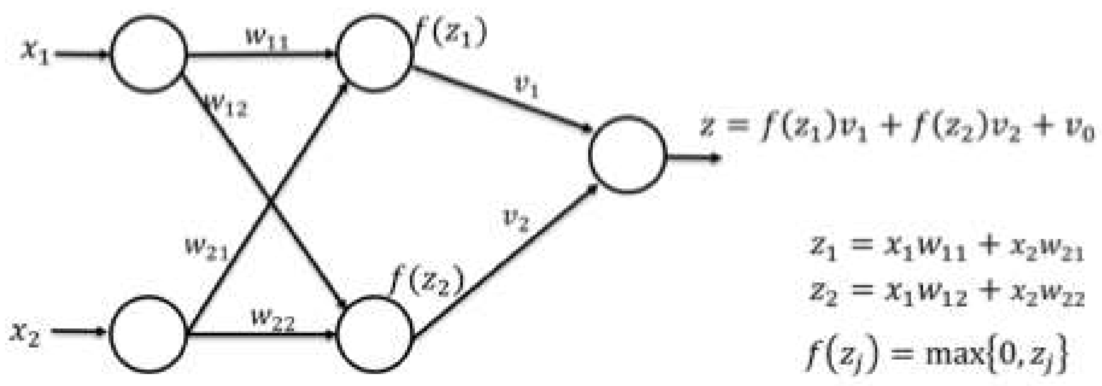
Problem 5

Midterm due Nov 10, 2020 05:29 IST *Past Due*

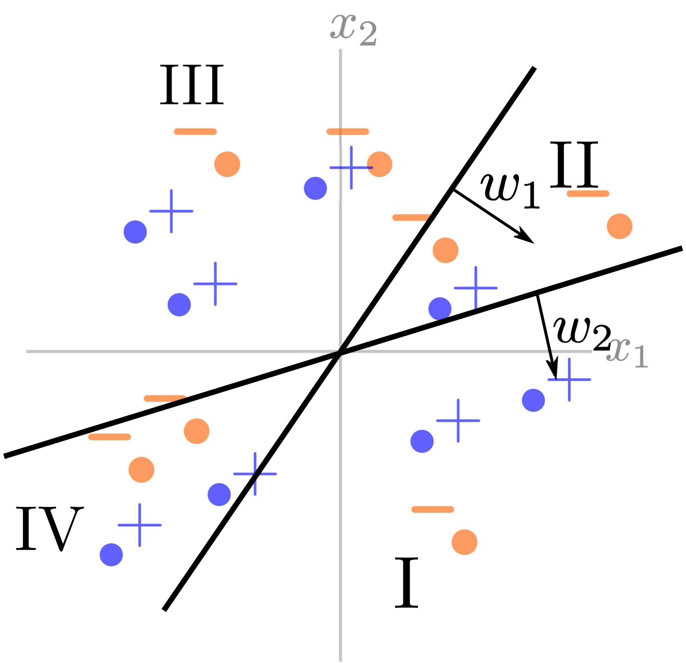
Consider a 2-layer feed-forward neural network that takes in $x \in \mathbb{R}^2$ and has two ReLU hidden units as defined in the figure below. **Note that hidden units have no offset parameters in this problem.**



5. (1)

4 points possible (graded)

The values of the weights in the hidden layer are set such that they result in the z_1 and z_2 “classifiers" as shown in the (x_1, x_2)**-space** in the figure below:



The z_1 “classifier" with the normal $w_1 = [w_{11} \quad w_{21}]^T$ is the line given by $z_1 = x \cdot w_1 = 0.$

Similarly, the z_2 “classifier" with the normal $w_2 = [w_{12} \quad w_{22}]^T$ is the line given by $z_2 = x \cdot w_2 = 0.$

The arrows labeled w_1 and w_2 point in the **positive** directions of the respective normal vectors.

The regions labeled I, II, III, IV are the 4 regions defined by these two lines not including the boundaries.

Choose the region(s) in (x_1, x_2) space which are mapped into each of the following regions in (f_1, f_2)-space, the 2-dimensional space of hidden unit activations $f(z_1)$ and $f(z_2)$. (For example, for the second column below, choose the region(s) in (x_1, x_2) space which are mapped into the f_1-axis in (f_1, f_2)-space.)

(Choose all that apply for each column.)

$\{(f_1, f_2) : f_1 > 0,\ f_2 > 0\}$:	f_1-axis:	f_2-axis:	the origin $(f_1, f_2) = (0,0)$:

(Choose all that apply.)

- [ ] I ✓
- [ ] II
- [ ] III
- [ ] IV
- [ ] None of the above

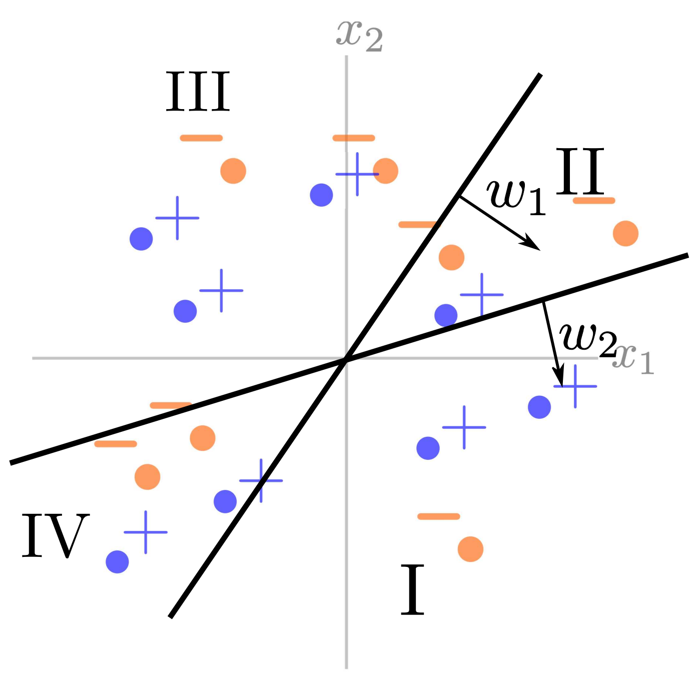
- [ ] I
- [ ] II ✓
- [ ] III ✓
- [ ] IV
- [ ] None of the above

- [ ] I
- [ ] II
- [ ] III ✓
- [ ] IV ✓
- [ ] None of the above

- [ ] I
- [ ] II
- [ ] III ✓
- [ ] IV
- [ ] None of the above

Grading Note: Regrading with partial credits.

Solution:



The regions $\mathrm{I}, \mathrm{II}, \mathrm{III}, \mathrm{IV}$ are defined by (but do not include) the lines $z_1 = x \cdot w_1 = 0$ and $z_2 = x \cdot w_2 = 0$. Hence,

$$z_1 \begin{cases} > 0 & \text{in } \mathrm{I}, \mathrm{II} \\ < 0 & \text{in } \mathrm{III}, \mathrm{IV} \end{cases}$$

$$z_2 \begin{cases} > 0 & \text{in I, IV} \\ < 0 & \text{in II, III} \end{cases}$$

Applying the reLu function, we get

$$f_1 = f(z_1) \begin{cases} > 0 & \text{in I, II} \\ = 0 & \text{in III, IV} \end{cases}$$
$$f_2 = f(z_2) \begin{cases} > 0 & \text{in I, IV} \\ = 0 & \text{in II, III} \end{cases}.$$

Hence

- The region I in (x_1, x_2)-space maps into the region $\{(f_1, f_2) : f_1 > 0,\ f_2 > 0\}$: in (f_1, f_2)-space ;
- The regions II,III maps into the region $\{(f_1, f_2) : f_1 \geq 0,\ f_2 = 0\},$ which is the f_1-axis in (f_1, f_2)-space ;
- The regions III,IV maps into the region $\{(f_1, f_2) : f_2 \geq 0,\ f_1 = 0\},$ which is the f_2-axis in (f_1, f_2)-space ;
- The regions III maps to $\{(f_1, f_2) : f_2 = 0,\ f_1 = 0\},$ the origin in (f_1, f_2)-space.

Submit You have used 0 of 3 attempts

Answers are displayed within the problem

5. (2)

2/2 points (graded)

If we keep the hidden layer parameters above fixed but add and train additional hidden layers (applied after this layer) to further transform the data, could the resulting neural network solve this classification problem?

- ◯ yes
- ◉ no

✔

Suppose we stick to the 2-layer architecture but add many more ReLU hidden units, all of them without offset parameters. Would it be possible to train such a model to perfectly separate these points?

Note : Assume that no 2 data points lie on the same line through the origin.

- [ ] yes ✔
- [x] no

✔

Solution:

- Since points with different labels, namely those in region III, are all mapped to the origin, it is impossible to classify these correctly by only adding more hidden layers.
- The points in regions I,II, IV can be assumed to be distinguished by either the f_1 or f_2 coordinates, so we only need to separate the points in region III, and adding more units, which correspond to more lines through the origin, will work.

Submit You have used 1 of 3 attempts

Answers are displayed within the problem

5. (3)

4/5 points (graded)

Which of the following statements is correct?

1. The gradient calculated in the backpropagation algorithm consists of the partial derivatives of the loss function with respect to each network weight.

- [x] True
- [ ] False

✔

2.

Initialization of the parameters is often important when training large feed-forward neural networks.

If weights in a neural network with sigmoid units are initialized to close to zero values, then during early stochastic gradient descent steps, the network represents a nearly linear function of the inputs.

- (x) True
- ( ) False

✔

3. On the other hand, if we randomly set all the weights to very large values, or don't scale them properly with the number of units in the layer below, then the sigmoid units would behave like sign units. Here, "behave like sign units" allows for shifting or rescaling of the sign function.

   (Note that a sign unit is a unit with activation function $\text{sign}\,(x) = 1$ if $x > 0$ and $\text{sign}\,(x) = -1$ if $x < 0$. For the purpose of this question, it does not matter what $\text{sign}\,(0)$ is.)

   - (x) True

   ✔

   Grading Note: (November 1) Since there is an error in this question, i.e. there is only 1 option in the multichoice, everyone will receive credit.

4. If we use only sign units in a feedforward neural network, then the stochastic gradient descent update will

   - (x) almost never change any of the weights
   - ( ) change the weights by large amounts at random

   ✔

5. Stochastic gradient descent differs from (true) gradient descent by updating only one network weight during each gradient descent step.

True

False ✔

✖

Correction note (Nov 7 03:00UTC):. In the earlier version, there was a grammatical error in question 2 above. The problem statement started with "If weights in a neural network with sigmoid units are initialized to close to zero values".

Solution:

- True, by definition of the backpropagation algorithm.
- True, because those activation functions are linear-like near zero.
- True, because far from zero the sigmoid function looks like a scaled and shifted version of the sign function.
- True, because of the gradient is zero except at one point.
- False. Stochastic gradient descent differs by considering the gradient with respect to just one training example on each step.

Submit You have used 1 of 3 attempts

Answers are displayed within the problem

5. (4)

2/3 points (graded)
There are many good reasons to use convolutional layers in CNNs as opposed to replacing them with fully connected layers. Please check T or F for each statement.

Since we apply the same convolutional filter throughout the image, we can learn to recognize the same feature wherever it appears.

True

False

A fully connected layer for a reasonably sized image would simply have too many parameters

True

False

Grading Note: The intended answer was true because it's a justification for using CNNs over FC layers, but in fact the FC net used in the mnist project did have quite good accuracy, and was trainable. Since the statement "simply have too many parameters" is debatable, full credit is given to all. (The intended answer will still show as the correct answer, but you will see the credit in your score.)

A fully connected layer can learn to recognize features anywhere in the image even if the features appeared preferentially in one location during training

True

False ✔

✖

Solution:

- True by definition
- True. For a fully connected layer of size n and a picture of size $w \times h$, the size of the layer would be whn, several times larger than the image itself. The intended answer was true because it's a justification for using CNNs over FC layers, but in fact the FC net used in the mnist project did have quite good accuracy, and was trainable. So it's not entirely true that FC layers for images "simply have too many parameters." I think the answer is debatable given that we were able to successfully train a FC on mnist in the project, so it might be fair to just give everyone credit for that part of the question.

- False.Only a convolutional layer can do this. A fully connected layer's weights depend on the location, so they would only be able to recogize features at locations that appeared preferentially in the training data.

Submit You have used 2 of 3 attempts

Answers are displayed within the problem

Error and Bug Reports/Technical Issues

Hide Discussion

Topic: Midterm Exam (1 week):Midterm Exam 1 / Problem 5

Add a Post

Show all posts by recent activity

?	Micromasters on DS&S now has two tracks (!) Probably the wrong place to ask this but I just found out looking for my record and am excited. A...	4
?	ATTN STAFF: Partial Scores for selecting options II & IV on Question 5(1) for points on f1 and f2 axis!!! I selected options II and IV for points that lie on f1 and f2 axis. I admit that I did not really notice t...	16
?	5. (4) Q1 Convolutional filter A clarification, is it more correct to say that polling instead of convolutional filter is dedicated to r...	3
?	[STAFF] 5. (3) - I got 4/5 correct. However grader is showing only 3/5. Hello Staff, I got 4/5 correct. However grader is showing only 3/5. could you please check and let ...	3
	MIDTERM RESULTS Hello, Does anyone know when we'll get our midterm results? Also, is everybody able to see their...	11
?	Staff--Missed the Submission Button Hey, I sent an email to the MIT.edu address explaining this a few days ago (before the test was d...	4
	Question 5.(1) region III Could someone please explain why region III, with f1 = f2 = 0, correspond not only to the origin, ...	6
?	[STAFF] Please consider my solution as correct for 5.1 column 2 and 3	5
	[Staff] Clarification needed on 5.1	8
?	[Staff] - Q5.4 ii - grader appears to be incorrect @Staff - can you please review / correct Q5.4ii to accept "False" as a valid answer, as per the com...	1

[Staff] Clarification on 5.1
The grader says that II and III are mapped to the f1 axis, but the 'show answer' (and my brain) sa… 2

[STAFF] clarification needed for 5. (3)
Hello Staff, I got 4/5 correct. However grader is showing only 3/5. could you please check and let … 2

[Staff] Will there be a webinar with a/the professor(s) for this course?
The course will be over in ~5 weeks, and I haven't seen any mention of it. 1