CS 6375 **Name (Print):** _____
Spring 2020
Final Exam
5/5/2020 - 5/8/2020

This exam contains 6 pages (including this cover page) and 4 problems.
You may **NOT** collaborate with other any other person on this exam. Examinees found to be collaborating with other students or copying solutions from electronic sources will receive a zero on the exam and face possible disciplinary action.

The following rules apply:

- **Organize your work**, in a reasonably neat and coherent way, in the space provided. Work scattered all over the page without a clear ordering will receive very little credit.
- **To ensure maximum credit** on short answer / algorithmic questions, be sure to **EXPLAIN** your solution.
- **Problems/subproblems** are not ordered by difficulty.
- **Do not** write in the table to the right.

Problem	Points	Score
1	20	
2	25	
3	30	
4	25	
Total:	100	

1. **Missing Entries:** Suppose that you are given a symmetric matrix $A \in \mathbb{R}^{n \times n}$ that is missing some entries, e.g., $A_{ij} =?$ for some indices $i, j \in \{1, \ldots, n\}$. To determine which entries are missing, we will use an index matrix $Q \in \{0, 1\}^{n \times n}$ such that $Q_{ij} = 1$ if $A_{ij} =?$ and $Q_{ij} = 0$ otherwise.

   (a) (3 points) Suppose $A = \begin{bmatrix} 3 & 1 & ? \\ 1 & 1 & -2 \\ ? & -2 & 6 \end{bmatrix}$. Find a matrix B such that B is positive semidefinite and that the Frobenius norm, $||A - B||_F^2 = \sum_{i,j}(A_{ij} - B_{ij})^2$, is as small as possible.

   (b) (3 points) Any symmetric positive semidefinite matrix $B \in \mathbb{R}^{n \times n}$ can be written as $B_{ij} = {v^{(i)}}^T v^{(j)}$ for some $v^{(1)}, \ldots, v^{(n)} \in \mathbb{R}^n$. Using this observation, formulate the problem of filling in the missing entries of A as using the same strategy as part (a) as an optimization problem.

   (c) (10 points) The optimization problem in part (b) is not convex. Describe, in detail, a block coordinate descent scheme for your objective in part (b) that has the property that the optimization problem over each block is convex (recall that the EM algorithm enjoys a similar property). For full credit, you should prove that your algorithm has the desired property.

   (d) (4 points) Explain why adding an ℓ_2 penalty to the optimization problem in part (b) might be important in practice. Rewrite the optimization problem to include it.

2. **Neural Networks :**

   (a) (10 points) Derive the backpropagation algorithm in the case of a feedforward neural network with softplus activation functions and a squared loss function. For full credit, you should explicitly compute the derivatives as we did in class.

   (b) (15 points) Suppose that we want to fit a conditional mixture model for a regression task. That is, given data points $x^{(1)}, \ldots, x^{(M)} \in \mathbb{R}^n$ with corresponding labels $y^{(1)}, \ldots, y^{(M)} \in \mathbb{R}$, we would like to fit a model of the form

$$p(y|x, \lambda, \theta_1, \ldots, \theta_K) = \sum_{k=1}^{K} \lambda_k p(y|x, \theta_k),$$

   where $p(y|x, \theta_k)$ is a conditional probability distribution parameterized by θ_k.
   Suppose we choose these distributions to be of the form

$$p(y|x, \theta_k) = \mathcal{N}(y; NN_k(x|\theta_k)),$$

   where $NN_k(x)$ is a neural network, whose parameters are given by θ_k, that takes as input x and returns a pair of outputs μ_k and $\sigma_k^2 > 0$, which act as the mean and variance of the k^{th} normal distribution.

   1. Explain how to apply the EM algorithm to fit this conditional mixture model to data. For full credit, your explanation should provide sufficient algorithmic details.
   2. Give an example to illustrate why you might want to apply models of this form in practice. Describe how you might structure the neural networks for this application.

3. **Rolling Dice:** Consider the following maximum likelihood estimation problems.

   (a) (5 points) Suppose you are given a collection of observed 4-sided die rolls (i.e., the numbers one through 4) that were generated by picking one of two die (A or B) with respect to some probability distribution and then rolling it. If, in addition to the observed die rolls, you are also told which die was rolled, write the formula for the log-likelihood, and compute the maximum likelihood estimate of the parameters using the following sequence of observations:

Outcome	Observed Die
1	A
2	A
3	A
2	A
4	A
1	B
4	B
4	B
4	B
4	B
3	B
2	B
3	B
1	B

   (b) Now suppose that you are given a sequence of data observations generated by rolling k, 6-sided dice for some $k \in \{1, 2, 3\}$ and summing up their pips, e.g., if $k = 2$ and die one is a 5 and die two is a six then the observation would be 11. Suppose that each of the k die are identical, i.e., the probability of observing n pips, denoted θ_n, for $n \in \{1, \ldots, 6\}$ is the same for each die. However, k is unknown and must be estimated from data.

      1. (10 points) Write the formula for the log-likelihood and compute the maximum likelihood estimate for k and θ given the following data.

Observed Sum
5
9
10
12
12
12
9
12
12
10

      2. (5 points) What would be a good choice of prior be for θ? What is the MAP estimate for the above data under this prior?

(c) (10 points) Finally, consider data generated by first flipping a coin with bias b. If the coin comes up heads, then a possibly loaded dice $D1$ is rolled twice. If the coin comes up tails, then a possibly loaded dice $D2$ is rolled twice. Given that you only observe the sum of whichever die is rolled, explain how to estimate the bias of the coin and the probability distribution of the outcomes of each of the loaded dice using the EM algorithm. For full credit, you should completely specify the E and the M steps.

4. **Dots and Boxes:** Consider the classic children's game in which, starting from a 9×9 grid of evenly spaced dots, two players alternatively take turns adding either a vertical or horizontal line between neighboring dots. Whenever a player places a line that completes a box around four neighboring dots, the player scores a single point and gets to take another turn. The game ends when no more lines can be placed. The aim is to be the player that completes the most boxes / scores the most points. For more details and a visualization, see the Wikipedia page for Dots-and-Boxes.

   (a) (15 points) Describe a Markov decision process for an agent acting as a player in this game whose aim is to win the game. You should provide the states, actions, transition function, and reward function.

   (b) (5 points) Given your MDP in part (a), does the choice of discount factor, i.e., $\gamma \in (0, 1)$, matter in terms of the policy that maximizes the cumulative reward function?

   (c) (5 points) Explain, in detail, how to apply deep Q-learning with a neural network to approximate the Q-value function. In particular, you should describe the neural network and how you can use it to find the "optimal" action, with respect to its Q-value, for a given state.