

HOMWORK 5: NEURAL NETWORKS

10-301/10-601 Introduction to Machine Learning (Spring 2021)

<https://www.cs.cmu.edu/~10601/>

DUE: Monday, March 29, 2021 11:59 PM

Summary In this assignment, you will build a handwriting 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 handwritten letter classification.

START HERE: Instructions

- **Collaboration Policy:** Please read the collaboration policy here: <https://www.cs.cmu.edu/~10601/>
- **Late Submission Policy:** See the late submission policy here: <https://www.cs.cmu.edu/~10601/>
- **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, we will be using Gradescope (<https://gradescope.com/>). Please use the provided template. Submissions must be written in LaTeX. Regrade requests can be made, however this gives the staff the opportunity to regrade your entire paper, meaning if additional mistakes are found then points will be deducted. Each derivation/proof should be completed in the boxes provided. For short answer questions you **should not** include your work in your solution. If you include your work in your solutions, 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 3.6.9, OpenJDK 11.0.5, g++ 7.4.0) and versions of permitted libraries (e.g. `numpy` 1.17.0 and `scipy` 1.4.1) match those used on Gradescope. You have **a total of 10 Gradescope programming submissions**. Use them wisely. In order to not waste Gradescope 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 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 0.38. When compiling and running your code, we will add the additional command line argument `-cp "linalg_lib/ejml-v0.38-libs/*:linalg_lib/nd4j-v1.0.0-beta7-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-beta7. When compiling and running your code, we will add the additional command line argument `-cp "linalg_lib/ejml-v0.38-libs/*:linalg_lib/nd4j-v1.0.0-beta7-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 3.3.7. The command line arguments above demonstrate how we will call your 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 Piazza Resources page 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://deeplearning4j.org/docs/latest/nd4j-overview>

^c<http://eigen.tuxfamily.org/>

Written Questions (44 points)

1 Convolutional Neural Networks

In this problem, consider only the convolutional layer of a standard implementation of a CNN as described in Lecture 12.

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 6×4 and filter size 2×2 , using no padding and a stride of 2, what would be the resulting output size? Write your answer in terms of height \times width.

Answer
3×2

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 black and white 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. Without

sharing, 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) Parameter sharing is not usually used for fully-connected layers, but is usually used for convolutional layers. In a sentence, describe a reason why parameter sharing is a good idea for a convolutional layer, besides reduction in problem complexity. Hint: think about applications of CNNs.

Answer

There are many potential reasons that makes parameter sharing a good idea. If we focus on the applications of CNNs, then since CNN usually deals with image recognition, for example the car image question from the previous assignment, thus we are trying to learn and recognize a specific filter on different positions on the image. Thus parameter sharing is more common for CNNs.

2 Backpropagation

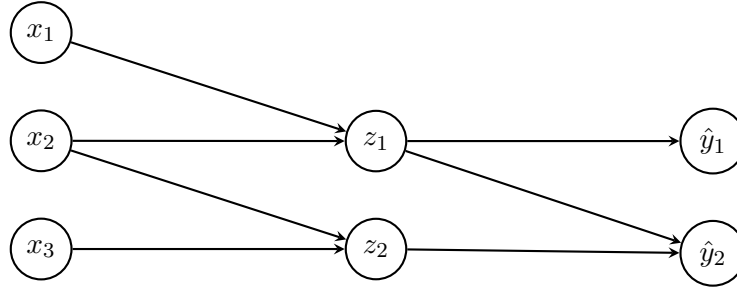


Figure 1: A Directed Acyclic Graph (DAG)

Consider this above Directed Acyclic Graph (DAG). Notice that it looks different than the fully-connected Neural Networks that we have seen before. Recall from lecture that you can perform back propagation on any DAG. We will work through back propagation on this graph.

Let (\mathbf{x}, \mathbf{y}) be the training example that is considered, where $\mathbf{x} = [x_1 \ x_2 \ x_3]^T$ and $\mathbf{y} = [y_1 \ y_2]^T$. All the other nodes in the graph are defined as:

$$\begin{aligned}
 z_1 &= \text{ReLU}(w_{1,1}x_1 + w_{2,1}x_2) \\
 z_2 &= w_{2,2}x_2^2 + w_{3,2}x_3 + b_2 \\
 \hat{y}_1 &= \sigma(m_{1,1}z_1^3 + c_1) \\
 \hat{y}_2 &= m_{1,2} \sin(z_1) + m_{2,2} \cos(z_2)
 \end{aligned}$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function, and $\text{ReLU}(x) = \max(0, x)$. Let θ be the set of all parameters to be learned in this graph. We have that

$$\theta = \{w_{1,1}, w_{2,1}, w_{2,2}, w_{3,2}, m_{1,1}, m_{1,2}, m_{2,2}, b_2, c_1\}$$

For every set of input $\mathbf{x} = (x_1, x_2, x_3)$, we will define objective of the problem as minimizing the loss function

$$J(\theta) = \log((y_1 - \hat{y}_1)^2) + \log((y_2 - \hat{y}_2)^2)$$

Assume that you have already gone through the forward pass with inputs $\mathbf{x} = (x_1, x_2, x_3)$ and stored all the relevant values. In the following questions, you will derive the backpropagation algorithm applied to the above DAG.

- (a) (1 point) First, we will derive the gradients with respect to the outputs. What are the expressions for $\frac{\partial J}{\partial \hat{y}_1}$? Write your solution in terms of \hat{y}_1 .

Answer

$$\frac{-2}{y_1 - \hat{y}_1}$$

- (b) Now, we will derive the gradients associated with the last layer, ie nodes y_1, y_2 . Note that for the full backpropagation algorithm, you would need to calculate the gradients of the loss function with respect to every parameter ($m_{1,1}, m_{1,2}, m_{2,2}, c_1$) as well as every input of the layer (z_1, z_2), but we are not asking for all of them in this part. For all of the questions in this part, you should use Chain Rule, and write your solution in terms of values from the forward pass, or **gradients with respect to the outputs of this layer**, $\frac{\partial J}{\partial \hat{y}_1}, \frac{\partial J}{\partial \hat{y}_2}$, because you have already calculated these values. In addition, use the sigmoid function $\sigma(x)$ in your answer instead of its explicit form.

- (a) (1 point) What is the expression for $\frac{\partial J}{\partial z_1}$?

Answer

$$-\frac{2}{y_1 - \hat{y}_1} \hat{y}_1 (1 - \hat{y}_1) 3m_{1,1} z_1^2 - \frac{2}{y_2 - \hat{y}_2} m_{1,2} \cos(z_1)$$

- (b) (1 point) What is the expression for $\frac{\partial J}{\partial z_2}$?

Answer

$$\frac{2}{y_2 - \hat{y}_2} m_{2,2} \sin(z_2)$$

- (c) (1 point) What is the expression for $\frac{\partial J}{\partial m_{1,1}}$?

Answer

$$\frac{-2}{y_1 - \hat{y}_1} \hat{y}_1 (1 - \hat{y}_1) z_1^3$$

- (d) (1 point) What is the expression for $\frac{\partial J}{\partial m_{1,2}}$?

Answer

$$\frac{-2}{y_2 - \hat{y}_2} \sin(z_1)$$

- (e) (1 point) What is the expression for $\frac{\partial J}{\partial c_1}$?

Answer

$$\frac{-2}{y_1 - \hat{y}_1} \hat{y}_1 (1 - \hat{y}_1)$$

- (f) (1 point) Lastly, we will derive the gradients associated with the second layer, ie nodes z_1, z_2 . Note that for the full backpropagation algorithm, you need to calculate the gradients of the loss function with respect to every parameter (every $w_{i,j}, b_2$). However, we do not need to calculate the gradients with respect to the inputs of this layer (x_1, x_2, x_3), because they are fixed inputs of the model. For all of the questions in this part, you should use Chain Rule, and write your solution in terms of values from the forward pass or **gradients with respect to the outputs of this layer**, $\frac{\partial J}{\partial z_1}, \frac{\partial J}{\partial z_2}$, because you have already calculated these values.

- (a) (1 point) What is the expression for $\frac{\partial J}{\partial w_{2,2}}$?

Answer

$$\frac{\partial J}{\partial \hat{y}_2} \frac{\partial \hat{y}_2}{\partial z_2} \frac{\partial z_2}{\partial w_{2,2}} = \frac{2}{y_2 - \hat{y}_2} m_{2,2} \sin(z_2) x_2^2$$

- (b) (1 point) What is the expression for $\frac{\partial J}{\partial b_2}$?

Answer

$$\frac{\partial J}{\partial \hat{y}_2} \frac{\partial \hat{y}_2}{\partial z_2} \frac{\partial z_2}{\partial b_2} = \frac{2}{y_2 - \hat{y}_2} m_{2,2} \sin(z_2)$$

- (c) (2 points) Recall that $\text{ReLU}(x) = \max(x, 0)$. The ReLU function is not differentiable at $x = 0$, but for backpropagation, we define its derivative as

$$\text{ReLU}'(x) = \begin{cases} 0, & x < 0 \\ 1, & \text{otherwise} \end{cases}$$

Now, what is the expression for $\frac{\partial J}{\partial w_{1,1}}$? Explicitly write out the cases.

Answer

$$\begin{aligned} & \left(\frac{\partial J}{\partial \hat{y}_1} \frac{\partial \hat{y}_1}{\partial z_1} + \frac{\partial J}{\partial \hat{y}_2} \frac{\partial \hat{y}_2}{\partial z_1} \right) \frac{\partial z_1}{\partial w_{1,1}} \\ &= \begin{cases} 0, & w_{1,1}x_1 + w_{2,1}x_2 < 0 \\ -\left[\frac{2}{y_1 - \hat{y}_1} \hat{y}_1 (1 - \hat{y}_1) 3m_{1,1}z_1^2 + \frac{2}{y_2 - \hat{y}_2} m_{1,2} \cos(z_1) \right] x_1, & w_{1,1}x_1 + w_{2,1}x_2 \geq 0 \end{cases} \end{aligned}$$

3 Backpropagation Through Time

Consider the following one-layer many-to-many Recurrent neural network (RNN),

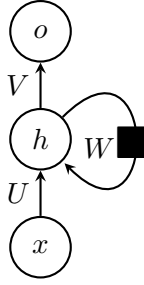


Figure 2: A one hidden layer many-to-many Recurrent Neural Network

Recall from lecture, this is equivalent to

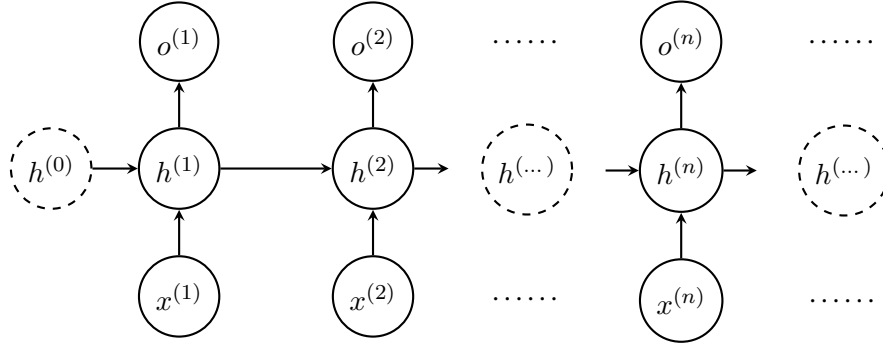


Figure 3: The unfolding of the above RNN in time of the computation involved in its forward computation

Assume $h^{(0)}$ is given. To keep things simple, we consider an RNN with identity activation function and no bias terms. We also assume that $W = I$, the identity matrix. For each timestep $n \geq 1$, the hidden state and output can be expressed as

$$\begin{aligned} h^{(n)} &= f(h^{(n-1)}, x^{(n)}, U) = Ux^{(n)} + h^{(n-1)} \\ o^{(n)} &= g(h^{(n)}, V) = Vh^{(n)} \end{aligned}$$

where $x^{(n)} \in \mathbb{R}^a$, $h^{(n)} \in \mathbb{R}^b$, $o^{(n)} \in \mathbb{R}^c$, $U \in \mathbb{R}^{b \times a}$, and $V \in \mathbb{R}^{c \times b}$. Assume finite time N . The loss can then be expressed as

$$J(\theta) = \sum_{t=1}^N l(y^{(t)}, o^{(t)})$$

where $l(y^{(t)}, o^{(t)})$ is the loss at time step t .

For the following questions, express your answer in terms of $x^{(n)}, h^{(n)}, o^{(n)}, U, V$ unless specified otherwise. Some rules of matrix calculus may be helpful: for $J \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times m}$,

$$\frac{\partial J}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial J}{\partial x_1} \\ \frac{\partial J}{\partial x_2} \\ \vdots \\ \frac{\partial J}{\partial x_n} \end{bmatrix} \quad \frac{\partial J}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial J}{\partial x_{11}} & \frac{\partial J}{\partial x_{12}} & \cdots & \frac{\partial J}{\partial x_{1m}} \\ \frac{\partial J}{\partial x_{21}} & \frac{\partial J}{\partial x_{22}} & \cdots & \frac{\partial J}{\partial x_{2m}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial J}{\partial x_{n1}} & \frac{\partial J}{\partial x_{n2}} & \cdots & \frac{\partial J}{\partial x_{nm}} \end{bmatrix}$$

1. Assume $N = 1$. For the following subparts (a)-(d), you may give your answer in terms of $\frac{\partial l}{\partial o^{(1)}}$.

(a) (1 point) What is $\frac{\partial J}{\partial h^{(1)}}$?

Answer

$$\frac{\partial J}{\partial l} \frac{\partial l}{\partial o^{(1)}} \frac{\partial o^{(1)}}{\partial h^{(1)}} = V^T \frac{\partial l}{\partial o^{(1)}}$$

(b) (1 point) What is $\frac{\partial J}{\partial V}$?

Answer

$$\frac{\partial J}{\partial l} \frac{\partial l}{\partial o^{(1)}} \frac{\partial o^{(1)}}{\partial V} = \frac{\partial l}{\partial o^{(1)}} (h^{(1)})^T$$

(c) (1 point) What is $\frac{\partial J}{\partial U}$?

Answer

$$\frac{\partial J}{\partial l} \frac{\partial l}{\partial o^{(1)}} \frac{\partial o^{(1)}}{\partial h^{(1)}} \frac{\partial h^{(1)}}{\partial U} = V^T \frac{\partial l}{\partial o^{(1)}} (x^{(1)})^T$$

2. Assume $N = 2$. For the following subparts (a)-(d), you may give your answer in terms of $\frac{\partial l}{\partial o^{(1)}}$ and $\frac{\partial l}{\partial o^{(2)}}$.

(a) (1 point) What is $\frac{\partial J}{\partial h^{(1)}}$?

Answer

$$V^T \frac{\partial l}{\partial o^{(1)}} + V^T \frac{\partial l}{\partial o^{(2)}}$$

(b) (1 point) What is $\frac{\partial J}{\partial V}$?

Answer

$$\frac{\partial l}{\partial o^{(1)}} (h^{(1)})^T + \frac{\partial l}{\partial o^{(2)}} (h^{(2)})^T$$

(c) (1 point) What is $\frac{\partial J}{\partial U}$?

Answer

$$V^T \frac{\partial l}{\partial o^{(1)}} (x^{(1)})^T + V^T \frac{\partial l}{\partial o^{(2)}} (x^{(2)})^T$$

3. Now let's generalize the result to an arbitrary time $N > 1$. For the following subparts (a)-(c), you may give your answer in terms of $\frac{\partial l}{\partial o^{(n)}}$, $n = 1, \dots, N$.

(a) (2 points) What is $\frac{\partial J}{\partial h^{(n)}}$?

Answer
$\sum_{t=n}^N V^T \frac{\partial l}{\partial o^{(N+n-t)}}$

(b) (2 points) What is $\frac{\partial J}{\partial V}$?

Answer
$\sum_{t=1}^N \frac{\partial l}{\partial o^{(t)}} (h^{(t)})^T$

(c) (2 points) What is $\frac{\partial J}{\partial U}$?

Answer
$\sum_{t=1}^N V^T \frac{\partial l}{\partial o^{(t)}} (x^{(t)})^T$

4 Empirical Questions

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

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

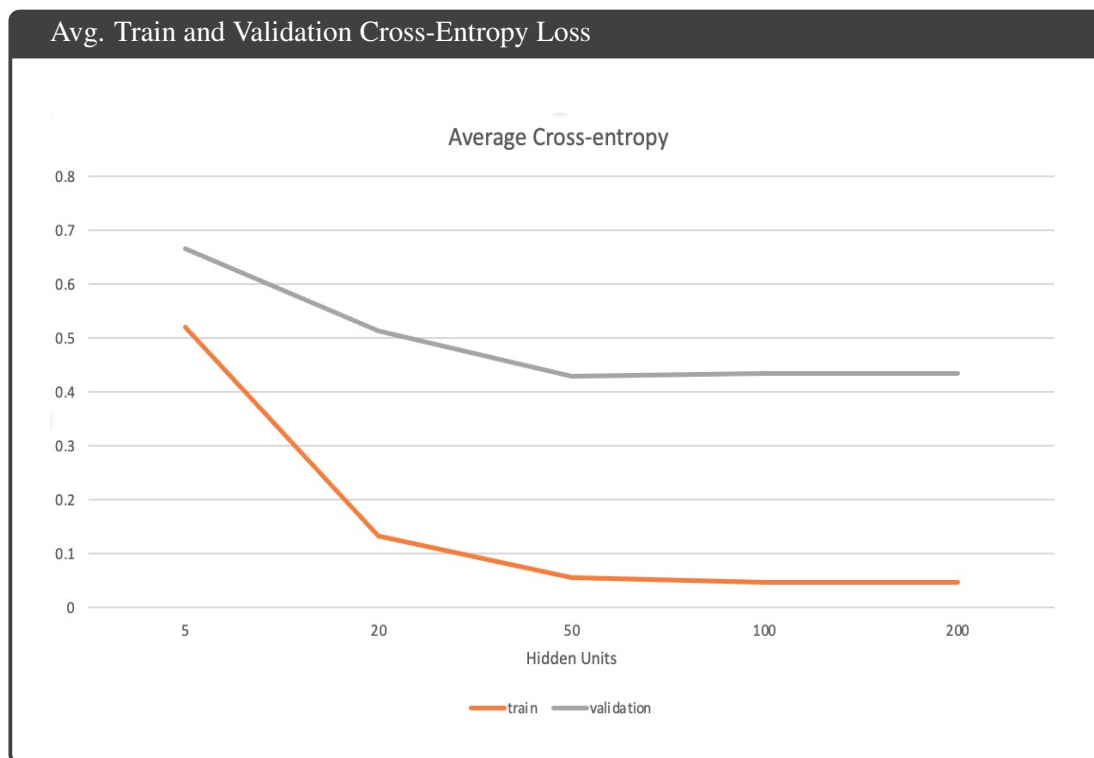
Parameter	Value
Number of Hidden Units	50
Weight Initialization	RANDOM
Learning Rate	0.01

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.



- (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?

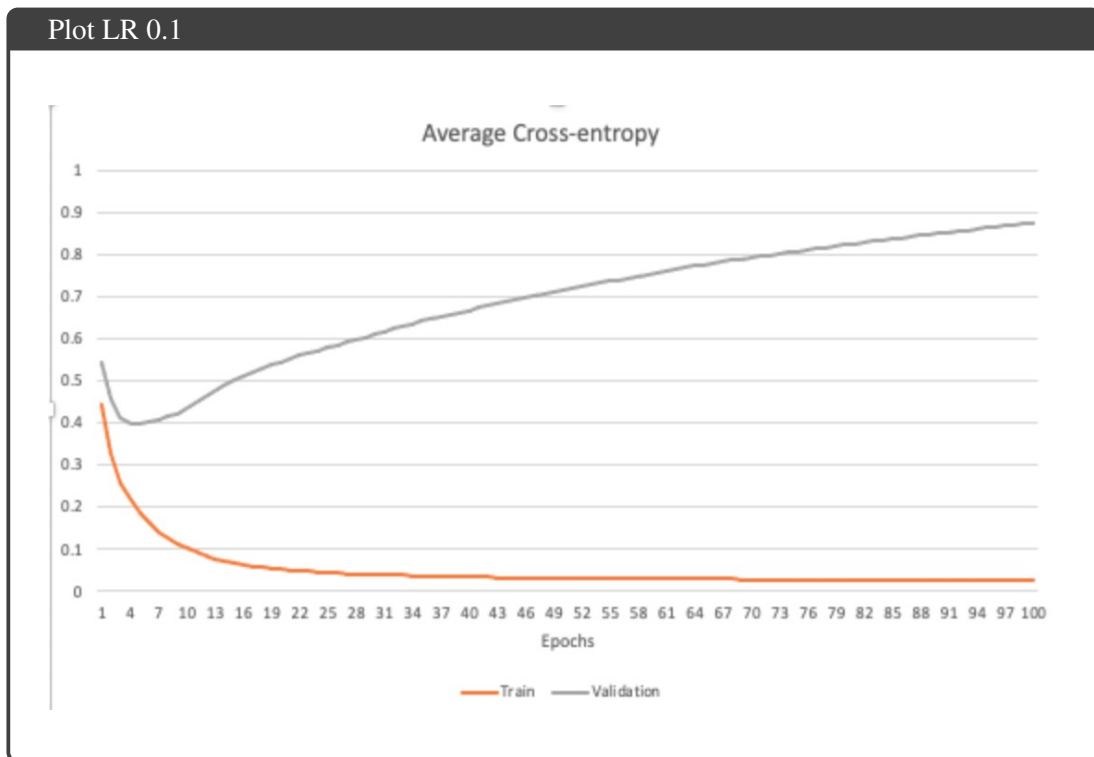
Answer

Both the average cross-entropy for training and validation sets decrease as we increase the number of hidden units. This is because as we increase the number of hidden units, we are making the model more complicated and thus achieve lower average cross-entropy. The two curves converges to flat line as we further increase the number of hidden units. Thus if we increase the number of hidden units, both curves first decrease, then converge to flat lines.

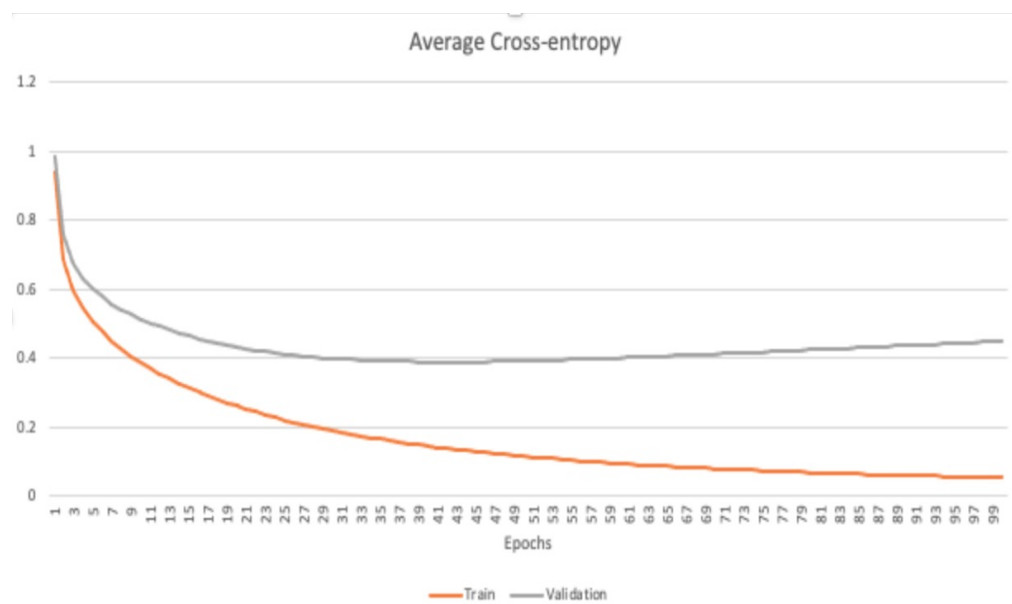
2. Learning Rate

- (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.

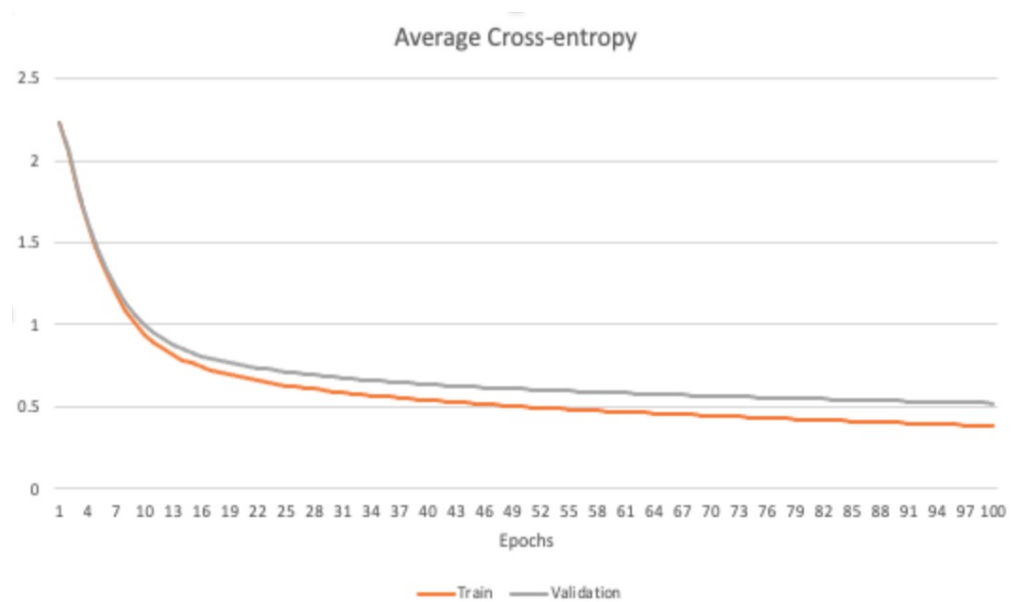
Plot the average training cross-entropy on the y-axis vs the number of epochs on the x-axis for the mentioned learning rates. In the **same figure**, plot the average validation cross-entropy loss. Make a separate figure for each learning rate.



Plot LR 0.01



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

The larger the learning rate, the faster it converges. If the learning rate is 0.1, then we can see that for the initial several epochs, both train and validation decreases, but then quickly the two diverges due to over-fitting. For learning rate with 0.01, it has similar pattern but the divergence happens late. When learning rate equals to 0.001, the convergence is quite slow.

3. Weight Initialization

- (a) (2 points) For this exercise, you can work on any data set. Initialize α and β 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 α and β . Comment on what you observed. Do you think it is reasonable to use zero initialization? Why or why not?

Answer

I ran the same command code as described above. After quite many SGD rounds, there are still many zeros in the alpha and beta if we initialize with all zeros. It is more reluctance to change, and the neurons try to memorize the same functions almost in each iteration due to the symmetry structure. Thus it isn't reasonable to use zero initialization.

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? Is so, include full details.
2. Did you give any help whatsoever to anyone in solving this assignment? Is 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

1. No.
2. No.
3. No.

Programming (56 points)

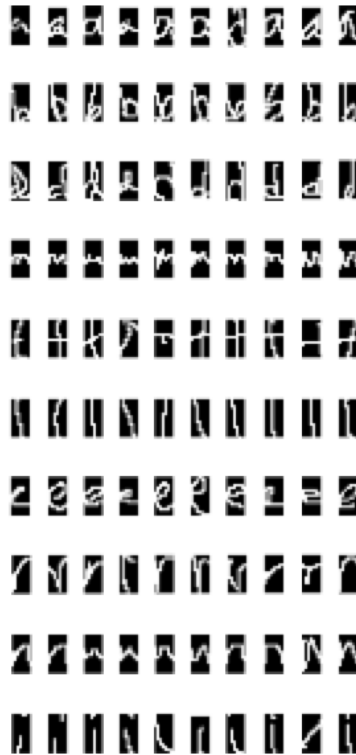


Figure 4: 10 Random Images of Each of 10 Letters in OCR

5 The Task

Your goal in this assignment is to label images of handwritten letters by implementing a Neural Network from scratch. You will implement all of the functions needed to initialize, train, evaluate, and make predictions with the network.

The programs you write will be automatically graded using the Gradescope system. You may write your programs in **Python, Java, or C++**. However, you should use the same language for all parts below.

6 The Datasets

Datasets We will be using a subset of an Optical Character Recognition (OCR) dataset. This data includes images of all 26 handwritten letters; our subset will include only the letters “a,” “e,” “g,” “i,” “l,” “n,” “o,” “r,” “t,” and “u.” The handout contains three datasets drawn from this data: a small dataset with 60 samples *per class* (50 for training and 10 for validation), a medium dataset with 600 samples per class (500 for training and 100 for validation), and a large dataset with 1000 samples per class (900 for training and 100 for validation). Figure 4 shows a random sample of 10 images of few letters from the dataset.

File Format Each dataset (small, medium, and large) consists of two csv files—train and validation. Each row contains 129 columns separated by commas. The first column contains the label and columns 2 to 129 represent the pixel values of a 16×8 image in a row major format. Label 0 corresponds to “a,” 1 to “e,” 2 to “g,” 3 to “i,” 4 to “l,” 5 to “n,” 6 to “o,” 7 to “r,” 8 to “t,” and 9 to “u.” Because the original images are black-and-white (not grayscale), the pixel values are either 0 or 1. However, you should write your code to

accept arbitrary pixel values in the range $[0, 1]$. The images in Figure 4 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 character 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 , the hidden layer \mathbf{z} consist of D hidden units, and 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 .

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

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. $\boldsymbol{\alpha}_{:,0}$ and $\boldsymbol{\beta}_{:,0}$) hold the bias parameters.

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

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

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

In Equation 1, 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 k th 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. Note that SGD has a slight impact on the objective function, where we are “summing” over the current point, i :

$$J_{SGD}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = - \sum_{k=1}^K y_k^{(i)} \log(\hat{y}_k^{(i)}) \quad (2)$$

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{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. $\{0, 1, \dots, 9\}$). 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:	<code>\$ python3 neuralnet.py [args...]</code>
For Java:	<code>\$ javac -cp "./lib/ejml-v0.38-libs/*:./" neuralnet.java</code> <code>\$ java -cp "./lib/ejml-v0.38-libs/*:./" neuralnet [args...]</code>
For C++:	<code>\$ g++ -g -std=c++11 -I./lib neuralnet.cpp; ./a.out [args...]</code>

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 backpropagation loops through all of the training data (e.g., if `<num_epoch>` equals 5, then each training example will be used in backpropagation 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 learning rate for SGD.

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 auto-grader 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). The first few lines of the predicted labels for the validation dataset is given below.

```
6
4
8
8
```

8.3 Output Metrics

Generate a file where you report the following metrics:

cross entropy After each Stochastic Gradient Descent (SGD) epoch, report mean cross entropy on the training data `crossentropy(train)` and validation data `crossentropy(validation)` (See Equation 1). 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): 2.18506276114
epoch=1 crossentropy(validation): 2.18827302588
```

```
epoch=2 crossentropy(train): 1.90103257727
epoch=2 crossentropy(validation): 1.91363803461
error(train): 0.77
error(validation): 0.78
```

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.

There is more information in the `README.md` file. Do read through the `README` file if you plan to use it for debugging. 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 i th training example (\mathbf{x}, \mathbf{y}) where \mathbf{y} is a one-hot encoding of the true label. Our neural network outputs $\hat{\mathbf{y}} = h_{\alpha, \beta}(\mathbf{x})$, where α and β are the parameters of the first and second layers respectively and $h_{\alpha, \beta}(\cdot)$ is a one-hidden layer neural network with a sigmoid activation and softmax output. The loss function is negative cross-entropy $J = \ell(\hat{\mathbf{y}}, \mathbf{y}) = -\mathbf{y}^T \log(\hat{\mathbf{y}})$. $J = J_{\mathbf{x}, \mathbf{y}}(\alpha, \beta)$ is actually a function of our training example (\mathbf{x}, \mathbf{y}) , and our model parameters α, β 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

```
1: procedure SGD(Training data  $\mathcal{D}$ , test data  $\mathcal{D}_t$ )
2:   Initialize parameters  $\alpha, \beta$  ▷ Use either RANDOM or ZERO from Section 7.1
3:   for  $e \in \{1, 2, \dots, E\}$  do ▷ For each epoch
4:     for  $(\mathbf{x}, \mathbf{y}) \in \mathcal{D}$  do ▷ For each training example (No shuffling)
5:       Compute neural network layers:
6:        $\mathbf{o} = \text{object}(\mathbf{x}, \mathbf{a}, \mathbf{b}, \mathbf{z}, \hat{\mathbf{y}}, J) = \text{NNFORWARD}(\mathbf{x}, \mathbf{y}, \alpha, \beta)$ 
7:       Compute gradients via backprop:
8:       
$$\left. \begin{array}{l} \mathbf{g}_\alpha = \frac{\partial J}{\partial \alpha} \\ \mathbf{g}_\beta = \frac{\partial J}{\partial \beta} \end{array} \right\} = \text{NNBACKWARD}(\mathbf{x}, \mathbf{y}, \alpha, \beta, \mathbf{o})$$

9:       Update parameters:
10:       $\alpha \leftarrow \alpha - \gamma \mathbf{g}_\alpha$ 
11:       $\beta \leftarrow \beta - \gamma \mathbf{g}_\beta$ 
12:      Evaluate training mean cross-entropy  $J_{\mathcal{D}}(\alpha, \beta)$ 
13:      Evaluate test mean cross-entropy  $J_{\mathcal{D}_t}(\alpha, \beta)$ 
14:   return parameters  $\alpha, \beta$ 
```

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

Algorithm 2 Prediction at Test Time

```
1: procedure PREDICT(Unlabeled train or test dataset  $\mathcal{D}'$ , Parameters  $\alpha, \beta$ )
2:   for  $\mathbf{x} \in \mathcal{D}'$  do
3:     Compute neural network prediction  $\hat{\mathbf{y}} = h(\mathbf{x})$ 
4:     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.

$$\alpha = \begin{bmatrix} \alpha_{10} & \alpha_{11} & \dots & \alpha_{1M} \\ \alpha_{20} & \alpha_{21} & \dots & \alpha_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{D0} & \alpha_{D1} & \dots & \alpha_{DM} \end{bmatrix} \quad \mathbf{g}_\alpha = \frac{\partial J}{\partial \alpha} = \begin{bmatrix} \frac{dJ}{d\alpha_{10}} & \frac{dJ}{d\alpha_{11}} & \dots & \frac{dJ}{d\alpha_{1M}} \\ \frac{dJ}{d\alpha_{20}} & \frac{dJ}{d\alpha_{21}} & \dots & \frac{dJ}{d\alpha_{2M}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{dJ}{d\alpha_{D0}} & \frac{dJ}{d\alpha_{D1}} & \dots & \frac{dJ}{d\alpha_{DM}} \end{bmatrix} \quad (3)$$

$$\beta = \begin{bmatrix} \beta_{10} & \beta_{11} & \dots & \beta_{1D} \\ \beta_{20} & \beta_{21} & \dots & \beta_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{K0} & \beta_{K1} & \dots & \beta_{KD} \end{bmatrix} \quad \mathbf{g}_\beta = \frac{\partial J}{\partial \beta} = \begin{bmatrix} \frac{dJ}{d\beta_{10}} & \frac{dJ}{d\beta_{11}} & \dots & \frac{dJ}{d\beta_{1D}} \\ \frac{dJ}{d\beta_{20}} & \frac{dJ}{d\beta_{21}} & \dots & \frac{dJ}{d\beta_{2D}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{dJ}{d\beta_{K0}} & \frac{dJ}{d\beta_{K1}} & \dots & \frac{dJ}{d\beta_{KD}} \end{bmatrix} \quad (4)$$

Observe that we have (in a rather *tricky* fashion) defined the matrices such that both α and \mathbf{g}_α are $D \times (M + 1)$ matrices. Likewise, β and \mathbf{g}_β are $K \times (D + 1)$ matrices. The $+1$ comes from the extra columns $\alpha_{\cdot,0}$ and $\beta_{\cdot,0}$ 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, \dots, 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* substitute 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 Java¹ 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+

¹Java 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 $\mathbf{o} = \text{NNFORWARD}(\mathbf{x}, \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \text{object}(\mathbf{x}, \mathbf{a}, \mathbf{b}, \mathbf{z}, \hat{\mathbf{y}}, J)$, 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 Torch², PyTorch³, and DyNet⁴—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 $\mathbf{b} = [b_1, \dots, b_B]$ given input $\mathbf{a} = [a_1, \dots, a_A]$ via some differentiable function f . That is $\mathbf{b} = f(\mathbf{a})$.
2. Backward computation of the gradient of the input $\mathbf{g}_\mathbf{a} = \frac{\partial J}{\partial \mathbf{a}} = [\frac{dJ}{da_1}, \dots, \frac{dJ}{da_A}]$ given the gradient of output $\mathbf{g}_\mathbf{b} = \frac{\partial J}{\partial \mathbf{b}} = [\frac{dJ}{db_1}, \dots, \frac{dJ}{db_B}]$, 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 $i \in \{1, \dots, A\}$.

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 $\mathbf{g}_\mathbf{a} = \text{*BACKWARD}(\mathbf{a}, \mathbf{b}, \mathbf{g}_\mathbf{b})$ 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.

- 1: **procedure** LINEARFORWARD(\mathbf{a}, α)
- 2: Compute \mathbf{b}
- 3: **return** \mathbf{b}
- 4: **procedure** LINEARBACKWARD($\mathbf{a}, \alpha, \mathbf{b}, \mathbf{g}_\mathbf{b}$)
- 5: Compute \mathbf{g}_α
- 6: Compute $\mathbf{g}_\mathbf{a}$
- 7: **return** $\mathbf{g}_\alpha, \mathbf{g}_\mathbf{a}$

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://dymnet.readthedocs.io>

⁵Static 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

```
1: procedure NNFORWARD(Training example  $(\mathbf{x}, \mathbf{y})$ , Parameters  $\alpha, \beta$ )
2:    $\mathbf{a} = \text{LINEARFORWARD}(\mathbf{x}, \alpha)$ 
3:    $\mathbf{z} = \text{SIGMOIDFORWARD}(\mathbf{a})$ 
4:    $\mathbf{b} = \text{LINEARFORWARD}(\mathbf{z}, \beta)$ 
5:    $\hat{\mathbf{y}} = \text{SOFTMAXFORWARD}(\mathbf{b})$ 
6:    $J = \text{CROSSENTROPYFORWARD}(\mathbf{y}, \hat{\mathbf{y}})$ 
7:    $\mathbf{o} = \text{object}(\mathbf{x}, \mathbf{a}, \mathbf{z}, \mathbf{b}, \hat{\mathbf{y}}, J)$ 
8:   return intermediate quantities  $\mathbf{o}$ 
```

Algorithm 4 Backpropagation

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

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:

$$\frac{\partial}{\partial \theta_i} J(\boldsymbol{\theta}) \approx \frac{(J(\boldsymbol{\theta} + \epsilon \cdot \mathbf{d}_i) - J(\boldsymbol{\theta} - \epsilon \cdot \mathbf{d}_i))}{2\epsilon} \quad (5)$$

where \mathbf{d}_i is a 1-hot vector consisting of all zeros except for the i th entry of \mathbf{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 finite 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).