1. **Linear algebra refresher.**

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

   i. Show that $\mathbf{Q^T}$ and $\mathbf{Q^{-1}}$ are also orthogonal.

   Proof: Suppose $\mathbf{Q}$ is orthogonal. Then $\mathbf{QQ^T} = \mathbf{Q^TQ} = \mathbf{I}$. Consider $\mathbf{Q^T}$and note that $\left(\mathbf{Q^T}\right)^\mathbf{T} = \mathbf{Q}$. Then

   $$\mathbf{Q^T}\left(\mathbf{Q^T}\right)^\mathbf{T} = \mathbf{Q^TQ} = \mathbf{I} = \mathbf{QQ^T} = \left(\mathbf{Q^T}\right)^\mathbf{T}\mathbf{Q^T}$$

   Note that if $\mathbf{Q}$ is orthogonal, then $\mathbf{Q^T} = \mathbf{Q^{-1}}$. Then, since $\mathbf{Q^T}$ is orthogonal, $\mathbf{Q^{-1}}$ is orthogonal. □

   ii. Show that $\mathbf{Q}$ has eigenvalues with norm 1.

   Proof: Suppose $\lambda \in \mathbb{R}$ is an eigenvalue of $\mathbf{Q}$. Then

   $$\begin{aligned}
   \mathbf{Qx} &= \lambda\mathbf{x} && \\
   (\mathbf{Qx})^\mathbf{T}\mathbf{Qx} &= (\mathbf{Qx})^\mathbf{T}\lambda\mathbf{x} && \\
   \mathbf{x^TQ^TQx} &= (\lambda\mathbf{x})^\mathbf{T}\lambda\mathbf{x} && \mathbf{Qx} = \lambda\mathbf{x} \\
   \mathbf{x^TIx} &= \mathbf{x^T}\lambda^\mathbf{T}\lambda\mathbf{x} && \mathbf{Q} \text{ is orthogonal} \\
   \mathbf{x^Tx} &= \lambda^2\mathbf{x^Tx} && \lambda^\mathbf{T} = \lambda \\
   \|\mathbf{x}\|^2 &= \lambda^2\|\mathbf{x}\|^2 && \mathbf{x^Tx} = \|\mathbf{x}\|^2 \\
   \lambda^2 &= 1 &&
   \end{aligned}$$

   This implies that $|\lambda| = 1$ because $\mathbf{Q}$ is real. □

   iii. Show that the determinant of $\mathbf{Q}$ is ± 1.

   Proof: Because $\mathbf{Q}$ is orthogonal, we have that $\mathbf{QQ^T} = \mathbf{I}$. Taking the determinant of both sides, we get

   $$\det\left(\mathbf{QQ^T}\right) = \det\left(\mathbf{Q}\right) \cdot \det\left(\mathbf{Q^T}\right) = \det\left(\mathbf{I}\right)$$

   Since $\det\left(\mathbf{I}\right) = 1$, we have $\det\left(\mathbf{Q}\right) \cdot \det\left(\mathbf{Q^T}\right) = 1$. Note that $\det\left(\mathbf{Q}\right) = \det\left(\mathbf{Q^T}\right)$. Then

   $$\det\left(\mathbf{Q}\right) \cdot \det\left(\mathbf{Q^T}\right) = \det\left(\mathbf{Q}\right) \cdot \det\left(\mathbf{Q}\right) = \left[\det\left(\mathbf{Q}\right)\right]^2 = 1$$

   so $\det\left(\mathbf{Q}\right) = \pm 1$. □

iv. Show that $\mathbf{Q}$ defines a length preserving transformation.

Proof: Consider a linear transformation $T : \mathbb{R}^n \to \mathbb{R}^n$. By assumption, $\mathbf{Q}$ is an orthogonal matrix, so $\mathbf{QQ^T} = \mathbf{Q^TQ} = \mathbf{I}$. We can represent the linear transformation T by $\mathbf{Q}$, so write $T\mathbf{x} = \mathbf{Qx}$. Then, taking the norm of both sides, we get

$$\begin{aligned}
\|T\mathbf{x}\|^2 &= \|\mathbf{Qx}\|^2 \\
&= (\mathbf{Qx})^{\mathbf{T}} \mathbf{Qx} \\
&= \mathbf{x^T Q^T Qx} \\
&= \mathbf{x^T Ix} && \mathbf{Q} \text{ is orthogonal} \\
&= \mathbf{x^T x} \\
\|T\mathbf{x}\|^2 &= \|\mathbf{x}\|^2 && \mathbf{x^T x} = \|\mathbf{x}\|^2
\end{aligned}$$

Taking the square root of both sides, we get $\|T\mathbf{x}\| = \|\mathbf{x}\|$, so $\mathbf{Q}$ is a length preserving transformation. □

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

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

Response: The singular value decomposition of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is $\mathbf{A} = \mathbf{U\Sigma V^{-1}}$, where $\mathbf{U} \in \mathbb{R}^{m\times m}, \Sigma \in \mathbb{R}^{m\times n}, \mathbf{V} \in \mathbb{R}^{n\times n}$. But because $\mathbf{U}$ and $\mathbf{V}$ are orthogonal, we have $\mathbf{A} = \mathbf{U\Sigma V^{-1}} = \mathbf{U\Sigma V^T}$. Then, we can write $\mathbf{AA^T} \in \mathbb{R}^{m\times m}$ as

$$\begin{aligned}
\mathbf{AA^T} &= \left(\mathbf{U\Sigma V^T}\right)\left(\mathbf{U\Sigma V^T}\right)^{\mathbf{T}} \\
&= \mathbf{U\Sigma V^T}\left(\mathbf{V^T}\right)^{\mathbf{T}} \Sigma^{\mathbf{T}} \mathbf{U^T} \\
&= \mathbf{U\Sigma V^T V\Sigma^T U^T} \\
&= \mathbf{U\Sigma I\Sigma^T U^T} && \mathbf{V} \text{ is orthogonal} \\
&= \mathbf{U\Sigma\Sigma^T U^T} \\
&= \mathbf{U}\Sigma^2\mathbf{U^T} && \Sigma \text{ is diagonal}
\end{aligned}$$

So $\mathbf{AA^T} = \mathbf{U}\Sigma^2\mathbf{U^T}$, where $\mathbf{U}$ are the eigenvectors of $\mathbf{AA^T}$. Then, the left singular vectors of $\mathbf{A}$ are the eigenvectors of $\mathbf{AA^T}$.

Similarly, we can write $\mathbf{A^TA} \in \mathbb{R}^{n\times n}$ as

$$\begin{aligned}
\mathbf{A^TA} &= \left(\mathbf{U\Sigma V^T}\right)^{\mathbf{T}}\left(\mathbf{U\Sigma V^T}\right) \\
&= \left(\mathbf{V^T}\right)^{\mathbf{T}} \Sigma^{\mathbf{T}} \mathbf{U^T U\Sigma V^T} \\
&= \mathbf{V\Sigma^T U^T U\Sigma V^T} \\
&= \mathbf{V\Sigma^T I\Sigma V^T} && \mathbf{U} \text{ is orthogonal} \\
&= \mathbf{V}\Sigma^2\mathbf{V^T} && \Sigma \text{ is diagonal}
\end{aligned}$$

So, $\mathbf{A^T A} = \mathbf{V}\Sigma^2\mathbf{V^T}$, where $\mathbf{V}$ are the eigenvectors of $\mathbf{A^T A}$. Then, the right singular vectors of $\mathbf{A}$ are the eigenvectors of $\mathbf{A^T A}$.

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

Response: From the above part, we have that $\mathbf{AA^T} = \mathbf{U}\Sigma^2\mathbf{U^T}$ and $\mathbf{A^T A} = \mathbf{V}\Sigma^2\mathbf{V^T}$. Then, the singular values of $\mathbf{A}$ are the square root of the eigenvalues of $\mathbf{AA^T}$ and $\mathbf{A^T A}$.

(c) 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.

Response: False. Every linear operator in an n-dimensional vector space has ***at most*** n distinct eigenvalues.

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

Response: Consider two eigenvectors $\mathbf{x}, \mathbf{y}$ of a matrix $\mathbf{A} \in \mathbb{R}^n$. There are two cases:

Case 1: If $\mathbf{x}, \mathbf{y}$ correspond to the same eigenvalue λ, the statement is True since
$$\mathbf{A}(\mathbf{x}+\mathbf{y}) = \mathbf{Ax} + \mathbf{Ay} = \lambda\mathbf{x} + \lambda\mathbf{y} = \lambda(\mathbf{x}+\mathbf{y})$$

Case 2: If $\mathbf{x}, \mathbf{y}$ correspond to unique eigenvalues $\lambda_\mathbf{x}, \lambda_\mathbf{y}$, the statement is False since
$$\mathbf{A}(\mathbf{x}+\mathbf{y}) = \mathbf{Ax} + \mathbf{Ay} = \lambda_\mathbf{x}\mathbf{x} + \lambda_\mathbf{y}\mathbf{y} \neq \lambda(\mathbf{x}+\mathbf{y})$$

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

Response: True. Suppose a matrix $\mathbf{A}$ has the positive semidefinite property; i.e. $\mathbf{x^T Ax} \geq 0$ for all $\mathbf{x}$. Consider an arbirtrary eigenvalue λ of $\mathbf{A}$. Then, $\mathbf{Ax} = \lambda\mathbf{x}$ for some eigenvector $\mathbf{x}$. Multiplying both sides by $\mathbf{x^T}$, we get

$$0 \leq \mathbf{x^T Ax} = \mathbf{x^T}\lambda\mathbf{x} = \lambda\mathbf{x^T x}$$

and since $\mathbf{x^T x} = \|\mathbf{x}\|^2 \geq 0$ for every $\mathbf{x}$, λ is non-negative.

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

Response: True. Consider a matrix $\mathbf{A}$ with $\text{rank}(A) = 2$ and an eigenvalue λ with algebraic multiplicity 2. Then, the rank of the matrix exceeds 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.

Response: True. Consider two eigenvectors $\mathbf{x}, \mathbf{y}$ of a matrix $\mathbf{A}$ and suppose $\mathbf{x}, \mathbf{y}$ correspond to the same eigenvalue λ. Then

$$\mathbf{A}(\mathbf{x}+\mathbf{y}) = \mathbf{A}\mathbf{x} + \mathbf{A}\mathbf{y} = \lambda\mathbf{x} + \lambda\mathbf{y} = \lambda(\mathbf{x}+\mathbf{y})$$

2. **Probability refresher.**

(a) A and B are involved in a duel. The rules of the duel are that they are to pick up their guns and shoot at each other simultaneously. If one or both are hit, then the duel is over. If both shots miss, then they repeat the process. Suppose that the results of the shots are independent and that each shot of A will hit B with probability p_A and each shot of B will hit A with probability p_B. What is:

i. the probability that A is not hit?

Response: The probability that A is not hit is

$$p_{(A \text{ is not hit})} = \frac{p_A\,(1-p_B)}{(p_A+p_B) - p_A p_B}$$

ii. the probability that both duelists are hit?

Response: The probability that both duelists are hit is

$$p_{(A \text{ and } B \text{ are hit})} = \frac{p_A p_B}{(p_A+p_B) - p_A p_B}$$

iii. the probability that the duel ends after the n^{th} round of shots?

Response: The probability that the duel ends is

$$p_{(\text{duel ends})} = (p_A+p_B) - (p_A \cdot p_B)$$

Then, the probability that the duel ends after the n^{th} round of shots is

$$p_{(X=n)} = ([p_A+p_B] - p_A p_B) \cdot ([1-p_A][1-p_B])^{n-1}$$

iv. the conditional probability that the duel ends after the n^{th} round of shots given that A is not hit?

Response: The conditional probability that the duel ends after the n^{th} round of shots given that A is not hit is

$$p_{(X=n|A \text{ is not hit})} = ([p_A+p_B] - p_A p_B) \cdot ([1-p_B]\,[1-p_A])^{n-1}$$

v. the conditional probability that the duel ends after the n^{th} round of shots given that both duelists are hit?

> **Response:** The conditional probability that the duel ends after the n^{th} round of shots given that both duelists are hit is
>
> $$p_{(X=n|A \text{ and } B \text{ are hit})} = ([p_A + p_B] - p_A p_B) \cdot ([1 - p_B]\,[1 - p_A])^{n-1}$$

(b) Consider a group of 18 UCLA faculty members containing 6 ECE faculty, 6 CSE faculty, and 6 Math faculty. We mix up the faculty members and place all 18 faculties around a circular table, with one faculty member per seat.

i. A faculty member is called "isolated" if his/her department does not agree with either of the nearby faculties (i.e., if he/she has a different department than the faculty to his/her right and a different department than the faculty to his/her left). Let X denote the number of isolated faculties. Find $E(X)$.

> **Response:**
>
> $$E(X) = 18 \cdot \frac{12}{17} \cdot \frac{11}{16} = 8.735$$

ii. A faculty member is called "semi-happy" if his/her department agrees with exactly one (but not both) of the nearby faculties (i.e., if his/her department agrees with the department of the faculty on his/her left or on his/her right, but not both). Let Y denote the number of semi-happy faculties. Find $E(Y)$.

> **Response:**
>
> $$E(Y) = 18 \cdot \left(\frac{5}{17} \cdot \frac{11}{16} + \frac{12}{17} \cdot \frac{5}{16} \right) = 7.61$$

iii. A faculty member is called "joyous" if his/her department agrees with both of the nearby faculties (i.e., if his/her department agrees with the department of the faculties on his/her left and on his/her right). Let Z denote the number of joyous faculties. Find $E(Z)$.

> **Response:**
>
> $$E(Z) = 18 \cdot \frac{5}{17} \cdot \frac{4}{16} = 1.324$$

(c) There is a screening test for lung cancer that looks at the level of LSA (lung specific antigen) in the blood. There are a number of reasons besides lung cancer that a man can have elevated LSA levels. In addition, many types of lung cancer develop so slowly that they are never a problem. Unfortunately, there is currently no test to distinguish the different types and using the test is controversial because it's hard to quantify the accuracy rates and the harm done by false positives. For this problem, we will call a positive test a true positive if it catches a dangerous type of lung cancer. Also, we will assume the following numbers:

- Rate of dangerous type of lung cancer among men over 30 = 0.0005
- True positive rate for the test = 0.9
- False positive rate for the test = 0.01

Suppose you randomly select a man over 30 and perform a screening test.

i. What is the probability that the man has a dangerous type of the disease given that he had a positive test?

Response: Given the following,

$$\begin{aligned} p_{(\text{disease})} &= 0.005 \\ p_{(\text{positive}|\text{disease})} &= 0.9 \\ p_{(\text{positive}|\text{no disease})} &= 0.01 \end{aligned}$$

the probability that the man has a dangerous type of the disease given that he had a positive test is

$$\begin{aligned} p &= \frac{p_{(\text{disease})} \cdot p_{(\text{positive}|\text{disease})}}{\left[p_{(\text{positive}|\text{disease})} \cdot p_{(\text{disease})}\right] + \left[p_{(\text{positive}|\text{no disease})} \cdot \left(1 - p_{(\text{disease})}\right)\right]} \\ &= \frac{0.0005 \cdot 0.9}{(0.9)(0.0005) + (0.01)(1 - 0.0005)} \\ p &= 0.0431 \end{aligned}$$

ii. What is the probability that the man has a dangerous type of the disease given that he had a negative test?

Response: Given the following,

$$\begin{aligned} p_{(\text{disease})} &= 0.005 \\ p_{(\text{negative}|\text{disease})} &= 1 - p_{(\text{positive}|\text{disease})} = 0.1 \\ p_{(\text{negative}|\text{no disease})} &= 1 - p_{(\text{positive}|\text{no disease})} = 0.99 \end{aligned}$$

the probability that the man has a dangerous type of the disease given that he had a negative test is

$$\begin{aligned} p &= \frac{p_{(\text{disease})} \cdot p_{(\text{negative}|\text{disease})}}{\left[p_{(\text{negative}|\text{disease})} \cdot p_{(\text{disease})}\right] + \left[p_{(\text{negative}|\text{no disease})} \cdot \left(1 - p_{(\text{disease})}\right)\right]} \\ &= \frac{0.0005 \cdot 0.1}{(0.1)(0.0005) + (0.99)(1 - 0.0005)} \\ p &= 0.0000505 \end{aligned}$$

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

Response: Note that if $\mathbf{A}$ and $\mathbf{b}$ are deterministic, they are independent of $\mathbf{x}$; i.e. $\mathbb{E}(\mathbf{A}) = \mathbf{A}$ and $\mathbb{E}(\mathbf{b}) = \mathbf{b}$.
Then,

$$\mathbb{E}(\mathbf{Ax} + \mathbf{b}) = \mathbb{E}(\mathbf{Ax}) + \mathbb{E}(\mathbf{b}) = \mathbf{A}\mathbb{E}(\mathbf{x}) + \mathbf{b}$$

(e) Let

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

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

Response: Note that

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

and because $\mathbf{A}$ and $\mathbf{b}$ are deterministic,

$$\mathbb{E}(\mathbf{Ax} + \mathbf{b}) = \mathbf{A}\mathbb{E}(\mathbf{x}) + \mathbf{b}$$

Then,

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

3. **Multivariate derivatives.**

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

Response: Setting $\mathbf{v} := \mathbf{Ay}$, we have

$$\nabla_{\mathbf{x}} \mathbf{x}^{\mathbf{T}} \mathbf{A} \mathbf{y} = \nabla_{\mathbf{x}} \mathbf{x}^{\mathbf{T}} \mathbf{v} = \mathbf{v} = \mathbf{Ay} \qquad \text{mat. cookbook (69)}$$

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

Response: Setting $\mathbf{v}^\mathbf{T} := \mathbf{x}^\mathbf{T} \mathbf{A}$, we have

$$\nabla_{\mathbf{y}} \mathbf{x}^\mathbf{T} \mathbf{A} \mathbf{y} = \nabla_{\mathbf{y}} \mathbf{v}^\mathbf{T} \mathbf{y} = \mathbf{v} = \mathbf{A}^\mathbf{T} \mathbf{x} \qquad \text{mat. cookbook (69)}$$

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

Response: Let

$$s := \mathbf{x}^\mathbf{T} \mathbf{A} \mathbf{y} = \sum_{i=1}^{n} \sum_{j=1}^{m} x_i \cdot a_{ij} \cdot y_j$$

where s is a scalar. Then

$$\nabla_{\mathbf{A}} \mathbf{x}^\mathbf{T} \mathbf{A} \mathbf{y} = \begin{bmatrix} \frac{\partial s}{\partial a_{11}} & \frac{\partial s}{\partial a_{12}} & \cdots & \frac{\partial s}{\partial a_{1m}} \\ \frac{\partial s}{\partial a_{21}} & \frac{\partial s}{\partial a_{22}} & \cdots & \frac{\partial s}{\partial a_{2m}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial s}{\partial a_{n1}} & \frac{\partial s}{\partial a_{n2}} & \cdots & \frac{\partial s}{\partial a_{nm}} \end{bmatrix} = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_m \\ x_2 y_1 & x_2 y_2 & \cdots & x_2 y_m \\ \vdots & \vdots & \ddots & \vdots \\ x_n y_1 & x_n y_2 & \cdots & x_n y_m \end{bmatrix} = \mathbf{x} \mathbf{y}^\mathbf{T}$$

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

Response: Note that $\mathbf{b}^\mathbf{T}$ is a constant with respect to $\mathbf{x}$. Then,

$$\nabla_{\mathbf{x}} f = (\mathbf{A} + \mathbf{A}^\mathbf{T}) \mathbf{x} + \mathbf{b} \qquad \nabla_{\mathbf{x}} \mathbf{x}^\mathbf{T} \mathbf{A} \mathbf{x} = (\mathbf{A} + \mathbf{A}^\mathbf{T}) \mathbf{x} \text{ from lecture}$$

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

Response: Note that

$$f = \operatorname{tr}(\mathbf{A}\mathbf{B}) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \cdot b_{ji}$$

Then,

$$\nabla_{\mathbf{A}} f = \begin{bmatrix} \frac{\partial f}{\partial a_{11}} & \frac{\partial f}{\partial a_{12}} & \cdots & \frac{\partial f}{\partial a_{1n}} \\ \frac{\partial f}{\partial a_{21}} & \frac{\partial f}{\partial a_{22}} & \cdots & \frac{\partial f}{\partial a_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial a_{n1}} & \frac{\partial f}{\partial a_{n2}} & \cdots & \frac{\partial f}{\partial a_{nn}} \end{bmatrix} = \begin{bmatrix} b_{11} & b_{21} & \cdots & b_{n1} \\ b_{12} & b_{22} & \cdots & b_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ b_{1n} & b_{2n} & \cdots & b_{nn} \end{bmatrix} = \mathbf{B}^\mathbf{T}$$

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

Response: Note that since trace is linear, $\operatorname{tr}(\mathbf{X}+\mathbf{Y}) = \operatorname{tr}(\mathbf{X}) + \operatorname{tr}(\mathbf{Y})$. Then

$$\operatorname{tr}\left(\mathbf{BA} + \mathbf{A^T B} + \mathbf{A}^2\mathbf{B}\right) = \operatorname{tr}\left(\mathbf{BA}\right) + \operatorname{tr}\left(\mathbf{A^T B}\right) + \operatorname{tr}\left(\mathbf{A}^2\mathbf{B}\right)$$

and consider the following properties of trace:

$$\operatorname{tr}\left(\mathbf{AB}\right) = \operatorname{tr}\left(\mathbf{BA}\right)$$
$$\operatorname{tr}\left(\mathbf{A^T}\right) = \operatorname{tr}\left(\mathbf{A}\right)$$

We can rewrite the equation as

$$\begin{aligned}\operatorname{tr}\left(\mathbf{BA} + \mathbf{A^T B} + \mathbf{A}^2\mathbf{B}\right) &= \operatorname{tr}\left(\mathbf{BA}\right) + \operatorname{tr}\left(\mathbf{A^T B}\right) + \operatorname{tr}\left(\mathbf{A}^2\mathbf{B}\right)\\ &= \operatorname{tr}\left(\mathbf{AB}\right) + \operatorname{tr}\left(\mathbf{A^T B}\right) + \operatorname{tr}\left(\mathbf{A}^2\mathbf{B}\right)\\ &= \operatorname{tr}\left(\mathbf{AB}\right) + \operatorname{tr}\left(\mathbf{A^T B}\right) + \operatorname{tr}\left(\mathbf{AAB}\right)\end{aligned}$$

Then

$$\begin{aligned}\nabla_{\mathbf{A}} f &= \nabla_{\mathbf{A}} \operatorname{tr}\left(\mathbf{AB}\right) + \nabla_{\mathbf{A}} \operatorname{tr}\left(\mathbf{A^T B}\right) + \nabla_{\mathbf{A}} \operatorname{tr}\left(\mathbf{AAB}\right)\\ &= \mathbf{B^T} + \mathbf{B} + \left(\mathbf{AB} + \mathbf{BA}\right)^{\mathbf{T}} && \text{mat. cookbook } (103), (107)\\ \nabla_{\mathbf{A}} f &= \mathbf{B^T} + \mathbf{B} + \mathbf{B^T A^T} + \mathbf{A^T B^T}\end{aligned}$$

(g) 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 $\nabla_{\mathbf{A}} f$?

Response: Note that

$$f = \|\mathbf{A} + \lambda\mathbf{B}\|_F^2 = \operatorname{tr}\left(\left[\mathbf{A} + \lambda\mathbf{B}\right]\left[\mathbf{A} + \lambda\mathbf{B}\right]^{\mathbf{H}}\right)$$

Then

$$\begin{aligned}\nabla_A \|\mathbf{A} + \lambda\mathbf{B}\|_F^2\| &= \nabla_{\mathbf{A}} \operatorname{tr}\left(\left[\mathbf{A} + \lambda\mathbf{B}\right]\left[\mathbf{A} + \lambda\mathbf{B}\right]^{\mathbf{H}}\right)\\ &= \nabla_{\mathbf{A}} \operatorname{tr}\left(\left[\mathbf{A} + \lambda\mathbf{B}\right]\left[\mathbf{A} + \lambda\mathbf{B}\right]^{\mathbf{T}}\right) && \mathbf{X} \in \mathbb{R} \to \mathbf{X^H} = \mathbf{A^T}\\ &= \nabla_{\mathbf{A}} \operatorname{tr}\left(\left[\mathbf{A} + \lambda\mathbf{B}\right]\left[\mathbf{A^T} + \lambda^{\mathbf{T}}\mathbf{B^T}\right]\right)\\ &= \nabla_{\mathbf{A}} \operatorname{tr}\left(\mathbf{AA^T} + \lambda\mathbf{BA^T} + \mathbf{A}\lambda\mathbf{B^T} + \lambda^2\mathbf{BB^T}\right) && \lambda^{\mathbf{T}} = \lambda\\ &= \nabla_{\mathbf{A}} \operatorname{tr}\left(\mathbf{AA^T}\right) + \nabla_{\mathbf{A}} \operatorname{tr}\left(\lambda\mathbf{BA^T}\right) + \nabla_{\mathbf{A}} \operatorname{tr}\left(\mathbf{A}\lambda\mathbf{B^T}\right) + \nabla_{\mathbf{A}} \operatorname{tr}\left(\lambda^2\mathbf{BB^T}\right)\\ &= \nabla_{\mathbf{A}} \operatorname{tr}\left(\mathbf{AA^T}\right) + \nabla_{\mathbf{A}}\lambda \operatorname{tr}\left(\left[\mathbf{BA^T}\right]^{\mathbf{T}}\right) + \nabla_{\mathbf{A}}\lambda \operatorname{tr}\left(\mathbf{AB^T}\right) + 0 && \operatorname{tr}\left(\mathbf{X}\right) = \operatorname{tr}\left(\mathbf{X^T}\right)\\ &= \nabla_{\mathbf{A}} \operatorname{tr}\left(\mathbf{AA^T}\right) + \nabla_{\mathbf{A}} \operatorname{tr}\left(\lambda\mathbf{BA^T}\right) + \nabla_{\mathbf{A}} \operatorname{tr}\left(\lambda\mathbf{BA^T}\right)\\ &= \nabla_{\mathbf{A}} \operatorname{tr}\left(\mathbf{AA^T}\right) + \nabla_{\mathbf{A}} 2\lambda \operatorname{tr}\left(\mathbf{BA^T}\right)\\ &= 2\mathbf{A} + 2\lambda\mathbf{B}\\ \nabla_A \|\mathbf{A} + \lambda\mathbf{B}\|_F^2\| &= 2\left(\mathbf{A} + \lambda\mathbf{B}\right)\end{aligned}$$

4. **Deriving least-squares with matrix derivatives.**
In least-squares, we seek to estimates 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 $\left(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}\right)$ from $i = 1, \ldots, n$. Least-squares is the following quadratic optimization problem:

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

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:

$$\frac{\partial \operatorname{tr}(\mathbf{W}\mathbf{A})}{\partial \mathbf{W}} = \mathbf{A}^{\mathbf{T}} \tag{1}$$

$$\frac{\partial \operatorname{tr}\left(\mathbf{W}\mathbf{A}\mathbf{W}^{\mathbf{T}}\right)}{\partial \mathbf{W}} = \mathbf{W}\mathbf{A}^{\mathbf{T}} + \mathbf{W}\mathbf{A} \tag{2}$$

Response: We can rewrite $\frac{1}{2} \sum_{i=1}^{n} \left\| \mathbf{y}^{(i)} - \mathbf{W}\mathbf{x}^{(i)} \right\|^2$ by writing $\mathbf{x}^{(i)}$ and $\mathbf{y}^{(i)}$ in matrix form to get

$$\mathbf{X} := \begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \\ \vdots \\ \mathbf{x}^{(n)} \end{bmatrix}, \mathbf{Y} := \begin{bmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \\ \vdots \\ \mathbf{y}^{(n)} \end{bmatrix}$$

Then

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

where

$$\|\mathbf{Y} + \mathbf{W}\mathbf{X}\|_F^2 = \operatorname{tr}\left([\mathbf{Y} + \mathbf{W}\mathbf{X}] [\mathbf{Y} + \mathbf{W}\mathbf{X}]^{\mathbf{H}} \right) \qquad \text{mat. cookbook (541)}$$

Taking the gradient with respect to $\mathbf{W}$, we get

$$\begin{aligned}
\nabla_{\mathbf{W}} \frac{1}{2} \|\mathbf{Y} - \mathbf{WX}\|^2 &= \nabla_{\mathbf{W}} \frac{1}{2} \operatorname{tr}\left([\mathbf{Y} - \mathbf{WX}] [\mathbf{Y} - \mathbf{WX}]^{\mathbf{H}} \right) \\
&= \nabla_{\mathbf{W}} \frac{1}{2} \operatorname{tr}\left([\mathbf{Y} - \mathbf{WX}] [\mathbf{Y} - \mathbf{WX}]^{\mathbf{T}} \right) && \mathbf{A} \in \mathbb{R} \to \mathbf{A}^{\mathbf{H}} = \mathbf{A}^{\mathbf{T}} \\
&= \nabla_{\mathbf{W}} \frac{1}{2} \operatorname{tr}\left([\mathbf{Y} - \mathbf{WX}] \left[\mathbf{Y}^{\mathbf{T}} - \mathbf{X}^{\mathbf{T}} \mathbf{W}^{\mathbf{T}} \right] \right) \\
&= \nabla_{\mathbf{W}} \frac{1}{2} \operatorname{tr}\left(\mathbf{Y}\mathbf{Y}^{\mathbf{T}} - \mathbf{WXY}^{\mathbf{T}} - \mathbf{YX}^{\mathbf{T}} \mathbf{W}^{\mathbf{T}} - \mathbf{WXX}^{\mathbf{T}} \mathbf{W}^{\mathbf{T}} \right) \\
&= \frac{1}{2} \left[\nabla_{\mathbf{W}} \operatorname{tr}\left(\mathbf{YY}^{\mathbf{T}} \right) - \nabla_{\mathbf{W}} \operatorname{tr}\left(\mathbf{WXY}^{\mathbf{T}} \right) - \nabla_{\mathbf{W}} \operatorname{tr}\left(\mathbf{YX}^{\mathbf{T}} \mathbf{W}^{\mathbf{T}} \right) + \nabla_{\mathbf{W}} \operatorname{tr}\left(\mathbf{WXX}^{\mathbf{T}} \mathbf{W}^{\mathbf{T}} \right) \right] \\
&= \frac{1}{2} \left[0 + \nabla_{\mathbf{W}} \operatorname{tr}\left(\mathbf{WXY}^{\mathbf{T}} \right) - \nabla_{\mathbf{W}} \operatorname{tr}\left(\mathbf{WXY}^{\mathbf{T}} \right) + \nabla_{\mathbf{W}} \operatorname{tr}\left(\mathbf{WXX}^{\mathbf{T}} \mathbf{W}^{\mathbf{T}} \right) \right] \\
&= \frac{1}{2} \left[-2 \nabla_{\mathbf{W}} \operatorname{tr}\left(\mathbf{WXY}^{\mathbf{T}} \right) + \nabla_{\mathbf{W}} \operatorname{tr}\left(\mathbf{WXX}^{\mathbf{T}} \mathbf{W}^{\mathbf{T}} \right) \right] \\
&= \frac{1}{2} \left[-2 \mathbf{YX}^{\mathbf{T}} + \nabla_{\mathbf{W}} \operatorname{tr}\left(\mathbf{WXX}^{\mathbf{T}} \mathbf{W}^{\mathbf{T}} \right) \right] && \mathbf{A} := \mathbf{XY}^{\mathbf{T}} \to (1) \\
&= \frac{1}{2} \left[-2 \mathbf{YX}^{\mathbf{T}} + \mathbf{W} \left(\mathbf{XX}^{\mathbf{T}} \right)^{\mathbf{T}} + \mathbf{W} \left(\mathbf{XX}^{\mathbf{T}} \right) \right] && \mathbf{A} := \mathbf{XX}^{\mathbf{T}} \to (2) \\
&= \frac{1}{2} \left[-2 \mathbf{YX}^{\mathbf{T}} + \mathbf{W} \left(\mathbf{X}^{\mathbf{T}} \right)^{\mathbf{T}} \mathbf{X}^{\mathbf{T}} + \mathbf{WXX}^{\mathbf{T}} \right] \\
&= \frac{1}{2} \left[-2 \mathbf{YX}^{\mathbf{T}} + \mathbf{WXX}^{\mathbf{T}} + \mathbf{WXX}^{\mathbf{T}} \right] \\
&= \frac{1}{2} \left(-2 \mathbf{YX}^{\mathbf{T}} + 2 \mathbf{WXX}^{\mathbf{T}} \right) \\
\nabla_{\mathbf{W}} \frac{1}{2} \|\mathbf{Y} - \mathbf{WX}\|^2 &= \left(-\mathbf{Y} + \mathbf{WX} \right) \mathbf{X}^{\mathbf{T}}
\end{aligned}$$

Then, setting $\nabla_{\mathbf{W}} \frac{1}{2} \|\mathbf{Y} - \mathbf{WX}\|^2 = 0$, we get

$$\begin{aligned}
0 &= \nabla_{\mathbf{W}} \frac{1}{2} \|\mathbf{Y} - \mathbf{WX}\|^2 \\
0 &= -\mathbf{YX}^{\mathbf{T}} + \mathbf{WXX}^{\mathbf{T}} \\
\mathbf{WXX}^{\mathbf{T}} &= \mathbf{YX}^{\mathbf{T}} \\
\mathbf{W} \left(\mathbf{XX}^{\mathbf{T}} \right) \left(\mathbf{XX}^{\mathbf{T}} \right)^{-1} &= \mathbf{YX}^{\mathbf{T}} \left(\mathbf{XX}^{\mathbf{T}} \right)^{-1} \\
\mathbf{W} &= \mathbf{YX}^{\mathbf{T}} \left(\mathbf{XX}^{\mathbf{T}} \right)^{-1}
\end{aligned}$$

5. **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^*.

Response: Note that from lecture,

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

Then define f to be

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

From the matrix cookbook,

$$\nabla_{\mathbf{x}}\mathbf{x}^{\mathbf{T}}\mathbf{a} = \nabla_{\mathbf{x}}\mathbf{a}^{\mathbf{T}}\mathbf{x} = \mathbf{a} \tag{69}$$

$$\nabla_{\mathbf{X}}\,\mathrm{tr}\left(\mathbf{X}^{\mathbf{T}}\mathbf{A}\mathbf{X}\right) = \mathbf{X}\mathbf{A}^{\mathbf{T}} + \mathbf{X}\mathbf{A} \tag{108}$$

Then,

$$\begin{aligned}
\nabla_\theta f &= \nabla_\theta \frac{1}{2}\left[(\mathbf{y}-\mathbf{X}\theta)^{\mathbf{T}}(\mathbf{y}-\mathbf{X}\theta)+\lambda\|\theta\|_2^2\right] \\
&= \nabla_\theta \frac{1}{2}\left[\left(\mathbf{y}^{\mathbf{T}}-\theta^{\mathbf{T}}X^{\mathbf{T}}\right)(\mathbf{y}-\mathbf{X}\theta)+\lambda\|\theta\|_2^2\right] \\
&= \nabla_\theta \frac{1}{2}\left[\mathbf{y}^{\mathbf{T}}\mathbf{y}-\theta^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{y}-\mathbf{y}^{\mathbf{T}}\mathbf{X}\theta+\theta^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{X}\theta+\lambda\|\theta\|_2^2\right] \\
&= \frac{1}{2}\left[\nabla_\theta\left(\mathbf{y}^{\mathbf{T}}\mathbf{y}\right)-\nabla_\theta\left(\theta^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{y}\right)-\nabla_\theta\left(\mathbf{y}^{\mathbf{T}}\mathbf{X}\theta\right)+\nabla_\theta\left(\theta^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{X}\theta\right)+\nabla_\theta\left(\lambda\theta^{\mathbf{T}}\theta\right)\right] \\
&= \frac{1}{2}\left[0-\mathbf{X}^{\mathbf{T}}\mathbf{y}-\nabla_\theta\left(\mathbf{y}^{\mathbf{T}}\mathbf{X}\theta\right)+\nabla_\theta\left(\theta^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{X}\theta\right)+\nabla_\theta\lambda\left(\theta^{\mathbf{T}}\theta\right)\right] && \mathbf{a}:=\mathbf{X}^{\mathbf{T}}\mathbf{y}\to(69) \\
&= \frac{1}{2}\left[-\mathbf{X}^{\mathbf{T}}\mathbf{y}-\mathbf{X}^{\mathbf{T}}\mathbf{y}+\nabla_\theta\left(\theta^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{X}\theta\right)+\nabla_\theta\lambda\left(\theta^{\mathbf{T}}\theta\right)\right] && \mathbf{a}^{\mathbf{T}}:=\mathbf{y}^{\mathbf{T}}\mathbf{X}\to(69) \\
&= \frac{1}{2}\left[-2\mathbf{X}^{\mathbf{T}}\mathbf{y}+\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{\mathbf{T}}\theta+\mathbf{X}^{\mathbf{T}}\mathbf{X}\theta+\nabla_\theta\lambda\left(\theta^{\mathbf{T}}\theta\right)\right] && \mathbf{a}:=\theta^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{X},\mathbf{a}':=\mathbf{X}^{\mathbf{T}}\mathbf{X}\theta\to(69) \\
&= \frac{1}{2}\left[-2\mathbf{X}^{\mathbf{T}}\mathbf{y}+2\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{\mathbf{T}}\theta+\mathbf{X}^{\mathbf{T}}\mathbf{X}\theta+2\lambda\theta\right] && (69) \\
&= \frac{1}{2}\left[-2\mathbf{X}^{\mathbf{T}}\mathbf{y}+2\mathbf{X}^{\mathbf{T}}\mathbf{X}\theta+\mathbf{X}^{\mathbf{T}}\mathbf{X}\theta+2\lambda\theta\right] \\
&= \frac{1}{2}\left[-2\mathbf{X}^{\mathbf{T}}\mathbf{y}+2\mathbf{X}^{\mathbf{T}}\mathbf{X}\theta+2\lambda\theta\right] \\
\nabla_\theta f &= -\mathbf{X}^{\mathbf{T}}\mathbf{y}+\mathbf{X}^{\mathbf{T}}\mathbf{X}\theta+\lambda\theta
\end{aligned}$$

Then, setting $\nabla_\theta f = -\mathbf{X}^{\mathbf{T}}\mathbf{y}+\mathbf{X}^{\mathbf{T}}\mathbf{X}\theta+\lambda\theta=0$, we get

$$\begin{aligned}
0 &= -\mathbf{X}^{\mathbf{T}}\mathbf{y}+\mathbf{X}^{\mathbf{T}}\mathbf{X}\theta+\lambda\theta \\
0 &= -\mathbf{X}^{\mathbf{T}}\mathbf{y}+\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}+\lambda\mathbf{I}\right)\theta \\
\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y} &= \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}+\lambda\mathbf{I}\right)\theta \\
\theta &= \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y}
\end{aligned}$$