Midterm for CSC321, Intro to Neural Networks
Winter 2015, afternoon section
Tuesday, Feb. 24, 1:10-2pm

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:

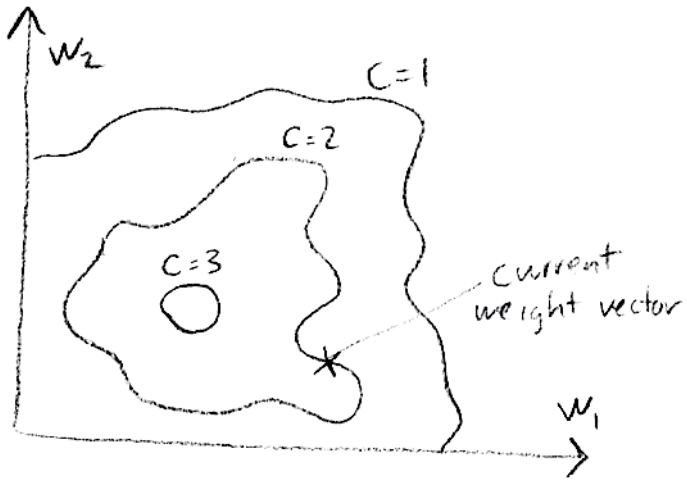
- The questions are NOT arranged in order of difficulty, so you should attempt every question.
- Questions that ask you to "briefly explain" something only require short (1-3 sentence) explanations. Don't write a full page of text. We're just looking for the main idea.
- None of the questions require 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

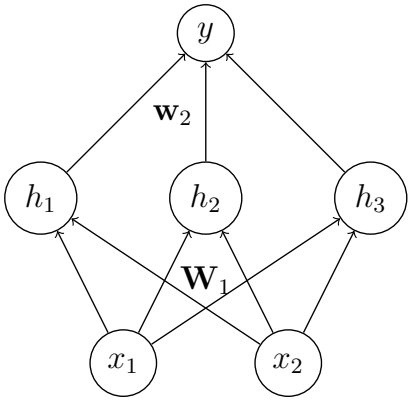
1. (2 marks) Briefly explain what is meant by overfitting. Is it true that if you choose the hyperparameters (e.g. number of hidden units) well, then there will be no overfitting? Why or why not? (Either YES or NO is acceptable, as long as you justify your answer.)

2. (1 mark) Recall our study of the weight space geometry of linear regression. For this question, assume there is no bias parameter. We saw that the set of weight vectors $\mathbf{w}$ which predict a given target exactly, *i.e.* $\mathbf{w}^T\mathbf{x}^{(i)} = t^{(i)}$, is a hyperplane in weight space. If all the hyperplanes for a given training set intersect at a single point $\mathbf{w}^\star$, then must $\mathbf{w}^\star$ be an optimal solution to the linear regression problem? Why or why not?

3. (1 mark) The following diagram shows the level curves in weight space of a cost function C which we are trying to *minimize*. The current weight vector is marked by an $\times$. Sketch the gradient descent update. (We haven't given you enough information to determine the magnitude, so we just want you to get the direction correct.)

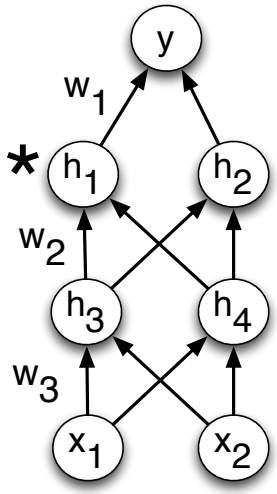


4. (2 marks) In the first week of class, we discussed how linear regression could be made more powerful using a basis function expansion, i.e. a function ϕ which maps each data point $\mathbf{x}$ to a feature vector $\phi(\mathbf{x})$. We later saw how this is analogous to fitting a feed-forward neural net with one hidden layer, where one set of weights is held fixed. (Such a network is shown in the following figure.) Which set of weights is held fixed? Briefly explain what the hidden activations and both sets of weights correspond to.



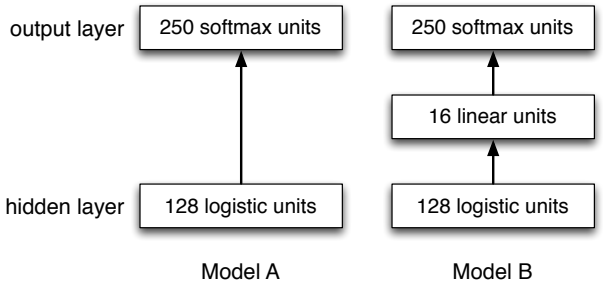
5. (3 marks) Consider the network shown in the figure. All of the hidden units use the linear rectification nonlinearity $h_i = \max(z_i, 0)$. We are trying to minimize a cost function C which depends only on the activation of the output unit y. The unit h_1 (marked with a $\star$) receives an input of -1 on a particular training case, so its output is 0. Based only on this information, which of the following weight derivatives are **guaranteed** to be 0 for this training case? Write YES or NO for each. Justify your answers informally. *Hint: don't work through the backprop computations. Instead think about what the partial derivatives really mean.*

$\partial C/\partial w_1$: _____ $\partial C/\partial w_2$: _____ $\partial C/\partial w_3$: _____



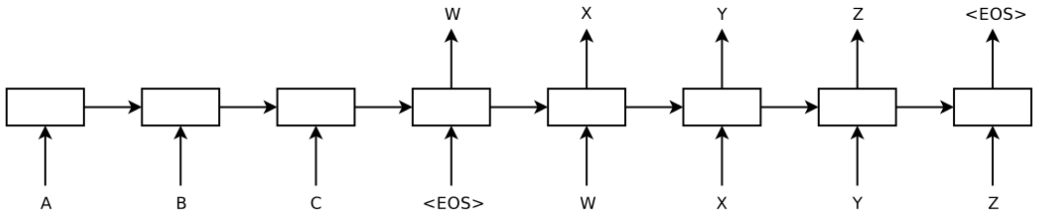
Note: Each of w_1, w_2, and w_3 refers to the weight on a *single* connection, not the whole layer.

6. (2 marks) Let's compare the following two models. Model A is the neural probabilistic language model from Assignment 1. Model B is the same as Model A, with the following modification: in between the hidden layer (of size 128) and the output layer (of size 250), we insert a layer consisting of 16 linear units. The top layers of both models are shown in the figure. Describe one advantage of Model A and one advantage of Model B.



7. (1 mark) Briefly explain one method for dealing with the problem of exploding and/or vanishing gradients in recurrent nets. Why does this method help?

8. (1 mark) We saw that we can apply a recurrent net to machine translation by feeding it an English sentence, and then having it generate the French sentence the same way an RNN language model generates text. This setup is shown in the figure. Would it work to use the neural probabilistic language model from Assignment 1 in the same way? Why or why not?



9. (1 mark) Design a finite state machine which determines if a given sequence of binary digits contains at least 2 zeros. Specify which state is the initial state and which state(s) correspond to an answer of YES. You do not need to justify your answer.

10. (1 mark) Suppose we compute the convolution $I * w$, where I is a grayscale image (white pixels correspond to values of 1 and black pixels correspond to 0) and

$$w = \begin{pmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{pmatrix}$$

is a convolution kernel. For what parts of the image will the output be farthest from zero? In the image shown, will the output at the location marked by $\times$ take a positive or a negative value?