UNIVERSITY OF TORONTO
Faculty of Arts and Science
APRIL 2017 EXAMINATIONS
CSC321H1S

Duration — 3 hours
No Aids Allowed

Name: _____

Student number: _____

This is a closed-book test. It is marked out of 35 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 "justify your answer" or "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.

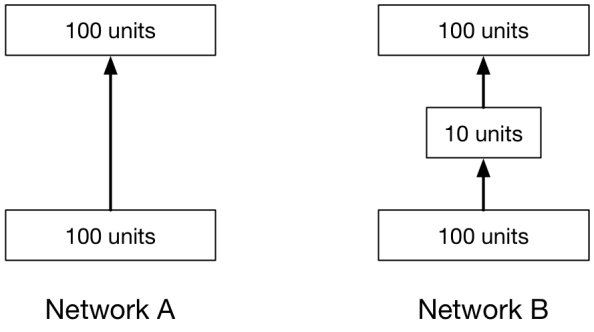
.

Q1: _____ / 2
Q2: _____ / 2
Q3: _____ / 2
Q4: _____ / 2
Q5: _____ / 1
Q6: _____ / 2
Q7: _____ / 2
Q8: _____ / 2
Q9: _____ / 2
Q10: _____ / 2
Q11: _____ / 2
Q12: _____ / 2
Q13: _____ / 2
Q14: _____ / 1
Q15: _____ / 1
Q16: _____ / 3
Q17: _____ / 1
Q18: _____ / 1
Q19: _____ / 1
Q20: _____ / 2

Final mark: _____ / 35

1. **[2pts]** Suppose you design a multilayer perceptron for classification with the following architecture. It has a single hidden layer with the hard threshold activation function. The output layer uses the softmax activation function with cross-entropy loss. What will go wrong if you try to train this network using gradient descent? Justify your answer in terms of the backpropagation rules.

2. Consider the following two multilayer perceptrons, where all of the layers use linear activation functions.



   (a) **[1pt]** Give one advantage of Network A over Network B.

   (b) **[1pt]** Give one advantage of Network B over Network A.

3. Suppose you train an ensemble of 5 networks for object recognition; all of the networks use the same architecture, but start from different random initializations. We saw that if you average the predictions of all five networks, you are guaranteed to do better in expectation (in terms of cross-entropy loss) than if you just use one of the networks.

   (a) **[1pt]** This guarantee depended on a particular property of cross-entropy loss. What is that property? (A one-word answer is sufficient, but an explanation may help you get partial credit.)

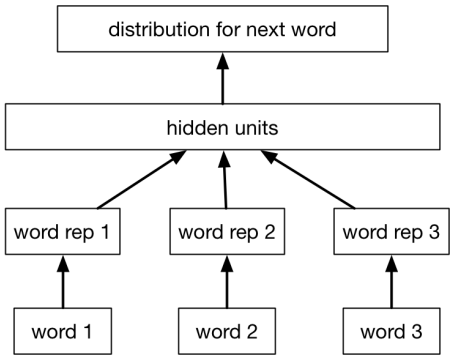
   (b) **[1pt]** Does this guarantee hold if you instead average the weights and biases of the networks? Why or why not?

4. Consider Bayesian optimization, where we are trying to minimize a function $f(\boldsymbol{\theta})$.

   (a) **[1pt]** Give one reason that Probability of Improvement is a better acquisition function than the negative predictive mean (i.e. $-\mathbb{E}[f(\boldsymbol{\theta})]$).

   (b) **[1pt]** Give one reason that Expected Improvement is a better acquisition function that Probability of Improvement.

5. **[1pt]** Recall the neural language model architecture:



Why is it that the learned word representations for "professor" and "teach" are likely to be far apart? (It's not sufficient to say they are "semantically dissimilar" – your answer should mention how the representations are used in the prediction task.)

6. Suppose you want to redesign the AlexNet architecture to reduce the number of arithmetic operations required for each backprop update.

   (a) **[1pt]** Would you try to cut down on the number of weights, units, or connections? Justify your answer.

   (b) **[1pt]** Would you modify the convolution layers or the fully connected layers? Justify your answer.

7. **[2pts]** Suppose you have a convolutional network with the following architecture:

   - The input is an RGB image of size 256×256.
   - The first layer is a convolution layer with 32 feature maps and filters of size 3×3. It uses a stride of 1, so it has the same width and height as the original image.
   - The next layer is a pooling layer with a stride of 2 (so it reduces the size of each dimension by a factor of 2) and pooling groups of size 3×3.

   Determine the size of the receptive field for a single unit in the pooling layer. (I.e., determine the size of the region of the input image which influences the activation of that unit.) You may assume the receptive field lies entirely within the image. *Hint: you may want to draw a one-dimensional conv net to reason about this problem.*

8. In this course, we've discussed five cases where you want to use backprop to compute the gradient of some function *with respect to the pixels of an image*. Describe two of these cases. For each one, explain what function we compute the gradient of and what the gradient is useful for. (A one-sentence verbal description is sufficient; you don't need to write equations.)

   (a) **[1pt]** The first example:

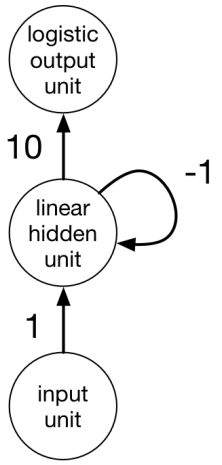
   (b) **[1pt]** The second example:

9. We considered two different models for binary images: the mixture of Bernoullis (MoB) and restricted Boltzmann machine (RBM).

   (a) **[1pt]** Give one advantage of an RBM over an MoB. (It's not sufficient to say that it gets higher likelihood or produces better samples — you should explain why it can model the data better.)

   (b) **[1pt]** Give one advantage of an MoB over an RBM.

10. **[2pts]** Determine what the following recurrent network computes. More precisely, determine the function computed by the output unit at the final time step; the other outputs are not important. All of the biases are 0. You may assume the inputs are integer valued and the length of the input sequence is even.



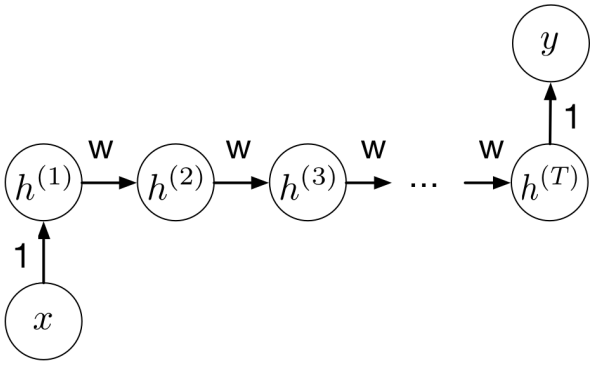
11. **[2pts]** Consider a residual network built out of residual units, where the residual function is given by an MLP with one hidden layer:

$$\begin{aligned}\mathbf{z} &= \mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)} \\ \mathbf{h} &= \phi(\mathbf{z}) \\ \mathbf{y} &= \mathbf{x} + \mathbf{W}^{(2)}\mathbf{h}\end{aligned}$$

Give a way of setting the weights and biases such that the derivatives will not explode or vanish. Briefly explain your answer, but you do not need to provide a detailed derivation.

12. Consider the following RNN, which has a scalar input at the first time step, makes a scalar prediction at the last time step, and uses a shifted logistic activation function:

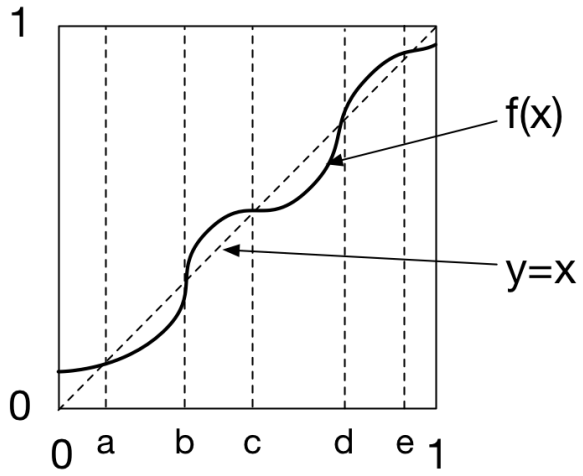
$$\phi(z) = \sigma(z) - 0.5.$$



(a) **[1pt]** Write the formula for the derivative $\overline{h_t}$ as a function of $\overline{h_{t+1}}$, for $t < T$. (Use z_t to denote the input to the activation function at time t. You may write your answer in terms of σ', i.e. you don't need to explicitly write out the derivative of σ.)

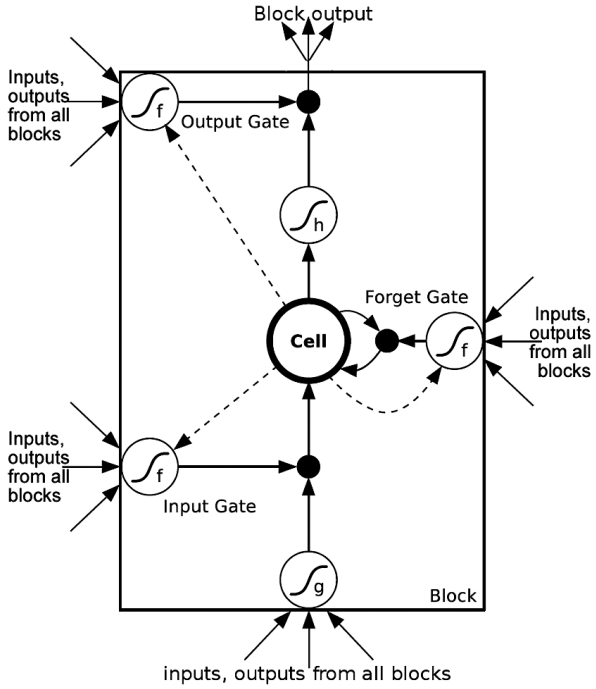
(b) **[1pt]** Suppose the input to the network is $x = 0$. Notice that $h_t = 0$ for all t. Based on your answer to part (a), determine the value α such that if $w < \alpha$, the gradient vanishes, while if $w > \alpha$, the gradient explodes. You may use the fact that $\sigma'(0) = 1/4$.

13. **[2pts]** Consider the following function $f(x)$:

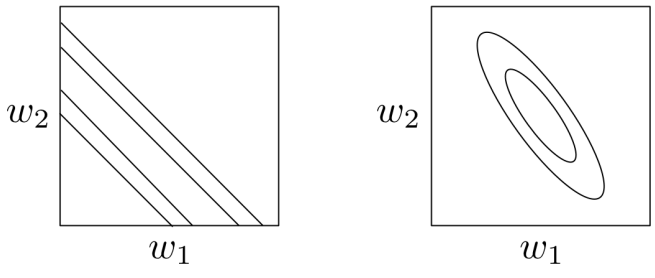


Here, a, b, c, d, e represent real values between 0 and 1. Sketch the function $f^{(100)}(x)$, i.e. f iterated 100 times. Label any relevant values on the axes.

14. **[1pt]** Recall the LSTM architecture. Suppose you want the memory cell to sum its inputs over time. What values should the input gate and forget gate take? You do not need to justify your answer.



15. **[1pt]** For linear regression with scalar-valued targets, which of the following contour plots in weight space best represents the cost function for a *single* training example? Justify your answer.



Plot A Plot B

16. Consider the problem of MAP estimation for the mean μ of a Gaussian distribution with known standard deviation σ. For the prior distribution, we will use a Gaussian distribution with mean 0 and standard deviation γ.

    (a) **[2pts]** Determine the function that we need to maximize. You do not need to determine the constant terms explicitly.

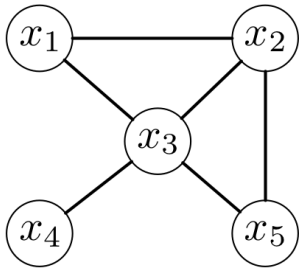
    (b) **[1pt]** Determine the optimal value of μ by setting the derivative to 0. (You do not need to justify why it is a maximum rather than a minimum.)

17. **[1pt]** Recall that for Bayesian parameter estimation for a Bernoulli random variable, we use the beta distribution as the prior for the mean parameter θ:

$$p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}.$$

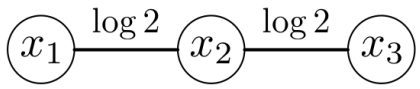
Give values of a and b for the beta prior which encode a strong prior belief that $\theta \approx 0.7$. You do not need to justify your answer.

18. **[1pt]** In the following Boltzmann machine, list all of the variables which are conditionally independent of x_1 given x_3. You do not need to justify your answer.



19. **[1pt]** In this course, we covered two network architectures whose targets are the same as their inputs. Pick one of these architectures. Name the network architecture and briefly explain why the task isn't trivial for the network to learn.

20. **[2pts]** Consider the following Boltzmann machine, where all the variables take values in $\{0, 1\}$.



(This figure indicates that both edges have weight $\log 2$ and the biases are all 0.) Determine the conditional probability $\Pr(x_1 = 1 \,|\, x_3 = 1)$. *Hint: make a table which lists the happiness values for four configurations.*