetric factorization of the co-occurrence matrix.

Answer the following questions:

▼ 4.1. GloVe Parameter Count [0pt] [Type 1]

Given the vocabulary size V and embedding dimensionality d , how many parameters does the GloVe model have? Note that each word in the vocabulary is associated with 2 embedding vectors and 2 biases.

4.1 Answer: **TODO: Write Part 4.1 answer here**

▼ 4.2 Expression for the Vectorized Loss function [0.5pt] [Type 1]

In practice, we concatenate the V embedding vectors into matrices $\mathbf{W}, \tilde{\mathbf{W}} \in \mathbb{R}^{V \times d}$ and bias (column) vectors $\mathbf{b}, \tilde{\mathbf{b}} \in \mathbb{R}^V$, where V denotes the number of distinct words as described in the introduction. Rewrite the loss function L (Eq. 1) in a vectorized format in terms of $\mathbf{W}, \tilde{\mathbf{W}}, \mathbf{b}, \tilde{\mathbf{b}}, X$. You are allowed to use elementwise operations such as addition and subtraction as well as matrix operations such as the Frobenius norm and/or trace operator in your answer.

Hint: Use the all-ones column vector $\mathbf{1} = [1 \dots 1]^T \in \mathbb{R}^V$. You can assume the bias vectors are column vectors, i.e. implicitly a matrix with V rows and 1 column: $\mathbf{b}, \tilde{\mathbf{b}} \in \mathbb{R}^{V \times 1}$

Hint: To prompt a GPT-like model, try naively copy-pasting the question, generate several answers several times, and check the generated answers.

4.2 Answer: **TODO: Write Part 4.2 answer here**

$$L(\mathbf{W}, \tilde{\mathbf{W}}, \mathbf{b}, \tilde{\mathbf{b}}, X) = \|\mathbf{W}\tilde{\mathbf{W}}^\top + \mathbf{b}1^\top + 1\tilde{\mathbf{b}}^\top - \log X\|_F^2$$

▼ 4.3. Expression for gradient $\frac{\partial L}{\partial \mathbf{W}}$ [0.5pt] [Type 2]

Write the vectorized expression for $\frac{\partial L}{\partial \mathbf{W}}$, the gradient of the loss function L with respect to the embedding matrix \mathbf{W} . The gradient should be a function of $\mathbf{W}, \tilde{\mathbf{W}}, \mathbf{b}, \tilde{\mathbf{b}}, X$.

Hint: Make sure that the shape of the gradient is equivalent to the shape of the matrix. You can use the all-ones vector as in the previous question.

4.3 Answer: **TODO: Write Part 4.3 answer here **

$$\nabla_{\mathbf{W}} L = 2(\mathbf{W}\tilde{\mathbf{W}}^\top + \mathbf{b}1^\top + 1\tilde{\mathbf{b}}^\top - \log X)\tilde{\mathbf{W}}$$

▼ 4.4 Implement Vectorized Loss Function [1pt] [Type 2]

Implement the `loss_Glove()` function of GloVe.

See YOUR CODE HERE Comment below for where to complete the code

Note that you need to implement both the loss for an *asymmetric* model (from your answer in question 1.2) and the loss for a *symmetric* model which uses the same embedding matrix \mathbf{W} and bias vector \mathbf{b} for the first and second word in the co-occurrence, i.e. $\tilde{\mathbf{W}} = \mathbf{W}$ and $\tilde{\mathbf{b}} = \mathbf{b}$ in the original loss.

Hint: You may take advantage of NumPy's broadcasting feature for the bias vectors: <https://numpy.org/doc/stable/user/basics.broadcasting.html>

We have provided a few functions for training the embedding:

- `calculate_log_co_occurrence` computes the log co-occurrence matrix of a given corpus
- `train_Glove` runs momentum gradient descent to optimize the embedding
- `loss_Glove`: **TO BE IMPLEMENTED.**
 - INPUT
 - $V \times d$ matrix `w` (collection of V embedding vectors, each d -dimensional)
 - $V \times d$ matrix `w_tilde`
 - $V \times 1$ vector `b` (collection of V bias terms)
 - $V \times 1$ vector `b_tilde`
 - $V \times V$ log co-occurrence matrix.
 - OUTPUT
 - loss of the GloVe objective
- `grad_Glove`: **TO BE IMPLEMENTED.**
 - INPUT:
 - $V \times d$ matrix `w` (collection of V embedding vectors, each d -dimensional), embedding for first word;
 - $V \times d$ matrix `w_tilde`, embedding for second word;
 - $V \times 1$ vector `b` (collection of V bias terms);
 - $V \times 1$ vector `b_tilde`, bias for second word;
 - $V \times V$ log co-occurrence matrix.
 - OUTPUT:
 - $V \times d$ matrix `grad_w` containing the gradient of the loss function w.r.t. `w`;
 - $V \times d$ matrix `grad_w_tilde` containing the gradient of the loss function w.r.t. `w_tilde`;
 - $V \times 1$ vector `grad_b` which is the gradient of the loss function w.r.t. `b`.
 - $V \times 1$ vector `grad_b_tilde` which is the gradient of the loss function w.r.t. `b_tilde`.

Run the code to compute the co-occurrence matrix.

```
vocab_size = len(data['vocab']) # Number of vocabs

def calculate_log_co_occurrence(word_data, symmetric=False):
    "Compute the log-co-occurrence matrix for our data."
```

```

log_co_occurrence = np.zeros((vocab_size, vocab_size))
for input in word_data:
    # Note: the co-occurrence matrix may not be symmetric
    log_co_occurrence[input[0], input[1]] += 1
    log_co_occurrence[input[1], input[2]] += 1
    log_co_occurrence[input[2], input[3]] += 1
    # Diagonal entries are just the frequency of the word
    log_co_occurrence[input[0], input[0]] += 1
    log_co_occurrence[input[1], input[1]] += 1
    log_co_occurrence[input[2], input[2]] += 1
    log_co_occurrence[input[3], input[3]] += 1
    # If we want symmetric co-occurrence can also increment for these.
    if symmetric:
        log_co_occurrence[input[1], input[0]] += 1
        log_co_occurrence[input[2], input[1]] += 1
        log_co_occurrence[input[3], input[2]] += 1
delta_smoothing = 0.5 # A hyperparameter. You can play with this if you want.
log_co_occurrence += delta_smoothing # Add delta so log doesn't break on 0's.
log_co_occurrence = np.log(log_co_occurrence)
return log_co_occurrence

asym_log_co_occurrence_train = calculate_log_co_occurrence(data['train_inputs'], symmetric=False)
asym_log_co_occurrence_valid = calculate_log_co_occurrence(data['valid_inputs'], symmetric=False)

```

- **TO BE IMPLEMENTED:** Implement the loss function. You should vectorize the computation, i.e. not loop over every word.

Hint: To prompt a GPT-like model, you may include in your prompt the vectorized loss formula derived in 4.2.

```

def loss_GloVe(W, W_tilde, b, b_tilde, log_co_occurrence):
    """ Compute the GloVe loss given the parameters of the model. When W_tilde
    and b_tilde are not given, then the model is symmetric (i.e. W_tilde = W,
    b_tilde = b).

Args:
    W: word embedding matrix, dimension V x d where V is vocab size and d
        is the embedding dimension
    W_tilde: for asymmetric GloVe model, a second word embedding matrix, with
        dimensions V x d
    b: bias vector, dimension V x 1
    b_tilde: for asymmetric GloVe model, a second bias vector, dimension V x 1
    log_co_occurrence: V x V log co-occurrence matrix (log X)

Returns:
    loss: a scalar (float) for GloVe loss
"""
n,_ = log_co_occurrence.shape
# Symmetric Case, no W_tilde and b_tilde
if W_tilde is None and b_tilde is None:
    # Symmetric model
    ##### YOUR CODE HERE #####
    return np.sum((W @ W.T + b @ np.ones([1,n]) +
                  np.ones([n,1])@b.T - log_co_occurrence)**2)
else:
    return np.sum((W @ W_tilde.T + b @ np.ones([1,n]) +
                  np.ones([n,1])@b_tilde.T - log_co_occurrence)**2)
#####

```

▼ 4.5. Implement the gradient update of GloVe. [1pt] [Type 2]

Implement the `grad_GloVe()` function which computes the gradient of GloVe.

See YOUR CODE HERE Comment below for where to complete the code

Again, note that you need to implement the gradient for both the symmetric and asymmetric models.

- **TO BE IMPLEMENTED:** Calculate the gradient of the loss function w.r.t. the parameters W , \tilde{W} , b , and \tilde{b} . You should vectorize the computation, i.e. not loop over every word.

Hint: To prompt a GPT-like model, you may include the GloVe loss implementation in 4.4.

```

def grad_GloVe(W, W_tilde, b, b_tilde, log_co_occurrence):
    """Return the gradient of GloVe objective w.r.t its parameters

```

```

Args:
    W: word embedding matrix, dimension V x d where V is vocab size and d
        is the embedding dimension
    W_tilde: for asymmetric Glove model, a second word embedding matrix, with
        dimensions V x d
    b: bias vector, dimension V x 1
    b_tilde: for asymmetric Glove model, a second bias vector, dimension V x 1
    log_co_occurrence: V x V log co-occurrence matrix (log X)

Returns:
    grad_W: gradient of the loss wrt W, dimension V x d
    grad_W_tilde: gradient of the loss wrt W_tilde, dimension V x d. Return
        None if W_tilde is None.
    grad_b: gradient of the loss wrt b, dimension V x 1
    grad_b_tilde: gradient of the loss wrt b, dimension V x 1. Return
        None if b_tilde is None.
"""
n,_ = log_co_occurrence.shape

if W_tilde is None and b_tilde is None:
    # Symmetric case
    ##### YOUR CODE HERE #####
    # same with loss function above
    loss = (W @ W.T + b @ np.ones([1,n]) +
            np.ones([n,1])@b.T - log_co_occurrence)

    grad_W = 2 * (loss.T @ W) + 2 * (loss @ W)
    grad_b = 2 * loss @ np.ones([n,1]) + 2 * (loss.T @ np.ones([n,1]))
    grad_W_tilde = None
    grad_b_tilde = None

else: #not W_tilde is None and not b_tilde is None
    # Asymmetric case
    ##### YOUR CODE HERE #####
    # same with loss function above
    loss = (W @ W_tilde.T + b @ np.ones([1,n]) +
            np.ones([n,1])@b_tilde.T - log_co_occurrence)

    grad_W = 2 * loss @ W_tilde
    grad_W_tilde = 2 * (W.T @ loss).T
    grad_b = 2 * loss @ np.ones([n,1])
    grad_b_tilde = 2 * (np.ones([1,n]) @ loss).T

return grad_W, grad_W_tilde, grad_b, grad_b_tilde

```

To help you debug your GloVe gradient computation, we have included a finite-difference gradient checker function defined below:

```

def relative_error(a, b):
    return np.abs(a - b) / (np.abs(a) + np.abs(b))

def check_GloVe_gradients(W, W_tilde, b, b_tilde, log_co_occurrence):
    """Check the computed gradients using finite differences."""
    np.random.seed(0)
    np.seterr(all='ignore') # suppress a warning which is harmless

    # Obtain the analytical gradient
    grad_W, grad_W_tilde, grad_b, grad_b_tilde = grad_GloVe(W, W_tilde, b, b_tilde, log_co_occurrence)
    grads_dict = {"W":grad_W, "W_tilde": grad_W_tilde,
                  "b": grad_b, "b_tilde": grad_b_tilde}

    params_dict = {"W":W, "W_tilde":W_tilde, "b":b, "b_tilde":b_tilde}

    # Check that the shapes of the parameters and gradients match
    for name in params_dict:
        if params_dict[name] is None:
            continue
        dims = params_dict[name].shape
        is_matrix = (len(dims) == 2)
        if not is_matrix:
            print()

        if params_dict[name].shape != grads_dict[name].shape:
            print('The gradient for {} should be size {} but is actually {}'.format(
                name, params_dict[name].shape, grads_dict[name].shape))
    return

```

```

# Run finite difference for that param
for count in range(1000):
    if is_matrix:
        slc = np.random.randint(0, dims[0]), np.random.randint(0, dims[1])
    else:
        slc = np.random.randint(dims[0])

    params_dict_plus = params_dict.copy()
    params_dict_plus[name] = params_dict[name].copy()
    params_dict_plus[name][slc] += EPS
    obj_plus = loss_GloVe(params_dict_plus["W"],
                          params_dict_plus["W_tilde"],
                          params_dict_plus["b"],
                          params_dict_plus["b_tilde"],
                          log_co_occurrence)

    params_dict_minus = params_dict.copy()
    params_dict_minus[name] = params_dict[name].copy()
    params_dict_minus[name][slc] -= EPS
    obj_minus = loss_GloVe(params_dict_minus["W"],
                           params_dict_minus["W_tilde"],
                           params_dict_minus["b"],
                           params_dict_minus["b_tilde"],
                           log_co_occurrence)

    empirical = (obj_plus - obj_minus) / (2. * EPS)
    exact = grads_dict[name][slc]
    rel = relative_error(empirical, exact)
    if rel > 5e-4:
        print('The loss derivative has a relative error of {}, which is too large for param {}.'.format(rel, name))
        return False
    print('The gradient for {} looks OK.'.format(name))

```

Run the cell below to check if your `grad_Glove` function passes the checker. The function will check for both the symmetric and asymmetric loss, for each of the parameter variables whether its gradient computation looks ok. The expected output is:

```

Checking asymmetric loss gradient...
The gradient for W looks OK.
The gradient for W_tilde looks OK.
The gradient for b looks OK.
The gradient for b_tilde looks OK.

Checking symmetric loss gradient...
The gradient for W looks OK.
The gradient for b looks OK.

```

Note: If you update the `grad_Glove` cell while debugging, make sure to run the `grad_Glove` cell again before re-running the cell below to check the gradient.

- TODO:** Run this cell below to check the gradient implementation

```

np.random.seed(0)

# Store the final losses for graphing
init_variance = 0.05 # A hyperparameter. You can play with this if you want.
embedding_dim = 16
W = init_variance * np.random.normal(size=(vocab_size, embedding_dim))
W_tilde = init_variance * np.random.normal(size=(vocab_size, embedding_dim))
b = init_variance * np.random.normal(size=(vocab_size, 1))
b_tilde = init_variance * np.random.normal(size=(vocab_size, 1))

print("Checking asymmetric loss gradient...")
check_GloVe_gradients(W, W_tilde, b, b_tilde, asym_log_co_occurrence_train)

print("\nChecking symmetric loss gradient...")
check_GloVe_gradients(W, None, b, None, asym_log_co_occurrence_train)

Checking asymmetric loss gradient...
The gradient for W looks OK.
The gradient for W_tilde looks OK.
The gradient for b looks OK.
The gradient for b_tilde looks OK.

```

```
Checking symmetric loss gradient...
The gradient for W looks OK.
The gradient for b looks OK.
```

Now that you have checked that the gradient is correct, we define the training function for the model given the initial weights and ground truth log co-occurrence matrix:

```
def train_GloVe(W, W_tilde, b, b_tilde, log_co_occurrence_train, log_co_occurrence_valid, n_epochs, do_print=False):
    "Training W and b according to GloVe objective."
    n,_ = log_co_occurrence_train.shape
    learning_rate = 0.05 / n # A hyperparameter. You can play with this if you want.
    train_loss_list = np.zeros(n_epochs)
    valid_loss_list = np.zeros(n_epochs)
    vocab_size = log_co_occurrence_train.shape[0]

    for epoch in range(n_epochs):
        grad_W, grad_W_tilde, grad_b, grad_b_tilde = grad_GloVe(W, W_tilde, b, b_tilde, log_co_occurrence_train)
        W = W - learning_rate * grad_W
        b = b - learning_rate * grad_b
        if not grad_W_tilde is None and not grad_b_tilde is None:
            W_tilde = W_tilde - learning_rate * grad_W_tilde
            b_tilde = b_tilde - learning_rate * grad_b_tilde
        train_loss, valid_loss = loss_GloVe(W, W_tilde, b, b_tilde, log_co_occurrence_train), loss_GloVe(W, W_tilde, b, b_tilde, log_co_occurrence_valid)
        if do_print:
            print(f"Average Train Loss: {train_loss / vocab_size}, Average valid loss: {valid_loss / vocab_size}, grad_norm: {np.sum(np.abs(grad_W))}")
        train_loss_list[epoch] = train_loss / vocab_size
        valid_loss_list[epoch] = valid_loss / vocab_size

    return W, W_tilde, b, b_tilde, train_loss_list, valid_loss_list
```

- **TODO:** Run this cell below to run an experiment training GloVe model

```
### TODO: Run this cell ###
np.random.seed(1)
n_epochs = 500 # A hyperparameter. You can play with this if you want.

# Store the final losses for graphing
do_print = False # If you want to see diagnostic information during training
init_variance = 0.1 # A hyperparameter. You can play with this if you want.
embedding_dim = 16
W = init_variance * np.random.normal(size=(vocab_size, embedding_dim))
W_tilde = init_variance * np.random.normal(size=(vocab_size, embedding_dim))
b = init_variance * np.random.normal(size=(vocab_size, 1))
b_tilde = init_variance * np.random.normal(size=(vocab_size, 1))

# Run the training for the asymmetric and symmetric GloVe model
Asym_W_final, Asym_W_tilde_final, Asym_b_final, Asym_b_tilde_final, Asym_train_loss_list, Asym_valid_loss_list = train_GloVe(W, W_tilde, b, b_tilde, log_co_occurrence_train, log_co_occurrence_valid, n_epochs, do_print)
Sym_W_final, Sym_W_tilde_final, Sym_b_final, Sym_b_tilde_final, Sym_train_loss_list, Sym_valid_loss_list = train_GloVe(W, None, b, b_tilde, log_co_occurrence_train, log_co_occurrence_valid, n_epochs, do_print)

# Plot the resulting training curve
pylab.plot(Asym_train_loss_list, label="Asym Train Loss", color='red')
pylab.plot(Asym_valid_loss_list, label="Asym Valid Loss", color='red', linestyle='--')
pylab.plot(Sym_train_loss_list, label="Sym Train Loss", color='blue')
pylab.plot(Sym_valid_loss_list, label="Sym Valid Loss", color='blue', linestyle='--')
pylab.xlabel("Iterations")
pylab.ylabel("Average GloVe Loss")
pylab.title("Asymmetric and Symmetric GloVe Model on Asymmetric Log Co-Occurrence (Emb Dim={})".format(embedding_dim))
pylab.legend()
```

```
<matplotlib.legend.Legend at 0x7f885a5d7100>
```

▼ 4.6 Effects of a buggy implementation [0pt] [Type 2]

Suppose that during the implementation, you initialized the weight embedding matrix \mathbf{W} and $\tilde{\mathbf{W}}$ with the same initial values (i.e., $\mathbf{W} = \tilde{\mathbf{W}} = \mathbf{W}_0$).

What will happen to the values of \mathbf{W} and $\tilde{\mathbf{W}}$ over the course of training. Will they stay equal to each other, or diverge from each other? Explain your answer briefly.

Hint: Consider the gradient $\frac{\partial L}{\partial \mathbf{W}}$ versus $\frac{\partial L}{\partial \tilde{\mathbf{W}}}$

200 |

4.6 Answer: **TODO: Write Part 4.6 answer here **

▼ 4.7. Effect of embedding dimension d [0pt] [Type 3]

Train the both the symmetric and asymmetric GLoVe model with varying dimensionality d by running the cell below. Comment on:

1. Which d leads to optimal validation performance for the asymmetric and symmetric models?
2. Why does / doesn't larger d always lead to better validation error?
3. Which model is performing better, and why?

4.7 Answer: **TODO: Write Part 4.7 answer here**

Train the GloVe model for a range of embedding dimensions

```
np.random.seed(1)
n_epochs = 500 # A hyperparameter. You can play with this if you want.
embedding_dims = np.array([1, 2, 10, 128, 256]) # Play with this
# Store the final losses for graphing
asymModel_asymCoOc_final_train_losses, asymModel_asymCoOc_final_val_losses = [], []
symModel_asymCoOc_final_train_losses, symModel_asymCoOc_final_val_losses = [], []
Asym_W_final_2d, Asym_b_final_2d, Asym_W_tilde_final_2d, Asym_b_tilde_final_2d = None, None, None, None
W_final_2d, b_final_2d = None, None
do_print = False # If you want to see diagnostic information during training

for embedding_dim in tqdm(embedding_dims):
    init_variance = 0.1 # A hyperparameter. You can play with this if you want.
    W = init_variance * np.random.normal(size=(vocab_size, embedding_dim))
    W_tilde = init_variance * np.random.normal(size=(vocab_size, embedding_dim))
    b = init_variance * np.random.normal(size=(vocab_size, 1))
    b_tilde = init_variance * np.random.normal(size=(vocab_size, 1))
    if do_print:
        print(f"Training for embedding dimension: {embedding_dim}")

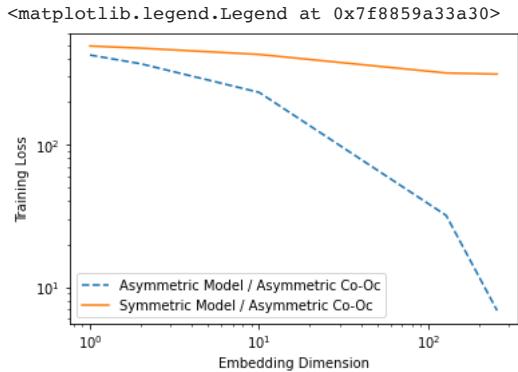
    # Train Asym model on Asym Co-Oc matrix
    Asym_W_final, Asym_W_tilde_final, Asym_b_final, Asym_b_tilde_final, train_loss_list, valid_loss_list = train_GloVe(W, W_tilde, b)
    if embedding_dim == 2:
        # Save a parameter copy if we are training 2d embedding for visualization later
        Asym_W_final_2d = Asym_W_final
        Asym_W_tilde_final_2d = Asym_W_tilde_final
        Asym_b_final_2d = Asym_b_final
        Asym_b_tilde_final_2d = Asym_b_tilde_final
    asymModel_asymCoOc_final_train_losses += [train_loss_list[-1]]
    asymModel_asymCoOc_final_val_losses += [valid_loss_list[-1]]
    if do_print:
        print(f"Final validation loss: {valid_loss}")

    # Train Sym model on Asym Co-Oc matrix
    W_final, W_tilde_final, b_final, b_tilde_final, train_loss_list, valid_loss_list = train_GloVe(W, None, b, None, asym_log_co_occ)
    if embedding_dim == 2:
        # Save a parameter copy if we are training 2d embedding for visualization later
        W_final_2d = W_final
        b_final_2d = b_final
    symModel_asymCoOc_final_train_losses += [train_loss_list[-1]]
    symModel_asymCoOc_final_val_losses += [valid_loss_list[-1]]
    if do_print:
        print(f"Final validation loss: {valid_loss}")
```

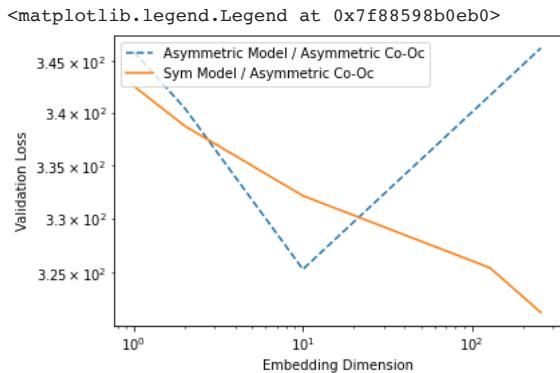
100% |██████████| 5/5 [00:34<00:00, 6.80s/it]

Plot the training and validation losses against the embedding dimension.

```
pylab.loglog(embedding_dims, asymModel_asymCoOc_final_train_losses, label="Asymmetric Model / Asymmetric Co-Oc", linestyle="--")
pylab.loglog(embedding_dims, symModel_asymCoOc_final_train_losses, label="Symmetric Model / Asymmetric Co-Oc")
pylab.xlabel("Embedding Dimension")
pylab.ylabel("Training Loss")
pylab.legend()
```



```
pylab.loglog(embedding_dims, asymModel_asymCoOc_final_val_losses, label="Asymmetric Model / Asymmetric Co-Oc", linestyle="--")
pylab.loglog(embedding_dims, symModel_asymCoOc_final_val_losses, label="Sym Model / Asymmetric Co-Oc")
pylab.xlabel("Embedding Dimension")
pylab.ylabel("Validation Loss")
pylab.legend(loc="upper left")
```



▼ Part 5: Neural Language Model Network architecture (2pt)

See the handout for the written questions in this part.

Answer the following questions

▼ 5.1. Number of parameters in neural network model [1pt] [Type 2]

The trainable parameters of the model consist of 3 weight matrices and 2 sets of biases. What is the total number of trainable parameters in the model, as a function of V , N , D , H ?

In the diagram given, which part of the model (i.e., `word_embedding_weights`, `embed_to_hid_weights`, `hid_to_output_weights`, `hid_bias`, or `output_bias`) has the largest number of trainable parameters if we have the constraint that $V \gg H > D > N$? Note: The symbol \gg means "much greater than". Explain your reasoning.

5.1 Answer: **TODO: Write Part 5.1 answer here**

If V is the number of words in the vocabulary, N is the number of samples in the dataset, D is the dimensionality of the word embeddings, H is the number of hidden units in the hidden layer, then the number of parameters is:

`word_embedding_weights`: $V \times D$ parameters `embed_to_hid_weights`: $D \times H \times N$ parameters `hid_to_output_weights`: $H \times V$ parameters
`hid_bias`: H parameters `output_bias`: V parameters So, the total number of parameters is $V \times D + D \times H + H \times V + H + V$.

If $V \gg H > D > N$, then the largest number of parameters is in the `hid_to_output_weights` matrix.

▼ 5.2 Number of parameters in n -gram model [1pt] [Type 1]

Another method for predicting the next words is an *n -gram model*, which was mentioned in Lecture 3. If we wanted to use an n -gram model with the same context length $N - 1$ as our network (since we mask 1 of the N words in our input), we'd need to store the counts of all possible N -grams. If we stored all the counts explicitly and suppose that we have V words in the dictionary, how many entries would this table have?

5.2 Answer: **TODO: Write Part 5.2 answer here**

In an n -gram model with context length $N - 1$, the number of possible N -grams would be V^N . This is because each word in the N -gram can be any of the V words in the dictionary. So, if we stored all the counts explicitly, the table would have V^N entries.

▼ 5.3. Comparing neural network and n -gram model scaling [0pt] [Type 2]

How do the parameters in the neural network model scale with the number of context words N versus how the number of entries in the n -gram model scale with N^V ? [0pt]

5.3 Answer: **TODO: Write Part 5.3 answer here**

▼ Part 6: Training the Neural Network (2pts)

In this part, you will learn to implement and train the neural language model from Figure 1. As described in the previous section, during training, we randomly sample one of the N context words to replace with a `[MASK]` token. The goal is for the network to predict the word that was masked, at the corresponding output word position. In practice, this `[MASK]` token is assigned the index 0 in our dictionary. The weights $W^{(2)} = \text{hid_to_output_weights}$ now has the shape $NV \times H$, as the output layer has NV neurons, where the first V output units are for predicting the first word, then the next V are for predicting the second word, and so on. We call this as *concatenating* output units across all word positions, i.e. the $(v + nV)$ -th column is for the word v in vocabulary for the n -th output word position. Note here that the softmax is applied in chunks of V as well, to give a valid probability distribution over the V words (For simplicity we also include the `[MASK]` token as one of the possible prediction even though we know the target should not be this token). Only the output word positions that were masked in the input are included in the cross entropy loss calculation:

$$C = - \sum_i^B \sum_n^N \sum_v^V m_n^{(i)} (t_{v+nV}^{(i)} \log y_{v+nV}^{(i)})$$

Where:

- $y_{v+nV}^{(i)}$ denotes the output probability prediction from the neural network for the i -th training example for the word v in the n -th output word. Denoting z as the logits output, we define the output probability y as a softmax on z over contiguous chunks of V units (see also Figure 1):

$$y_{v+nV}^{(i)} = \frac{e^{z_{v+nV}^{(i)}}}{\sum_l^V e^{z_{l+nV}^{(i)}}}$$

- $t_{v+nV}^{(i)} \in \{0, 1\}$ is 1 if for the i -th training example, the word v is the n -th word in context
- $m_n^{(i)} \in \{0, 1\}$ is a mask that is set to 1 if we are predicting the n -th word position for the i -th example (because we had masked that word in the input), and 0 otherwise

There are three classes defined in this part: `Params`, `Activations`, `Model`. You will make changes to `Model`, but it may help to read through `Params` and `Activations` first.

```
class Params(object):
    """A class representing the trainable parameters of the model. This class has five fields:
    word_embedding_weights, a matrix of size V x D, where V is the number of words in the vocabulary
    and D is the embedding dimension.
    embed_to_hid_weights, a matrix of size H x ND, where H is the number of hidden units. The first D
    columns represent connections from the embedding of the first context word, the next D columns
```

```

        for the second context word, and so on. There are N context words.
hid_bias, a vector of length H
hid_to_output_weights, a matrix of size NV x H
output_bias, a vector of length NV"""

def __init__(self, word_embedding_weights, embed_to_hid_weights, hid_to_output_weights,
            hid_bias, output_bias):
    self.word_embedding_weights = word_embedding_weights
    self.embed_to_hid_weights = embed_to_hid_weights
    self.hid_to_output_weights = hid_to_output_weights
    self.hid_bias = hid_bias
    self.output_bias = output_bias

def copy(self):
    return self.__class__(self.word_embedding_weights.copy(), self.embed_to_hid_weights.copy(),
                         self.hid_to_output_weights.copy(), self.hid_bias.copy(), self.output_bias.copy())

@classmethod
def zeros(cls, vocab_size, context_len, embedding_dim, num_hid):
    """A constructor which initializes all weights and biases to 0."""
    word_embedding_weights = np.zeros((vocab_size, embedding_dim))
    embed_to_hid_weights = np.zeros((num_hid, context_len * embedding_dim))
    hid_to_output_weights = np.zeros((vocab_size * context_len, num_hid))
    hid_bias = np.zeros(num_hid)
    output_bias = np.zeros(vocab_size * context_len)
    return cls(word_embedding_weights, embed_to_hid_weights, hid_to_output_weights,
               hid_bias, output_bias)

@classmethod
def random_init(cls, init_wt, vocab_size, context_len, embedding_dim, num_hid):
    """A constructor which initializes weights to small random values and biases to 0."""
    word_embedding_weights = np.random.normal(0., init_wt, size=(vocab_size, embedding_dim))
    embed_to_hid_weights = np.random.normal(0., init_wt, size=(num_hid, context_len * embedding_dim))
    hid_to_output_weights = np.random.normal(0., init_wt, size=(vocab_size * context_len, num_hid))
    hid_bias = np.zeros(num_hid)
    output_bias = np.zeros(vocab_size * context_len)
    return cls(word_embedding_weights, embed_to_hid_weights, hid_to_output_weights,
               hid_bias, output_bias)

##### The functions below are Python's somewhat oddball way of overloading operators, so that
##### we can do arithmetic on Params instances. You don't need to understand this to do the assignment.

def __mul__(self, a):
    return self.__class__(a * self.word_embedding_weights,
                         a * self.embed_to_hid_weights,
                         a * self.hid_to_output_weights,
                         a * self.hid_bias,
                         a * self.output_bias)

def __rmul__(self, a):
    return self * a

def __add__(self, other):
    return self.__class__(self.word_embedding_weights + other.word_embedding_weights,
                         self.embed_to_hid_weights + other.embed_to_hid_weights,
                         self.hid_to_output_weights + other.hid_to_output_weights,
                         self.hid_bias + other.hid_bias,
                         self.output_bias + other.output_bias)

def __sub__(self, other):
    return self + -1. * other

class Activations(object):
    """A class representing the activations of the units in the network. This class has three fields:

embedding_layer, a matrix of B x ND matrix (where B is the batch size, D is the embedding dimension,
and N is the number of input context words), representing the activations for the embedding
layer on all the cases in a batch. The first D columns represent the embeddings for the
first context word, and so on.
hidden_layer, a B x H matrix representing the hidden layer activations for a batch
output_layer, a B x V matrix representing the output layer activations for a batch"""

def __init__(self, embedding_layer, hidden_layer, output_layer):
    self.embedding_layer = embedding_layer
    self.hidden_layer = hidden_layer
    self.output_layer = output_layer

```

```

def get_batches(inputs, batch_size, shuffle=True):
    """Divide a dataset (usually the training set) into mini-batches of a given size. This is a
    'generator', i.e. something you can use in a for loop. You don't need to understand how it
    works to do the assignment."""

    if inputs.shape[0] % batch_size != 0:
        raise RuntimeError('The number of data points must be a multiple of the batch size.')
    num_batches = inputs.shape[0] // batch_size

    if shuffle:
        idxs = np.random.permutation(inputs.shape[0])
        inputs = inputs[idxs, :]

    for m in range(num_batches):
        yield inputs[m * batch_size:(m + 1) * batch_size, :]

```

In this part of the assignment, you implement a method which computes the gradient using backpropagation. To start you out, the *Model* class contains several important methods used in training:

- `compute_activations` computes the activations of all units on a given input batch
- `compute_loss_derivative` computes the gradient with respect to the output logits $\frac{\partial C}{\partial z}$
- `evaluate` computes the average cross-entropy loss for a given set of inputs and targets

You will need to complete the implementation of two additional methods to complete the training, and print the outputs of the gradients.

6.1 Implement gradient with respect to output layer inputs [0.5pt] [Type 2]

Implement a vectorized `compute_loss` function, which computes the total cross-entropy loss on a mini-batch according to Eq. 2. Look for the `## YOUR CODE HERE ##` comment for where to complete the code. The docstring provides a description of the inputs to the function.

▼ 6.2 Implement gradient with respect to parameters [1pt] [Type 2]

`back_propagate` is the function which computes the gradient of the loss with respect to model parameters using backpropagation. It uses the derivatives computed by `compute_loss_derivative`. Some parts are already filled in for you, but you need to compute the matrices of derivatives for `embed_to_hid_weights`, `hid_bias`, `hid_to_output_weights`, and `output_bias`. These matrices have the same sizes as the parameter matrices (see previous section). These matrices have the same sizes as the parameter matrices. Look for the `## YOUR CODE HERE ##` comment for where to complete the code.

In order to implement backpropagation efficiently, you need to express the computations in terms of matrix operations, rather than *for* loops. You should first work through the derivatives on pencil and paper. First, apply the chain rule to compute the derivatives with respect to individual units, weights, and biases. Next, take the formulas you've derived, and express them in matrix form. You should be able to express all of the required computations using only matrix multiplication, matrix transpose, and elementwise operations --- no *for* loops! If you want inspiration, read through the code for `Model.compute_activations` and try to understand how the matrix operations correspond to the computations performed by all the units in the network.

Hint: Your implementations should also be similar to `hid_to_output_weights_grad`, `hid_bias_grad` in the same function call

Hint: To prompt a GPT-like model, you may only include functions that are relevant to the implementation in your prompt.

```

class Model(object):
    """A class representing the language model itself. This class contains
    the model and visualizing the learned representations. It has two fields

    params, a Params instance which contains the model parameters
    vocab, a list containing all the words in the dictionary; vocab[0]
    is 0, and so on."""

    def __init__(self, params, vocab):
        self.params = params
        self.vocab = vocab

        self.vocab_size = len(vocab)
        self.embedding_dim = self.params.word_embedding_weights.shape[1]
        self.embedding_layer_dim = self.params.embed_to_hid_weights.shape[0]
        self.context_len = self.embedding_layer_dim // self.embedding_dim
        self.num_hid = self.params.embed_to_hid_weights.shape[1]

    def copy(self):
        return self.__class__(self.params.copy(), self.vocab[:])

```

```

@classmethod
def random_init(cls, init_wt, vocab, context_len, embedding_dim, num_h
    """Constructor which randomly initializes the weights to Gaussians
    and initializes the biases to all zeros."""
    params = Params.random_init(init_wt, len(vocab), context_len, embe
    return Model(params, vocab)

def indicator_matrix(self, targets, mask_zero_index=True):
    """Construct a matrix where the (v + n*V)th entry of row i is 1 if
    for example i is v, and all other entries are 0.

    Note: if the n-th target word index is 0, this corresponds to the
    and we set the entry to be 0.
    """

    batch_size, context_len = targets.shape
    expanded_targets = np.zeros((batch_size, context_len * len(self.vo
    offset = np.repeat((np.arange(context_len) * len(self.vocab))[np.n
    targets_offset = targets + offset

    for c in range(context_len):
        expanded_targets[np.arange(batch_size), targets_offset[:,c]] = 1
        if mask_zero_index:
            # Note: Set the targets with index 0, V, 2V to be zero since i
            expanded_targets[np.arange(batch_size), offset[:,c]] = 0.
    return expanded_targets

def compute_loss_derivative(self, output_activations, expanded_target_
    """Compute the gradient of cross-entropy loss wrt output logits z

    For example:

    [y_{0} .... y_{V-1}] [y_{V}, ..., y_{2*V-1}] [y_{2*V} ... y_{i,3

    Where for column v + n*V,
    y_{v + n*V} = e^{z_{v + n*V}} / \sum_{m=0}^{V-1} e^{z_{m + n*V}

    This function should return a dC / dz matrix of size [batch_size x
    where each row i in dC / dz has columns 0 to V-1 containing the gr
    context word from i-th training example, then columns vocab_size t
    output context word of the i-th training example, etc.

    C is the loss function summed acrossed all examples as well:

    C = -\sum_{i,j,n} mask_{i,n} (t_{i, j + n*V} log y_{i, j + n*V

    where mask_{i,n} = 1 if the i-th training example has n-th context
    otherwise mask_{i,n} = 0.

    Args:
        output_activations: A [batch_size x (context_len * vocab_size)]
            for the activations of the output layer, i.e. the y_j's.
        expanded_target_batch: A [batch_size x (context_len * vocab_size
            where expanded_target_batch[i,n*V:(n+1)*V] is the indicator
            the n-th context target word position, i.e. the (i, j + n*V)
            i'th example, the context word at position n is j, and 0 oth
        target_mask: A [batch_size x context_len x 1] tensor, where targ
            if for the i'th example the n-th context word is a target po

    Outputs:
        loss_derivative: A [batch_size x (context_len * vocab_size)] mat
            where loss_derivative[i,0:vocab_size] contains the gradient
            dC / dz_0 for the i-th training example gradient for 1st out
            context word, and loss_derivative[i,vocab_size:2*vocab_size]
            the 2nd output context word of the i-th training example, et
    """

    # Reshape output_activations and expanded_target_batch and use bro
    output_activations_reshape = output_activations.reshape(-1, self.c
    expanded_target_batch_reshape = expanded_target_batch.reshape(-1,
    gradient_masked_reshape = target_mask * (output_activations_resha
    gradient_masked = gradient_masked_reshape.reshape(-1, self.context
    return gradient_masked

def compute_loss(self, output_activations, expanded_target_batch, targ
    """Compute the total cross entropy loss over a mini-batch.

    Args:
        ...

```

```

        output_activations: [batch_size x (context_len * vocab_size)] ma
            for the activations of the output layer, i.e. the y_j's.
        expanded_target_batch: [batch_size (context_len * vocab_size)] m
            where expanded_target_batch[i,n*V:(n+1)*V] is the indicato
            the n-th context target word position, i.e. the (i, j + n*
            i'th example, the context word at position n is j, and 0 o
        target_mask: A [batch_size x context_len x 1] tensor, where targ
            if for the i'th example the n-th context word is a target

    Returns:
        loss: a scalar for the total cross entropy loss over the batch,
            defined in Part 3
    """
    ##### YOUR CODE HERE #####
    C = self.context_len
    V = self.vocab_size
    expanded_target_batch_update = expanded_target_batch.reshape(-1, C
    output_activations_update = output_activations.reshape(-1, C, V)
    temp = target_mask * (expanded_target_batch_update
        * np.log(output_activations_update))
    loss = -temp.reshape(-1, C * V).sum()

    #####
    return loss

def compute_activations(self, inputs):
    """Compute the activations on a batch given the inputs. Returns an
    You should try to read and understand this function, since this wi
    how to implement back_propagate."""

    batch_size = inputs.shape[0]
    if inputs.shape[1] != self.context_len:
        raise RuntimeError('Dimension of the input vectors should be {
            self.context_len, inputs.shape[1])}

    # Embedding layer
    # Look up the input word indices in the word_embedding_weights mat
    embedding_layer_state = self.params.word_embedding_weights[inputs.

    # Hidden layer
    inputs_to_hid = np.dot(embedding_layer_state, self.params.embed_
        self.params.hid_bias
    # Apply logistic activation function
    hidden_layer_state = 1. / (1. + np.exp(-inputs_to_hid))

    # Output layer
    inputs_to_softmax = np.dot(hidden_layer_state, self.params.hid_to_
        self.params.output_bias

    # Subtract maximum.
    # Remember that adding or subtracting the same constant from each
    # softmax unit does not affect the outputs. So subtract the maximu
    # make all inputs <= 0. This prevents overflows when computing the
    inputs_to_softmax -= inputs_to_softmax.max(1).reshape((-1, 1))

    # Take softmax along each V chunks in the output layer
    output_layer_state = np.exp(inputs_to_softmax)
    output_layer_state_shape = output_layer_state.shape
    output_layer_state = output_layer_state.reshape((-1, self.context_
    output_layer_state /= output_layer_state.sum(axis=-1, keepdims=True)
    output_layer_state = output_layer_state.reshape(output_layer_state

    return Activations(embedding_layer_state, hidden_layer_state, outp

def back_propagate(self, input_batch, activations, loss_derivative):
    """Compute the gradient of the loss function with respect to the t
    of the model.

    Part of this function is already completed, but you need to fill in
    computations for hid_to_output_weights_grad, output_bias_grad, emb
    and hid_bias_grad. See the documentation for the Params class for
    these matrices represent.

    Args:
        input_batch: A [batch_size x context_length] matrix containing t
            indices of the context words
        activations: an Activations object representing the output of

```

```

    Model.compute_activations
    loss_derivative: A [batch_size x (context_len * vocab_size)] ma
        where loss_derivative[i,0:vocab_size] contains the gradient
        dC / dz_0 for the i-th training example gradient for 1st out
        context word, and loss_derivative[i,vocab_size:2*vocab_size]
        the 2nd output context word of the i-th training example, et
        Obtained from calling compute_loss_derivative()

    Returns:
        Params object containing the gradient for word_embedding_weights
        embed_to_hid_weights_grad, hid_to_output_weights_grad,
        hid_bias_grad, output_bias_grad
    """

    # The matrix with values dC / dz_j, where dz_j is the input to the
    # i.e. h_j = 1 / (1 + e^{-z_j})
    hid_deriv = np.dot(loss_derivative, self.params.hid_to_output_weig
                        * activations.hidden_layer * (1. - activations.hidden_

    hid_to_output_weights_grad = np.dot(loss_derivative.T, activations

    ##### YOUR CODE HERE #####
    output_bias_grad = np.sum(loss_derivative, 0)
    embed_to_hid_weights_grad = np.dot(hid_deriv.T,
                                       activations.embedding_layer)

#####



    hid_bias_grad = hid_deriv.sum(0)

    # The matrix of derivatives for the embedding layer
    embed_deriv = np.dot(hid_deriv, self.params.embed_to_hid_weights)

    # Word Embedding Weights gradient
    word_embedding_weights_grad = np.dot(self.indicator_matrix(input_b
                                                               embed_deriv.reshape([-1,

    return Params(word_embedding_weights_grad, embed_to_hid_weights_gr
                  hid_bias_grad, output_bias_grad)

def sample_input_mask(self, batch_size):
    """Samples a binary mask for the inputs of size batch_size x conte
    For each row, at most one element will be 1.
    """
    mask_idx = np.random.randint(self.context_len, size=(batch_size,))
    mask = np.zeros((batch_size, self.context_len), dtype=np.int)# Con
    mask[np.arange(batch_size), mask_idx] = 1
    return mask

def evaluate(self, inputs, batch_size=100):
    """Compute the average cross-entropy over a dataset.

    inputs: matrix of shape D x N"""
    ndata = inputs.shape[0]

    total = 0.
    for input_batch in get_batches(inputs, batch_size):
        mask = self.sample_input_mask(batch_size)
        input_batch_masked = input_batch * (1 - mask)
        activations = self.compute_activations(input_batch_masked)
        expanded_target_batch = self.indicator_matrix(input_batch)
        target_mask = np.expand_dims(mask, axis=2)
        cross_entropy = self.compute_loss(activations.output_layer, ex
        total += cross_entropy

    return total / float(ndata)

def display_nearest_words(self, word, k=10):
    """List the k words nearest to a given word, along with their dist

    if word not in self.vocab:
        print('Word "{}" not in vocabulary.'.format(word))

```

```

        return

    # Compute distance to every other word.
    idx = self.vocab.index(word)
    word_rep = self.params.word_embedding_weights[idx, :]
    diff = self.params.word_embedding_weights - word_rep.reshape((1, -
    distance = np.sqrt(np.sum(diff ** 2, axis=1))

    # Sort by distance.
    order = np.argsort(distance)
    order = order[1:l + k] # The nearest word is the query word itself
    for i in order:
        print(' {}: {}'.format(self.vocab[i], distance[i]))

def word_distance(self, word1, word2):
    """Compute the distance between the vector representations of two

    if word1 not in self.vocab:
        raise RuntimeError('Word "{}" not in vocabulary.'.format(word1)
    if word2 not in self.vocab:
        raise RuntimeError('Word "{}" not in vocabulary.'.format(word2

    idx1, idx2 = self.vocab.index(word1), self.vocab.index(word2)
    word_repl = self.params.word_embedding_weights[idx1, :]
    word_rep2 = self.params.word_embedding_weights[idx2, :]
    diff = word_repl - word_rep2
    return np.sqrt(np.sum(diff ** 2))

```

▼ 6.3 Print the gradients [0.5pt] [Type 4]

To make your life easier, we have provided the routine `check_gradients`, which checks your gradients using finite differences. You should make sure this check passes before continuing with the assignment. Once `check_gradients()` passes, call `print_gradients()` and include its output in your write-up.

```

def relative_error(a, b):
    return np.abs(a - b) / (np.abs(a) + np.abs(b))

def check_output_derivatives(model, input_batch, target_batch, mask):
    def softmax(z):
        z = z.copy()
        z -= z.max(-1, keepdims=True)
        y = np.exp(z)
        y /= y.sum(-1, keepdims=True)
        return y

    batch_size = input_batch.shape[0]
    z = np.random.normal(size=(batch_size, model.context_len, model.vocab_size))
    y = softmax(z).reshape((batch_size, model.context_len * model.vocab_size))
    z = z.reshape((batch_size, model.context_len * model.vocab_size))

    expanded_target_batch = model.indicator_matrix(target_batch)
    target_mask = np.expand_dims(mask, axis=2)
    loss_derivative = model.compute_loss_derivative(y, expanded_target_batch, target_mask)

    if loss_derivative is None:
        print('Loss derivative not implemented yet.')
        return False

    if loss_derivative.shape != (batch_size, model.vocab_size * model.context_len):
        print('Loss derivative should be size {} but is actually {}'.format(
            (batch_size, model.vocab_size), loss_derivative.shape))
        return False

    def obj(z):
        z = z.reshape((-1, model.context_len, model.vocab_size))
        y = softmax(z).reshape((batch_size, model.context_len * model.vocab_size))
        return model.compute_loss(y, expanded_target_batch, target_mask)

    for count in range(1000):
        i, j = np.random.randint(0, loss_derivative.shape[0]), np.random.randint(0, loss_derivative.shape[1])

        z_plus = z.copy()
        z_plus[i, j] += EPS

```

```

obj_plus = obj(z_plus)

z_minus = z.copy()
z_minus[i, j] -= EPS
obj_minus = obj(z_minus)

empirical = (obj_plus - obj_minus) / (2. * EPS)
rel = relative_error(empirical, loss_derivative[i, j])
if rel > 1e-4:
    print('The loss derivative has a relative error of {}, which is too large.'.format(rel))
    return False

print('The loss derivative looks OK.')
return True

def check_param_gradient(model, param_name, input_batch, target_batch, mask):
    activations = model.compute_activations(input_batch)
    expanded_target_batch = model.indicator_matrix(target_batch)
    target_mask = np.expand_dims(mask, axis=2)
    loss_derivative = model.compute_loss_derivative(activations.output_layer, expanded_target_batch, target_mask)
    param_gradient = model.back_propagate(input_batch, activations, loss_derivative)

def obj(model):
    activations = model.compute_activations(input_batch)
    return model.compute_loss(activations.output_layer, expanded_target_batch, target_mask)

dims = getattr(model.params, param_name).shape
is_matrix = (len(dims) == 2)

if getattr(param_gradient, param_name).shape != dims:
    print('The gradient for {} should be size {} but is actually {}'.format(
        param_name, dims, getattr(param_gradient, param_name).shape))
    return

for count in range(1000):
    if is_matrix:
        slc = np.random.randint(0, dims[0]), np.random.randint(0, dims[1])
    else:
        slc = np.random.randint(dims[0])

    model_plus = model.copy()
    setattr(model_plus.params, param_name)[slc] += EPS
    obj_plus = obj(model_plus)

    model_minus = model.copy()
    setattr(model_minus.params, param_name)[slc] -= EPS
    obj_minus = obj(model_minus)

    empirical = (obj_plus - obj_minus) / (2. * EPS)
    exact = getattr(param_gradient, param_name)[slc]
    rel = relative_error(empirical, exact)
    if rel > 5e-4:
        print('The loss derivative has a relative error of {}, which is too large for param {}.'.format(rel, param_name))
        return False

print('The gradient for {} looks OK.'.format(param_name))

def load_partially_trained_model():
    obj = pickle.load(open(PARTIALLY_TRAINED_MODEL, 'rb'))
    params = Params(obj['word_embedding_weights'], obj['embed_to_hid_weights'],
                    obj['hid_to_output_weights'], obj['hid_bias'],
                    obj['output_bias'])
    vocab = obj['vocab']
    return Model(params, vocab)

def check_gradients():
    """Check the computed gradients using finite differences."""
    np.random.seed(0)

    np.seterr(all='ignore') # suppress a warning which is harmless

    model = load_partially_trained_model()
    data_obj = pickle.load(open(data_location, 'rb'))
    train_inputs = data_obj['train_inputs']

```

```

input_batch = train_inputs[:100, :]
mask = model.sample_input_mask(input_batch.shape[0])
input_batch_masked = input_batch * (1 - mask)

if not check_output_derivatives(model, input_batch_masked, input_batch, mask):
    return

for param_name in ['word_embedding_weights', 'embed_to_hid_weights', 'hid_to_output_weights',
                   'hid_bias', 'output_bias']:
    check_param_gradient(model, param_name, input_batch_masked, input_batch, mask)

def print_gradients():
    """Print out certain derivatives for grading."""
    np.random.seed(0)

    model = load_partially_trained_model()
    data_obj = pickle.load(open(data_location, 'rb'))
    train_inputs = data_obj['train_inputs']
    input_batch = train_inputs[:100, :]

    mask = model.sample_input_mask(input_batch.shape[0])
    input_batch_masked = input_batch * (1 - mask)
    activations = model.compute_activations(input_batch_masked)
    expanded_target_batch = model.indicator_matrix(input_batch)
    target_mask = np.expand_dims(mask, axis=2)
    loss_derivative = model.compute_loss_derivative(activations.output_layer, expanded_target_batch, target_mask)
    param_gradient = model.back_propagate(input_batch, activations, loss_derivative)

    print('loss_derivative[46, 785]', loss_derivative[46, 785])
    print('loss_derivative[46, 766]', loss_derivative[46, 766])
    print('loss_derivative[5, 42]', loss_derivative[5, 42])
    print('loss_derivative[5, 31]', loss_derivative[5, 31])
    print()
    print('param_gradient.word_embedding_weights[27, 2]', param_gradient.word_embedding_weights[27, 2])
    print('param_gradient.word_embedding_weights[43, 3]', param_gradient.word_embedding_weights[43, 3])
    print('param_gradient.word_embedding_weights[22, 4]', param_gradient.word_embedding_weights[22, 4])
    print('param_gradient.word_embedding_weights[2, 5]', param_gradient.word_embedding_weights[2, 5])
    print()
    print('param_gradient.embed_to_hid_weights[10, 2]', param_gradient.embed_to_hid_weights[10, 2])
    print('param_gradient.embed_to_hid_weights[15, 3]', param_gradient.embed_to_hid_weights[15, 3])
    print('param_gradient.embed_to_hid_weights[30, 9]', param_gradient.embed_to_hid_weights[30, 9])
    print('param_gradient.embed_to_hid_weights[35, 21]', param_gradient.embed_to_hid_weights[35, 21])
    print()
    print('param_gradient.hid_bias[10]', param_gradient.hid_bias[10])
    print('param_gradient.hid_bias[20]', param_gradient.hid_bias[20])
    print()
    print('param_gradient.output_bias[0]', param_gradient.output_bias[0])
    print('param_gradient.output_bias[1]', param_gradient.output_bias[1])
    print('param_gradient.output_bias[2]', param_gradient.output_bias[2])
    print('param_gradient.output_bias[3]', param_gradient.output_bias[3])

# Run this to check if your implement gradients matches the finite difference within tolerance
# Note: this may take a few minutes to go through all the checks
check_gradients()

<ipython-input-18-13f30d93e8e6>:226: DeprecationWarning: `np.int` is a deprecated alias for the builtin `int`. To silence th:
  Deprecated in NumPy 1.20; for more details and guidance: https://numpy.org/devdocs/release/1.20.0-notes.html#deprecations
  mask = np.zeros((batch_size, self.context_len), dtype=np.int) # Convert to one hot B x N, B batch size, N context len
The loss derivative looks OK.
The gradient for word_embedding_weights looks OK.
The gradient for embed_to_hid_weights looks OK.
The gradient for hid_to_output_weights looks OK.
The gradient for hid_bias looks OK.
The gradient for output_bias looks OK.

# Run this to print out the gradients
print_gradients()

loss_derivative[46, 785] 0.7137561447745507
loss_derivative[46, 766] -0.9661570033238931
loss_derivative[5, 42] -0.0
loss_derivative[5, 31] 0.0

param_gradient.word_embedding_weights[27, 2] 0.0
param_gradient.word_embedding_weights[43, 3] 0.011596892511489458
param_gradient.word_embedding_weights[22, 4] -0.0222670623817297

```

```

param_gradient.word_embedding_weights[2, 5] 0.0

param_gradient.embed_to_hid_weights[10, 2] 0.3793257091930164
param_gradient.embed_to_hid_weights[15, 3] 0.01604516132110917
param_gradient.embed_to_hid_weights[30, 9] -0.4312854367997419
param_gradient.embed_to_hid_weights[35, 21] 0.06679896665436337

param_gradient.hid_bias[10] 0.023428803123345148
param_gradient.hid_bias[20] -0.024370452378874197

param_gradient.output_bias[0] 0.000970106146902794
param_gradient.output_bias[1] 0.16868946274763222
param_gradient.output_bias[2] 0.0051664774143909235
param_gradient.output_bias[3] 0.15096226471814364
<ipython-input-18-13f30d93e8e6>:226: DeprecationWarning: `np.int` is a deprecated alias for the builtin `int`. To silence th:
Depreciated in NumPy 1.20; for more details and guidance: https://numpy.org/devdocs/release/1.20.0-notes.html#deprecations
mask = np.zeros((batch_size, self.context_len), dtype=np.int)# Convert to one hot B x N, B batch size, N context len

```

▼ 6.4 Run model training [Opt]

Once you've implemented the gradient computation, you'll need to train the model. The function *train* implements the main training procedure. It takes two arguments:

- `embedding_dim`: The number of dimensions in the distributed representation.
- `num_hid`: The number of hidden units

As the model trains, the script prints out some numbers that tell you how well the training is going. It shows:

- The cross entropy on the last 100 mini-batches of the training set. This is shown after every 100 mini-batches.
- The cross entropy on the entire validation set every 1000 mini-batches of training.

At the end of training, this function shows the cross entropies on the training, validation and test sets. It will return a *Model* instance.

```

_train_inputs = None
_train_targets = None
_vocab = None

DEFAULT_TRAINING_CONFIG = {'batch_size': 100, # the size of a mini-batch
                           'learning_rate': 0.1, # the learning rate
                           'momentum': 0.9, # the decay parameter for the momentum vector
                           'epochs': 50, # the maximum number of epochs to run
                           'init_wt': 0.01, # the standard deviation of the initial random weights
                           'context_len': 4, # the number of context words used
                           'show_training_CE_after': 100, # measure training error after this many mini-batches
                           'show_validation_CE_after': 1000, # measure validation error after this many mini-batches
                           }

def find_occurrences(word1, word2, word3):
    """Lists all the words that followed a given tri-gram in the training set and the number of
    times each one followed it."""

    # cache the data so we don't keep reloading
    global _train_inputs, _train_targets, _vocab
    if _train_inputs is None:
        data_obj = pickle.load(open(data_location, 'rb'))
        _vocab = data_obj['vocab']
        _train_inputs, _train_targets = data_obj['train_inputs'], data_obj['train_targets']

    if word1 not in _vocab:
        raise RuntimeError('Word "{}" not in vocabulary.'.format(word1))
    if word2 not in _vocab:
        raise RuntimeError('Word "{}" not in vocabulary.'.format(word2))
    if word3 not in _vocab:
        raise RuntimeError('Word "{}" not in vocabulary.'.format(word3))

    idx1, idx2, idx3 = _vocab.index(word1), _vocab.index(word2), _vocab.index(word3)
    idxs = np.array([idx1, idx2, idx3])

    matches = np.all(_train_inputs == idxs.reshape((1, -1)), 1)

    if np.any(matches):
        counts = collections.defaultdict(int)
        for m in np.where(matches)[0]:
            counts[_vocab[_train_targets[m]]] += 1

```

```

word_counts = sorted(list(counts.items()), key=lambda t: t[1], reverse=True)
print('The tri-gram "{} {} {}" was followed by the following words in the training set:'.format(
    word1, word2, word3))
for word, count in word_counts:
    if count > 1:
        print('    {} ({} times)'.format(word, count))
    else:
        print('    {} (1 time)'.format(word))
else:
    print('The tri-gram "{} {} {}" did not occur in the training set.'.format(word1, word2, word3))

def train(embedding_dim, num_hid, config=DEFAULT_TRAINING_CONFIG):
    """This is the main training routine for the language model. It takes two parameters:

        embedding_dim, the dimension of the embedding space
        num_hid, the number of hidden units."""

    # For reproducibility
    np.random.seed(123)

    # Load the data
    data_obj = pickle.load(open(data_location, 'rb'))
    vocab = data_obj['vocab']
    train_inputs = data_obj['train_inputs']
    valid_inputs = data_obj['valid_inputs']
    test_inputs = data_obj['test_inputs']

    # Randomly initialize the trainable parameters
    model = Model.random_init(config['init_wt'], vocab, config['context_len'], embedding_dim, num_hid)

    # Variables used for early stopping
    best_valid_CE = np.infty
    end_training = False

    # Initialize the momentum vector to all zeros
    delta = Params.zeros(len(vocab), config['context_len'], embedding_dim, num_hid)

    this_chunk_CE = 0.
    batch_count = 0
    for epoch in range(1, config['epochs'] + 1):
        if end_training:
            break

        print()
        print('Epoch', epoch)

        for m, (input_batch) in enumerate(get_batches(train_inputs, config['batch_size'])):
            batch_count += 1

            # For each example (row in input_batch), select one word to mask out
            mask = model.sample_input_mask(config['batch_size'])
            input_batch_masked = input_batch * (1 - mask) # We only zero out one word per row

            # Forward propagate
            activations = model.compute_activations(input_batch_masked)

            # Compute loss derivative
            expanded_target_batch = model.indicator_matrix(input_batch)
            loss_derivative = model.compute_loss_derivative(activations.output_layer, expanded_target_batch, mask[:, :, np.newaxis])
            loss_derivative /= config['batch_size']

            # Measure loss function
            cross_entropy = model.compute_loss(activations.output_layer, expanded_target_batch, np.expand_dims(mask, axis=2)) / config['batch_size']
            this_chunk_CE += cross_entropy
            if batch_count % config['show_training_CE_after'] == 0:
                print('Batch {} Train CE {:.3f}'.format(
                    batch_count, this_chunk_CE / config['show_training_CE_after']))
            this_chunk_CE = 0.

            # Backpropagate
            loss_gradient = model.back_propagate(input_batch, activations, loss_derivative)

            # Update the momentum vector and model parameters
            delta = config['momentum'] * delta + loss_gradient
            model.params -= config['learning_rate'] * delta

            # Validate

```

```

if batch_count % config['show_validation_CE_after'] == 0:
    print('Running validation...')
    cross_entropy = model.evaluate(valid_inputs)
    print('Validation cross-entropy: {:.3f}'.format(cross_entropy))

    if cross_entropy > best_valid_CE:
        print('Validation error increasing! Training stopped.')
        end_training = True
        break

    best_valid_CE = cross_entropy

print()
train_CE = model.evaluate(train_inputs)
print('Final training cross-entropy: {:.3f}'.format(train_CE))
valid_CE = model.evaluate(valid_inputs)
print('Final validation cross-entropy: {:.3f}'.format(valid_CE))
test_CE = model.evaluate(test_inputs)
print('Final test cross-entropy: {:.3f}'.format(test_CE))

return model

```

Run the training.

```

embedding_dim = 16
num_hid = 128
trained_model = train(embedding_dim, num_hid)

```

```

Epoch 1
<ipython-input-18-13f30d93e8e6>:226: DeprecationWarning: `np.int` is a deprecated alias for the builtin `int`. To silence th:
Deprecation in NumPy 1.20; for more details and guidance: https://numpy.org/devdocs/release/1.20.0-notes.html#deprecations
    mask = np.zeros((batch_size, self.context_len), dtype=np.int)# Convert to one hot B x N, B batch size, N context len
Batch 100 Train CE 4.793
Batch 200 Train CE 4.645
Batch 300 Train CE 4.649
Batch 400 Train CE 4.629
Batch 500 Train CE 4.633
Batch 600 Train CE 4.648
Batch 700 Train CE 4.617
Batch 800 Train CE 4.607
Batch 900 Train CE 4.606
Batch 1000 Train CE 4.615
Running validation...
Validation cross-entropy: 4.615
Batch 1100 Train CE 4.615
Batch 1200 Train CE 4.624
Batch 1300 Train CE 4.608
Batch 1400 Train CE 4.595
Batch 1500 Train CE 4.611
Batch 1600 Train CE 4.598
Batch 1700 Train CE 4.577
Batch 1800 Train CE 4.578
Batch 1900 Train CE 4.568
Batch 2000 Train CE 4.589
Running validation...
Validation cross-entropy: 4.589
Batch 2100 Train CE 4.573
Batch 2200 Train CE 4.611
Batch 2300 Train CE 4.562
Batch 2400 Train CE 4.587
Batch 2500 Train CE 4.589
Batch 2600 Train CE 4.587
Batch 2700 Train CE 4.561
Batch 2800 Train CE 4.544
Batch 2900 Train CE 4.521
Batch 3000 Train CE 4.524
Running validation...
Validation cross-entropy: 4.496
Batch 3100 Train CE 4.504
Batch 3200 Train CE 4.449
Batch 3300 Train CE 4.384
Batch 3400 Train CE 4.352
Batch 3500 Train CE 4.324
Batch 3600 Train CE 4.261
Batch 3700 Train CE 4.267

Epoch 2
Batch 3800 Train CE 4.208
Batch 3900 Train CE 4.168
Batch 4000 Train CE 4.117

```

```
Running validation...
Validation cross-entropy: 4.112
Batch 4100 Train CE 4.105
Batch 4200 Train CE 4.049
Batch 4300 Train CE 4.008
```

To convince us that you have correctly implemented the gradient computations, please include the following with your assignment submission:

- You will submit `a1-code.ipynb` through MarkUs. You do not need to modify any of the code except the parts we asked you to implement.
- In your writeup, include the output of the function `print_gradients`. This prints out part of the gradients for a partially trained network which we have provided, and we will check them against the correct outputs. **Important:** make sure to give the output of `print_gradients`, **not** `check_gradients`.

▼ Part 7: Bias in Word Embeddings (2pts)

Unfortunately, stereotypes and prejudices are often reflected in the outputs of natural language processing algorithms. For example, Google Translate is more likely to translate a non-English sentence to "*He* is a doctor" than "*She* is a doctor when the sentence is ambiguous. In this section, you will explore how bias enters natural language processing algorithms by implementing and analyzing a popular method for measuring bias in word embeddings.

Note: In AI and machine learning, **bias** generally refers to prior information, a necessary prerequisite for intelligent action. However, bias can be problematic when it is derived from aspects of human culture known to lead to harmful behaviour, such as stereotypes and prejudices.

▼ 7.1 WEAT method for detecting bias [1pt] [Type 2]

Word embedding models such as GloVe attempt to learn a vector space where semantically similar words are clustered close together. However, they have been shown to learn problematic associations, e.g. by embedding "man" more closely to "doctor" than "woman" (and vice versa for "nurse"). To detect such biases in word embeddings, "[Semantics derived automatically from language corpora contain human-like biases](#)" introduced the Word Embedding Association Test (WEAT). The WEAT test measures whether two *target* word sets (e.g., {programmer, engineer, scientist, ...} and {nurse, teacher, librarian, ...}) have the same relative association to two *attribute* word sets (e.g., man, male, ... and woman, female ...).

There is an excellent blog on bias in word embeddings and the WEAT test [here](#).

In the following section, you will run a WEAT test for a given set of target and attribute words. Specifically, you must implement the function `weat_association_score` and then run the remaining cells to compute the p-value and effect size. Before you begin, make sure you understand the formal definition of the WEAT test given in section 4.1 of the handout.

Run the following cell to download pretrained GloVe embeddings.

```
import gensim.downloader as api

glove = api.load("glove-wiki-gigaword-50")
num_words, num_dims = glove.vectors.shape
print(f"Downloaded {num_words} word embeddings of dimension {num_dims}.")
```

Downloaded 400000 word embeddings of dimension 50.

Before proceeding, you should familiarize yourself with the `similarity` method, which computes the cosine similarity between two words. You will need this method to implement `weat_association_score`. Some examples are given below.

Can you spot the gender bias between occupations in the examples below?

```
print(glove.similarity("man", "scientist"))
print(glove.similarity("man", "nurse"))
print(glove.similarity("woman", "scientist"))
print(glove.similarity("woman", "nurse"))

0.49226817
0.5718704
0.43883628
0.715502
```

Below, we define our target words (`occupations`) and attribute words (`A` and `B`). Our target words consist of `occupations`, and our attribute words are `gendered`. We will use the WEAT test to determine if the word embeddings contain gender biases for certain occupations.

```
# Target words (occupations)
occupations = ["programmer", "engineer", "scientist", "nurse", "teacher", "librarian"]
# Two sets of gendered attribute words, A and B
A = ["man", "male", "he", "boyish"]
B = ["woman", "female", "she", "girlish"]
```

- **TODO:** Implement the following function, `weat_association_score` which computes the association of a word w with the attribute:

$$s(w, A, B) = \text{mean}_{a \in A} \cos(w, a) - \text{mean}_{b \in B} \cos(w, b)$$

Hint: To prompt a GPT-like model, think about how to provide the information that the glove has an method `similarity` which can be called.

```
def weat_association_score(w, A, B, glove):
    """Given a target word w, the set of attribute words A and B,
    and the GloVe embeddings, returns the association score s(w, A, B).
    """
    ##### YOUR CODE HERE #####
    cos_A = 0
    cos_B = 0
    for a in A:
        cos_A += glove.similarity(a, w)
    for b in B:
        cos_B += glove.similarity(b, w)
    return cos_A/len(A)-cos_B/len(B)

#####
```

Use the following code to check your implementation:

```
np.isclose(weat_association_score("programmer", A, B, glove), 0.019615129)
True
```

Now, compute the WEAT association score for each element of `occupations` and the attribute sets A and B. Include the printed out association scores in your pdf.

```
# TODO: Print out the weat association score for each occupation
##### YOUR CODE HERE #####
for title in occupations:
    print(title, ":", weat_association_score(title, A, B, glove))

#####
programmer : 0.01961511862464249
engineer : 0.053647358901798725
scientist : 0.06795816496014595
nurse : -0.09486919268965721
teacher : -0.01893029361963272
librarian : -0.024141337256878614
```

▼ 7.2 Reasons for bias in word embeddings [0pt] [Type 1]

Based on these WEAT association scores, do the pretrained word embeddings associate certain occupations with one gender more than another? What might cause word embedding models to learn certain stereotypes and prejudices? How might this be a problem in downstream applications?

7.2 Answer: **TODO: Write Part 7.2 answer here**

▼ 7.3 Analyzing WEAT [1pt]

While WEAT makes intuitive sense by asserting that closeness in the embedding space indicates greater similarity, more recent work ([Ethayarajh et al. \[2019\]](#)) has further analyzed the mathematical assertions and found some flaws with this method. Analyzing edge cases is a good way to find logical inconsistencies with any algorithm, and WEAT in particular can behave strangely when A and B contain just one word each.

▼ 7.3.1 1-word subsets [0.5 pts] [Type 4]

Find 1-word subsets of the original A and B that reverse the sign of the association score for at least some of the occupations

```
## Original sets provided here for convenience - try commenting out all but one word from each set
# Two sets of gendered attribute words, C and D

#Choose "boyish" from C and "she" from D
C = [#"man",
      #'male',
      #'he',
      "boyish"
    ]
D = [#"woman",
      "female",
      #'she',
      #'girlish"
    ]

# TODO: Print out the weat association score for each word in occupations, with regards to C and D
#####
# YOUR CODE HERE #####
#####

score = [weat_association_score(title, A, B, glove) for title in occupations]
score_new = [weat_association_score(title, C, D, glove) for title in occupations]
print("Gendered attribute words chosen:", C, D)
for i in range(len(occupations)):
    print(occupations[i],':',score[i],'==>',score_new[i])

#####
Gendered attribute words chosen: ['boyish'] ['female']
programmer : 0.01961511862464249 ==> -0.19506452232599258
engineer : 0.053647358901798725 ==> -0.3192642293870449
scientist : 0.06795816496014595 ==> -0.24408689141273499
nurse : -0.09486919268965721 ==> -0.41440044343471527
teacher : -0.01893029361963272 ==> -0.3508371561765671
librarian : -0.024141337256878614 ==> -0.14386512339115143
```

▼ 7.3.2 How word frequency affects embedding similarity [0.5 pts] [Type 3] [EC]

Consider the fact that the squared norm of a word embedding is linear in the log probability of the word in the training corpus. In other words, the more common a word is in the training corpus, the larger the norm of its word embedding. (See handout for more thorough description)

Briefly explain how this fact might contribute to the results from the previous section when using different attribute words. Provide your answers in no more than three sentences.

Hint 2: The paper cited above is a great resource if you are stuck.

7.3.2 Answer: **TODO: Write Part 7.3.2 answer here**

For example, if the word from A is more common in the training corpus, the larger the norm of its word embedding, the score is more likely to be positive. We can choose the words in above step according to the word frequency in the training corpus to decide whether the signs of the association scores for some particular occupations are reversed.

▼ 7.3.3 Relative association between two sets of target words [0 pts] [Type 3]

In the original WEAT paper, the authors do not examine the association of individual words with attributes, but rather compare the relative association of two sets of target words. For example, are insect words more associated with positive attributes or negative attributes than flower words.

Formally, let X and Y be two sets of target words of equal size. The WEAT test statistic is given by:

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

Will the same technique from the previous section work to manipulate this test statistic as well? Provide your answer in no more than 3 sentences.

7.3.3 Answer: TODO: Write 7.3.3 answer here

▼ What you have to submit

Refer to the handout for the checklist



3 Programming Assignment –LLM program trace, ChatGPT

Interact with ChatGPT:

For Q4.2 Expression for the Vectorized Loss function:



$\$ \$ L(\{\mathbf{w}_i\}_{i=1}^V) = \sum_{i,j=1}^V (\mathbf{w}_i^\top \tilde{\mathbf{w}}_j + b_i - \log X_{ij})^2$. Rewrite the loss function L (Eq. 1) in a vectorized format in terms of \mathbf{W} , $\tilde{\mathbf{w}}$, \mathbf{b} , X . You are allowed to use elementwise operations such as addition and subtraction as well as matrix operations such as the Frobenius norm and/or trace operator in your answer.

Hint: Use the all-ones column vector $\mathbf{1} = [1 \dots 1]^T$ in \mathbf{R}^V . You can assume the bias vectors are column vectors, i.e. implicitly a matrix with V rows and 1 column: $\mathbf{b}, \tilde{\mathbf{b}} \in \mathbf{R}^V$ \times V times 1



We can rewrite the loss function L in the vectorized format as follows:



$$\$ \$ L(\mathbf{W}, \tilde{\mathbf{w}}, \mathbf{b}, \tilde{\mathbf{b}}, X) = \left\| \mathbf{W}^\top \tilde{\mathbf{w}} + \mathbf{b} \mathbf{1}^\top + \tilde{\mathbf{b}} \mathbf{1}^\top - \log X \right\|_F^2$$

where $\mathbf{1} = [1 \dots 1]^T$ in \mathbf{R}^V is the all-ones column vector and $\|\cdot\|_F$ denotes the Frobenius norm. The term $\mathbf{b} \mathbf{1}^\top$ is a matrix with all its elements equal to b_i , and $\tilde{\mathbf{b}} \mathbf{1}^\top$ is a matrix with all its elements equal to \tilde{b}_j . The Frobenius norm squared computes the mean squared error between the predicted log-counts and the actual log-counts, as given in the original equation.

Original output for Q4.2:

We can rewrite the loss function L in the vectorized format as follows:

$$L(\mathbf{W}, \tilde{\mathbf{w}}, \mathbf{b}, \tilde{\mathbf{b}}, X) = \left\| (\mathbf{W}^\top \tilde{\mathbf{w}} + \mathbf{b} \mathbf{1}^\top + \tilde{\mathbf{b}} \mathbf{1}^\top - \log X)^2 \right\|_F^2$$

where $\mathbf{1} = [1 \dots 1]^T \in \mathbf{R}^V$ is the all-ones column vector and $\|\cdot\|_F$ denotes the Frobenius norm. The term $\mathbf{b} \mathbf{1}^\top$ is a matrix with all its elements equal to b_i , and $\tilde{\mathbf{b}} \mathbf{1}^\top$ is a matrix with all its elements equal to \tilde{b}_j . The Frobenius norm squared computes the mean squared error between the predicted log-counts

and the actual log-counts, as given in the original equation.

For Q4.3 Expression for gradient:

A small, square profile picture of a person with long dark hair, wearing a white shirt and a dark jacket.

Given \$L (\mathbf{W}, \mathbf{b}) = \frac{1}{N} \sum_{i=1}^N \log(\|\mathbf{W}\mathbf{x}_i + \mathbf{b}\|_2) + \lambda \|\mathbf{W}\|_F^2\$. Write the vectorized expression for \$\frac{\partial L}{\partial \mathbf{W}}\$, the gradient of the loss function \$L\$ with respect to the embedding matrix \$\mathbf{W}\$. The gradient should be a function of \$\mathbf{W}\$, \$\mathbf{x}\$, and \$\mathbf{b}\$.

*Hint: Make sure that the shape of the gradient is equivalent to the shape of the matrix.
You can use the all-ones vector as in the previous question.*

/ 2 >

The gradient of the loss function L with respect to the matrix \mathbf{W} can be obtained as follows:

$$\frac{\partial}{\partial \mathbf{W}} \log X = \mathbf{F}^T$$

Using the chain rule and the fact that the Frobenius norm is a differentiable function, we can write:

$$\frac{\partial}{\partial W} \tilde{L}(W) = 2(\mathbf{W})^T + (\mathbf{W})^T + \tilde{L}'(\mathbf{W}) - \log X$$

Since the only term in the partial derivative involving \mathbf{W} is $\tilde{\mathbf{W}}^T$, we can write:

$$\$ \$ \frac{\partial L}{\partial \mathbf{W}} = 2(\mathbf{W}^T \tilde{\mathbf{W}} + \mathbf{b}^T \mathbf{b} - \log X) \tilde{\mathbf{W}} \$ \$$$

Thus, the gradient of the loss function L with respect to the embedding matrix \mathbf{W} is given by:

Original output for Q4.3:

The gradient of the loss function L with respect to the matrix \mathbf{W} can be obtained as follows:

$$\frac{\partial L}{\partial \mathbf{W}} = \frac{\partial}{\partial \mathbf{W}} ||\mathbf{W}\tilde{\mathbf{W}}^{+\mathbf{b}1+1\tilde{\mathbf{b}}} - \log X||_F^2$$

Using the chain rule and the fact that the Frobenius norm is a differentiable function, we can write:

$$\frac{\partial L}{\partial \mathbf{W}} = 2(\mathbf{W}\tilde{\mathbf{W}}^{+\mathbf{b}1+1\tilde{\mathbf{b}}^{-\log X})\frac{\partial}{\partial \mathbf{W}}(\mathbf{W}\tilde{\mathbf{W}})}$$

Since the only term in the partial derivative involving \mathbf{W} is $\mathbf{W}\tilde{\mathbf{W}}$, we can write:

$$\frac{\partial L}{\partial \mathbf{W}} = 2(\mathbf{W}\tilde{\mathbf{W}}^{+\mathbf{b}1+1\tilde{\mathbf{b}}^{-\log X})\tilde{\mathbf{W}}$$

Thus, the gradient of the loss function L with respect to the embedding matrix \mathbf{W} is given by:

$$\frac{\partial L}{\partial \mathbf{W}} = 2(\mathbf{W}\tilde{\mathbf{W}}^{+\mathbf{b}1+1\tilde{\mathbf{b}}^{-\log X})\tilde{\mathbf{W}}$$

For Q4.4 Implement Vectorized Loss Function:



Given the vectorized loss formula is $L(\mathbf{W}, \tilde{\mathbf{W}}, \mathbf{b}, \tilde{\mathbf{b}}, \mathbf{X}) = \|\mathbf{W}\tilde{\mathbf{W}} + \mathbf{b}\tilde{\mathbf{b}}^T - \mathbf{X}\|_F^2$, implement Vectorized Loss Function using python code.

Note that you need to implement both the loss for an *asymmetric* model and the loss for a *symmetric* model which uses the same embedding matrix \mathbf{W} and bias vector \mathbf{b} for the first and second word in the co-occurrence, i.e. $\tilde{\mathbf{b}} = \mathbf{b}$ and $\tilde{\mathbf{b}} = -\mathbf{b}$ in the original loss.

*Hint: You may take advantage of NumPy's broadcasting feature for the bias vectors



Here's the python code for both the asymmetric and symmetric model:



```
python
import numpy as np

def asymmetric_loss(W, tW, b, tb, X):
    L = np.linalg.norm(W @ tW.T + b[:, np.newaxis] + tb[np.newaxis, :] - X) ** 2
    return L

def symmetric_loss(W, b, X):
    L = np.linalg.norm(W @ W.T + b[:, np.newaxis] + b[np.newaxis, :] - X) ** 2
    return L
```

Note that `np.newaxis` is used to add an additional dimension to the bias vectors `b` and `tb` to allow broadcasting with the matrix `W @ tW.T`.

For Q4.5 Implement the gradient update of GloVe: program trace:

```
When W_tilde  
and b_tilde are not given, then the model is symmetric (i.e. W_tilde = W,  
b_tilde = b). Here is the loss_Glove function which Compute the GloVe loss given  
the parameters of the model:\ def loss_GloVe(W, W_tilde, b, b_tilde,  
log_co_occurrence):  
    """ Compute the GloVe loss given the parameters of the model.  
  
Args:  
    W: word embedding matrix, dimension V x d where V is vocab size and d  
        is the embedding dimension  
    W_tilde: for asymmetric GloVe model, a second word embedding matrix, with  
        dimensions V x d  
    b: bias vector, dimension V x 1  
    b_tilde: for asymmetric GloVe model, a second bias vector, dimension V x 1  
    log_co_occurrence: V x V log co-occurrence matrix (log X)  
  
Returns:  
    loss: a scalar (float) for GloVe loss  
    """  
    n,_ = log_co_occurrence.shape  
    if W_tilde is None and b_tilde is None:  
        return np.sum((W @ W.T + b @ np.ones([1,n]) +  
                      np.ones([n,1])@b.T - log_co_occurrence)**2)  
    else:  
        return np.sum((W @ W_tilde.T + b @ np.ones([1,n]) +  
                      np.ones([n,1])@b_tilde.T - log_co_occurrence)**2)
```

```

def grad_GloVe(W, W_tilde, b, b_tilde, log_co_occurrence):
    """ Compute the gradient of the GloVe loss w.r.t. the parameters

    Args:
        W: word embedding matrix, dimension V x d where V is vocab size
            is the embedding dimension
        W_tilde: for asymmetric GloVe model, a second word embedding matrix
            dimensions V x d
        b: bias vector, dimension V x 1
        b_tilde: for asymmetric GloVe model, a second bias vector, dimension V x 1
        log_co_occurrence: V x V log co-occurrence matrix (log X)

    Returns:
        grad_W: gradient of the loss w.r.t W, dimension V x d
        grad_W_tilde: gradient of the loss w.r.t W_tilde, dimension V x d
            None if W_tilde is None.
        grad_b: gradient of the loss w.r.t b, dimension V x 1
        grad_b_tilde: gradient of the loss w.r.t b_tilde, dimension V x 1. Returns
            None if b_tilde is None.
    """
    n, _ = log_co_occurrence.shape
    ones = np.ones([n, 1])

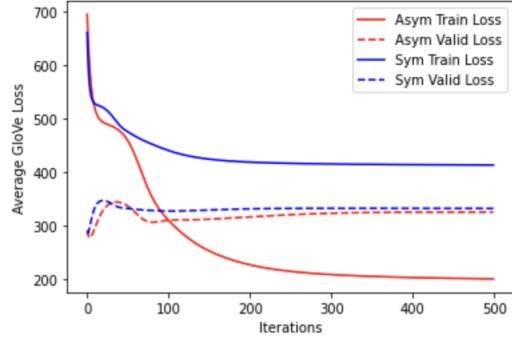
    if W_tilde is None and b_tilde is None:
        # Symmetric model
        X_pred = W @ W.T + b @ ones.T + ones @ b.T
        error = X_pred - log_co_occurrence
        grad_W = 2 * error @ W
        grad_b = 2 * np.sum(error, axis=0).reshape(-1, 1)
        grad_W_tilde = None
        grad_b_tilde = None
    else:
        # Asymmetric model
        X_pred = W @ W_tilde.T + b @ ones.T + ones @ b_tilde.T
        error = X_pred - log_co_occurrence
        grad_W = 2 * error @ W_tilde
        grad_W_tilde = 2 * error.T @ W
        grad_b = 2 * np.sum(error, axis=0).reshape(-1, 1)
        grad_b_tilde = 2 * np.sum(error, axis=1).reshape(-1, 1)

    return grad_W, grad_W_tilde, grad_b, grad_b_tilde

```

output graph:

Asymmetric and Symmetric GloVe Model on Asymmetric Log Co-Occurrence (Emb Dim=16)



For Q5.1 Number of parameters in neural network model: program trace:

The trainable parameters of the model consist of 3 weight matrices and 2 sets of biases. What is the total number of trainable parameters in the model, as a function of V, N, D, H ?

In the diagram given, which part of the model (i.e., `word_embedding_weights`, `embed_to_hid_weights`, `hid_to_output_weights`, `hid_bias`, or `output_bias`) has the largest number of trainable parameters if we have the constraint that $V \gg H > D > N$?
Note: The symbol \gg means ``much greater than'' Explain your reasoning.

The total number of trainable parameters in the model is given by the number of elements in the 3 weight matrices and 2 sets of biases. If V is the number of words in the vocabulary, N is the number of samples in the dataset, D is the dimensionality of the word embeddings, H is the number of hidden units in the hidden layer, then the number of parameters is:

- `word_embedding_weights`: $V \times D$ parameters
- `embed_to_hid_weights`: $D \times H$ parameters
- `hid_to_output_weights`: $H \times V$ parameters
- `hid_bias`: H parameters
- `output_bias`: V parameters

So, the total number of parameters is $V \times D + D \times H + H \times V + H + V$.

If $V \gg H > D > N$, then the largest number of parameters is in the `word_embedding_weights` matrix, which has $V \times D$ parameters. This is because the number of parameters in this matrix depends directly on the size of the vocabulary (V), which is usually much greater than the number of hidden units (H) or the number of samples (N).

For Q5.2 Number of parameters in n-gram model: program trace:



Another method for predicting the next words is an n-gram model. If we wanted to use [this](#) an n-gram model with the same context length $N-1$ as our network (since we mask 1 of the N words in our input), we'd need to store the counts of all possible N -grams. If we stored all the counts explicitly and suppose that we have V words in the dictionary, how many entries would this table have?



In an n-gram model with context length $N-1$, the number of possible N -grams would [like](#) [edit](#) be V^N . This is because each word in the N -gram can be any of the V words in the dictionary. So, if we stored all the counts explicitly, the table would have V^N entries.

For Q6.1 Implement Vectorized Loss: program trace for compute_loss:



Implement a vectorized `compute_loss` function, which computes the total cross-entropy loss on a mini-batch.
`\def compute_loss(self, output_activations, expanded_target_batch, target_mask):`

`"""Compute the total cross entropy loss over a mini-batch.`

Args:

`output_activations: [batch_size x (context_len * vocab_size)] matrix,
 for the activations of the output layer, i.e. the y_j 's.
 expanded_target_batch: [batch_size (context_len * vocab_size)] matrix,
 where expanded_target_batch[i,n*V:(n+1)*V] is the indicator vector for
 the n-th context target word position, i.e. the $(i, j + n*V)$ entry is 1 if the
 i'th example, the context word at position n is j, and 0 otherwise. matrix
 obtained`

`target_mask: A [batch_size x context_len x 1] tensor, where target_mask[i,n,0]
 = 1
 if for the i'th example the n-th context word is a target position,
 otherwise
 0`

Returns:

`loss: a scalar for the total cross entropy loss over the batch
 """`

```
# Get the number of examples in the batch
batch_size = output_activations.shape[0]
# Compute the log softmax activation for the target positions
log_softmax = - np.log(np.exp(output_activations)) / np.sum(np.exp(output_activations), axis=1, keepdims=True)
# Compute the cross entropy loss for each target position
cross_entropy = np.sum(expanded_target_batch * log_softmax, axis=1)
# Multiply the cross entropy loss by the target mask to only keep the loss for the target positions
masked_cross_entropy = cross_entropy * target_mask.flatten()
# Compute the total loss by summing the masked cross entropy loss over all examples
total_loss = np.sum(masked_cross_entropy) / batch_size
return total_loss
```

For Q6.3 Print the gradients:

output of print_gradients:

```
The loss derivative looks OK.  
The gradient for word_embedding_weights looks OK.  
The gradient for embed_to_hid_weights looks OK.  
The gradient for hid_to_output_weights looks OK.  
The gradient for hid_bias looks OK.  
The gradient for output_bias looks OK.
```

```
# Run this to print out the gradients  
print_gradients()  
  
loss_derivative[46, 785] 0.7137561447745507  
loss_derivative[46, 766] -0.9661570033238931  
loss_derivative[5, 42] -0.0  
loss_derivative[5, 31] 0.0  
  
param_gradient.word_embedding_weights[27, 2] 0.0  
param_gradient.word_embedding_weights[43, 3] 0.011596892511489458  
param_gradient.word_embedding_weights[22, 4] -0.0222670623817297  
param_gradient.word_embedding_weights[2, 5] 0.0  
  
param_gradient.embed_to_hid_weights[10, 2] 0.3793257091930164  
param_gradient.embed_to_hid_weights[15, 3] 0.01604516132110917  
param_gradient.embed_to_hid_weights[30, 9] -0.4312854367997419  
param_gradient.embed_to_hid_weights[35, 21] 0.06679896665436337  
  
param_gradient.hid_bias[10] 0.023428803123345148  
param_gradient.hid_bias[20] -0.024370452378874197  
  
param_gradient.output_bias[0] 0.000970106146902794  
param_gradient.output_bias[1] 0.16868946274763222  
param_gradient.output_bias[2] 0.0051664774143909235  
param_gradient.output_bias[3] 0.15096226471814364
```

For Q7.1 WEAT method for detecting bias:

program trace for weat_association_score:



Word embedding models such as GloVe attempt to learn a vector space where semantically similar words are clustered close together. However, they have been shown to learn problematic associations, e.g. by embedding `\say{man}` more closely to `\say{doctor}` than `\say{woman}` (and vice versa for `\say{nurse}`). To detect such biases in word embeddings, [\cite{caliskan2017semantics}](#) introduced the Word Embedding Association Test (WEAT). The WEAT test measures whether two `\textit{target}` word sets (e.g. `\{programmer, engineer, scientist, ...\}`) and `\{nurse, teacher, librarian, ...\}` have the same relative association to two `\textit{attribute}` word sets (e.g. `\{man, male, ...\}` and `\{woman, female ...\}`).[\footnote{There is an excellent blog on bias in word embeddings and the WEAT test at \url{https://developers.googleblog.com/2018/04/text-embedding-models-contain-bias.html}}](#)



Formally, let `\(A)`, `\(B)` be two sets of attribute words. Then

```
\begin{align}
s(w, A, B) = \text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b})
\end{align}
```

measures the association of a target word `\(w)` with the attribute sets - for convenience, we will call this the WEAT association score. A positive score means that the word `\w` is more associated with `A`, while a negative score means the opposite. For example, a WEAT association score of 1 in the following test `\$s(``programmer", \{man\}, \{woman\})=1$`, implies the ``programmer'' has a stronger association to `\{man\}`. For reference, the cosine similarity between two word vectors `\(|\vec{a}|)` and `\(|\vec{b}|)` is given by:

```
\begin{align}
\cos(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| |\vec{b}|}
\end{align}
```

\end{align}. Implement the function `\verb+weat_association_score()` and compute the WEAT association score for each target word.

output of weat_association_score from ChatGPT:



The function `weat_association_score` takes as input the target word, the two attribute sets `A` and `B`, and the word embeddings `embeddings`. It returns the WEAT association score for the target word:

Copy code

```
python

import numpy as np

def cosine_similarity(a, b):
    return np.dot(a, b) / (np.linalg.norm(a) * np.linalg.norm(b))

def weat_association_score(word, A, B, embeddings):
    target_vec = embeddings[word]
    mean_A = np.mean([cosine_similarity(target_vec, embeddings[a]) for a in A])
    mean_B = np.mean([cosine_similarity(target_vec, embeddings[b]) for b in B])
    return mean_A - mean_B

target_word = "programmer"
A = {"man", "male"}
B = {"woman", "female"}
embeddings = {
    "programmer": np.array([0.1, 0.2, 0.3]),
    "man": np.array([0.2, 0.3, 0.4]),
    "male": np.array([0.3, 0.4, 0.5]),
    "woman": np.array([-0.1, -0.2, -0.3]),
    "female": np.array([-0.2, -0.3, -0.4]),
}
weat_association_score(target_word, A, B, embeddings)
```

This code computes the WEAT association score for the target word `'"programmer"'` with respect to the attribute sets `A` and `B`. The cosine similarity between two word vectors is computed using the `cosine_similarity` function. The `weat_association_score` function first computes the mean cosine similarity between the target word and each word in the attribute set `A`, and then computes the mean cosine similarity between the target word and each word in the attribute set `B`. The difference between these two means gives the WEAT association score.

For Q7.3.1 1-word subsets:
output :

```
Gendered attribute words chosen: ['boyish'] ['female']
programmer : 0.01961511862464249 ==> -0.19506452232599258
engineer : 0.053647358901798725 ==> -0.3192642293870449
scientist : 0.06795816496014595 ==> -0.24408689141273499
nurse : -0.09486919268965721 ==> -0.41440044343471527
teacher : -0.01893029361963272 ==> -0.3508371561765671
librarian : -0.024141337256878614 ==> -0.14386512339115143
```

For Q7.3.2 How word frequency affects embedding similarity:

For example, if the word from A is more common in the training corpus, the larger the norm of its word embedding, the score is more likely to be positive. We can choose the words in above step according to the word frequency in the training corpus to decide whether the signs of the association scores for some particular occupations are reversed.