Assignment 1

CMSC 726: Machine Learning
August 27[th], 2019 **Name:**

1. Problem 9.1 from the text book.

2. Let $J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(\theta^T\mathbf{x}^{(i)} - y^{(i)})^2$

   (a) Compute $\nabla_\theta J(\theta)$ and $\nabla_\theta^2 J(\theta)$.

   (b) Show that $J(\theta)$ is convex.

   (c) Under what conditions on input samples, $J(\theta)$ is strictly convex?

3. Let $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}\}$ be m i.i.d samples drawn from a Gaussian distribution $\mathcal{N}(\mu_{\text{true}}, \sigma_{\text{true}}^2\mathbf{I})$ where parameters μ_{true} and σ_{true} are unknown. A common approach to estimate model parameters is maximum likelihood estimation (MLE).

   (a) The likelihood function $L(\mu, \sigma)$ is defined as the probability of observing samples $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}\}$ from the distribution $\mathcal{N}(\mu, \sigma^2\boldsymbol{I})$. Write down the likelihood function in this case.

   (b) Argue that $\text{argmax}_{\mu,\sigma}\ L(\mu, \sigma) = \text{argmax}_{\mu,\sigma}\ \log L(\mu, \sigma)$.

   (c) By maximizing the log likelihood function, compute MLE estimates of model parameters.

4. Compute $\nabla_{\boldsymbol{X}} \text{Tr}\Big[\boldsymbol{AXBX^TCXD}\Big]$ =?

   - **Hint 1**: if $dy = \text{Tr}[\boldsymbol{A}(d\boldsymbol{X})]$, then the $\frac{dy}{d\boldsymbol{X}} = \boldsymbol{A^T}$.
   - **Hint 2**: The trace is invariant to cyclic permutations. For example, $\text{Tr}[\boldsymbol{A_1A_2A_3}] = \text{Tr}[\boldsymbol{A_3A_1A_2}] = \text{Tr}[\boldsymbol{A_2A_3A_1}]$. In general, $\text{Tr}[\boldsymbol{A_1A_2\dots A_n}] = \text{Tr}[\boldsymbol{A_kA_{k+1}\dots A_nA_1\dots A_{k-1}}]$, $1 \le k \le n$.
   - **Hint 3**: $\text{Tr}[\boldsymbol{A}] = \text{Tr}[\boldsymbol{A^T}]$.
   - **Hint 4**: The following is a solution for a simplified version of the problem. To compute $\nabla_{\boldsymbol{X}} \text{Tr}\Big[\boldsymbol{AXBX^T}\Big]$, we can write

     $$d\,\text{Tr}\Big[\boldsymbol{AXBX^T}\Big] = \text{Tr}\Big[d(\boldsymbol{AXBX^T})\Big]$$
     $$= \text{Tr}\Big[\boldsymbol{A}d(\boldsymbol{X})\boldsymbol{BX^T}\Big] + \text{Tr}\Big[\boldsymbol{AXB}d(\boldsymbol{X^T})\Big] \quad \text{(using the product rule of derivatives)}$$
     $$= \text{Tr}\Big[\boldsymbol{BX^TA}d(\boldsymbol{X})\Big] + \text{Tr}\Big[\boldsymbol{AXB}d(\boldsymbol{X^T})\Big] \quad \text{(using the cyclic permutation property for the first term)}$$
     $$= \text{Tr}\Big[\boldsymbol{BX^TA}d(\boldsymbol{X})\Big] + \text{Tr}\Big[d(\boldsymbol{X})\boldsymbol{B^TX^TA^T}\Big] \quad \text{(using the transpose invarience property for the second term)}$$
     $$= \text{Tr}\Big[\boldsymbol{BX^TA}d(\boldsymbol{X})\Big] + \text{Tr}\Big[\boldsymbol{B^TX^TA^T}d(\boldsymbol{X})\Big] \quad \text{(using the cyclic permutation property for the second term)}$$
     $$= \text{Tr}\Big[(\boldsymbol{BX^TA} + \boldsymbol{B^TX^TA^T})d(\boldsymbol{X})\Big].$$

     Therefore, using **Hint 1**, we have $\nabla_{\boldsymbol{X}} \text{Tr}\Big[\boldsymbol{AXBX^T}\Big] = \boldsymbol{A^TXB^T} + \boldsymbol{AXB}$.

5. (Programming Assignment) Let $\mathbf{x} \in \mathbb{R}^n$ and $z \in \mathbb{R}$ be zero-mean independent Gaussian random variables with covariance matrices $\mathbf{I}$ and σ^2, respectively. That is, $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I})$ and $z \sim \mathcal{N}(0, \sigma^2)$. Define $y = \theta^{\mathbf{T}}\mathbf{x} + \theta_0 + z$. In this assignment, we want to use stochastic gradient descent (SGD) to compute a linear regression model between $\mathbf{x}$ and y. Write a Python code to do the following:

   (a) Let $n = 4$, $\sigma^2 = 1/4$, $\theta = [1, 1/2, 1/4, 1/8]^T$ and $\theta_0 = 2$. Generate $m = 10,000$ i.i.d. *training* samples from $\mathbb{P}_{X,Y}$. That is $\{(\mathbf{x}^{(1)}, y^{(1)}), ..., (\mathbf{x}^{(m)}, y^{(m)})\}$.

   (b) Use SGD with a batch size of 10 to estimate model parameters. Plot the Mean-Squared Error (MSE) vs. the number of iterations.

   (c) Generate m new i.i.d. *test* samples from $\mathbb{P}_{X,Y}$. Use estimated parameters to compute the MSE on the test set.

   (d) Repeat parts (a)-(c) using $m = 10$. How do training and test errors change? Why?