HOMEWORK 5: NEURAL NETWORKS

10-301/10-601 Introduction to Machine Learning (Fall 2021)
http://www.cs.cmu.edu/~mgormley/courses/10601/
OUT: Monday, October 11, 2021
DUE: Thursday, October 21, 2021 11:59 PM
TAs: Sana, Abhi, Sami, Helena, Chi

Summary In this assignment, you will build an image recognition system using a neural network. In the Written component, you will walk through an on-paper example of how to implement a neural network. Then, in the Programming component, you will implement an end-to-end system that learns to perform image classification.

START HERE: Instructions

- **Collaboration Policy**: Please read the collaboration policy here: http://www.cs.cmu.edu/~mgormley/courses/10601/syllabus.html
- **Late Submission Policy:** See the late submission policy here: http://www.cs.cmu.edu/~mgormley/courses/10601/syllabus.html
- **Submitting your work:** You will use Gradescope to submit answers to all questions and code. Please follow instructions at the end of this PDF to correctly submit all your code to Gradescope.
  - **Written:** For written problems such as short answer, multiple choice, derivations, proofs, or plots, please use the provided template. Submissions can be handwritten onto the template, but should be labeled and clearly legible. If your writing is not legible, you will not be awarded marks. If your scanned submission misaligns the template, there will be a 5% penalty. Alternatively, submissions can be written in LaTeX. Each derivation/proof should be completed in the boxes provided. If you do not follow the template, your assignment may not be graded correctly by our AI assisted grader.
  - **Programming:** You will submit your code for programming questions on the homework to Gradescope (https://gradescope.com). After uploading your code, our grading scripts will autograde your assignment by running your program on a virtual machine (VM). When you are developing, check that the version number of the programming language environment (e.g. Python 3.9.6, OpenJDK 11.0.11, g++ 7.5.0) and versions of permitted libraries (e.g. `numpy` 1.21.2 and `scipy` 1.7.1) match those used on Gradescope. You have 10 Gradescope programming submissions. We recommend debugging your implementation on your local machine (or the Linux servers) and making sure your code is running correctly first before submitting you code to Gradescope.
- **Materials:** The data that you will need in order to complete this assignment is posted along with the writeup and template on Piazza.

Linear Algebra Libraries When implementing machine learning algorithms, it is often convenient to have a linear algebra library at your disposal. In this assignment, Java users may use EJML[a] or ND4J[b] and C++ users may use Eigen[c]. Details below. (As usual, Python users have NumPy.)

EJML for Java EJML is a pure Java linear algebra package with three interfaces. We strongly recommend using the SimpleMatrix interface. The autograder will use EJML version 0.41. When compiling and running your code, we will add the additional command line argument `-cp "linalg_lib/ejml-v0.41-libs/*:linalg_lib/nd4j-v1.0.0-M1.1-libs/*:./"` to ensure that all the EJML jars are on the classpath as well as your code.

ND4J for Java ND4J is a library for multidimensional tensors with an interface akin to Python's NumPy. The autograder will use ND4J version 1.0.0-M1.1. When compiling and running your code, we will add the additional command line argument `-cp "linalg_lib/ejml-v0.41-libs/*:linalg_lib/nd4j-v1.0.0-M1.1-libs/*:./"` to ensure that all the ND4J jars are on the classpath as well as your code.

Eigen for C++ Eigen is a header-only library, so there is no linking to worry about—just `#include` whatever components you need. The autograder will use Eigen version 3.4.0. The command line arguments above demonstrate how we will call you code. When compiling your code we will include, the argument `-I./linalg_lib` in order to include the `linalg_lib/Eigen` subdirectory, which contains all the headers.

We have included the correct versions of EJML/ND4J/Eigen in the `linalg_lib.zip` posted on the Coursework page of the course website for your convenience. It contains the same `linalg_lib/` directory that we will include in the current working directory when running your tests. Do **not** include EJML, ND4J, or Eigen in your homework submission; the autograder will ensure that they are in place.

---

[a] https://ejml.org
[b] https://javadoc.io/doc/org.nd4j/nd4j-api/latest/index.html
[c] http://eigen.tuxfamily.org/

Instructions for Specific Problem Types

For "Select One" questions, please fill in the appropriate bubble completely:

Select One: Who taught this course?

● Matt Gormley / Henry Chai

○ Marie Curie

○ Noam Chomsky

If you need to change your answer, you may cross out the previous answer and bubble in the new answer:

Select One: Who taught this course?

● Matt Gormley / Henry Chai

○ Marie Curie

● Noam Chomsky

For "Select all that apply" questions, please fill in all appropriate squares completely:

Select all that apply: Which are scientists?

□ Stephen Hawking

■ **Albert Einstein**

□ Isaac Newton

□ None of the above

Again, if you need to change your answer, you may cross out the previous answer(s) and bubble in the new answer(s):

Select all that apply: Which are scientists?

■ Stephen Hawking

■ Albert Einstein

■ Isaac Newton

■ I don't know

For questions where you must fill in a blank, please make sure your final answer is fully included in the given space. You may cross out answers or parts of answers, but the final answer must still be within the given space.

Fill in the blank: What is the course number?

10-601

10-7601

Written Questions (53 points)

1 Example Feed Forward and Backpropagation

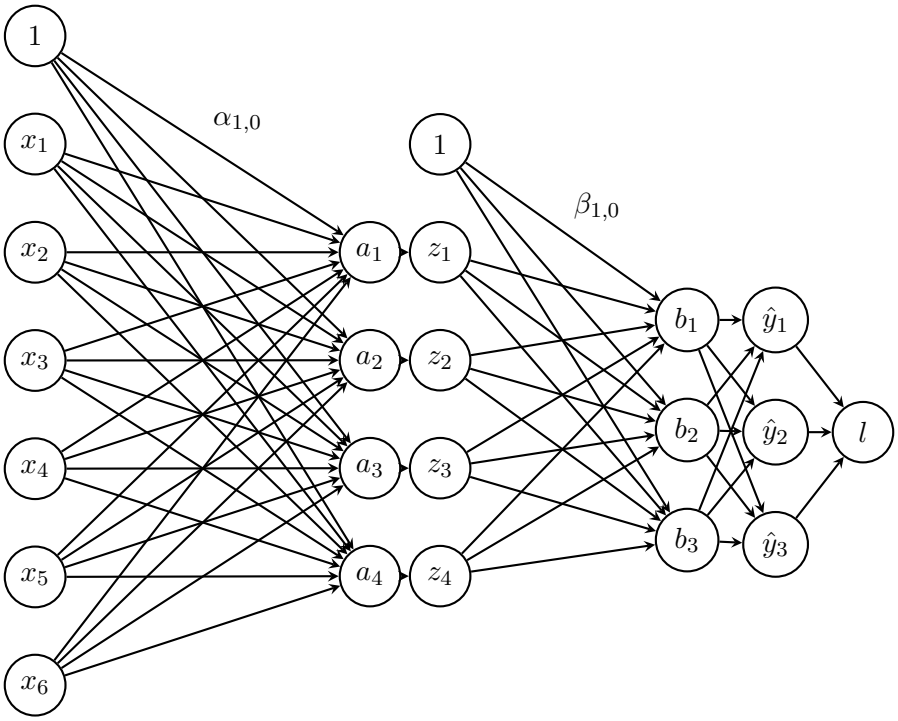


Figure 1: A One Hidden Layer Neural Network

Network Overview Consider the neural network with one hidden layer shown in Figure 1. The input layer consists of 6 features $\mathbf{x} = [x_1, ..., x_6]^T$, the hidden layer has 4 nodes $\mathbf{z} = [z_1, ..., z_4]^T$, and the output layer is a probability distribution $\mathbf{y} = [y_1, y_2, y_3]^T$ over 3 classes (**1-indexed** such that y_i is the probability of label i). We also allow for a bias term by adding a dimension with constant value one to the input ($x_0 = 1$) and to the hidden layer ($z_0 = 1$).

$\boldsymbol{\alpha}$ is the matrix of weights from the inputs to the hidden layer and $\boldsymbol{\beta}$ is the matrix of weights from the hidden layer to the output layer.

$\alpha_{j,i}$ represents the weight going *to* the node z_j in the hidden layer *from* the node x_i in the input layer (e.g. $\alpha_{1,2}$ is the weight from x_2 to z_1), and $\boldsymbol{\beta}$ is defined similarly. We will use a sigmoid activation function for the hidden layer and a softmax for the output layer.

Network Details Equivalently, we define each of the following.

The input:

$$\mathbf{x} = [x_1, x_2, x_3, x_4, x_5, x_6]^T \tag{1}$$

Linear combination at the first (hidden) layer:

$$a_j = \alpha_{j,0} + \sum_{i=1}^{6} \alpha_{j,i} \cdot x_i, \ \forall j \in \{1, \ldots, 4\} \tag{2}$$

Activation at the first (hidden) layer:

$$z_j = \sigma(a_j) = \frac{1}{1 + \exp(-a_j)}, \ \forall j \in \{1, \ldots, 4\} \tag{3}$$

Equivalently, we can write this as vector operation where the sigmoid activation is applied individually to each element of the vector $\mathbf{a}$:

$$\mathbf{z} = \sigma(\mathbf{a}) \tag{4}$$

Linear combination at the second (output) layer:

$$b_k = \beta_{k,0} + \sum_{j=1}^{4} \beta_{k,j} \cdot z_j, \ \forall k \in \{1, \ldots, 3\} \tag{5}$$

Activation at the second (output) layer:

$$\hat{y}_k = \frac{\exp(b_k)}{\sum_{l=1}^{3} \exp(b_l)}, \ \forall k \in \{1, \ldots, 3\} \tag{6}$$

Note that the linear combination equations can be written equivalently as the product of the weight matrix with the input vector. We can even fold in the bias term α_0 by thinking of $x_0 = 1$, and fold in $\beta_{j,0}$ by thinking of $z_0 = 1$.

Loss We will use cross entropy loss, $\ell(\hat{\mathbf{y}}, \mathbf{y})$. If $\mathbf{y}$ represents our target output, which will be a one-hot vector representing the correct class, and $\hat{\mathbf{y}}$ represents the output of the network, the loss is calculated by:

$$\ell(\hat{\mathbf{y}}, \mathbf{y}) = -\sum_{i=1}^{3} y_i \log(\hat{y}_i) \tag{7}$$

For the below questions use natural log in the equation.

Prediction When doing prediction, we will predict the argmax of the output layer. For example, if $\hat{y}_1 = 0.3, \hat{y}_2 = 0.2, \hat{y}_3 = 0.5$ we would predict class 3. If the true class from the training data was 2 we would have a one-hot vector $\mathbf{y}$ with values $y_1 = 0$, $y_2 = 1$, $y_3 = 0$.

1. In the following questions you will derive the matrix and vector forms of the previous equations which define our neural network. These are what you should hope to program in order to keep your program under the Gradescope time-out.
   When working these out it is important to keep a note of the vector and matrix dimensions in order for you to easily identify what is and isn't a valid multiplication. Suppose you are given an training example: $\mathbf{x}^{(1)} = [x_1, x_2, x_3, x_4, x_5, x_6]^T$ with **label class 2**, so $\mathbf{y}^{(1)} = [0, 1, 0]^T$. We initialize the network weights as:

$$\boldsymbol{\alpha}^* = \begin{bmatrix} \alpha_{1,1} & \alpha_{1,2} & \alpha_{1,3} & \alpha_{1,4} & \alpha_{1,5} & \alpha_{1,6} \\ \alpha_{2,1} & \alpha_{2,2} & \alpha_{2,3} & \alpha_{2,4} & \alpha_{2,5} & \alpha_{2,6} \\ \alpha_{3,1} & \alpha_{3,2} & \alpha_{3,3} & \alpha_{3,4} & \alpha_{3,5} & \alpha_{3,6} \\ \alpha_{4,1} & \alpha_{4,2} & \alpha_{4,3} & \alpha_{4,4} & \alpha_{4,5} & \alpha_{4,6} \end{bmatrix}$$

$$\boldsymbol{\beta}^* = \begin{bmatrix} \beta_{1,1} & \beta_{1,2} & \beta_{1,3} & \beta_{1,4} \\ \beta_{2,1} & \beta_{2,2} & \beta_{2,3} & \beta_{2,4} \\ \beta_{3,1} & \beta_{3,2} & \beta_{3,3} & \beta_{3,4} \end{bmatrix}$$

We want to also consider the bias term and the weights on the bias terms ($\alpha_{j,0}$ and $\beta_{k,0}$). To account for these we can add a new column to the beginning of our initial weight matrices to represent biases.

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_{1,0} & \alpha_{1,1} & \alpha_{1,2} & \alpha_{1,3} & \alpha_{1,4} & \alpha_{1,5} & \alpha_{1,6} \\ \alpha_{2,0} & \alpha_{2,1} & \alpha_{2,2} & \alpha_{2,3} & \alpha_{2,4} & \alpha_{2,5} & \alpha_{2,6} \\ \alpha_{3,0} & \alpha_{3,1} & \alpha_{3,2} & \alpha_{3,3} & \alpha_{3,4} & \alpha_{3,5} & \alpha_{3,6} \\ \alpha_{4,0} & \alpha_{4,1} & \alpha_{4,2} & \alpha_{4,0} & \alpha_{4,4} & \alpha_{4,5} & \alpha_{4,6} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{1,0} & \beta_{1,1} & \beta_{1,2} & \beta_{1,3} & \beta_{1,4} \\ \beta_{2,0} & \beta_{2,1} & \beta_{2,2} & \beta_{2,3} & \beta_{2,4} \\ \beta_{3,0} & \beta_{3,1} & \beta_{3,2} & \beta_{3,3} & \beta_{3,4} \end{bmatrix}$$

We then add a corresponding new first dimension to our input vectors, always set to 1 ($x_0^{(i)} = 1$), so our input becomes:

$$\mathbf{x}^{(1)} = [1, x_1, x_2, x_3, x_4, x_5, x_6]^T$$

(a) (1 point) By examining the shapes of the initial weight matrices, how many neurons do we have in the first hidden layer of the neural network? (Not including the bias neuron)

Answer

4

(b) (1 point) How many output neurons will our neural network have?

Answer

3

(c) (1 point) What is the vector $\mathbf{a}$ whose elements are made up of the entries a_j in Equation 2 (using $x_i^{(1)}$ in place of x_i). Write your answer in terms of α and $\mathbf{x}^{(1)}$.

Answer

$a = \alpha x^{(1)}$

(d) (1 point) **Select one:** We cannot take the matrix multiplication of our weights $\boldsymbol{\beta}$ and our vector $\mathbf{z}$ since they are not compatible shapes. Which of the following would allow us to take the matrix multiplication of $\boldsymbol{\beta}$ and $\mathbf{z}$ such that the entries of the vector $\mathbf{b} = \boldsymbol{\beta}\mathbf{z}$ are equivalent to the values of b_k in Equation 5?

- ○ A) Remove the last column of $\boldsymbol{\beta}$
- ○ B) Remove the first row of $\mathbf{z}$
- ● C) Append a value of 1 to be the first entry of $\mathbf{z}$
- ○ D) Append an additional column of 1's to be the first column of $\boldsymbol{\beta}$
- ○ E) Append a row of 1's to be the first row of $\boldsymbol{\beta}$
- ○ F) Take the transpose of $\boldsymbol{\beta}$

(e) (1 point) What are the entries of the output vector $\hat{\mathbf{y}}$? Your answer should be written in terms of b_1, b_2, b_3.

$\hat{\mathbf{y}}$

$\hat{y_1} = \frac{exp(b_1)}{exp(b_1)+exp(b_2)+exp(b_3)}$ $\hat{y_2} = \frac{exp(b_2)}{exp(b_1)+exp(b_2)+exp(b_3)}$ $\hat{y_3} = \frac{exp(b_3)}{exp(b_1)+exp(b_2)+exp(b_3)}$

2. We will now derive the matrix and vector forms for the backpropagation algorithm.

$$\frac{\partial \ell}{\partial \boldsymbol{\alpha}} = \begin{bmatrix} \frac{\partial \ell}{\partial \alpha_{10}} & \frac{\partial \ell}{\partial \alpha_{11}} & \cdots & \frac{\partial \ell}{\partial \alpha_{16}} \\ \frac{\partial \ell}{\partial \alpha_{20}} & \frac{\partial \ell}{\partial \alpha_{21}} & \cdots & \frac{\partial \ell}{\partial \alpha_{26}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \ell}{\partial \alpha_{40}} & \frac{\partial \ell}{\partial \alpha_{41}} & \cdots & \frac{\partial \ell}{\partial \alpha_{46}} \end{bmatrix}$$

The mathematics which you have to derive in this section jump significantly in difficulty, you should always be examining the shape of the matrices and vectors and making sure that you are comparing your matrix elements with calculations of individual derivatives to make sure they match (e.g., the element of the matrix $(\frac{\partial \ell}{\partial \boldsymbol{\alpha}})_{2,1}$ should be equal to $\frac{\partial \ell}{\partial \alpha_{2,1}}$). Recall that ℓ is our loss function defined in Equation 7:

(a) (3 points) The derivative of the softmax function with respect to b_k is as follows:

$$\frac{\partial \hat{y}_l}{\partial b_k} = \hat{y}_l(\mathbb{I}[k = l] - \hat{y}_k)$$

where $\mathbb{I}[k = l]$ is an indicator function such that if $k = l$ then it returns value 1 and 0 otherwise. Using this, write the derivative $\frac{\partial \ell}{\partial b_k}$ in a smart way such that you do not need this indicator function. Write your solutions in terms of $\hat{y}_k, y_k$. Show your work below.

HINT: Recall that $\frac{\partial \ell}{\partial b_k} = \sum_l \frac{\partial \ell}{\partial \hat{y}_l} \frac{\partial \hat{y}_l}{\partial b_k}$.

$\partial \ell / \partial b_k$

$\frac{\partial \ell}{\partial b_k} = \sum_l \frac{\partial \ell}{\partial \hat{y}_l} \frac{\partial \hat{y}_l}{\partial b_k} = \sum_l -\frac{y_l}{\hat{y}_l} \hat{y}_l(\mathbb{I}[k = l] - \hat{y}_k) = -y_k + \hat{y_k} \sum_l y_l = \hat{y_k} - y_k$

(b) (1 point) What are the elements of the vector $\frac{\partial \ell}{\partial \mathbf{b}}$? Use the convention that $\frac{\partial \ell}{\partial \mathbf{b}}$ is a row vector. (Recall that $\mathbf{y}^{(1)} = [0, 1, 0]^T$)

$\partial \ell / \partial \mathbf{b}$

$\partial \ell / \partial \mathbf{b} = c(\partial \ell / \partial b_1, \partial \ell / \partial b_2, \partial \ell / \partial b_3) = (\hat{y_1} - y_1, \hat{y_2} - y_2, \hat{y_3} - y_3) = (\hat{y_1}, \hat{y_2} - 1, \hat{y_3})$

(c) (2 points) What is the derivative $\frac{\partial \ell}{\partial \boldsymbol{\beta}}$? Your answer should be in terms of $\frac{\partial \ell}{\partial \mathbf{b}}$ and $\mathbf{z}$. Use the convention that $\frac{\partial \ell}{\partial \mathbf{b}}$ is a row vector and $\mathbf{z}$ is a column vector.

You should first consider a single entry in this matrix: $\frac{\partial \ell}{\partial \beta_{kj}}$.

$\partial \ell / \partial \boldsymbol{\beta}$

$\partial \ell / \partial \boldsymbol{\beta} = (\mathbf{z} \frac{\partial \ell}{\partial \mathbf{b}})^T$

(d) (1 point) Explain in one short sentence why we use the matrix $\boldsymbol{\beta}^*$ (the matrix $\boldsymbol{\beta}$ without the first column of ones) when calculating the derivative matrix $\frac{\partial \ell}{\partial \alpha}$?

Answer

Because α does not contribute to z_0, and subsequently we do not need to evaluate $\frac{\partial \mathbf{b}}{\partial z_0}$. Therefore, we use the matrix $\boldsymbol{\beta}^*$.

(e) (1 point) What is the derivative $\frac{\partial \ell}{\partial \mathbf{z}}$? Your answer should be in terms of $\frac{\partial \ell}{\partial \mathbf{b}}$ and $\boldsymbol{\beta}^*$. Use the convention that $\frac{\partial \ell}{\partial \mathbf{b}}$ is a row vector.

$\partial \ell / \partial \mathbf{z}$

$\partial \ell / \partial \mathbf{z} = (\frac{\partial \ell}{\partial \mathbf{b}} \beta^*)^T$

(f) (1 point) What is the derivative $\frac{\partial \ell}{\partial a_j}$ in terms of $\frac{\partial \ell}{\partial z_j}$ and z_j?

$\partial \ell / \partial a_j$

$\partial \ell / \partial a_j = \frac{\partial \ell}{\partial z_j} \frac{\partial z_j}{\partial a_j} = \frac{\partial \ell}{\partial z_j} z_j (1 - z_j)$.

(g) (1 point) What is the matrix $\frac{\partial \ell}{\partial \boldsymbol{\alpha}}$? Your answer should be in terms of $\frac{\partial \ell}{\partial \mathbf{a}}$ and $\mathbf{x}^{(1)}$. Use the convention that $\frac{\partial \ell}{\partial \mathbf{a}}$ is a row vector and $\mathbf{x}^{(1)}$ is a column vector.

$\partial \ell / \partial \boldsymbol{\alpha}$

$\partial \ell / \partial \boldsymbol{\alpha} = (\mathbf{x}^{(1)} \frac{\partial \ell}{\partial \mathbf{a}})^T$.

3. Now you will put these equations to use in an example with numerical values. **You should use the answers you get here to debug your code.**
You are given a training example $\mathbf{x}^{(1)} = [1, 1, 0, 0, 1, 1]^T$ with **label class 2**, so $\mathbf{y}^{(1)} = [0, 1, 0]^T$. We initialize the network weights as:

$$\boldsymbol{\alpha}^* = \begin{bmatrix} 1 & 2 & -3 & 0 & 1 & -3 \\ 3 & 1 & 2 & 1 & 0 & 2 \\ 2 & 2 & 2 & 2 & 2 & 1 \\ 1 & 0 & 2 & 1 & -2 & 2 \end{bmatrix}$$

$$\boldsymbol{\beta}^* = \begin{bmatrix} 1 & 2 & -2 & 1 \\ 1 & -1 & 1 & 2 \\ 3 & 1 & -1 & 1 \end{bmatrix}$$

We want to also consider the bias term and the weights on the bias terms ($\alpha_{j,0}$ and $\beta_{j,0}$). Lets say they are all initialized to 1. To account for this we can add a column of 1's to the beginning of our initial weight matrices.

$$\boldsymbol{\alpha} = \begin{bmatrix} 1 & 1 & 2 & -3 & 0 & 1 & -3 \\ 1 & 3 & 1 & 2 & 1 & 0 & 2 \\ 1 & 2 & 2 & 2 & 2 & 2 & 1 \\ 1 & 1 & 0 & 2 & 1 & -2 & 2 \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} 1 & 1 & 2 & -2 & 1 \\ 1 & 1 & -1 & 1 & 2 \\ 1 & 3 & 1 & -1 & 1 \end{bmatrix}$$

And we can set our first value of our input vectors to always be 1 ($x_0^{(i)} = 1$), so our input becomes:

$$\mathbf{x}^{(1)} = [1, 1, 1, 0, 0, 1, 1]^T$$

Using the initial weights, run the feed forward of the network over this example (rounding to 4 decimal places during the calculation) and then answer the following questions.

(a) (1 point) What is a_1?

a_1	Work
2.0000	$a_1 = \boldsymbol{\alpha}_1 \mathbf{x}^{(1)} = 2$

(b) (1 point) What is z_1?

z_1	Work
0.8808	$z_1 = \frac{1}{1+e^{-a_1}} = 0.8808$

(c) (1 point) What is b_2? We have computed z_2 = 0.9991, z_3 = 0.9997, z_4 = 0.8808 for you.

b_2	Work
3.6430	$b_2 = 1 + 0.8808 - 0.9991 + 0.9997 + 2 * 0.8808 = 3.6430$

(d) (1 point) What is $\hat{y}_2$? We have computed $b_1 = 2.7604, b_3 = 4.5226$ for you.

$\hat{y}_2$	Work
0.2615	$\hat{y_2} = \frac{exp(b_2)}{exp(b_1)+exp(b_2)+exp(b_3)} = 0.2615$

(e) (1 point) Which class would we predict on this example? Your answer should just be an integer $\in \{1, 2, 3\}$.

Class	Work
3	$\hat{y_1} = \frac{exp(b_1)}{exp(b_1)+exp(b_2)+exp(b_3)} = 0.1082$ $\hat{y_3} = \frac{exp(b_3)}{exp(b_1)+exp(b_2)+exp(b_3)} = 0.6303$

(f) (1 point) What is the total loss on this example?

Loss	Work
1.3412	$-y_2 log(\hat{y_2}) = -log(\hat{y_2}) = 1.3412$

4. Now use the results of the previous question to run backpropagation over the network and update the weights. Use learning rate $\eta = 1$.

   Do your backpropagation calculations rounding to 4 decimal places then answer the following questions:

   (a) (1 point) What is the value of $\frac{\partial \ell}{\partial \beta_{1,0}}$?

$\partial \ell / \partial \beta_{1,0}$	Work
0.1082	$\frac{\partial \ell}{\partial \beta_{1,0}} = \sum_l \frac{\partial \ell}{\partial \hat{y}_l} \frac{\partial \hat{y}_l}{\partial b_1} \frac{\partial \hat{b_1}}{\partial \beta_{1,0}} = (-y_1 + \hat{y_1}) * z_0 = \hat{y_1} = 0.1082$

   (b) (1 point) What is the updated value of the weight $\beta_{1,0}$?

$\beta_{1,0}$	Work
0.8918	$1 - (\hat{y_1} - y_1) * \eta = 1 - \hat{y_1} = 0.8918$

   (c) (1 point) What is the value of $\frac{\partial \ell}{\partial \alpha_{3,4}}$?

$\partial \ell / \partial \alpha_{3,4}$	Work
0.0000	$\sum_l \sum_j \frac{\partial \ell}{\partial \hat{y}_l} \frac{\partial \hat{y}_l}{\partial b_j} \frac{\partial b_j}{\partial z_3} \frac{\partial z_3}{\partial a_3} \frac{\partial a_3}{\partial \alpha_{3,4}} = 0$

(d) (1 point) What is the updated value of the weight $\alpha_{3,4}$?

$\alpha_{3,4}$	Work
2.0000	$\alpha_{3,4} - \eta * \frac{\partial \ell}{\partial \alpha_{3,4}} = \alpha_{3,4} = 2$

(e) (2 points) What is the updated weight of the input layer bias term applied to z_2 (i.e. $\alpha_{2,0}$)?

$\alpha_{2,0}$	Work
0.9986	$\alpha_{2,0} - \eta * \frac{\partial \ell}{\alpha_{2,0}} = 1 - 0.0014 = 0.9986$

2 Convolutional Neural Networks

In this problem, consider only the convolutional layer of a standard implementation of a CNN as described in Lecture 13.

1. We are given image X and filter F below.

$X =$

1	0	-2	3	4	1
2	9	5	6	0	-1
0	-3	1	3	4	4
6	5	2	0	6	8
-5	4	-3	1	3	-2
4	1	2	8	9	7

$F =$

-1	-1	-1
-1	8	-1
-1	-1	-1

$Y =$

a	b	c	d
e	f	g	h
i	j	k	l
m	n	o	p

(a) (1 point) Let X be convolved with F using no padding and a stride of 1 to produce an output Y. What is value of j in the output Y?

Answer

8

(b) (1 point) Suppose you had an input feature map of size (height $\times$ width) 6x4 and filter size 2x2, using no padding and a stride of 2, what would be the resulting output size? Write your answer in the format height $\times$ width.

Answer

3x2

2. Parameter sharing is a very important concept for CNN because it drastically reduces the complexity of the learning problem. For the following questions, assume that there is no bias term in our convolutional layer.

   (a) (1 point) Which of the following are parameters of a convolutional layer?
   Select all that apply:

   - ☐ stride size
   - ☐ padding size
   - ☐ image size
   - ☐ filter size
   - ☒ weights in the filter
   - ☐ None of above.

   (b) (1 point) Which of the following are hyperparameters of a convolutional layer?
   Select all that apply:

   - ☒ stride size
   - ☒ padding size
   - ☐ image size
   - ☒ filter size
   - ☐ weights in the filter
   - ☐ None of above.

   (c) (1 point) Suppose for the convolutional layer, we are given grayscale images of size 22×22. Using one single 4×4 filter with a stride of 2 and no padding, what is the number of parameters you are learning in this layer?

   Answer: 16

   (d) (1 point) Suppose instead of sharing the same filter for the entire image, you learn a new filter each time you move across the image. Using 4×4 filters with a stride of 2 and no padding, what is the number of parameters you are learning in this layer?

   Answer: 1600

   (e) (1 point) Now suppose you are given a 40×40 colored image, which consists of 3 channels (so your input is a $40 \times 40 \times 3$ tensor), each representing the intensity of one primary color. Suppose

you learn a new filter each time you move across the image. Using 4×4 filters with a stride of 2 and no padding, what is the number of parameters you are learning in this layer?

Answer

17328

(f) (1 point) In a sentence, describe a reason why parameter sharing is a good idea for a convolutional layer applied to image data, besides the reduction in number of learned parameters.

Answer

We can take advantage of the translation invariance to detect a feature regardless of where in the image the feature occurs.

3 Empirical Questions

The following questions should be completed after you work through the programming portion of this assignment.

For these questions, **use the small dataset**. Use the following values for the hyperparameters unless otherwise specified:

Parameter	Value
Number of Hidden Units	50
Weight Initialization	RANDOM
Learning Rate	0.01

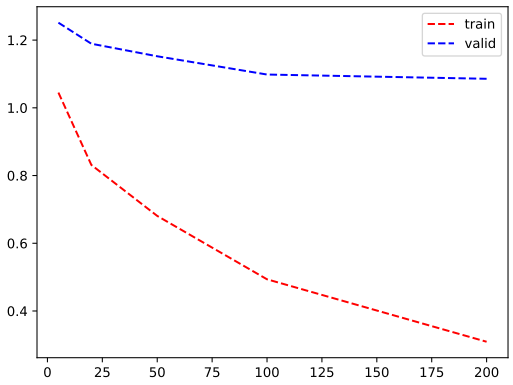
Please submit computer-generated plots for (a)i and (b)i. Note: we expect it to take about **5 minutes** to train each of these networks.

1. Hidden Units

   (a) (2 points) Train a single hidden layer neural network using the hyperparameters mentioned in the table above, except for the number of hidden units which should vary among 5, 20, 50, 100, and 200. Run the optimization for 100 epochs each time.

   Plot the average training cross-entropy (sum of the cross-entropy terms over the training dataset divided by the total number of training examples) on the y-axis vs number of hidden units on the x-axis. In the **same figure**, plot the average validation cross-entropy.

Avg. Train and Validation Cross-Entropy Loss



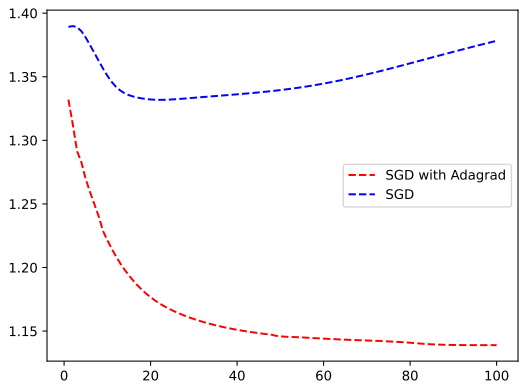
(b) (2 points) Examine and comment on the the plots of training and validation cross-entropy. What is the effect of changing the number of hidden units?

Answer

First of all, the training cross entropy loss is always smaller than the validation cross entropy loss across all the number of hidden units. In addition, as the number of hidden units increases, the training cross entropy loss drops faster, while the validation cross entropy gradually converges, so the gap between training loss and validation loss becomes larger and larger. This is not surprised since large number of hidden units may lead to overfitting where the training loss is very small but the validation loss is not.

(c) (2 points) In the handout folder, we provide `val_loss_sgd_small.txt`, a text file with the validation cross-entropy loss values for SGD performed using 100 epochs, 50 hidden units, random init, and 0.01 learning rate. In the **same figure**, plot them against your validation results for SGD **with** Adagrad using the same set of parameters and the small dataset.

Avg. Validation Cross-Entropy Loss of SGD with and without AdaGrad



(d) (2 points) Examine and compare the two results. What do you observe?

Answer

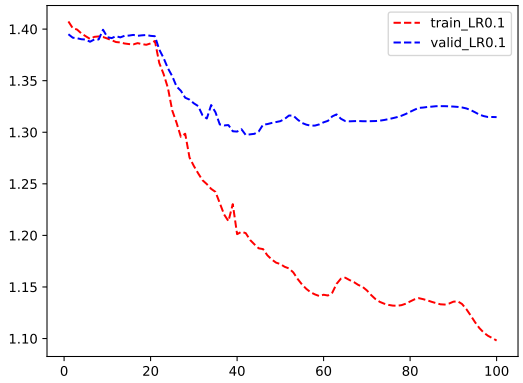
The validation cross entropy loss by SGD with Adagrad is always smaller than the validation loss by SGD. In addition, with Adagrad, the validation loss is monotonically decreasing as the number of epoch increases, while with SGD, the validation loss first decreases then increases. Because the SGD with Adagrad actually changes the step size in each iteration, to be more specifically, reduces the step size each time, so this is helpful to avoid stepping over the minimum point. Therefore, as the number of epoch/iteration increases, the validation loss by SGD may increases but not by SGD with Adagrad.

2. Learning Rate

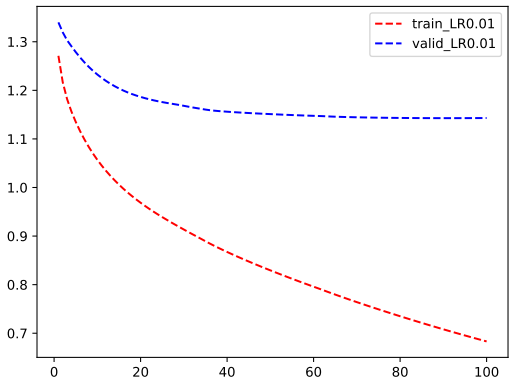
(a) (6 points) Train a single hidden layer neural network using the hyperparameters mentioned in the table above, except for the learning rate which should vary among 0.1, 0.01, and 0.001. Run the optimization for 100 epochs each time.

Plot the average training cross-entropy on the y-axis vs the number of epochs on the x-axis for the mentioned learning rates. In the **same figure**, plot the average validation cross-entropy loss. Make a separate figure for each learning rate.

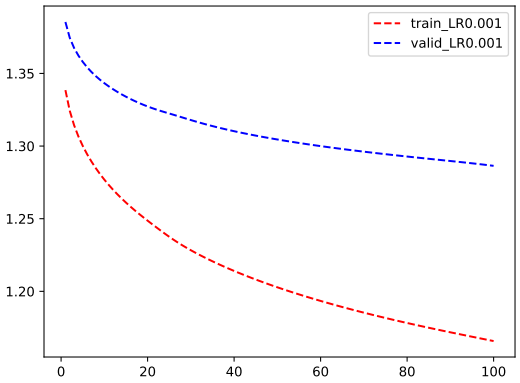
Plot LR 0.1



Plot LR 0.01



Plot LR 0.001



(b) (2 points) Examine and comment on the plots of training and validation cross-entropy. How does adjusting the learning rate affect the convergence of cross-entropy on the datasets?

Answer

First of all, the largest learning rate causes the fluctuated validation and training loss lines over the number of epoch, which indicated that the cross entropy loss is not monotonically decreasing over epoches and some updates may step over or go the the opposite direction. For both the moderate and smallest learning rate, the validation and training loss lines are smooth and monotonically non-increasing. However, for the smallest learning rate, both the training loss and the validation loss converge very slowly. Moreover, the moderate learning rate has the smallest training loss and validation loss compared to learning rate of 0.1 and 0.001 after 100 epoches.

3. Weight Initialization

(a) (2 points) For this exercise, you can work on any data set. Initialize $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to zero and print them out after the first few updates. For example, you may use the following command to begin:

```
$ python neuralnet.py smallTrain.csv smallValidation.csv \
smallTrain_out.labels smallValidation_out.labels \
smallMetrics_out.txt 1 4 2 0.1
```

Compare the values across rows and columns in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Comment on what you observed. Do you think it is reasonable to use zero initialization? Why or why not?

Answer

After the first 15 updates, the rows of α are the same and the columns of β (except the first column) are the same. This results indicate that the hidden neuron units actually learn the same things from the features, and subsequently each neuron (except z_0) has the same contribution to $\hat{y}_i$. It may not be good to use zero initialization, because this will lead to the same gradient, and consequently all neurons will learn the same things from the features and in each iterative update.

4 Collaboration Questions

After you have completed all other components of this assignment, report your answers to these questions regarding the collaboration policy. Details of the policy can be found here.

1. Did you receive any help whatsoever from anyone in solving this assignment? If so, include full details.
2. Did you give any help whatsoever to anyone in solving this assignment? If so, include full details.
3. Did you find or come across code that implements any part of this assignment ? If so, include full details.

Your Answer

None

Programming (61 points)

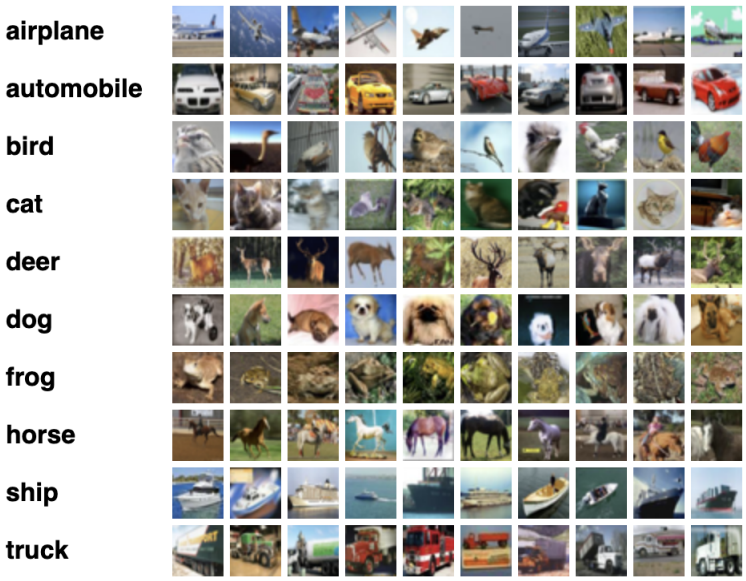


Figure 2: 10 random images from each of 10 image categories in CIFAR-10. You will be using a simplified version of this dataset.

5 The Task

Your goal in this assignment is to implement a neural network to classify images using a single hidden layer neural network. In addition, you will implement Adagrad, a variant of stochastic gradient descent.

6 The Datasets

Datasets We will be using a subset of a standard Computer Vision dataset, CIFAR-10. This data includes color images of various vehicles and animals; our subset will include black and white images of the 4 classes `automobile`, `bird`, `frog`, `ship`. The handout contains one dataset drawn from this data with 500 samples for training and 100 for validation.

File Format Each dataset (small, medium, and large) consists of two csv files—train and validation. Each row contains 1025 columns separated by commas. The first column contains the label and columns 2 to 1025 represent the pixel values of a 32×32 image in a row major format. Label 0 corresponds to `automobile`, 1 to `bird`, 2 to `frog`, and 3 to `ship`.

You should write your code to accept arbitrary pixel values in the range [0, 1]. The images in Figure 2 were produced by converting these pixel values into .png files for visualization. Observe that no feature engineering has been done here; instead the neural network you build will *learn* features appropriate for the task of image recognition.

7 Model Definition

In this assignment, you will implement a single-hidden-layer neural network with a sigmoid activation function for the hidden layer, and a softmax on the output layer. Let the input vectors $\mathbf{x}$ be of length M, and the hidden layer $\mathbf{z}$ consist of D hidden units. In addition, let the output layer $\hat{\mathbf{y}}$ be a probability distribution over K classes. That is, each element $\hat{y}_k$ of the output vector represents the probability of $\mathbf{x}$ belonging to the class k.

$$\hat{y}_k = \frac{\exp(b_k)}{\sum_{l=1}^{K} \exp(b_l)}$$
$$b_k = \beta_{k,0} + \sum_{j=1}^{D} \beta_{kj} z_j$$
$$z_j = \frac{1}{1 + \exp(-a_j)}$$
$$a_j = \alpha_{j,0} + \sum_{m=1}^{M} \alpha_{jm} x_m$$

We can compactly express this model by assuming that $x_0 = 1$ is a bias feature on the input and that $z_0 = 1$ is also fixed. In this way, we have two parameter matrices $\boldsymbol{\alpha} \in \mathbb{R}^{D \times (M+1)}$ and $\boldsymbol{\beta} \in \mathbb{R}^{K \times (D+1)}$. The extra 0th column of each matrix (i.e. $\boldsymbol{\alpha}_{\cdot,0}$ and $\boldsymbol{\beta}_{\cdot,0}$) hold the bias parameters.

$$\hat{y}_k = \frac{\exp(b_k)}{\sum_{l=1}^{K} \exp(b_l)}$$
$$b_k = \sum_{j=0}^{D} \beta_{kj} z_j$$
$$z_j = \frac{1}{1 + \exp(-a_j)}$$
$$a_j = \sum_{m=0}^{M} \alpha_{jm} x_m$$

The objective function we will use for training the neural network is the average cross entropy over the training dataset $\mathcal{D} = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}$:

$$J(\boldsymbol{\alpha}, \boldsymbol{\beta}) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} y_k^{(i)} \log(\hat{y}_k^{(i)}) \tag{8}$$

In Equation 8, J is a function of the model parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ because $\hat{y}_k^{(i)}$ is implicitly a function of $\mathbf{x}^{(i)}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ since it is the output of the neural network applied to $\mathbf{x}^{(i)}$. Of course, $\hat{y}_k^{(i)}$ and $y_k^{(i)}$ are the kth components of $\hat{\mathbf{y}}^{(i)}$ and $\mathbf{y}^{(i)}$ respectively.

To train, you should optimize this objective function using stochastic gradient descent (SGD), where the gradient of the parameters for each training example is computed via backpropagation. You should **not** shuffle the training points when performing SGD. Note that SGD has a slight impact on the objective function, where we are "summing" over the current point, i:

$$J_{SGD}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = -\sum_{k=1}^{K} y_k^{(i)} \log(\hat{y}_k^{(i)}) \tag{9}$$

Lastly, let's take a look at the Adagrad update that you will be performing. For parameters θ_t, you will first compute an intermediate value $\mathbf{s}_t$, and then use this to compute θ_{t+1}. $\mathbf{s}_t$ will contain the element-wise sums (denoted by $\odot$) of all the element-wise squared gradients. Therefore, $\mathbf{s}_t$ should have the same shape as $\frac{\partial J(\boldsymbol{\theta}_t)}{\partial \boldsymbol{\theta}_t}$. $\mathbf{s}_t$ should be initialized once, before the first epoch, to a zero vector. The update equations for $\mathbf{s}$ and θ are below.

$$\mathbf{s}_{t+1} = \mathbf{s}_t + \frac{\partial J(\boldsymbol{\theta}_t)}{\partial \boldsymbol{\theta}_t} \odot \frac{\partial J(\boldsymbol{\theta}_t)}{\partial \boldsymbol{\theta}_t}. \tag{10}$$

Then, we use $\mathbf{s}_t$ to scale the gradient for the update:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \frac{\eta}{\sqrt{\mathbf{s}_{t+1} + \epsilon}} \odot \frac{\partial J(\boldsymbol{\theta}_t)}{\partial \boldsymbol{\theta}_t}. \tag{11}$$

Here, η is the learning rate, and $\epsilon = \texttt{1e-5}$.

7.1 Initialization

In order to use a deep network, we must first initialize the weights and biases in the network. This is typically done with a random initialization, or initializing the weights from some other training procedure. For this assignment, we will be using two possible initialization:

RANDOM The weights are initialized randomly from a uniform distribution from -0.1 to 0.1. The bias parameters are initialized to zero.

ZERO All weights are initialized to 0.

You must support both of these initialization schemes.

8 Implementation

Write a program `neuralnet.{py|java|cpp|m}` that implements an optical character recognizer using a one hidden layer neural network with sigmoid activations. Your program should learn the parameters of the model on the training data, report the cross-entropy at the end of each epoch on both train and validation data, and at the end of training write out its predictions and error rates on both datasets.

Your implementation must satisfy the following requirements:

- Use a **sigmoid** activation function on the hidden layer and **softmax** on the output layer to ensure it forms a proper probability distribution.
- Number of **hidden units** for the hidden layer should be determined by a command line flag.
- Support two different **initialization strategies**, as described in Section 7.1, selecting between them via a command line flag.
- Use stochastic gradient descent (SGD) to optimize the parameters for one hidden layer neural network. The number of **epochs** will be specified as a command line flag.
- Set the **learning rate** via a command line flag.
- Perform stochastic gradient descent updates on the training data in the order that the data is given in the input file. Although you would typically shuffle training examples when using stochastic gradient descent, in order to autograde the assignment, we ask that you **DO NOT** shuffle trials in this assignment.

- In case there is a tie in the output layer $\hat{\mathbf{y}}$, predict the smallest index to be the label.
- You may assume that the input data will always have the same output label space (i.e. $\{0, 1, \ldots, 3\}$). Other than this, do not hard-code any aspect of the datasets into your code. We will autograde your programs on multiple data sets that include different examples.
- Do *not* use any machine learning libraries. You may use supported linear algebra packages. See Section 8.1 for more details.

Implementing a neural network can be tricky: the parameters are not just a simple vector, but a collection of many parameters; computational efficiency of the model itself becomes essential; the initialization strategy dramatically impacts overall learning quality; other aspects which we will *not* change (e.g. activation function, optimization method) also have a large effect. These *tips* should help you along the way:

- Try to "vectorize" your code as much as possible—this is particularly important for Python. For example, in Python, you want to avoid for-loops and instead rely on `numpy` calls to perform operations such as matrix multiplication, transpose, subtraction, etc. over an entire `numpy` array at once. Why? Because these operations are actually implemented in fast C code, which won't get bogged down the way a high-level scripting language like Python will.
- For low level languages such as Java/C++, the use of primitive arrays and for-loops would not pose any computational efficiency problems—however, it is still helpful to make use of a linear algebra library to cut down on the number of lines of code you will write.
- Implement a finite difference test to check whether your implementation of backpropagation is correctly computing gradients. If you choose to do this, comment out this functionality once your backward pass starts giving correct results and before submitting to Gradescope—since it will otherwise slow down your code.

8.1 Command Line Arguments

The autograder runs and evaluates the output from the files generated, using the following command:

For Python:
```
$ python3 neuralnet.py [args...]
```

For Java:
```
$ javac -cp "./lib/ejml-v0.38-libs/*:./" neuralnet.java
$ java -cp "./lib/ejml-v0.38-libs/*:./" neuralnet [args...]
```

For C++:
```
$ g++ -g -std=c++11 -I./lib neuralnet.cpp; ./a.out [args...]
```

Where above `[args...]` is a placeholder for nine command-line arguments: `<train_input> <validation_input> <train_out> <validation_out> <metrics_out> <num_epoch> <hidden_units> <init_flag> <learning_rate>`. These arguments are described in detail below:

1. `<train_input>`: path to the training input `.csv` file (see Section 6)
2. `<validation_input>`: path to the validation input `.csv` file (see Section 6)
3. `<train_out>`: path to output `.labels` file to which the prediction on the *training* data should be written (see Section 8.2)
4. `<validation_out>`: path to output `.labels` file to which the prediction on the *validation* data should be written (see Section 8.2)

5. <metrics_out>: path of the output .txt file to which metrics such as train and validation error should be written (see Section 8.3)
6. <num_epoch>: integer specifying the number of times backpropogation loops through all of the training data (e.g., if <num_epoch> equals 5, then each training example will be used in backpropogation 5 times).
7. <hidden_units>: positive integer specifying the number of hidden units.
8. <init_flag>: integer taking value 1 or 2 that specifies whether to use RANDOM or ZERO initialization (see Section 7.1 and Section 8)—that is, if init_flag==1 initialize your weights randomly from a uniform distribution over the range [-0.1, 0.1] (i.e. RANDOM), if init_flag==2 initialize all weights to zero (i.e. ZERO). For both settings, **always initialize bias terms to zero**.
9. <learning_rate>: float value specifying the base learning rate for SGD with Adagrad.

As an example, if you implemented your program in Python, the following command line would run your program with 4 hidden units on the small data provided in the handout for 2 epochs using zero initialization and a learning rate of 0.1.

```
$ python3 neuralnet.py smallTrain.csv smallValidation.csv \
smallTrain_out.labels smallValidation_out.labels smallMetrics_out.txt \
2 4 2 0.1
```

8.2 Output: Labels Files

Your program should write two output .labels files containing the predictions of your model on training data (<train_out>) and validation data (<validation_out>). Each should contain the predicted labels for each example printed on a new line. Use \n to create a new line.

Your labels should exactly match those of a reference implementation – this will be checked by the autograder by running your program and evaluating your output file against the reference solution.

Note: You should output your predicted labels using the same *integer* identifiers as the original training data. You should also insert an empty line (again using '\n') at the end of each sequence (as is done in the input data files).

8.3 Output Metrics

Generate a file where you report the following metrics:

cross entropy After each epoch, report mean cross entropy on the training data `crossentropy(train)` and validation data `crossentropy(validation)` (See Equation 8). These two cross-entropy values should be reported at the end of each epoch and prefixed by the epoch number. For example, after the second pass through the training examples, these should be prefixed by `epoch=2`. The total number of train losses you print out should equal `num_epoch`—likewise for the total number of validation losses.

error After the final epoch (i.e. when training has completed fully), report the final training error `error(train)` and validation error `error(validation)`.

A sample output is given below. It contains the train and validation losses for the first 2 epochs and the final error rate when using the command given above.

```
epoch=1 crossentropy(train): 1.37990286268
epoch=1 crossentropy(validation): 1.40340064552
epoch=2 crossentropy(train): 1.37986222014
epoch=2 crossentropy(validation): 1.40252226818
error(train): 0.728
error(validation): 0.74
```

Take care that your output has the exact same format as shown above. There is an equal sign = between the word `epoch` and the epoch number, but no spaces. There should be a single space after the epoch number (e.g. a space after `epoch=1`), and a single space after the colon preceding the metric value (e.g. a space after `epoch=1 likelihood(train):`). Each line should be terminated by a Unix line ending `\n`.

8.4 Tiny Data Set

To help you with this assignment, we have also included a tiny data set, `tinyTrain.csv` and `tinyValidation.csv`, and a reference output file `tinyOutput.txt` for you to use. The tiny dataset is in a format similar to the other datasets, but it only contains two samples with five features. The reference file contains outputs from each layer of one correctly implemented neural network, for both forward and back-propagation steps. We advise you to use this set to help you debug in case your implementation doesn't produce the same results as in the written part.

For your reference, `tinyOutput.txt` is generated from the following command line specifications:

```
$ python3 neuralnet.py tinyTrain.csv tinyValidation.csv \
tinyTrain_out.labels tinyValidation_out.labels tinyMetrics_out.txt \
1 4 2 0.1
```

The specific output file names are not important, but be sure to keep the other arguments exactly as they are shown above.

9 Gradescope Submission

You should submit your `neuralnet.{py|java|cpp}` to Gradescope. Please do not use any other file name for your implementation. This will cause problems for the autograder to correctly detect and run your code.

Some additional tips: Make sure to read the autograder output carefully. The autograder for Gradescope prints out some additional information about the tests that it ran. For this programming assignment we've specially designed some buggy implementations that you might implement and will try our best to detect those and give you some more useful feedback in Gradescope's autograder. Make wise use of autograder's output for debugging your code.

Note: For this assignment, you may make up to 10 submissions to Gradescope before the deadline, but only your last submission will be graded.

A Implementation Details for Neural Networks

This section provides a variety of suggestions for how to efficiently and succinctly implement a neural network and backpropagation.

A.1 SGD for Neural Networks

Consider the neural network described in Section 8 applied to the ith training example $(\mathbf{x}, \mathbf{y})$ where $\mathbf{y}$ is a one-hot encoding of the true label. Our neural network outputs $\hat{\mathbf{y}} = h_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbf{x})$, where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the parameters of the first and second layers respectively and $h_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\cdot)$ is a one-hidden layer neural network with a sigmoid activation and softmax output. The loss function is negative cross-entropy $J = \ell(\hat{\mathbf{y}}, \mathbf{y}) = -\mathbf{y}^T \log(\hat{\mathbf{y}})$. $J = J_{\mathbf{x},\mathbf{y}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is actually a function of our training example $(\mathbf{x}, \mathbf{y})$, and our model parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}$ though we write just J for brevity.

In order to train our neural network, we are going to apply stochastic gradient descent. Because we want the behavior of your program to be deterministic for testing on Gradescope, we make a few simplifications: (1) you should *not* shuffle your data and (2) you will use a fixed learning rate. In the real world, you would *not* make these simplifications.

SGD proceeds as follows, where E is the number of epochs and γ is the learning rate.

Algorithm 1 Stochastic Gradient Descent (SGD) without Shuffle

procedure SGD(Training data $\mathcal{D}$, test data $\mathcal{D}_t$)
  Initialize parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}$ ▷ Use either RANDOM or ZERO from Section 7.1
  for $e \in \{1, 2, \ldots, E\}$ **do** ▷ For each epoch
    for $(\mathbf{x}, \mathbf{y}) \in \mathcal{D}$ **do** ▷ For each training example (No shuffling)
      Compute neural network layers:
      $\mathbf{o} = \texttt{object}(\mathbf{x}, \mathbf{a}, \mathbf{b}, \mathbf{z}, \hat{\mathbf{y}}, J) = \text{NNFORWARD}(\mathbf{x}, \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\beta})$
      Compute gradients via backprop:
      $$\left.\begin{aligned} \mathbf{g}_{\boldsymbol{\alpha}} &= \frac{\partial J}{\partial \boldsymbol{\alpha}} \\ \mathbf{g}_{\boldsymbol{\beta}} &= \frac{\partial J}{\partial \boldsymbol{\beta}} \end{aligned}\right\} = \text{NNBACKWARD}(\mathbf{x}, \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{o})$$
      Update parameters with Adagrad updates $\mathbf{g}'_{\boldsymbol{\alpha}}, \mathbf{g}'_{\boldsymbol{\beta}}$:
      $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} - \gamma \mathbf{g}'_{\boldsymbol{\alpha}}$
      $\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} - \gamma \mathbf{g}'_{\boldsymbol{\beta}}$
    Evaluate training mean cross-entropy $J_{\mathcal{D}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$
    Evaluate test mean cross-entropy $J_{\mathcal{D}_t}(\boldsymbol{\alpha}, \boldsymbol{\beta})$
  return parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}$

At test time, we output the most likely prediction for each example:

Algorithm 2 Prediction at Test Time

procedure PREDICT(Unlabeled train or test dataset $\mathcal{D}'$, Parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}$)
  for $\mathbf{x} \in \mathcal{D}'$ **do**
    Compute neural network prediction $\hat{\mathbf{y}} = h(\mathbf{x})$
    Predict the label with highest probability $l = \text{argmax}_k \, \hat{y}_k$

The gradients we need above are themselves matrices of partial derivatives. Let M be the number of input

features, D the number of hidden units, and K the number of outputs.

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_{10} & \alpha_{11} & \dots & \alpha_{1M} \\ \alpha_{20} & \alpha_{21} & \dots & \alpha_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{D0} & \alpha_{D1} & \dots & \alpha_{DM} \end{bmatrix} \qquad \mathbf{g}_{\boldsymbol{\alpha}} = \frac{\partial J}{\partial \boldsymbol{\alpha}} = \begin{bmatrix} \frac{dJ}{d\alpha_{10}} & \frac{dJ}{d\alpha_{11}} & \dots & \frac{dJ}{d\alpha_{1M}} \\ \frac{dJ}{d\alpha_{20}} & \frac{dJ}{d\alpha_{21}} & \dots & \frac{dJ}{d\alpha_{2M}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{dJ}{d\alpha_{D0}} & \frac{dJ}{d\alpha_{D1}} & \dots & \frac{dJ}{d\alpha_{DM}} \end{bmatrix} \tag{12}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{10} & \beta_{11} & \dots & \beta_{1D} \\ \beta_{20} & \beta_{21} & \dots & \beta_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{K0} & \beta_{K1} & \dots & \beta_{KD} \end{bmatrix} \qquad \mathbf{g}_{\boldsymbol{\beta}} = \frac{\partial J}{\partial \boldsymbol{\beta}} = \begin{bmatrix} \frac{dJ}{d\beta_{10}} & \frac{dJ}{d\beta_{11}} & \dots & \frac{dJ}{d\beta_{1D}} \\ \frac{dJ}{d\beta_{20}} & \frac{dJ}{d\beta_{21}} & \dots & \frac{dJ}{d\beta_{2D}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{dJ}{d\beta_{K0}} & \frac{dJ}{d\beta_{K1}} & \dots & \frac{dJ}{d\beta_{KD}} \end{bmatrix} \tag{13}$$

Observe that we have (in a rather *tricky* fashion) defined the matrices such that both $\boldsymbol{\alpha}$ and $\mathbf{g}_{\boldsymbol{\alpha}}$ are $D \times (M+1)$ matrices. Likewise, $\boldsymbol{\beta}$ and $\mathbf{g}_{\boldsymbol{\beta}}$ are $K \times (D+1)$ matrices. The $+1$ comes from the extra columns $\boldsymbol{\alpha}_{\cdot,0}$ and $\boldsymbol{\beta}_{\cdot,0}$ which are the bias parameters for the first and second layer respectively. We will always assume $x_0 = 1$ and $z_0 = 1$. This should greatly simplify your implementation as you will see in Section A.3.

A.2 Recursive Derivation of Backpropagation

In class, we described a very general approach to differentiating arbitrary functions: backpropagation. One way to understand *how* we go about deriving the backpropagation algorithm is to consider the natural consequence of recursive application of the chain rule.

In practice, the partial derivatives that we need for learning are $\frac{dJ}{d\alpha_{ij}}$ and $\frac{dJ}{d\beta_{kj}}$.

A.2.1 Symbolic Differentiation

Note In this section, we motivate backpropagation via a strawman: that is, we will work through the *wrong* approach first (i.e. symbolic differentiation) in order to see why we want a more efficient method (i.e. backpropagation). Do **not** use this symbolic differentiation in your code.

Suppose we wanted to find $\frac{dJ}{d\alpha_{ij}}$ using the method we know from high school calculus. That is, we will analytically solve for an equation representing that quantity.

1. Considering the computational graph for the neural network, we observe that α_{ij} has exactly one child $a_i = \sum_{m=0}^{M} \alpha_{im} x_m$. That is a_i is the *first and only* intermediate quantity that uses α_{ij}. Applying the chain rule, we obtain
$$\frac{dJ}{d\alpha_{ij}} = \frac{dJ}{da_i}\frac{da_i}{d\alpha_{ij}} = \frac{dJ}{da_i}x_j$$

2. So far so good, now we just need to compute $\frac{dJ}{da_i}$. Not a problem! We can just apply the chain rule again. a_i just has exactly one child as well, namely $z_i = \sigma(a_i)$. The chain rule gives us that $\frac{dJ}{da_i} = \frac{dJ}{dz_i}\frac{dz_i}{da_i} = \frac{dJ}{dz_i}z_i(1-z_i)$. Substituting back into the equation above we find that
$$\frac{dJ}{d\alpha_{ij}} = \frac{dJ}{dz_i}(z_i(1-z_i))x_j$$

3. How do we get $\frac{dJ}{dz_i}$? You guessed it: apply the chain rule yet again. This time, however, there are *multiple* children of z_i in the computation graph; they are $b_1, b_2, \ldots b_K$. Applying the chain rule gives us that $\frac{dJ}{dz_i} = \sum_{k=1}^{K} \frac{dJ}{db_k} \frac{\partial b_k}{\partial z_i} = \sum_{k=1}^{K} \frac{dJ}{db_k} \beta_{ki}$. Substituting back into the equation above gives:

$$\frac{dJ}{d\alpha_{ij}} = \sum_{k=1}^{K} \frac{dJ}{db_k} \beta_{ki}(z_i(1 - z_i))x_j$$

4. Next we need $\frac{dJ}{db_k}$, which we again obtain via the chain rule: $\frac{dJ}{db_k} = \sum_{l=1}^{K} \frac{dJ}{d\hat{y}_l} \frac{\partial \hat{y}_l}{\partial b_k} = \sum_{l=1}^{K} \frac{dJ}{d\hat{y}_l} \hat{y}_l(\mathbb{I}[k = l] - \hat{y}_k)$. Substituting back in above gives:

$$\frac{dJ}{d\alpha_{ij}} = \sum_{k=1}^{K} \sum_{l=1}^{K} \frac{dJ}{d\hat{y}_l} \hat{y}_l(\mathbb{I}[k = l] - \hat{y}_k)\beta_{ki}(z_i(1 - z_i))x_j$$

5. Finally, we know that $\frac{dJ}{d\hat{y}_l} = -\frac{y_l}{\hat{y}_l}$ which we can again substitute back in to obtain our final result:

$$\frac{dJ}{d\alpha_{ij}} = \sum_{k=1}^{K} \sum_{l=1}^{K} -\frac{y_l}{\hat{y}_l} \hat{y}_l(\mathbb{I}[k = l] - \hat{y}_k)\beta_{ki}(z_i(1 - z_i))x_j$$

   Although we have successfully derived the partial derivative w.r.t. α_{ij}, the result is far from satisfying. It is overly complicated and requires deeply nested for-loops to compute.

The above is an example of **symbolic differentiation**. That is, at the end we get an equation representing the partial derivative w.r.t. α_{ij}. At this point, you should be saying to yourself: What a mess! Isn't there a better way? Indeed there is and its called backpropagation. The algorithm works just like the above symbolic differentiation except that we *never* subsitute the partial derivative from the previous step back in. Instead, we work "backwards" through the steps above computing partial derivatives in a top-down fashion.

A.3 Matrix / Vector Operations for Neural Networks

Some programming languages are fast and some are slow. Below is a simple benchmark to show this concretely. The task is to compute a dot-product $\mathbf{a}^T\mathbf{b}$ between two vectors $\mathbf{a} \in \mathbb{R}^{500}$ and $\mathbf{b} \in \mathbb{R}^{500}$ one thousand times. Table 1 shows the time taken for several combinations of programming language and data structure.

language	data structure	time (ms)
Python	list	200.99
Python	`numpy` array	1.01
Java	`float[]`	4.00
C++	`vector<float>`	0.81

Table 1: Computation time required for dot-product in various languages.

Notice that Java[1] and C++ with standard data structures are quite efficient. By contrast, Python differs dramatically depending on which data structure you use: with a standard list object
(e.g. `a = [float(i) for x in range(500)]`) the computation time is an appallingly slow 200+

[1] Java would approach the speed of C++ if we had given the just-in-time (JIT) compiler inside the JVM time to "warm-up".

milliseconds. Simply by switching to a numpy array (e.g. `a = np.arange(500, dtype=float)`) we obtain a 200x speedup. This is because a numpy array is actually carrying out the dot-product computation in pure C, which is just as fast as our C++ benchmark, modulo some Python overhead.

Thus, for this assignment, Java and C++ programmers could easily implement the entire neural network using standard data structures and some for-loops. However, Python programmers would find that their code is simply too slow if they tried to do the same. As such, particularly for Python users, one must convert all the deeply nested for-loops into efficient "vectorized" math via `numpy`. Doing so will ensure efficient code. Java and C++ programmers can also benefit from linear algebra packages since it can cut down on the total number of lines of code you need to write.

A.4 Procedural Method of Implementation

Perhaps the simplest way to implement a 1-hidden-layer neural network is procedurally. Note that this approach has some drawbacks that we'll discuss below (Section A.4.1).

The procedural method: one function computes the outputs of the neural network and all intermediate quantities $\mathbf{o} = \text{NNFORWARD}(\mathbf{x}, \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \texttt{object}(\mathbf{x}, \mathbf{a}, \mathbf{b}, \mathbf{z}, \hat{\mathbf{y}}, J)$, where the object is just some struct. Then another function computes the gradients of our parameters $\mathbf{g}_{\boldsymbol{\alpha}}, \mathbf{g}_{\boldsymbol{\beta}} = \text{NNBACKWARD}(\mathbf{x}, \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{o})$, where $\mathbf{o}$ is a data structure that stores all the forward computation.

One must be careful to ensure that functions are vectorized. For example, your Sigmoid function should be able to take a vector input and return a vector output with the Sigmoid function applied to all of its elements. All of these operations should avoid for-loops when working in a high-level language like Python. We can compute the softmax function in a similar vectorized manner.

A.4.1 Drawbacks to Procedural Method

As noted in Section A.6, it is possible to use a finite difference method to check that the backpropagation algorithm is correctly computing the gradient of its corresponding forward computation. We *strongly* encourage you to do this.

There is a big problem however: what if your finite difference check informs you that the gradient is *not* being computed correctly. How will you know *which* part of your NNFORWARD() or NNBACKWARD() functions has a bug? There are two possible solutions here:

1. As usual, you can (and should) work through a tiny example dataset on paper. Compute each intermediate quantity and each gradient. Check that your code reproduces each number. The one that does not should indicate where to find the bug.
2. Replace your procedural implementation with a module-based one (as described in Section A.5) and then run a finite-difference check on *each* layer of the model individually. The finite-difference check that fails should indicate where to find the bug.

Of course, rather than waiting until you have a bug in your procedural implementation, you could jump straight to the module-based version—though it increases the complexity slightly (i.e. more lines of code) it *might* save you some time in the long run.

A.5 Module-based Method of Implementation

Module-based automatic differentiation (AD) is a technique that has long been used to develop libraries for deep learning. Dynamic neural network packages are those that allow a specification of the computation

graph dynamically at runtime, such as Torch[2], PyTorch[3], and DyNet[4]—these all employ module-based AD in the sense that we will describe here.[5]

The key idea behind module-based AD is to componentize the computation of the neural-network into layers. Each layer can be thought of as consolidating numerous nodes in the computation graph (a subset of them) into one *vector-valued* node. Such a vector-valued node should be capable of the following and we call each one a **module**:

1. Forward computation of output $\mathbf{b} = [b_1, \ldots, b_B]$ given input $\mathbf{a} = [a_1, \ldots, a_A]$ via some differentiable function f. That is $\mathbf{b} = f(\mathbf{a})$.
2. Backward computation of the gradient of the input $\mathbf{g_a} = \frac{\partial J}{\partial \mathbf{a}} = [\frac{dJ}{da_1}, \ldots, \frac{dJ}{da_A}]$ given the gradient of output $\mathbf{g_b} = \frac{\partial J}{\partial \mathbf{b}} = [\frac{dJ}{db_1}, \ldots, \frac{dJ}{db_B}]$, where J is the final real-valued output of the entire computation graph. This is done via the chain rule $\frac{dJ}{da_i} = \sum_{j=1}^{M} \frac{dJ}{db_j} \frac{\partial b_j}{\partial a_i}$ for all $i \in \{1, \ldots, A\}$.

A.5.1 Module Definitions

The modules we would define for our neural network would correspond to a Linear layer, a Sigmoid layer, a Softmax layer, and a Cross-Entropy layer. Each module defines a forward function $\mathbf{b} = *\text{FORWARD}(\mathbf{a})$, and a backward function $\mathbf{g_a} = *\text{BACKWARD}(\mathbf{a}, \mathbf{b}, \mathbf{g_b})$ method. These methods accept parameters if appropriate. You'll want to pay close attention to the dimensions that you pass into and return from your modules.

Linear Module The linear layer has two inputs: a vector $\mathbf{a}$ and parameters $\omega \in \mathbb{R}^{B \times A}$. The output $\mathbf{b}$ is not used by LINEARBACKWARD, but we pass it in for consistency of form.

procedure LINEARFORWARD($\mathbf{a}$, $\boldsymbol{\alpha}$)
    Compute $\mathbf{b}$
    return $\mathbf{b}$
procedure LINEARBACKWARD($\mathbf{a}$, α, $\mathbf{b}$, $\mathbf{g_b}$)
    Compute $\mathbf{g}_{\boldsymbol{\alpha}}$
    Compute $\mathbf{g_a}$
    return $\mathbf{g}_{\boldsymbol{\alpha}}, \mathbf{g_a}$

It's also quite common to combine the Cross-Entropy and Softmax layers into one. The reason for this is the cancelation of numerous terms that result from the zeros in the cross-entropy backward calculation. (Said trick is *not* required to obtain a sufficiently fast implementation for Gradescope.)

A.5.2 Module-based AD for Neural Network

Using these modules, we can re-define our functions NNFORWARD and NNBACKWARD as follows.

[2] http://torch.ch/
[3] http://pytorch.org/
[4] https://dynet.readthedocs.io
[5] Static neural network packages are those that require a static specification of a computation graph which is subsequently compiled into code. Examples include Theano, Tensorflow, and CNTK. These libraries are also module-based but the particular form of implementation is different from the dynamic method we recommend here.

Algorithm 3 Forward Computation

procedure NNFORWARD(Training example $(\mathbf{x}, \mathbf{y})$, Parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$)
  $\mathbf{a} = \text{LINEARFORWARD}(\mathbf{x}, \boldsymbol{\alpha})$
  $\mathbf{z} = \text{SIGMOIDFORWARD}(\mathbf{a})$
  $\mathbf{b} = \text{LINEARFORWARD}(\mathbf{z}, \boldsymbol{\beta})$
  $\hat{\mathbf{y}} = \text{SOFTMAXFORWARD}(\mathbf{b})$
  $J = \text{CROSSENTROPYFORWARD}(\mathbf{y}, \hat{\mathbf{y}})$
  $\mathbf{o} = \texttt{object}(\mathbf{x}, \mathbf{a}, \mathbf{z}, \mathbf{b}, \hat{\mathbf{y}}, J)$
  return intermediate quantities $\mathbf{o}$

Algorithm 4 Backpropagation

procedure NNBACKWARD(Training example $(\mathbf{x}, \mathbf{y})$, Parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, Intermediates $\mathbf{o}$)
  Place intermediate quantities $\mathbf{x}, \mathbf{a}, \mathbf{z}, \mathbf{b}, \hat{\mathbf{y}}, J$ in $\mathbf{o}$ in scope
  $g_J = \frac{\partial J}{\partial J} = 1$ ▷ Base case
  $\mathbf{g}_{\hat{\mathbf{y}}} = \text{CROSSENTROPYBACKWARD}(\mathbf{y}, \hat{\mathbf{y}}, g_J)$
  $\mathbf{g}_{\mathbf{b}} = \text{SOFTMAXBACKWARD}(\mathbf{b}, \hat{\mathbf{y}}, \mathbf{g}_{\hat{\mathbf{y}}})$
  $\mathbf{g}_{\boldsymbol{\beta}}, \mathbf{g}_{\mathbf{z}} = \text{LINEARBACKWARD}(\mathbf{z}, \boldsymbol{\beta}, \mathbf{g}_{\mathbf{b}})$
  $\mathbf{g}_{\mathbf{a}} = \text{SIGMOIDBACKWARD}(\mathbf{a}, \mathbf{z}, \mathbf{g}_{\mathbf{z}})$
  $\mathbf{g}_{\boldsymbol{\alpha}}, \mathbf{g}_{\mathbf{x}} = \text{LINEARBACKWARD}(\mathbf{x}, \boldsymbol{\alpha}, \mathbf{g}_{\mathbf{a}})$ ▷ We discard $\mathbf{g}_{\mathbf{x}}$
  return parameter gradients $\mathbf{g}_{\boldsymbol{\alpha}}, \mathbf{g}_{\boldsymbol{\beta}}$

Here's the big takeaway: we can actually view these two functions as themselves defining another module! It is a 1-hidden layer neural network module. That is, the cross-entropy of the neural network for a single training example is *itself* a differentiable function and we know how to compute the gradients of its inputs, given the gradients of its outputs.

A.6 Testing Backprop with Numerical Differentiation

Numerical differentiation provides a convenient method for testing gradients computed by backpropagation. The *centered* finite difference approximation is:

$$\frac{\partial}{\partial \theta_i} J(\boldsymbol{\theta}) \approx \frac{(J(\boldsymbol{\theta} + \epsilon \cdot \boldsymbol{d}_i) - J(\boldsymbol{\theta} - \epsilon \cdot \boldsymbol{d}_i))}{2\epsilon} \tag{14}$$

where $\boldsymbol{d}_i$ is a 1-hot vector consisting of all zeros except for the ith entry of $\boldsymbol{d}_i$, which has value 1. Unfortunately, in practice, it suffers from issues of floating point precision. Therefore, it is typically only appropriate to use this on small examples with an appropriately chosen ϵ.

In order to apply this technique to test the gradients of your backpropagation implementation, you will need to ensure that your code is appropriately factored. Any of the modules including NNFORWARD and NNBACKWARD could be tested in this way.

For example, you could use two functions: `forward(x,y,theta)` computes the cross-entropy for a training example. `backprop(x,y,theta)` computes the gradient of the cross-entropy for a training example via backpropagation. Finally, `finite_diff` as defined below approximates the gradient by the centered finited difference method. The following pseudocode provides an overview of the entire procedure.

```
def finite_diff(x, y, theta):
    epsilon = 1e-5
```

```
    grad = zero_vector(theta.length)
    for m in [1, ..., theta.length]:
        d = zero_vector(theta.length)
        d[m] = 1
        v = forward(x, y, theta + epsilon * d)
        v -= forward(x, y, theta - epsilon * d)
        v /= 2*epsilon
        grad[m] = v

# Compute the gradient by backpropagation
grad_bp = backprop(x, y, theta)
# Approximate the gradient by the centered finite difference method
grad_fd = finite_diff(x, y, theta)

# Check that the gradients are (nearly) the same
diff = grad_bp - grad_fd  # element-wise difference of two vectors
print l2_norm(diff) # this value should be small (e.g. < 1e-7)
```

A.6.1 Limitations

This does *not* catch all bugs—the only thing it tells you is whether your backpropagation implementation is correctly computing the gradient for the forward computation. Suppose your *forward* computation is incorrect, e.g. you are always computing the cross-entropy of the wrong label. If your *backpropagation* is also using the same wrong label, then the check above will not expose the bug. Thus, you always want to *separately* test that your forward implementation is correct.

A.6.2 Finite Difference Checking of Modules

Note that the above would test the gradient for the entire end-to-end computation carried output by the neural network. However, if you implement a module-based automatic differentiation method (as in Section A.5), then you can test each individual component for correctness. The only difference is that you need to run the finite-difference check for each of the output values (i.e. a double for-loop).

A.7 Why AdaGrad?

So far, the loss functions we have discussed make it quite easy to find a global optimum. In these cases, a larger step size makes it quick and easy to reach convergence. This isn't always the case with neural networks. Nonconvex loss functions are much harder to optimize over. Here we want step sizes that will adapt to the domain in which they optimize. We want to take larger steps where possible, but smaller steps where we are in danger of overshooting the optima. Adagrad implicitly changes the step size based on the shape of the function inferred from the gradients. Interested? Read all about it here.