CS230: Deep Learning

Winter Quarter 2019

Stanford University

Midterm Examination

180 minutes

	Problem	Full Points	Your Score
1	Multiple Choice Questions	13	
2	Short Answer Questions	23	
3	Convolutional Neural Networks	28	
4	Adversarial Attacks	10	
5	Loss comparisons	23	
6	The Optimizer	20	
Total		117	

The exam contains 29 pages including this cover page.

- This exam is **closed book i.e. no laptops, notes, textbooks, etc. during the exam**. However, you may use one A4 sheet (front and back) of notes as reference.
- In all cases, and especially if you're stuck or unsure of your answers, **explain your work, including showing your calculations and derivations!** We'll give partial credit for good explanations of what you were trying to do.

Name: _____

SUNETID: _____@stanford.edu

The Stanford University Honor Code:
I attest that I have not given or received aid in this examination, and that I have done my share and taken an active part in seeing to it that others as well as myself uphold the spirit and letter of the Honor Code.

Signature: _____

Question 1 (Multiple Choice Questions, 13 points)

For each of the following questions, circle the letter of your choice. There is only ONE correct choice unless explicitly mentioned. No explanation is required. There is no penalty for a wrong answer.

(a) **(1 point)** Consider a Generative Adversarial Network (GAN) which successfully produces images of apples. Which of the following propositions is **false**?

(i) The generator aims to learn the distribution of apple images.

(ii) The discriminator can be used to classify images as apple vs. non-apple.

(iii) After training the GAN, the discriminator loss eventually reaches a constant value.

(iv) The generator can produce unseen images of apples.

(b) **(1 point)** Which of the following activation functions can lead to vanishing gradients?

(i) ReLU

(ii) Tanh

(iii) Leaky ReLU

(iv) None of the above

(c) **(1 point)** Consider a univariate regression $\hat{y} = wx$ where $w \in \mathbb{R}$, and $x \in \mathbb{R}^{1\times m}$. The cost function is the squared-error cost $\mathcal{J} = \frac{1}{m} \parallel \hat{y} - y \parallel^2$. Which of the following equations is **true**?

(i) $\frac{\partial \mathcal{J}}{\partial w} = \frac{1}{m}(\hat{y} - y)x^T$

(ii) $\frac{\partial \mathcal{J}}{\partial w} = \frac{1}{m}(\hat{y} - y)x$

(iii) $\frac{\partial \mathcal{J}}{\partial w} = \frac{2}{m}(\hat{y} - y)x^T$

(iv) $\frac{\partial \mathcal{J}}{\partial w} = \frac{2}{m}(\hat{y} - y)x$

(d) **(1 point)** Which of the following costs is the non-saturating generator cost for GANs (G is the generator and D is the discriminator)?

(i) $J^{(G)} = \frac{1}{m}\sum_{i=1}^{m} \log\big(1 - D(G(z^{(i)}))\big)$

(ii) $J^{(G)} = -\frac{1}{m}\sum_{i=1}^{m} \log\big(D(G(z^{(i)}))\big)$

(iii) $J^{(G)} = \frac{1}{m}\sum_{i=1}^{m} \log\big(1 - G(D(z^{(i)}))\big)$

(iv) $J^{(G)} = -\frac{1}{m}\sum_{i=1}^{m} \log\big(G(D(z^{(i)}))\big)$

(e) **(1 point)** After training a neural network, you observe a large gap between the training accuracy (100%) and the test accuracy (42%). Which of the following methods is commonly used to reduce this gap?

(i) Generative Adversarial Networks

(ii) Dropout

(iii) Sigmoid activation

(iv) RMSprop optimizer

(f) **(1 point)** Which of the following is **true** about Batchnorm?

(i) Batchnorm is another way of performing dropout.

(ii) Batchnorm makes training faster.

(iii) In Batchnorm, the mean is computed over the features.

(iv) Batchnorm is a non-linear transformation to center the dataset around the origin

(g) **(1 point)** Which of the following statements is **true** about Xavier Initialization?

(i) It is only used in fully connected neural networks.

(ii) It applies a scaling factor to the mean of the random weights.

(iii) It is commonly used in logistic regression.

(iv) The assumptions made are only valid at the beginning of training.

(h) **(1 point)** When should multi-task learning be used?

(i) When your problem involves more than one class label.

(ii) When two tasks have the same dataset.

(iii) When you have a small amount of data for a particular task that would benefit from the large dataset of another task.

(iv) When the tasks have datasets of different formats (text and images).

(i) **(1 point)** Which of the following is an advantage of end-to-end learning? **(Check all that apply.)**

(i) It usually requires less data.

(ii) It doesn't need hand crafted features.

(iii) It generally leads to lower bias.

(iv) None of the above.

(j) **(2 points)** Which of the following propositions are true about a CONV layer? **(Check all that apply.)**

(i) The number of weights depends on the depth of the input volume.

(ii) The number of biases is equal to the number of filters.

(iii) The total number of parameters depends on the stride.

(iv) The total number of parameters depends on the padding.

(k) **(1 point)** What is Error Analysis?

(i) The process of analyzing the performance of a model through metrics such as precision, recall or F1-score.

(ii) The process of scanning mis-classified examples to identify weaknesses of a model.

(iii) The process of tuning hyperparameters to reduce the loss function during training.

(iv) The process of identifying which parts of your model contributed to the error.

(l) **(1 point)** Which of the following is a non-iterative method to generate adversarial examples?

(i) Non-Saturating Cost Method

(ii) Input Optimization Method

(iii) Adversarial Training

(iv) Logit Pairing

(v) Fast Gradient Sign Method

(vi) Real-time Cryptographic Dropout Method

Question 2 (Short Answer Questions, 23 points)

Please write concise answers.

(a) **(2 points)** How does splitting a dataset into train, dev and test sets help identify overfitting?

(b) **(2 points)** Which regularization method leads to weight sparsity? Explain why.

(c) **(2 points)** You are designing a deep learning system to detect driver fatigue in cars. It is crucial that that your model detects fatigue, to prevent any accidents. Which of the following is the most appropriate evaluation metric: Accuracy, Precision, Recall, Loss Value. Explain your choice.

(d) **(4 points)** You have a single hidden-layer neural network for a binary classification task. The input is $X \in \mathbb{R}^{n \times m}$, output $\hat{y} \in \mathbb{R}^{1 \times m}$ and true label $y \in \mathbb{R}^{1 \times m}$. The forward propagation equations are:

$$
\begin{aligned}
z^{[1]} &= W^{[1]} X + b^{[1]} \\
a^{[1]} &= \sigma(z^{[1]}) \\
\hat{y} &= a^{[1]} \\
\mathcal{J} &= -\sum_{i=1}^{m} y^{(i)} \log(\hat{y}^{[i]}) + (1 - y^{(i)}) \log(1 - \hat{y}^{[i]})
\end{aligned}
$$

Write the expression for $\frac{\partial \mathcal{J}}{\partial W^{[1]}}$ as a matrix product of two terms.

(e) **(3 points)** You want to solve a classification task. You first train your network on 20 samples. Training converges, but the training loss is very high. You then decide to train this network on 10,000 examples. Is your approach to fixing the problem correct? If yes, explain the most likely results of training with 10,000 examples. If not, give a solution to this problem.

(f) **(2 points)** Give two benefits of using convolutional layers instead of fully connected ones for visual tasks.

(g) **(2 points)** You have a dataset $\mathcal{D}_1$ with 1 million labelled training examples for classification, and dataset $\mathcal{D}_2$ with 100 labelled training examples. Your friend trains a model from scratch on dataset $\mathcal{D}_2$. You decide to train on $\mathcal{D}_1$, and then apply transfer learning to train on $\mathcal{D}_2$. State one problem your friend is likely to find with his approach. How does your approach address this problem?

(h) **(2 points)** You are solving the binary classification task of classifying images as cat vs. non-cat. You design a CNN with a single output neuron. Let the output of this neuron be z. The final output of your network, $\hat{y}$ is given by:

$$\hat{y} = \sigma(ReLU(z))$$

You classify all inputs with a final value $\hat{y} \geq 0.5$ as cat images. What problem are you going to encounter?

(i) **(2 points)** You are given a content image X_C and a style image, X_S. You would like to apply neural style transfer to obtain an output image Y, with the content of X_C and the style of X_S, as discussed in section. You are told that you need a pretrained VGG-16 network to do this. What is the function of this pretrained network?

(j) **(2 points)** You are given the following piece of code for forward propagation through a single hidden layer in a neural network. This layer uses the sigmoid activation. Identify and correct the error.

```
import numpy as np

def forward_prop(W, a_prev, b):
    z = W*a_prev + b
    a = 1/(1+np.exp(-z)) #sigmoid
    return a
```

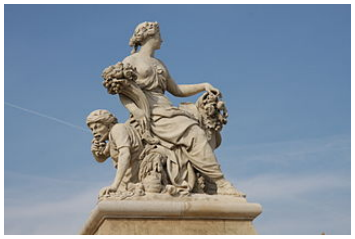
Question 3 (Convolutional Neural Networks, 28 points)

Two historians approach you for your deep learning expertise. They want to classify images of historical objects into 3 classes depending on the time they were created:

- Antiquity ($y = 0$)
- Middle Ages ($y = 1$)
- Modern Era ($y = 2$)

(A) Class: Antiquity

(B) Class: Middle Ages



(C) Class: Modern Era

Figure 1: Example of images found in the dataset along with their classes

(a) Over the last few years, the historians have collected nearly 5,000 hand-labelled RGB images.

(i) **(2 points)** Before training your model, you want to decide the image resolution to be used. Why is the choice of image resolution important?

(ii) **(1 point)** If you had 1 hour to choose the resolution to be used, what would you do?

(b) You have now figured out a good image resolution to use.

(i) **(2 points)** How would you partition your dataset? Formulate your answer in percentages.

(ii) **(3 points)** After visually inspecting the dataset, you realize that the training set only contains pictures taken during the day, whereas the dev set only has pictures taken at night. Explain what is the issue **and** how you would correct it.

(c) **(3 points)** As you train your model, you realize that you do not have enough data. Cite 3 data augmentation techniques that can be used to overcome the shortage of data.

(d) **(8 points)** You come up with a CNN classifier. For each layer, calculate the number of weights, number of biases and the size of the associated feature maps.

The notation follows the convention:

- CONV-K-N denotes a convolutional layer with N filters, each them of size $K \times K$, Padding and stride parameters are always 0 and 1 respectively.
- POOL-K indicates a $K \times K$ pooling layer with stride K and padding 0.
- FC-N stands for a fully-connected layer with N neurons.

Layer	Activation map dimensions	Number of weights	Number of biases
INPUT	$128 \times 128 \times 3$	0	0
CONV-9-32			
POOL-2			
CONV-5-64			
POOL-2			
CONV-5-64			
POOL-2			
FC-3			

(e) **(2 points)** Why is it important to place non-linearities between the layers of neural networks?

(f) **(3 points)** Following the last FC-3 layer of your network, what activation must be applied? Given a vector $a = [0.3, 0.3, 0.3]$, what is the result of using your activation on this vector?

(g) You find online that the exact same network has already been trained on 1,000,000 historical objects from a slightly different time period.

(i) **(1 point)** What is the name of the method that could reuse these pretrained weights for the task at hand?

(ii) **(3 points)** What are the new hyperparameters to tune for this method?

Question 4 (Adversarial Attacks, 10 points)

Alice and Bob work on a self-driving car project. They want to classify various traffic signs among 10 different classes. Bob has trained a deep convolutional neural network (CNN), f, on a dataset with 100,000 samples. Given an input image $\mathbf{x}$, his model predicts $\hat{y} = f(x)$. Overall, it achieves 95.6% test accuracy.

(a) **(2 points)** Alice has recently heard about adversarial attacks and is worried about the problems they could cause. To show Bob the potential dangers of adversarial attacks, she decides to design an input $\mathbf{x}$ which is classified as a "STOP" sign by Bob's CNN. Propose a loss function for this task, and explicitly state the parameter(s) being optimized. You are not allowed to use any images other than $\mathbf{x}$ for this optimization.

(b) **(2 points)** You run the optimization in part (a). Will the generated image look like a real image? Explain why.

(c) **(3 points)** Alice looks for better evidence to convince Bob that his trained CNN is not a robust classifier. She decides to take the image $\mathbf{x}_{\text{no park}}$, which is a real image of a "No Parking" sign, and finds an input $\mathbf{x}$ such that:

- $\mathbf{x}$ looks similar to $\mathbf{x}_{\text{no park}}$

Figure 2: Input $\mathbf{x}_{\text{no park}}$

- $\mathbf{x}$ is classified by Bob's network as a "STOP" sign. i.e. $f(\mathbf{x}) = \hat{y}_{\text{STOP}}$

Give the cost function for an iterative method which will achieve the above two objectives.

(d) **(3 points)** After seeing the results of Alice's experiments, Bob decides to retrain the deep convolutional network in a way that the trained classifier would be robust to adversarial attacks. Suggest two different solutions for improving the robustness of his CNN classifier.

Question 5 (Loss comparisons, 23 points)

Part I. You want to perform a regression task with the following dataset: $x^{(i)} \in \mathbb{R}$ and $y^{(i)} \in \mathbb{R}$, $i = 1, \ldots, m$ are the the ith example and output in the dataset, respectively. Denote the prediction for example i by $f(x^{(i)})$. Remember that for a given loss $\mathcal{L}$ we minimize the following cost function

$$\mathcal{J} = \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(f(x^{(i)}), y^{(i)}).$$

In this part we are deciding between using loss 1 and loss 2, given by:

$$\mathcal{L}_1(f(x^{(i)}), y^{(i)}) = |y^{(i)} - f(x^{(i)})|,$$
$$\mathcal{L}_2(f(x^{(i)}), y^{(i)}) = (y^{(i)} - f(x^{(i)}))^2.$$

(a) **(4 points)** Draw $\mathcal{L}_1(x, 0)$ and $\mathcal{L}_2(x, 0)$ versus $x \in \mathbb{R}$ on the same plot.

(b) **(2 points)** An outlier is a datapoint which is very different from other datapoints of the same class. Based on your plots, which method do you think works better when there is a large number of outliers in your dataset? Hint: Contributions of outliers to gradient calculations should be as small as possible.

(c) **(3 points)** "Using $\mathcal{L}_1$ loss enforces sparsity on the weights of the network." Do you agree with this statement? Why/Why not?

(d) **(3 points)** "Using $\mathcal{L}_2$ loss forces the weights of the network to end up small." Do you agree with this statement? Why/Why not?

Part II. You want to perform a classification task. You are hesitant between two choices: Approach A and Approach B. The only difference between these two approaches is the loss function that is minimized. Assume that $x^{(i)} \in \mathbb{R}$ and $y^{(i)} \in \{+1, -1\}$, $i = 1, \ldots, m$ are the ith example and output label in the dataset, respectively. $f(x^{(i)})$ denotes the output of the classifier for the ith example. Recall that for a given loss $\mathcal{L}$ you minimize the following cost function:

$$\mathcal{J} = \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(f(x^{(i)}), y^{(i)}).$$

As we mentioned, the only difference between approach A and approach B is the choice of the loss function:

$$\mathcal{L}_A(f(x^{(i)}), y^{(i)}) = \max\{0, 1 - y^{(i)} f(x^{(i)})\},$$
$$\mathcal{L}_B(f(x^{(i)}), y^{(i)}) = \log_2 \left(1 + \exp\left(-y^{(i)} f(x^{(i)})\right)\right).$$

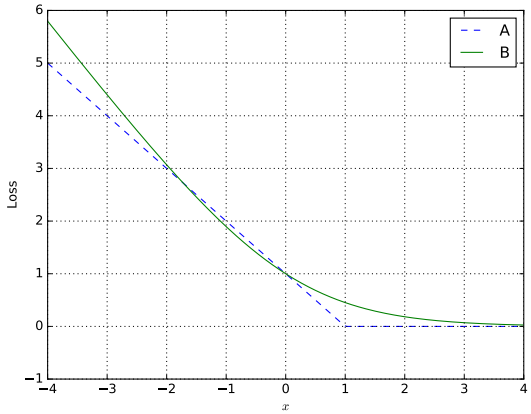
(e) Consider $\mathcal{L}_{\mathcal{B}}$.

(i) **(2 points)** Rewrite $\mathcal{L}_{\mathcal{B}}$ in terms of the sigmoid function.

(ii) **(2 points)** You are given an example with $y^{(i)} = -1$. What value of $f(x^{(i)})$ will minimize $\mathcal{L}_{\mathcal{B}}$?

(iii) **(2 points)** You are given an example with $y^{(i)} = -1$. What is the greatest value of $f(x^{(i)})$ that will minimize $\mathcal{L}_{\mathcal{A}}$?

The following is the plot of $L_A(x, 1)$ and $L_B(x, 1)$ versus $x \in \mathbb{R}$.



(f) **(2 points)** Based on these plots, which method would work better when there is a large number of outliers in the dataset? Explain why.

(g) **(3 points)** You would like a classifier whose output can be interpreted as a probability. Which loss function is better to optimize? Justify your choice.

Question 6 (The Optimizer, 20 points)

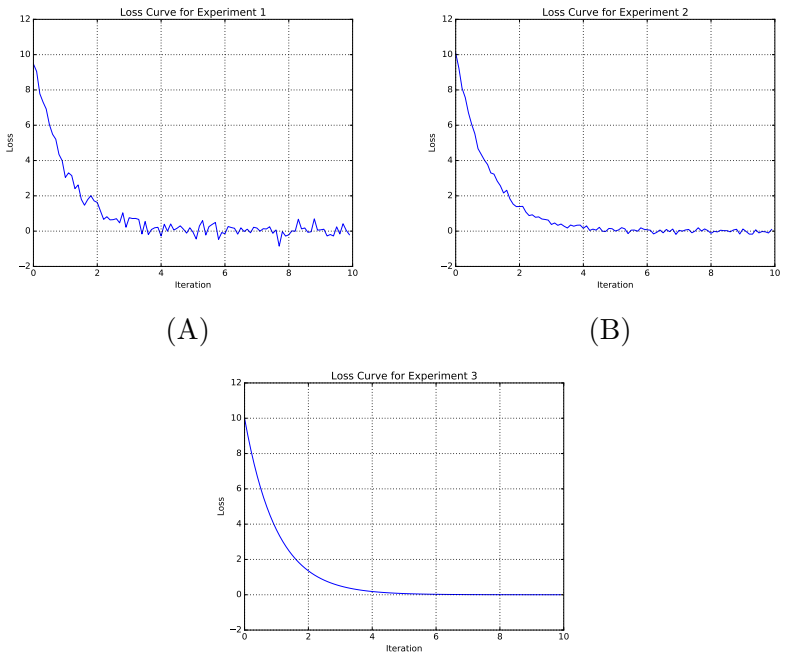


Figure 3: Loss curves for Experiment A, Experiment B, and Experiment C

You are given a black box optimizer which produces the loss curve shown in Figure 3A. You see a big red button on the optimizer and decide to push it. After doing this, you notice the loss curve shown in Figure 3B. You press the button one more time and finally notice the loss curve shown in Figure 3C.

(a) The red button modifies a single hyperparameter.

(i) **(1 point)** Which hyperparameter is likely to be modified by pressing the button?

(ii) **(1 point)** Of experiments A, B and C, which corresponds to largest magnitude of the hyperparameter?

(iii) **(2 points)** The loss curve for experiment C seems to be the most desirable. Despite this, give two reasons why you would choose the hyperparameter in experiment B for training your model.

You're tired of using the black box and decide instead to use an Adam Optimizer.

For the following questions, assume that each experiment begins with the same initial weights. You keep the default values of β_1 (0.9) and β_2 (0.999), and run the optimization for 100 updates. The distribution of gradient updates $\left(\frac{m_t}{\sqrt{s_t}+\epsilon}\right)$ is shown in Figure 4.

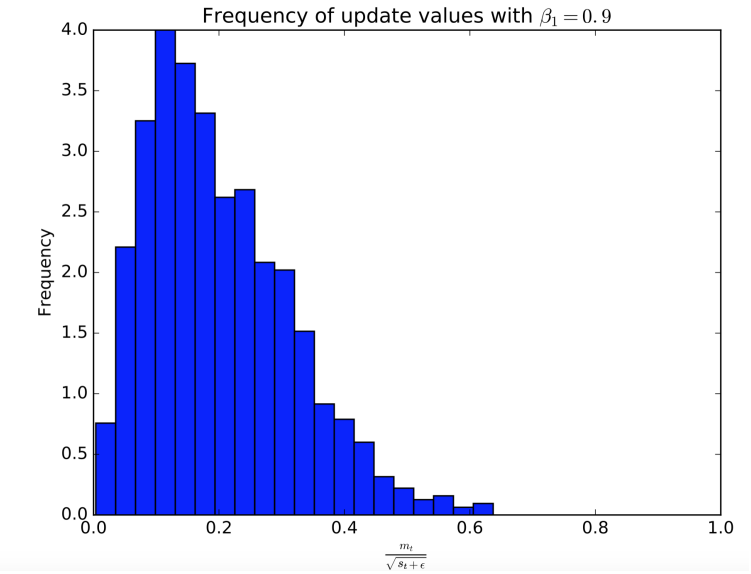


Figure 4: Distribution of updates with $\beta_1 = 0.9$.

Note that the area under the curve for a distribution like this is always 1.

(b) **(3 points)** You change the value of β_1 from the default 0.9 to 0.4, while keeping β_2 at 0.999 and run the optimization for 100 steps.
Which of the following are likely attributes of the curve to change in this new experiment?

- The peak value
- The width (the spread)

Justify.

(c) Given the gradient calculated at a point, the Adam optimizer has three distinct steps. First, update the moving averages. Second, apply the bias correction. Third, update the parameters.
Consider the moving average of the square of the gradients. It is given by the recursive formula:

$$s_t = \beta_2 s_{t-1} + (1 - \beta_2) g_t^2$$

(i) **(4 points)** Write down the expression for s_t only in terms of the gradients $g_0, g_1, \ldots, g_t$

(ii) **(5 points)** Given your expression in part (i), what is $\mathrm{E}[s_t]$ in terms of $\mathrm{E}[g_t^2]$ and β_2? You may assume that g_i's are independent and identically distributed. The following formulas might be useful:

$$\sum_{i=0}^{n-1}(a+id) = \frac{n}{2}\left(2a+(n-1)d\right)$$

$$\sum_{i=0}^{n-1} ar^i = \frac{a(r^n-1)}{r-1}$$

(iii) **(2 points)** Using your result in the previous part, explain what would happen if you did not perform the bias correction step.

(d) **(2 points)** You have two weights w_1 and w_2 within a model such that initially, $|w_1| >> |w_2|$. What can you say about the magnitude of $\frac{m_t}{\sqrt{s_t+\epsilon}}|_{w_1}$ compared to the magnitude of $\frac{m_t}{\sqrt{s_t+\epsilon}}|_{w_2}$ for the first update (greater than, equal to, less than)? Explain.

Extra Page 1/3

Extra Page 2/3

Extra Page 3/3

END OF PAPER