ECE C147/C247, Winter 2023 **Homework #1 Solutions**
Neural Networks & Deep Learning Prof. J.C. Kao
UCLA ECE TAs: T.M, P.L, R.G, K.K, N.V, S.R, S.P, M.E

Due Monday, 23 Jan 2023, by 11:59pm to Gradescope.
Covers material up to Introduction to machine learning refresher 1.
100 points total.

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

   (a) (12 points) Let $\mathbf{Q}$ be a real orthogonal matrix.

   i. (3 points) Show that $\mathbf{Q^T}$ and $\mathbf{Q^{-1}}$ are also orthogonal.
   ii. (3 points) Show that $\mathbf{Q}$ has eigenvalues with norm 1.
   iii. (3 points) Show that the determinant of $\mathbf{Q}$ is either +1 or -1.
   iv. (3 points) Show that $\mathbf{Q}$ defines a length preserving transformation.

   Solution:

   i. We will first show that $\mathbf{Q^T}$ is orthogonal:

$$\begin{aligned}(\mathbf{Q^T})^T\mathbf{Q^T} &= \mathbf{QQ^T} \\ &= \mathbf{I}\end{aligned}$$

   Now, we show that $\mathbf{Q^{-1}}$ is orthogonal:

$$\begin{aligned}(\mathbf{Q^{-1}})^T\mathbf{Q^{-1}} &= (\mathbf{Q^T})^T\mathbf{Q^{-1}} \\ &= \mathbf{I}\end{aligned}$$

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

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

   iii.

$$\begin{aligned}det(\mathbf{Q^TQ}) &= det(\mathbf{Q^T})det(\mathbf{Q}) \\ &= det(\mathbf{Q})^2\end{aligned}$$

   Since $\mathbf{Q^TQ} = \mathbf{I}$, so

$$det(\mathbf{Q})^2 = 1$$

   and hence,

$$det(\mathbf{Q}) = +1, \text{or} -1$$

iv. Since

$$\begin{aligned}\|\mathbf{Q}\mathbf{v}\|_2^2 &= \mathbf{v}^T\mathbf{Q}^{\mathbf{T}}\mathbf{Q}\mathbf{v} \\ &= \|\mathbf{v}\|_2^2\end{aligned}$$

so $\mathbf{Q}$ defines a length preserving transformation.

(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 $\mathbf{A}\mathbf{A}^T$? 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 $\mathbf{A}\mathbf{A}^T$? 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

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

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 distinct 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 , Consider a matrix a with positive semi-definite property $x^T Ax \geq 0 \forall x$. This must be true for an eigenvector v. This implies

$$v^T Av \geq 0$$
$$v^T \lambda v \geq 0$$
$$\text{Where } \lambda \text{ is the eigenvalue is the corresponding eigenvector v}$$
$$\lambda ||v||_2^2 \geq 0$$
$$\lambda \geq 0$$

Here , λ is any eigenvalue of matrix A.

iv. True, an example is I identity matrix which is full rank (rank is 2) but with eigenvalue 1 and multiplicity 2.

v. True. Consider a matrix A with 2 eigenvectors v_1, v_2 corresponding to the same eigenvalue λ. Now,

$$Av_1 = \lambda v_1, Av_2 = \lambda v_2$$
$$\text{Consider linear combination of}, v_1 \text{and } \ v_2, \ \ av_1 + bv_2$$
$$A(av_1 + bv_2) = aAv_1 + bAv_2$$
$$aAv_1 + bAv_2 = a\lambda v_1 + b\lambda v_2$$
$$\text{Since } \ v_1, v_2 \text{are eigenvectors}$$
$$A(av_1 + bv_2) = \lambda(av_1 + bv_2)$$

This completes the proof.

2. (25 points) **Probability refresher.**

(a) (5 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. (1 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. (2 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. (2 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.

$$\begin{aligned}
p_{X|Y}(H50|E) &= \frac{p_{Y,X}(E,H50)}{p_{X,Y}(H50,E)+p_{X,Y}(H55,E)+p_{X,Y}(H60,E)} \\
&= \frac{p_{Y|X}(E|H50)p_X(H50)}{p_{Y|X}(E|H50)p_X(H50)+p_{Y|X}(E|H55)p_X(H55)+p_{Y|X}(E|H60)p_X(H60)} \\
&= \frac{\binom{10}{1} \cdot 0.3333 \cdot (0.5 \cdot 0.5^9)}{\binom{10}{1} \cdot 0.3333 \cdot (0.5 \cdot 0.5^9) + \binom{10}{1} \cdot 0.333 \cdot (0.45 \cdot 0.55^9) + \binom{10}{1} \cdot 0.3333 \cdot (0.4 \cdot 0.6^9)} \\
&\approx 0.1379
\end{aligned}$$

$$\begin{aligned}
p_{X|Y}(H55|E) &= \frac{p_{Y,X}(E,H55)}{p_{X,Y}(H50,E)+p_{X,Y}(H55,E)+p_{X,Y}(H60,E)} \\
&= \frac{p_{Y|X}(E|H55)p_X(H55)}{p_{Y|X}(E|H50)p_X(H50)+p_{Y|X}(E|H55)p_X(H55)+p_{Y|X}(E|H60)p_X(H60)} \\
&= \frac{\binom{10}{1} \cdot 0.3333 \cdot (0.45 \cdot 0.55^9)}{\binom{10}{1} \cdot 0.3333 \cdot (0.5 \cdot 0.5^9) + \binom{10}{1} \cdot 0.333 \cdot (0.45 \cdot 0.55^9) + \binom{10}{1} \cdot 0.3333 \cdot (0.4 \cdot 0.6^9)} \\
&\approx 0.2927
\end{aligned}$$

$$\begin{aligned} p_{X|Y}(H60|E) &= \frac{p_{Y,X}(E,H60)}{p_{X,Y}(H60,E)+p_{X,Y}(H55,E)+p_{X,Y}(H60,E)} \\ &= \frac{p_{Y|X}(E|H60)p_X(H60)}{p_{Y|X}(E|H50)p_X(H50)+p_{Y|X}(E|H55)p_X(H55)+p_{Y|X}(E|H60)p_X(H60)} \\ &= \frac{\binom{10}{1}\cdot 0.3333\cdot(0.4\cdot 0.6^9)}{\binom{10}{1}\cdot 0.3333\cdot(0.5\cdot 0.5^9)+\binom{10}{1}\cdot 0.333\cdot(0.45\cdot 0.55^9)+\binom{10}{1}\cdot 0.3333\cdot(0.4\cdot 0.6^9} \\ &\approx 0.5964 \end{aligned}$$

(b) (5 points) Students at UCLA are from these disciplines: 15% Science, 21% Healthcare, 24% Liberal Arts, and 40% Engineering. (Each student belongs to a unique discipline.) The students attend a lecture and give feedback. Suppose 90% of the Science students liked the lecture, 18% of the Healthcare students liked it, none of the Liberal Arts students liked it, and 10% of the Engineering students liked it. If a student is randomly chosen, and the student liked the lecture, what is the conditional probability that the student is from Science?

Solution: Suppose we define the following events:

- S: The student is from Science
- H: The student is from Healthcare
- A: The student is from Liberal Arts
- E: The student is from Engineering
- L: The student liked the lecture

Since the events S, H, A, E forms a partition of the sample space so by bayes law,

$$\begin{aligned} P(S|L) &= \frac{P(L|S)P(S)}{P(L|S)P(S)+P(L|H)P(H)+P(L|A)P(A)+P(L|E)P(E)} \\ &= \frac{0.9\times 0.15}{0.9\times 0.15+0.18\times 0.21+0\times 0.24+0.1\times 0.4} \\ &= \frac{0.135}{0.2128} \\ &= 0.634 \end{aligned}$$

(c) (5 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.

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

Solution:

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

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

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

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

Solution:

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

3. (10 points) **Multivariate derivatives.**

   (a) (1 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) (1 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 $\nabla_{\mathbf{y}}\mathbf{x}^T\mathbf{Ay}$?

   (c) (1 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 $\nabla_{\mathbf{A}}\mathbf{x}^T\mathbf{Ay}$?

   (d) (1 points) Let $\mathbf{x} \in \mathbb{R}^n$,$\mathbf{A} \in \mathbb{R}^{n\times n}$, and let $f = \mathbf{x}^T\mathbf{Ax} + \mathbf{b}^T\mathbf{x}$. What is $\nabla_{\mathbf{x}}f$?

   (e) (1 points) Let $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times n}$ and $f = \mathrm{tr}(\mathbf{AB})$. What is $\nabla_{\mathbf{A}}f$?

   (f) (2 points) Let $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times n}$ and $f = \mathrm{tr}(\mathbf{BA} + \mathbf{A}^T\mathbf{B} + \mathbf{A}^2\mathbf{B})$. What is $\nabla_{\mathbf{A}}f$?

   (g) (3 points) Let $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times n}$ and $f = \|\mathbf{A} + \lambda\mathbf{B}\|_F^2$. 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)

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

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

(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)

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

Then,

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

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

(f) Since trace is a linear operator, so

$$f = \mathrm{tr}(\mathbf{BA})+\mathrm{tr}(\mathbf{A}^T\mathbf{B})+\mathrm{tr}(\mathbf{A}^2\mathbf{B})$$

Then using the gradients from matrix cookbook,

$$\nabla_{\mathbf{A}}f = \mathbf{B}^T+\mathbf{B}+(\mathbf{AB}+\mathbf{BA})^T$$

(g)

$$\begin{aligned}f &= \mathrm{tr}[(\mathbf{A}^T+\lambda\mathbf{B}^T)(\mathbf{A}+\lambda\mathbf{B})]\\ &= \mathrm{tr}[\mathbf{A}^T\mathbf{A}+\lambda\mathbf{A}^T\mathbf{B}+\lambda\mathbf{B}^T\mathbf{A}+\lambda^2\mathbf{B}^T\mathbf{B}]\end{aligned}$$

Since trace is a linear operator and using gradients from matrix cookbook,

$$\nabla_{\mathbf{A}}f = 2(\mathbf{A}+\lambda\mathbf{B})$$

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 $\mathbf{W}$.
Where $\mathbf{W}$ is a matrix, and for each example in the training set, both $\mathbf{x}^{(i)}$ and $\mathbf{y}^{(i)}$ $\forall i = 1, \ldots n$ are vectors .

Hint: you may find the following derivatives useful:

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

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.

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

Using the derivatives in the hint, we get to:

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

Solving this, we get that:

$$\mathbf{W} = \mathbf{Y}\mathbf{X}^T\left(\mathbf{X}\mathbf{X}^T\right)^{-1}$$

5. (10 points) **Regularized least squares**

In lecture, we worked through the following least squares problem

$$\arg\min_{\theta} \ \frac{1}{2}\sum_{i=1}^{N}(y^{(i)} - \theta^T\hat{\mathbf{x}}^{(i)})^2$$

However, the least squares has a tendency to overfit the training data. One common technique used to address the overfitting problem is regularization. In this problem, we work through one of the regularization techniques namely ridge regularization which is also known as the regularized least squares problem. In the regularized least squares we solve the following optimization problem

$$\arg\min_{\theta} \ \frac{1}{2}\sum_{i=1}^{N}(y^{(i)} - \theta^T\hat{\mathbf{x}}^{(i)})^2 + \frac{\lambda}{2}\|\theta\|_2^2$$

where λ is a tunable regularization parameter. From the above cost function it can be observed that we are seeking least squares solution with a smaller 2-norm. Derive the solution to the regularized least squares problem, i.e Find θ^*.

Solution: From lecture we know that

$$\frac{1}{2}\sum_{i=1}^{N}(y^{(i)} - \theta^T\hat{\mathbf{x}}^{(i)})^2 = \frac{1}{2}(\mathbf{Y} - \mathbf{X}\theta)^T(\mathbf{Y} - \mathbf{X}\theta)$$

where $\mathbf{Y}$ is a vector and $\mathbf{X}$ is a matrix. Then, we can rewrite the cost function for regularized least squares as

$$\mathcal{L}(\theta) = \frac{1}{2}(\mathbf{Y} - \mathbf{X}\theta)^T(\mathbf{Y} - \mathbf{X}\theta) + \frac{\lambda}{2}\|\theta\|_2^2.$$

Since the cost function is convex, so we can solve for θ^* by setting the derivative with respect to θ to zero:

$$\nabla_\theta\mathcal{L} = -\mathbf{X}^T\mathbf{Y} + (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\theta.$$

Setting the gradient to zero and solving for θ we get

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

6. (20 points) **Linear regression.**
   Complete the Jupyter notebook `linear_regression.ipynb`. Print out the Jupyter notebook as a PDF and submit it to Gradescope.

   Solution: As part of the policy, we don't release solution to the coding component.