Problem Set 3

CS 4277: Deep Learning

Name (print clearly): _____ Section: (e.g., 01) _____

Signature: _____

Student account username (e.g., msmith3): _____

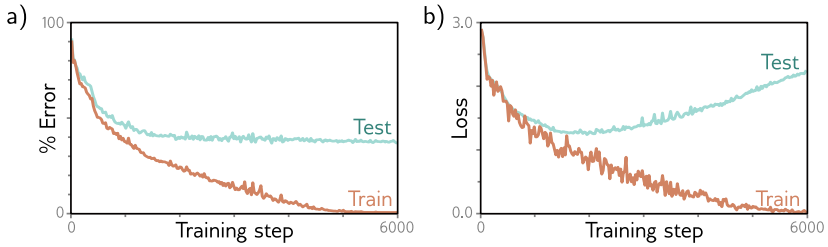
Signing signifies that you agree to comply with the **Academic Honor Code** and course policies stated in the syllabus.

Choose one of these two options for turn-in:

1. Print this document, write or answers, scan your finished homework to a PDF, name the PDF `cs4277-ps3-<your-student-account-username>.pdf`, e.g., `cs4277-ps3-msmith3.pdf` and submit the PDF to the assignment on D2L.
2. While viewing this document in your web browser, in the address bar change `.pdf` to `.tex`, save the LaTeX source as a text file, add your answers in appropriate LaTeX markup in the appropriate spaces, compile to a PDF named as in the instructions above, and submit the PDF file to the assignment on D2L.

Question:	1	2	3	4	Total
Points:	20	20	30	30	100
Score:					

1. (20 points) Problem 8.1 Will the multi-class cross-entropy training loss in Figure 8.2 (reproduced below) ever reach zero? Explain your reasoning.



2. (20 points) Problem 8.5 Consider the case where the model capacity exceeds the number of training data points, and the model is flexible enough to reduce the training loss to zero. What are the implications of this for fitting a heteroscedastic model? Propose a method to resolve any problems that you identify.

3. (30 points) Problem 9.1 Consider a model where the prior distribution over the parameters is a normal distribution with mean zero and variance σ_ϕ^2 so that

$$Pr(\phi) = \prod_{i=1}^{J} \text{Norm}_{\phi_j}(0, \sigma_\phi^2)$$

where j indexes the model parameters. When we apply a prior, we maximize $\prod_{i=1}^{I} Pr(\boldsymbol{y}_i|\boldsymbol{x}_i, \phi)Pr(\phi)$. Show that the associated loss function of this model is equivalent to L2 regularization (Equation 9.5, shown in the next question below).

4. (30 points) 9.2 How do the gradients of the loss function change when L2 regularization (Equation 9.5, reproduced below for convenience) is added?

$$\hat{\phi} = \underset{\phi}{\operatorname{argmin}} \Big(\sum_{i=1}^{I} \ell_i(\boldsymbol{x}_i, \boldsymbol{y}_i) + \lambda \sum_{j} \phi_j^2 \Big) \qquad \text{(Equation 9.5)}$$