Midterm for CSC321, Intro to Neural Networks
Winter 2015, night section
Tuesday, Feb. 24, 6:10-7pm

Name: _____

Student number: _____

This is a closed-book test. It is marked out of 15 marks. Please answer ALL of the questions. Here is some advice:

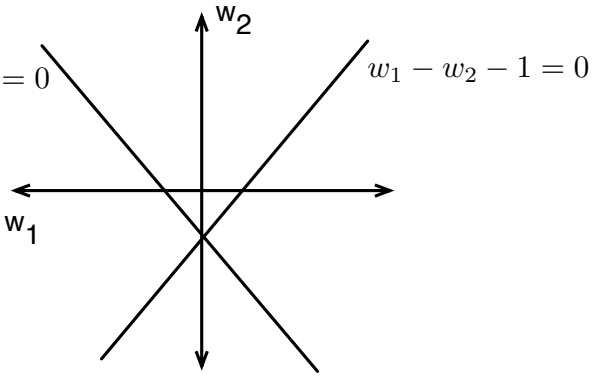
- The questions are NOT arranged in order of difficulty, so you should attempt every question.
- Questions that ask you to "briefly explain" something only require short (1-3 sentence) explanations. Don't write a full page of text. We're just looking for the main idea.
- None of the questions require long derivations. If you find yourself plugging through lots of equations, consider giving less detail or moving on to the next question.
- Many questions have more than one right answer.

Final mark: _____ / 15

1. (1 mark) Suppose we want to train a perceptron with weights w_1 and w_2 and a fixed bias $b = -1$. Sketch the constraints in weight space corresponding to the following training cases. (The decision boundaries have already been drawn for you, so you only need to draw arrows to indicate the half-spaces.) Shade the feasible region or indicate that none exists. You do not need to justify your answer.

$\mathbf{x} = (1, -1)$, $t = 1$

$\mathbf{x} = (-1, -1)$, $t = 0$



2. (1 mark) Suppose we have a fully connected, feed-forward network with no hidden layer, and 5 input units connected directly to 3 output units. Briefly explain why adding a hidden layer with 8 *linear* units does not make the network any more powerful (as opposed to not having a hidden layer).

3. (1 mark) Suppose we have a network with linear hidden units, so that each hidden unit computes its activation h_i as

$$h_i = \sum_j w_{ij} x_j,$$

where the x_j are the input values and the w_{ij} are the weights. Let the matrix $\mathbf{X}$ represent the input values for a mini-batch of training examples (rows correspond to training examples and columns correspond to input dimensions). Let $\mathbf{W}$ represent the weight matrix, where the (i, j) entry connects input j to hidden unit i. Write a matrix expression which computes the hidden activations on this mini-batch, and specify what the rows and columns of the result correspond to. You do not need to justify your answer.

4. (1 mark) In stochastic gradient descent, each pass over the dataset requires the same number of arithmetic operations, whether we use minibatches of size 1 or size 1000. Why can it nevertheless be more computationally efficient to use minibatches of size 1000?

5. (2 marks) In class, we saw that using squared error loss $C = (y - t)^2$ with a logistic output unit can make optimization difficult because the unit can saturate, leading to a small gradient. Cross-entropy loss doesn't have this problem. Suppose that we instead use the absolute loss $C = |y - t|$ (but keep the logistic output unit). Does this have the same problem with saturation that squared error does? Justify your answer algebraically and/or by drawing a figure.

6. (2 marks) Briefly explain a way in which the neural probabilistic language model from Assignment 1 was doing supervised learning, and a way in which it was doing unsupervised learning.

7. (1 mark) Briefly explain one thing you would use a validation set for, and why you can't just do it using the test set.

8. (2 marks) Fill in weights, biases, and initial activations for the following RNN so that it initially outputs 1, but as soon as it receives an input of 0, it switches to outputting 0 for all subsequent time steps. For instance, the input `1110101` produces the output `1110000`. All units are binary threshold units with a threshold of 0. The hidden unit has an initial value of 0. You don't need to provide an explanation, but doing so may help you receive partial credit.

   Hint: in one possible solution, the hidden unit has an activation $h_t = 0$ until there's an input $x_t = 0$, at which point it switches to maintaining an activation of 1 forever. The output unit always predicts the opposite of the hidden unit, i.e. $y = 1 - h$.

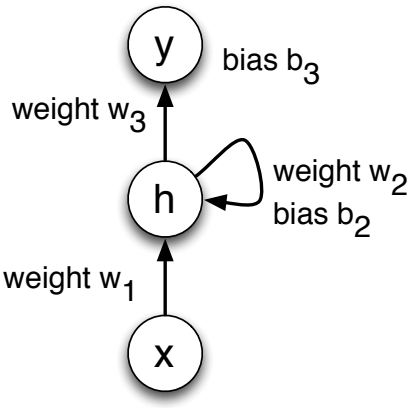
$w_1 =$ _____

$w_2 =$ _____

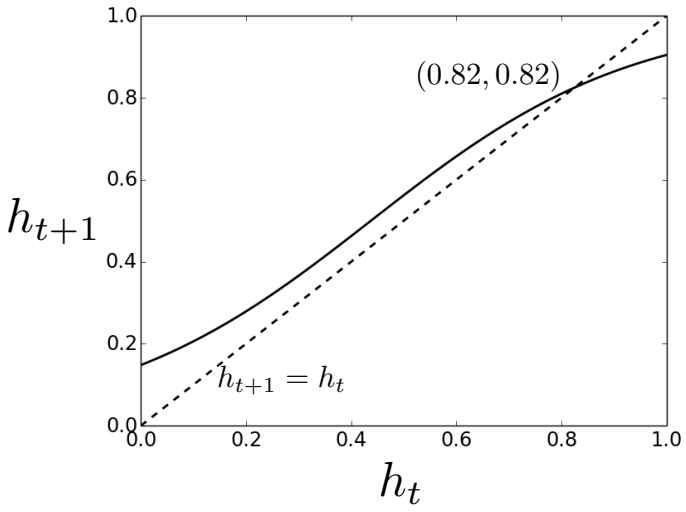
$b_2 =$ _____

$w_3 =$ _____

$b_3 =$ _____



9. (2 marks) Suppose we have an RNN with one hidden unit with a logistic nonlinearity, and no inputs or outputs. The unit has a bias of -1.75 and is connected to itself with a weight of 4. The figure shows the activation h_{t+1} as a function of the previous activation h_t. Draw a phase plot that summarizes the behavior of this system, and label any sources or sinks. For what values of the initial activation $r = h_0$ will we encounter exploding or vanishing gradients (with respect to r)?



10. (2 marks) Alice and Bob have implemented two neural networks for recognizing handwritten digits from 16×16 grayscale images. Each network has a single hidden layer, and makes predictions using a softmax output layer with 10 units, one for each digit class.

    - Alice's network is a convolutional net. The hidden layer consists of three 16×16 convolutional feature maps, each with filters of size 5×5, and uses the logistic nonlinearity. All of the hidden units are connected to all of the output units.
    - Bob's network is a fully connected network with no weight sharing. The hidden layer consists of 768 logistic units (the same number of units as in Alice's convolutional layer).

    Briefly explain one advantage of Alice's approach and one advantage of Bob's approach.