ENM 531 Midterm

Instructions: Please submit your answers on Canvas by 12pm ET, Saturday, April 3, 2021. Acceptable formats include Jupyter notebook, or LaTeX compiled .pdf. Handwritten submissions will not be accepted! The exam contains 9 questions adding up to 100 points + plus 10 bonus points. You can use Piazza for clarifications.

The exam is designed in a way that you don't need to consult any external resources (during normal times this would be an in-class exam with no access to books or notes), but now you are free to use any resources you want (since there is no possible way for us to enforce this rule). However, no plagiarism or collaboration is permitted and we'll be sure to keep an eye for those.

Question 1 (10pts)

Please answer the following True or False questions with some explanation.

0.1 (a) (3pts)

"After your training is done, the testing error appears less than your training error."

0.2 (b) (3pts)

"Increase the number of training examples in logistic regression will eventually decrease the Bias and increase the Variance of the predictions."

0.3 (c) (4pts)

"L_2 regularization is having better property to impose sparsity than L_1 regularization."

Question 2 (10pts)

A metric on a set X is a function (called the distance function or simply distance):

$$d : \mathcal{X} \times \mathcal{X} \to [0, +\infty)$$

where $[0, +\infty)$ is the set of non-negative real numbers and for all $x, y, z \in \mathcal{X}$ the following conditions are satisfied:

- $d(x, y) \geq 0$
- $d(x, y) = 0 \Leftrightarrow x = y$

- $d(x, y) = d(y, x)$
- $d(x, y) \leq d(x, z) + d(y, z)$

(a) (3pts) Show that $d(x, y) = \sqrt{|x - y|}$ defines a metric in the set of real numbers.
(b) (3pts) Does $d(x, y) = (x - y)^2$ define a metric on the set of real numbers? Why?
(c) (4pts) Does KL divergence define a metric in probability space? Why?

Question 3 (15pts)

Your training set is $\mathcal{D} = \{(x^{(1)}, y^{(1)}), (x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\}$. Assume that your model for the data is

$$y^{(i)} \sim \text{Laplace}(\theta^T x^{(i)}, 1).$$

The probability density function of the Laplace distribution with mean μ and scale parameter b is given by:

$$p(x|\mu, b) = \frac{1}{2b} \exp(\frac{-|x - \mu|}{b}).$$

(1) (5pts) Derive the loss function you would use to train this model using maximum likelihood estimation (MLE) on the training data-set $\mathcal{D}$.
(2) (5pts) Considering a zero-mean Laplace prior for the model parameters, i.e. $p(\theta) = \lambda \exp(-\lambda|\theta|)$, derive the loss function you would use to perform maximum a-posteriori (MAP) estimation.
(3) (5pts) Derive the Gradient Descent update rule for the loss corresponding to the MAP estimate.

Question 4 (15pts)

Consider a $\{0/1\}$ problem, where you have totally n_B observations and each them should take value from $\{0, 1\}$ (like throwing a coin). The number of positive sides y_B (that you get 1 in the trial) could be represented by Binomial distribution $p(y_B|\theta_B)$ where θ_B is the Bernoulli probability. You can consider a Beta distribution as the prior of the θ_B.
(1) (3pts) Make some comments on why Beta distribution may be a good choice for this problem.
(2) (3pts) Write down the expression for the conditional distribution of $p(y_B|\theta_B)$ and the prior $p(\theta_B)$.
(3) (4pts) Derive the posterior distribution of θ_B.
(4) (5pts) Consider a special case: $n_B = 40$ an $y_B = 0$. What is your estimation of θ_B if you are about to use maximum likelihood estimation? What is the confidence interval of that estimation? Do you see the problem?

Hint: Some useful formulas for the Gamma and Beta functions:

$$\Gamma(x) = \int_0^\infty t^{x-1}\exp(-t)dt, \tag{1}$$

$$\Gamma(x+1) = x! \quad \text{when} \quad x \in N^+, \tag{2}$$

$$\mathrm{B}(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt, \tag{3}$$

$$\mathrm{B}(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} \tag{4}$$

Hint: Some useful distributions:

$$\text{Beta distribution pdf Beta(x; a, b)=} \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} \tag{5}$$

$$\text{Gamma distribution pdf Gamma(x; a, b)=} \frac{b^a}{\Gamma(a)}x^{a-1}\exp(-bx) \tag{6}$$

Question 5 (10pts)

(10 pts) Please list out some different techniques you know for regularizing the training of deep neural networks (at least 4 and with some comments).

Question 6 (10pts)

Consider a pair of random variables X and Y whose joint distribution is as follows:
(1) (5pts) Compute the joint entropy $\mathcal{H}(X,Y)$.

	$Y=0$	$Y=1$
$X=0$	0.4	0.3
$X=1$	0.2	0.1

(2) (5pts) Compute the conditional entropy $\mathcal{H}(Y|X)$.

Question 7 (10pts)

Consider the following convolutional layer in a conv net. The input has a spatial dimension 12×12, and has 10 channels. The convolution kernels are 3×3, the stride is 2, and we use "valid" convolution, which means that each output neuron only looks at image regions that lie entirely within the spatial bounds of the input. The output dimension is 5×5, with 20 channels.
For this question, you don't need to show your work or justify your answer, but doing so may help you get partial credit.
(1) (5pts) How many weights are required for this convolution layer?

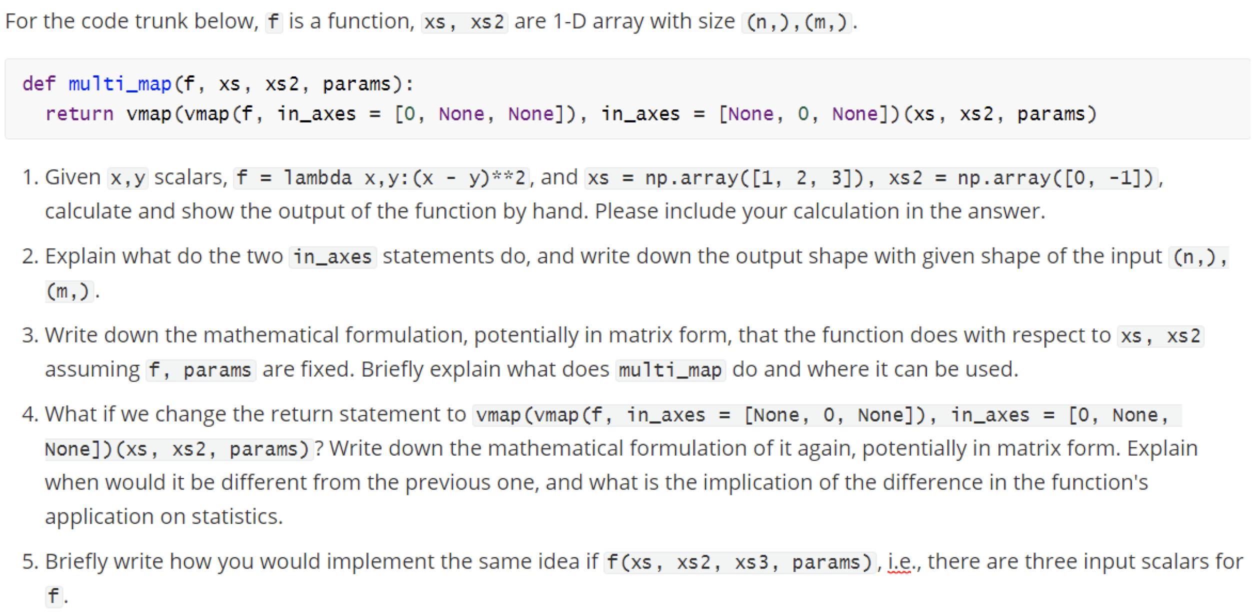
(2) (5pts) Suppose we instead make this a locally connected layer, i.e. don't use weight sharing. How many weights are required?

Question 8 (10pts)

Suppose if we want to perform variational inference on an arbitrary probability distribution $p(\boldsymbol{x})$ with and simple unit variance Gaussian distribution $q(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I})$ in R^n.
(1) (5pts) Please derive the KL divergence between $q(\boldsymbol{x})$ and $p(\boldsymbol{x})$ these two distributions and compute the differentiation with respect to μ.
(2) (5pts) Please prove that the optimal parameter is $\boldsymbol{\mu}^* = E(\boldsymbol{x})$ where the expectation of x is from $p(\boldsymbol{x})$.

Question 9 (20pts)

Please answer the questions in figure 1.



For the code trunk below, `f` is a function, `xs, xs2` are 1-D array with size `(n,),(m,)`.

```
def multi_map(f, xs, xs2, params):
  return vmap(vmap(f, in_axes = [0, None, None]), in_axes = [None, 0, None])(xs, xs2, params)
```

1. Given `x,y` scalars, `f = lambda x,y:(x - y)**2`, and `xs = np.array([1, 2, 3]), xs2 = np.array([0, -1])`, calculate and show the output of the function by hand. Please include your calculation in the answer.
2. Explain what do the two `in_axes` statements do, and write down the output shape with given shape of the input `(n,),` `(m,)`.
3. Write down the mathematical formulation, potentially in matrix form, that the function does with respect to `xs, xs2` assuming `f, params` are fixed. Briefly explain what does `multi_map` do and where it can be used.
4. What if we change the return statement to `vmap(vmap(f, in_axes = [None, 0, None]), in_axes = [0, None, None])(xs, xs2, params)`? Write down the mathematical formulation of it again, potentially in matrix form. Explain when would it be different from the previous one, and what is the implication of the difference in the function's application on statistics.
5. Briefly write how you would implement the same idea if `f(xs, xs2, xs3, params)`, i.e., there are three input scalars for `f`.

Figure 1: Coding question.