EE239AS, Winter 2018 **Homework #1**
Neural Networks & Deep Learning Prof. J.C. Kao
University of California, Los Angeles; Department of ECE TAs: T. Xing & C. Zheng

Due Monday, 22 Jan 2017, by 11:59pm to Gradescope.
Covers material up to Introduction to machine learning.
100 points total.

1. (25 points) **Linear algebra refresher.**

   (a) (12 points) Let $\mathbf{A}$ be a square matrix, and further let $\mathbf{A}\mathbf{A}^T = \mathbf{I}$.

   i. (3 points) Construct a 2×2 example of $\mathbf{A}$ and derive the eigenvalues and eigenvectors of this example. Show all work (i.e., do not use a computer's eigenvalue decomposition capabilities). You may not use a diagonal matrix as your 2×2 example. What do you notice about the eigenvalues and eigenvectors?

   ii. (3 points) Show that $\mathbf{A}$ has eigenvalues with norm 1.

   iii. (3 points) Show that the eigenvectors of $\mathbf{A}$ corresponding to distinct eigenvalues are orthogonal.

   iv. (3 points) In words, describe what may happen to a vector $\mathbf{x}$ under the transformation $\mathbf{A}\mathbf{x}$.

Solution:

i. Let

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

so that

$$\mathbf{A}\mathbf{A}^T = \begin{bmatrix} a^2 + b^2 & ac + bd \\ ca + db & c^2 + d^2 \end{bmatrix}$$

By inspection, we can pick $a = -\frac{1}{\sqrt{2}}$, $b = c = d = \frac{1}{\sqrt{2}}$ so that $\mathbf{A}\mathbf{A}^T = \mathbf{I}$. To solve for the eigenvalues and eigenvectors of $\mathbf{A}$, we realize that because $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$, then $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$. We have:

$$\mathbf{A} - \lambda\mathbf{I} = \begin{bmatrix} -\frac{1}{\sqrt{2}} - \lambda & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} - \lambda \end{bmatrix}$$

Setting the determinant equal to zero, we get that

$$\left(-\frac{1}{\sqrt{2}} - \lambda\right)\left(\frac{1}{\sqrt{2}} - \lambda\right) - \frac{1}{2} = 0$$

This gives $\lambda^2 = 1$. The eigenvalues are $+1$ and -1.

Next we solve for the eigenvector corresponding to $\lambda = 1$. We have that $(\mathbf{A}-\mathbf{I})\mathbf{x} = 0$. This gives:

$$\mathbf{A} - \mathbf{I} = \begin{bmatrix} -\frac{1}{\sqrt{2}} - 1 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} - 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Setting $x_1 = 1$ because we can normalize the eigenvectors later on, we can solve the equation to get that $x_2 = 1 + \sqrt{2}$. We then normalize by calculating a such that $\sqrt{(ax_1)^2 + (ax_2)^2} = \sqrt{a^2 + (1 + 2\sqrt{2} + 2)a^2} = 1$. Solving this, we have that $a = 0.3827$ and thus the eigenvector corresponding to the eigenvalue $+1$ is

$$\mathbf{x} = \begin{bmatrix} 0.3827 \\ 0.9239 \end{bmatrix}$$

The same algebra for the eigenvector corresponding to the eigenvalue -1 yields:

$$\mathbf{x} = \begin{bmatrix} -0.9239 \\ 0.3827 \end{bmatrix}$$

We notice the eigenvalues are 1 and that the eignevectors are orthogonal.

ii. Let the λ be the eigenvalue of $\mathbf{A}$ and $\mathbf{Ax} = \lambda\mathbf{x}$:

$$\begin{aligned} (\lambda\mathbf{x})^T(\lambda\mathbf{x}) &= (\mathbf{Ax})^T(\mathbf{Ax}) \\ |\lambda|^2\mathbf{x}^T\mathbf{x} &= \mathbf{x}^T\mathbf{A}^T\mathbf{Ax} \\ |\lambda|^2 &= 1 \end{aligned}$$

iii. Let the λ_i, λ_j be the $\mathbf{A}$'s eigenvalue corresponding to eigenvector $\mathbf{x}_i, \mathbf{x}_j$ respectively, $\mathbf{Ax}_i = \lambda_i\mathbf{x}_i$, $\mathbf{Ax}_j = \lambda_j\mathbf{x}_j$ for $i \neq j$:

$$\begin{aligned} (\mathbf{Ax}_i)^T(\mathbf{Ax}_j) &= (\lambda_i\mathbf{x}_i)^T(\lambda_j\mathbf{x}_j) \\ \mathbf{x}_i^T\mathbf{A}^T\mathbf{Ax}_j &= \lambda_i\lambda_j\mathbf{x}_i^T\mathbf{x}_j \\ (\lambda_i\lambda_j - 1)\mathbf{x}_i^T\mathbf{x}_j &= 0 \end{aligned}$$

Since $\lambda_i \neq \lambda_j$ and $|\lambda_{i,j}| = 1$, the only solution is $\mathbf{x}_i^T\mathbf{x}_j = 0$

iv. The length of vector $\mathbf{x}$ will not change. It will be rotated or reflected.

(b) (8 points) Let $\mathbf{A}$ be a matrix.

i. (4 points) What is the relationship between the singular vectors of $\mathbf{A}$ and the eigenvectors of $\mathbf{AA}^T$? What about $\mathbf{A}^T\mathbf{A}$?

ii. (4 points) What is the relationship between the singular values of $\mathbf{A}$ and the eigenvalues of $\mathbf{AA}^T$? What about $\mathbf{A}^T\mathbf{A}$?

Solution:

i. The singular value decomposition of matrix $\mathbf{A}$ is $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T$. So

$$\begin{aligned} \mathbf{AA}^T &= \mathbf{U}\Sigma^T\mathbf{V}^T\mathbf{V}^\Sigma\mathbf{U}^T \\ &= \mathbf{U}\Sigma^T\Sigma\mathbf{U}^T \\ \mathbf{A}^T\mathbf{A} &= \mathbf{V}\Sigma^T\mathbf{U}^T\mathbf{U}^\Sigma\mathbf{V}^T \\ &= \mathbf{V}\Sigma^T\Sigma\mathbf{V}^T \end{aligned}$$

The left singular vectors of $\mathbf{A}$ is the eigenvectors of $\mathbf{A}\mathbf{A}^T$. The right singular vectors of $\mathbf{A}$ is the eigenvectors of $\mathbf{A}^T\mathbf{A}$.

ii. The singular value of $\mathbf{A}$ is square roots of the eigenvalues of $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$.

(c) (5 points) True or False. Partial credit on an incorrect solution may be awarded if you justify your answer.

i. Every linear operator in an n-dimensional vector space has n distinct eigenvalues.

ii. A non-zero sum of two eigenvectors of a matrix $\mathbf{A}$ is an eigenvector.

iii. If a matrix $\mathbf{A}$ has the positive semidefinite property, i.e., $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$ for all $\mathbf{x}$, then its eigenvalues must be non-negative.

iv. The rank of a matrix can exceed the number of non-zero eigenvalues.

v. A non-zero sum of two eigenvectors of a matrix $\mathbf{A}$ corresponding to the same eigenvalue λ is always an eigenvector.

Solution:

i. False: The 2×2 identity matrix has all eigenvalues 1.

ii. False: Let $\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$; then $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ is an eigenvector of eigenvalue 1 and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ is an eigenvector of eigenvalue 2, but their sum is not an eigenvector of $\mathbf{A}$.

iii. True

iv. True

v. True

2. (22 points) **Probability refresher.**

(a) (9 points) A jar of coins is equally populated with two types of coins. One is type "H50" and comes up heads with probability 0.5. Another is type "H60" and comes up heads with probability 0.6.

i. (3 points) You take one coin from the jar and flip it. It lands tails. What is the posterior probability that this is an H50 coin?

ii. (3 points) You put the coin back, take another, and flip it 4 times. It lands T, H, H, H. How likely is the coin to be type H50?

iii. (3 points) A new jar is now equally populated with coins of type H50, H55, and H60 (with probabilities of coming up heads 0.5, 0.55, and 0.6 respectively. You take one coin and flip it 10 times. It lands heads 9 times. How likely is the coin to be of each possible type?

Solution:

i. We let X denote type of the coin H50(H50) or H60(H60), and Y denote the outcome

of the flip, head (H) or tail (T).

$$\begin{aligned} p_{X|Y}(H50|T) &= \frac{p_{Y|X}(T|H50)p_X(H50)}{p_{X,Y}(H50,T)+p_{X,Y}(H60,T)} \\ &= \frac{p_{Y|X}(T|H50)p_X(H50)}{p_{Y|X}(T|H50)p_X(H50)+p_{Y|X}(T|H60)p_X(H60)} \\ &= \frac{0.5\cdot 0.5}{0.5\cdot 0.5+0.6\cdot 0.4} \\ &= \frac{5}{9} \end{aligned}$$

ii.

$$\begin{aligned} p_{X|Y}(H50|THHH) &= \frac{p_{Y,X}(THHH,H50)}{p_{X,Y}(H50,THHH)+p_{X,Y}(H60,THHH)} \\ &= \frac{p_{Y|X}(THHH|H50)p_X(H50)}{p_{Y|X}(THHH|H50)p_X(H50)+p_{Y|X}(THHH|H60)p_X(H60)} \\ &= \frac{0.5\cdot(0.5\cdot 0.5\cdot 0.5\cdot 0.5)}{0.5\cdot(0.5\cdot 0.5\cdot 0.5\cdot 0.5)+0.5\cdot(0.4\cdot 0.6\cdot 0.6\cdot 0.6)} \\ &\approx 0.4197 \end{aligned}$$

iii. We let E denote the event that in 10 flips, there are 9 heads and 1 tail. The order doesn't matter. X can be H50, H55, H60.

$$\begin{aligned} p_{X|Y}(H50|E) &= \frac{p_{Y,X}(E,H50)}{p_{X,Y}(H50,E)+p_{X,Y}(H55,E)+p_{X,Y}(H60,E)} \\ &= \frac{p_{Y|X}(E|H50)p_X(H50)}{p_{Y|X}(E|H50)p_X(H50)+p_{Y|X}(E|H55)p_X(H55)+p_{Y|X}(E|H60)p_X(H60)} \\ &= \frac{0.3333\cdot(0.5\cdot 0.5^9)}{0.3333\cdot(0.5\cdot 0.5^9)+0.333\cdot(0.45\cdot 0.55^9)+0.3333\cdot(0.4\cdot 0.6^9)} \\ &\approx 0.1379 \end{aligned}$$

$$\begin{aligned} p_{X|Y}(H55|E) &= \frac{p_{Y,X}(E,H55)}{p_{X,Y}(H50,E)+p_{X,Y}(H55,E)+p_{X,Y}(H60,E)} \\ &= \frac{p_{Y|X}(E|H55)p_X(H55)}{p_{Y|X}(E|H50)p_X(H50)+p_{Y|X}(E|H55)p_X(H55)+p_{Y|X}(E|H60)p_X(H60)} \\ &= \frac{0.3333\cdot(0.45\cdot 0.55^9)}{0.3333\cdot(0.5\cdot 0.5^9)+0.333\cdot(0.45\cdot 0.55^9)+0.3333\cdot(0.4\cdot 0.6^9)} \\ &\approx 0.2927 \end{aligned}$$

$$\begin{aligned} p_{X|Y}(H60|E) &= \frac{p_{Y,X}(E,H60)}{p_{X,Y}(H60,E)+p_{X,Y}(H55,E)+p_{X,Y}(H60,E)} \\ &= \frac{p_{Y|X}(E|H60)p_X(H60)}{p_{Y|X}(E|H50)p_X(H50)+p_{Y|X}(E|H55)p_X(H55)+p_{Y|X}(E|H60)p_X(H60)} \\ &= \frac{0.3333\cdot(0.4\cdot 0.6^9)}{0.3333\cdot(0.5\cdot 0.5^9)+0.333\cdot(0.45\cdot 0.55^9)+0.3333\cdot(0.4\cdot 0.6^9)} \\ &\approx 0.5964 \end{aligned}$$

(b) (3 points) Consider a pregnancy test with the following statistics.

- If the woman is pregnant, the test returns "positive" (or 1, indicating the woman is pregnant) 99% of the time.
- If the woman is not pregnant, the test returns "positive" 10% of the time.
- At any given point in time, 99% of the female population is not pregnant.

What is the probability that a woman is pregnant given she received a positive test? The answer should make intuitive sense; given an explanation of the result that you find.

Solution: We let X denote whether the woman is pregnant (1) or not (0), and Y denote the outcome of the test, positive (1) or not (0).

$$\begin{aligned} p_{X|Y}(1|1) &= \frac{p_{Y|X}(1|1)p_X(1)}{p_{X,Y}(0,1)+p_{X,Y}(1,1)} \\ &= \frac{p_{Y|X}(1|1)p_X(1)}{p_X(0)p_{Y|X}(1|0)+p_X(1)p_{Y|X}(1,1)} \\ &= \frac{0.99\cdot 0.01}{0.99\cdot 0.1+0.01\cdot 0.99} \\ &= 0.09 \end{aligned}$$

This is an awful test. This makes sense because a huge proportion of the female population is not pregnant, and if fails on 10% of this huge population, that's many more false detections than there are pregnant women.

(c) (5 points) Let $x_1, x_2, \ldots, x_n$ be identically distributed random variables. A random vector, $\mathbf{x}$, is defined as

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

What is $\mathbb{E}(\mathbf{A}\mathbf{x}+\mathbf{b})$ in terms of $\mathbb{E}(\mathbf{x})$, given that $\mathbf{A}$ and $\mathbf{b}$ are deterministic?

Solution:

$$\begin{aligned} \mathbb{E}(\mathbf{A}\mathbf{x}_i) &= \mathbb{E}\left(\sum_{j=1}^{n}\mathbf{A}_{i,j}\mathbf{x}_j\right) \\ &= \left(\sum_{j=1}^{n}\mathbf{A}_{i,j}\mathbb{E}(\mathbf{x}_j)\right) \\ &= \left(\sum_{j=1}^{n}\mathbf{A}_{i,j}\mathbb{E}(\mathbf{x})_j\right) \\ &= [\mathbf{A}\cdot\mathbb{E}(\mathbf{x})]_i \end{aligned}$$

so we have $\mathbb{E}(\mathbf{Ax}) = \mathbf{A}\mathbb{E}(\mathbf{x})$

$$\begin{aligned}\mathbb{E}(\mathbf{Ax}+\mathbf{b}) &= \mathbb{E}(\mathbf{Ax})+\mathbf{b}\\ &= \mathbf{A}\mathbb{E}(\mathbf{x})+\mathbf{b}\end{aligned}$$

(d) (5 points) Let

$$\mathbf{cov}(\mathbf{x}) = \mathbb{E}\left((\mathbf{x}-\mathbb{E}\mathbf{x})(\mathbf{x}-\mathbb{E}\mathbf{x})^T\right)$$

What is $\mathbf{cov}(\mathbf{Ax}+\mathbf{b})$ in terms of $\mathbf{cov}(\mathbf{x})$, given that $\mathbf{A}$ and $\mathbf{b}$ are deterministic?

Solution:

$$\begin{aligned}\mathbf{cov}(\mathbf{Ax}+\mathbf{b}) &= \mathbb{E}\left((\mathbf{Ax}+\mathbf{b}-\mathbf{A}\mathbb{E}(\mathbf{x})-\mathbf{b})(\mathbf{Ax}+\mathbf{b}-(\mathbf{A}\mathbb{E}(\mathbf{x})+\mathbf{b})^T\right)\\ &= \mathbb{E}\left((\mathbf{Ax}-\mathbf{A}\mathbb{E}(\mathbf{x}))(\mathbf{Ax}-(\mathbf{A}\mathbb{E}(\mathbf{x}))^T\right)\\ &= \mathbb{E}\left(\mathbf{A}(\mathbf{x}-\mathbb{E}(\mathbf{x}))(\mathbf{A}(\mathbf{x}-\mathbb{E}(\mathbf{x})))^T\right)\\ &= \mathbb{E}\left(\mathbf{A}(\mathbf{x}-\mathbb{E}(\mathbf{x}))((\mathbf{x}-\mathbb{E}(\mathbf{x})))^T\mathbf{A}^T\right)\\ &= \mathbf{A}\mathbb{E}\left((\mathbf{x}-\mathbb{E}(\mathbf{x}))((\mathbf{x}-\mathbb{E}(\mathbf{x})))^T\right)\mathbf{A}^T\\ &= \mathbf{A}\mathbf{cov}(\mathbf{x})\mathbf{A}^T\end{aligned}$$

3. (13 points) **Multivariate derivatives.**

(a) (2 points) Let $\mathbf{x}\in\mathbb{R}^n$, $\mathbf{y}\in\mathbb{R}^m$, and $\mathbf{A}\in\mathbb{R}^{n\times m}$. What is $\nabla_{\mathbf{x}}\mathbf{x}^T\mathbf{Ay}$?

(b) (2 points) What is $\nabla_{\mathbf{y}}\mathbf{x}^T\mathbf{Ay}$?

(c) (3 points) What is $\nabla_{\mathbf{A}}\mathbf{x}^T\mathbf{Ay}$?

(d) (3 points) Let $f = \mathbf{x}^T\mathbf{Ax}+\mathbf{b}^T\mathbf{x}$. What is $\nabla_{\mathbf{x}}f$?

(e) (3 points) Let $f = \mathrm{tr}(\mathbf{AB})$. What is $\nabla_{\mathbf{A}}f$?

Solution:

(a)

$$\nabla_{\mathbf{x}}\mathbf{x}^T\mathbf{Ay} = \mathbf{Ay}$$

(b)

$$\nabla_{\mathbf{y}}\mathbf{x}^T\mathbf{Ay} = \mathbf{A}^T\mathbf{x}$$

(c)

$$\nabla_{\mathbf{A}} \mathbf{x}^T \mathbf{A} \mathbf{y} = \begin{bmatrix} \frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{y}}{\partial a_{1,1}} & \frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{y}}{\partial a_{1,2}} & \cdots & \frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{y}}{\partial a_{1,m}} \\ \frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{y}}{\partial a_{2,1}} & \frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{y}}{\partial a_{2,2}} & \cdots & \frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{y}}{\partial a_{2,m}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{y}}{\partial a_{n,1}} & \frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{y}}{\partial a_{n,2}} & \cdots & \frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{y}}{\partial a_{n,m}} \end{bmatrix}$$

This is equal to $\mathbf{x}\mathbf{y}^T$.

(d)

$$\begin{aligned} \nabla_{\mathbf{x}} \left(\mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x} \right) &= \nabla_{\mathbf{x}} \left(\mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x} \right) \\ &= \nabla_{\mathbf{x}} \left(\mathbf{x}^T \mathbf{A} \mathbf{x} \right) + \nabla_{\mathbf{x}} \left(\mathbf{b}^T \mathbf{x} \right) \\ &= \mathbf{A}\mathbf{x} + \mathbf{A}^T \mathbf{x} + b \end{aligned}$$

(e)

$$\begin{aligned} \mathrm{tr}(\mathbf{A}\mathbf{B}) &= \mathrm{tr} \left(\begin{bmatrix} -\mathbf{a}_1 - \\ -\mathbf{a}_2 - \\ \vdots \\ -\mathbf{a}_M - \end{bmatrix} \begin{bmatrix} | & | & & | \\ \mathbf{b}_1 & \mathbf{b}_2 & \ldots & \mathbf{b}_P \\ | & | & & | \end{bmatrix} \right) \\ &= \mathrm{tr} \left(\begin{bmatrix} \mathbf{a}_1^T \mathbf{b}_1 & \mathbf{a}_1^T \mathbf{b}_2 & \ldots & \mathbf{a}_1^T \mathbf{b}_P \\ \mathbf{a}_2^T \mathbf{b}_1 & \mathbf{a}_2^T \mathbf{b}_2 & \ldots & \mathbf{a}_2^T \mathbf{b}_P \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{a}_M^T \mathbf{b}_1 & \mathbf{a}_M^T \mathbf{b}_2 & \ldots & \mathbf{a}_M^T \mathbf{b}_P \end{bmatrix} \right) \\ &= \sum_{i=1}^{N} a_{1i} b_{i1} + \sum_{i=1}^{N} a_{2i} b_{i2} + \ldots \end{aligned}$$

Then,

$$\frac{\partial \mathrm{tr}(\mathbf{A}\mathbf{B})}{\partial a_{ij}} = bji$$

and hence $\nabla_{\mathbf{A}} \mathrm{tr}(\mathbf{A}\mathbf{B}) = \mathbf{B}^T$.

4. (10 points) **Deriving least-squares with matrix derivatives.**
In least-squares, we seek to estimate some multivariate output $\mathbf{y}$ via the model

$$\hat{\mathbf{y}} = \mathbf{W}\mathbf{x}$$

In the training set we're given paired data examples $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$ from $i = 1, \ldots, n$. Least-squares is the following quadratic optimization problem:

$$\min_{\mathbf{W}} \quad \frac{1}{2} \sum_{i=1}^{n} \left\| \mathbf{y}^{(i)} - \mathbf{W}\mathbf{x}^{(i)} \right\|^2$$

Derive the optimal $\mathbf{W}$.

Hint: you may find the following derivatives useful:

$$\begin{aligned}\frac{\partial \text{tr}(\mathbf{WA})}{\partial \mathbf{W}} &= \mathbf{A}^T \\ \frac{\partial \text{tr}(\mathbf{WAW}^T)}{\partial \mathbf{W}} &= \mathbf{WA}^T + \mathbf{WA}\end{aligned}$$

Solution: We differentiate the objective function with respect to $\mathbf{W}$. We denote the $\mathbf{x}^{(i)}$ as $\mathbf{x}_i$ for convenience. To do this, we first note that:

$$\frac{1}{2}\sum_{k=1}^{K}\|\mathbf{y}_i - \mathbf{W}\mathbf{x}_i\|^2 = \frac{1}{2}\sum_{k=1}^{K}(\mathbf{y}_i - \mathbf{W}\mathbf{x}_i)^T(\mathbf{y}_i - \mathbf{W}\mathbf{x}_i)$$

At this point, we're only going to take out terms with $\mathbf{W}$, since that's what we want to take the derivative with respect to.

$$\begin{aligned}f(\mathbf{W}) &= \frac{1}{2}\sum_{k=1}^{K}\left(-2\mathbf{y}_i^T\mathbf{W}\mathbf{x}_i + \mathbf{x}_i^T\mathbf{W}^T\mathbf{W}\mathbf{x}_i\right) \\ &= \sum_{k=1}^{K}\left[-\text{tr}\left(\mathbf{y}_i^T\mathbf{W}\mathbf{x}_i\right) + \frac{1}{2}\text{tr}\left(\mathbf{x}_i^T\mathbf{W}^T\mathbf{W}\mathbf{x}_i\right)\right] \\ &= \sum_{k=1}^{K}\left[-\text{tr}\left(\mathbf{W}\mathbf{x}_i\mathbf{y}_i^T\right) + \frac{1}{2}\text{tr}\left(\mathbf{W}\mathbf{x}_i\mathbf{x}_i^T\mathbf{W}^T\right)\right] \\ &= -\text{tr}\left(\mathbf{W}\sum_{k=1}^{K}\mathbf{x}_i\mathbf{y}_i^T\right) + \frac{1}{2}\text{tr}\left(\mathbf{W}\sum_{k=1}^{K}\left(\mathbf{x}_i\mathbf{x}_i^T\right)\mathbf{W}^T\right) \\ &= -\text{tr}\left(\mathbf{WXY}^T\right) + \frac{1}{2}\text{tr}\left(\mathbf{WXX}^T\mathbf{W}^T\right)\end{aligned}$$

Using the derivatives in the hint, we get to:

$$\begin{aligned}\frac{\partial f(\mathbf{W})}{\partial \mathbf{W}} &= -\mathbf{YX}^T + \frac{1}{2}\left(\mathbf{WXX}^T + \mathbf{WXX}^T\right) \\ &= -\mathbf{YX}^T + \mathbf{WXX}^T \\ &[=] \ \mathbf{0}\end{aligned}$$

Solving this, we get that:

$$\mathbf{W} = \mathbf{YX}^T\left(\mathbf{XX}^T\right)^{-1}$$

5. (30 points) **Hello World in Jupyer.**
Complete the Jupyter notebook `linear_regression.ipynb`. Print out the Jupyter notebook and submit it to Gradescope.