ECE 239AS, Winter 2018 **Practice Midterm Questions**

Department of Electrical Engineering Prof. J.C. Kao
University of California, Los Angeles TAs T. Xing and C. Zheng

UCLA True Bruin academic integrity principles apply.
Open: book, notes, handout, computer.
Closed: internet, except for downloading / viewing from ECE239AS website / CCLE.
6:00 - 7:50pm.
Wednesday, 21 Feb 2017.

State your assumptions and reasoning.
No credit without reasoning.
Show all work on these pages.

Name: _____

Signature: _____

ID#: _____

Problem 1 _____ / X
Problem 2 _____ / X
Problem 3 _____ / X
Problem 4 _____ / X
Problem 5 _____ / X

Total _____ / X points

1. (X points) **Short answer on Machine Learning basics.**

   (a) (X points) **SVM basics.** In multi-class SVM, the minimization problem is:

   $$\arg\min \frac{1}{m} \sum_{i=1}^{m} \sum_{j \neq y^{(i)}} max(0, 1 + a_j(x^{(i)}) - a_{y^{(i)}}(x^{(i)}))$$

   Explain every term in the hinge loss. When does it equal to zero?

   (b) (X points) **Generalization.** What is the standard test for generalization? Given a dataset, how do you use it to train and test your machine learning model? Whats the risk with tuning hyperparameters using a test dataset?

   (c) (X points) **Bias and variance.** What is the bias and variance of a statistical estimator? What is the bias variance trade-off in machine learning? How does it relate with overfitting and underfitting?

2. (X points) **Mean-square error optimization with L2 regularization.** Consider the L2 regularized least-squares minimization problem.

   $$\mathcal{L} = \min_{\mathbf{W}} \sum_{i=1}^{N} (\mathbf{y}^{(i)} - \mathbf{W}\mathbf{x}^{(i)})^2 + \frac{1}{2}\lambda \left\|\mathbf{W}\right\|_F^2$$

   Find the $\mathbf{W}$ that minimizes this loss.

3. (X points) **Backpropagation.** Consider a neural network with the hyperbolic tangent activation function: $\mathbf{h} = \tanh(\mathbf{W}\mathbf{x} + \mathbf{b})$ with $\mathbf{x} \in \mathbb{R}^D$ as input and $\mathbf{h} \in \mathbb{R}^H$ as output. Square-loss is given by $\mathcal{L} = (\mathbf{h} - \mathbf{t})^2$.

   (a) (X points) Please write the dimension of following variables $\mathbf{W}, \mathbf{b}, \mathbf{t}$.

   (b) (X points) Please calculate $\frac{\partial \mathcal{L}}{\partial \mathbf{W}}$ and $\frac{\partial \mathcal{L}}{\partial \mathbf{x}}$. You can use the following formula without proof.

   $$\frac{\partial \tanh x}{\partial x} = 1 - \tanh^2 x$$
   $$\frac{\partial \mathcal{L}}{\partial \mathbf{X}} = \mathbf{W}^T \frac{\partial \mathcal{L}}{\partial (\mathbf{W}\mathbf{X})}$$
   $$\frac{\partial \mathcal{L}}{\partial \mathbf{W}} = \frac{\partial \mathcal{L}}{\partial (\mathbf{W}\mathbf{x})} \mathbf{x}^T$$

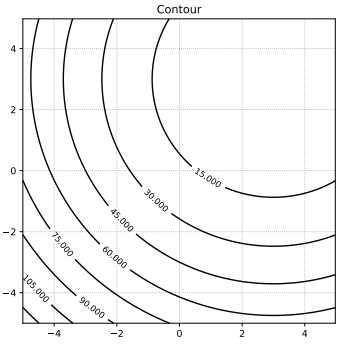
4. (X points) Optimization (momentum, RMSprop, adam).

   (a) (X points) **Momentum.** What is the difference between momentum and Nesterov momentum?

   (b) (X points) **Optimization example.** The following figure is the contour plot of $\mathcal{L} = (\mathbf{X} - 3)^2 + (\mathbf{Y} - 3)^2$.

   If we are trying to find global minimum of this function with optimizers *Momentum*, *RMSprop*, *Adam*. The initialization of parameters is $(\mathbf{X}, \mathbf{Y}) = (0, 0)$. Optimizer hyperparameters are:

   i. Momentum: $\alpha = 0.9, \epsilon = 0.1$



ii. RMSprop: $\beta = 0.9, \epsilon = 0.1$

iii. Adam: $\beta_1 = \beta_2 = 0.9, \epsilon = 0.1$

Please give two step of parameter update for optimizer *Momentum*, *RMSprop* and one step of parameter update for optimizer *Adam*.

5. (a) (X points) **Ensemble averaging.**

   i. (X points) Your colleague wants to use ensemble averaging to improve her validation accuracy. She trains 7 models and reports her validation accuracy vs a single model she trained with the same data. The models are all trained appropriately. The model prediction errors are identically distributed with zero mean. Surprisingly, she finds the 7 models performs worse than the 1 model according to her validation metric. She accuses you saying: "You told me that the average validation error should go down even if the models were partially correlated!" In light of her results, was your statement truthful? Justify your answer (*at most* three sentences).

   ii. (X points) What may be one reason why your colleague observed a poorer validation accuracy by averaging the results of 7 models compared to using 1 model? Justify your answer (*at most* three sentences).

   (b) (X points) **Dropout.** Consider a feedforward neural network using the sigmoid activation function. Imagine that it was trained via dropout (note: NOT inverted dropout) at $p = 0.5$, but at test time, you do not perform any modifications to the testing phase. Will the network perform well, and if not, what might be a problem with this approach?

   (c) (X points) **Batchnorm.** What is the purpose of the γ and β parameters in batch normalization?

1.

Solution:

(a) $a_i(x)$ is the score of the input data x being in class i. The 1 term is the margin in which, even if the classification is correct, we'll incur a small loss due to being within a margin distance to the separating hyperplane. It equals to zero only when there is enough margin between score of correct label and other labels.

(b) In practice, we can check the generalization by cross validation (n-fold); We split the dataset into 3 parts: training set, validation set, and testing set. We train the network model on training set and use the validation to pick our best hyperparameters as

well as monitor generalization error. After training and validation, we could then test our model on testing set. Tuning model hyperparameters to a test set means that the hyperparameters may overfit to that test set. If the same test set is used to estimate performance, it will produce an overestimate. Using a separate validation set for tuning and test set for measuring performance provides unbiased, realistic measurement of performance.

(c) The bias is the expected deviation of the estimated parameters from the true parameters, and the variance is the distribution of the estimated parameters. In the bias-variance tradeoff, we want to make both of these as small so as to not underfit the data (large bias) or overfit the data (large variance).

2.

Solution: We already know the derivative of the non-L2 regularizer component of the loss from HW #1, which resulted in $-\mathbf{Y}\mathbf{X}^T + \mathbf{W}\mathbf{X}\mathbf{X}^T$. (For notation, look at HW #1 solutions). With the L2 regularizer, we add to the derivative $\lambda\mathbf{W}$. Setting this equal to zero, we get:

$$-\mathbf{Y}\mathbf{X}^T + \mathbf{W}\mathbf{X}\mathbf{X}^T + \lambda\mathbf{W} = 0$$

Rearranging, we have that:

$$\mathbf{Y}\mathbf{X}^T = \mathbf{W}(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})$$

and hence,

$$\mathbf{W} = \mathbf{Y}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}$$

3.

Solution:

(a) $\mathbf{W} \in \mathbb{R}^{H\times D}, \mathbf{b} \in \mathbb{R}^H, \mathbf{t} \in \mathbb{R}^H$

(b) let $\mathbf{a} = \mathbf{h} - \mathbf{t}$ and $\mathbf{m} = \mathbf{W}\mathbf{x} + \mathbf{b}$. So

$$\frac{\partial \mathcal{L}}{\partial \mathbf{a}} = 2\mathbf{a}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{h}} = 2\mathbf{a}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{m}} = (1 - \tanh^2 \mathbf{m}) \odot 2\mathbf{a}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{x}} = \mathbf{W}^T[(1 - \tanh^2 \mathbf{m}) \odot 2\mathbf{a}]$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}} = [(1 - \tanh^2 \mathbf{m}) \odot 2\mathbf{a}]\mathbf{x}^T$$

4.

Solution:

(a) The gradient w.r.t. θ is different for these two methods. The Nesterov momentum uses gradient calculated at the parameter setting after taking a step along the direction of the momentum.

(b) First, the gradients respective to parameters are $\frac{\partial \mathcal{L}}{\partial \mathbf{X}} = 2(\mathbf{X} - 3)$ and $\frac{\partial \mathcal{L}}{\partial \mathbf{Y}} = 2(\mathbf{Y} - 3)$.

i. *Momentum* $\mathbf{v} \leftarrow \alpha\mathbf{v} - \epsilon\mathbf{g}$ and $\theta \leftarrow \theta + \mathbf{v}$

A. Step 1: $\mathbf{g} = (-6, -6), \mathbf{v} = (0.6, 0.6)$, so $\theta = (0.6, 0.6)$

B. Step 2: $\mathbf{g} = (-4.8, -4.8), \mathbf{v} = (1.02, 1.02)$, so $\theta = (1.62, 1.62)$

ii. *RMSprop* $\mathbf{a} \leftarrow \beta\mathbf{a} + (1-\beta)\mathbf{g} \odot \mathbf{g}$ and $\theta \leftarrow \theta - \frac{\epsilon}{\sqrt{\tilde{\mathbf{a}}+\nu}} \odot \mathbf{g}$

A. Step 1: $\mathbf{g} = (-6, -6), \mathbf{a} = (3.6, 3.6)$, so $\theta = (0.316, 0.316)$

B. Step 2: $\mathbf{g} = (-5.37, -5.37), \mathbf{a} = (6.12, 6.12)$, so $\theta = (0.533, 0.533)$

iii. *Adam* $t \leftarrow t+1, \mathbf{v} \leftarrow \beta_1\mathbf{v} + (1-\beta_1)\mathbf{g}, \mathbf{a} \leftarrow \beta_2\mathbf{a} + (1-\beta_2)\mathbf{g} \odot \mathbf{g}$ Bias correction in moments: $\tilde{\mathbf{v}} = \frac{1}{1-\beta_1^t}\mathbf{v}, \tilde{\mathbf{a}} = \frac{1}{1-\beta_2^t}\mathbf{a}, \theta \leftarrow \theta - \frac{\epsilon}{\sqrt{\tilde{\mathbf{a}}+\nu}} \odot \tilde{\mathbf{v}}$

A. Step 1: $\mathbf{g} = (-6, -6), \mathbf{v} = (-0.6, -0.6), \mathbf{a} = (3.6, 3.6), \tilde{\mathbf{v}} = (-6, -6), \tilde{\mathbf{a}} = (36, 36)$, so $\theta = (0.6, 0.6)$

5.

Solution:

(a) i. Yes, your answer is still correct, since for the outputs with this distribution, on average, the error is reduced by the amount the variables are independent. From the class slides, the error should be:

$$\frac{1}{7}\mathbb{E}\varepsilon_i^2 + \frac{6}{7}\mathbb{E}\varepsilon_i\varepsilon_j$$

ii. Your colleague may not have been using a sample size large enough to capture the average error. Your colleauge may also be using a validation set with very skewed statistics (e.g., they're all cats) and perhaps the 1 model is really good at predicting cats. However, this validation data isn't an appropriate reflection of the statistics of the data.

(b) No, the network will not perform well, as the activations at each step will be effectively twice as large as they were during training time. Because they are twice as large, and sigmoid saturates for large inputs (or large negative inputs, if they're twice as large in the negative direction), the network will likely have many units that have saturated.

(c) These parameters are used in case the neural network performs better by not having the distribution of activations be mean 0 and variance 1. In these scenarios, the network can set the mean to β and the variance to γ^2.