Linear Algebra Review

Akhil Vasvani

February 2019

1 Questions

Exercise 1. What is broadcasting in connection to Linear Algebra?

Proof. Typically for matrix addition, each matrix must be of the same dimension. However, in the context of deep learning, there is a shorthand notation for adding a matrix and a vector (the vector in this case is represented as a scalar). This vector ($\boldsymbol{b}$ for instance) is added to each row of the matrix, which eliminates the need to define a matrix with $\boldsymbol{b}$ copied into each row before doing the addition. This implicit copying of $\boldsymbol{b}$ to many locations is called **broadcasting**.

Example.

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + 1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

In this notation, the $\boldsymbol{b}$ is portrayed as a scalar 1, but in actuality it is a matrix of of 1's. This shorthand is broadcasting. □

Exercise 2. What are scalars, vectors, matrices, and tensors?

Proof. A **scalar** is just a single number or a matrix with a single entry.

Example. Let $s \in \mathbb{R}$ be the slope of the line, 4, or $\begin{bmatrix} 3 \end{bmatrix}$

A **vector** is a 1D array of numbers arranged in a particular order (a matrix with only one column). Another way to think of vectors is identifying points in space with each element giving the coordinate along a different axis.

Example. $\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix}^\top$

A matrix is a 2D array of numbers where each element is identified by two indices (ROW then COLUMN).

Example. $\boldsymbol{A} = \begin{bmatrix} 3 & 1 \\ 1 & 4 \end{bmatrix}$, where A has a height of two rows and a width of two columns. Hence, it is a shape of 2×2. The element 3 is found at $A_{1,1}$ and element 4 is found at $A_{2,2}$.

A tensor is a multi-dimensional array with more than two axes denoted as A where elements are at coordinates $\mathsf{A}_{i,j,k}$. □

Exercise 3. What is the Hadamard product of two matrices?

Proof. The Hadamard product (or element-wise product) is a matrix containing the product of the individual elements. It has a special symbol $\odot$. NOT TO BE CONFUSED WITH MATRIX MULTIPLICATION.

Example. $\boldsymbol{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ and $\boldsymbol{B} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$, so

$$\boldsymbol{AB} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} = \begin{bmatrix} 1 \cdot (2) + 1 \cdot (2) & 1 \cdot (2) + 1 \cdot (2) \\ 1 \cdot (2) + 1 \cdot (2) & 1 \cdot (2) + 1 \cdot (2) \end{bmatrix} = \begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix}, \textit{while}$$

$$\boldsymbol{A} \odot \boldsymbol{B} = \begin{bmatrix} 1 \cdot 2 & 1 \cdot 2 \\ 1 \cdot 2 & 1 \cdot 2 \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$$

Note the difference? □

Exercise 4. What is an inverse matrix?

Proof. The inverse matrix of $\boldsymbol{A}$ (denoted as $\boldsymbol{A^{-1}}$) is defined such that:

$$\boldsymbol{A A^{-1}} = \boldsymbol{A^{-1} A} = \boldsymbol{I_n},$$

where $\boldsymbol{I_n}$ is the identity matrix. □

Exercise 5. If the inverse of a matrix exists, how to calculate it?

Proof. There are a couple ways to calculate the inverse matrix (if it exists).

Option 1: Create an augmented matrix with $\boldsymbol{A}$ and the identity matrix $\boldsymbol{I_n}$ (represented as $\boldsymbol{A}|\boldsymbol{I_n}$). The goal would be to turn $\boldsymbol{A}$ into $\boldsymbol{I_n}$ using elementary row operations. In turn, whatever we do to $\boldsymbol{A}$, we do the same to $\boldsymbol{I_n}$ thereby turning $\boldsymbol{I_n}$ into $\boldsymbol{A^{-1}}$.

Example. Let's say $\boldsymbol{A} = \begin{bmatrix} 2 & -1 & 0 \\ 1 & 3 & -1 \\ -3 & 0 & 1 \end{bmatrix}$, so

$$\boldsymbol{A}|\boldsymbol{I_n} = \left[\begin{array}{ccc|ccc} 2 & -1 & 0 & 1 & 0 & 0 \\ 1 & 3 & -1 & 0 & 1 & 0 \\ -3 & 0 & 1 & 0 & 0 & 1 \end{array}\right]$$

Row Operation: $R_2 + R_3 \to R_3$

$$\Rightarrow \left[\begin{array}{ccc|ccc} 2 & -1 & 0 & 1 & 0 & 0 \\ 1 & 3 & -1 & 0 & 1 & 0 \\ -2 & 3 & 0 & 0 & 1 & 1 \end{array}\right]$$

Row Operation: $R_3 + R_1 \to R_1$

$$\Rightarrow \left[\begin{array}{ccc|ccc} 0 & 2 & 0 & 1 & 1 & 1 \\ 1 & 3 & -1 & 0 & 1 & 0 \\ -2 & 3 & 0 & 0 & 1 & 1 \end{array}\right]$$

Row Operations: $\frac{3}{2}R_1 - R_2 \to R_2$, $\frac{3}{2}R_1 - R_3 \to R_3$

$$\Rightarrow \left[\begin{array}{ccc|ccc} 0 & 3 & 0 & \frac{3}{2} & \frac{3}{2} & \frac{3}{2} \\ -1 & 0 & 1 & \frac{3}{2} & \frac{1}{2} & \frac{3}{2} \\ 2 & 0 & 0 & \frac{3}{2} & \frac{1}{2} & \frac{1}{2} \end{array}\right]$$

Row Operations: $\frac{1}{3}R_1 \to R_1$, $\frac{1}{2}R_3 \to R_3$

$$\Rightarrow \left[\begin{array}{ccc|ccc} 0 & 1 & 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ -1 & 0 & 1 & \frac{3}{2} & \frac{1}{2} & \frac{3}{2} \\ 1 & 0 & 0 & \frac{3}{4} & \frac{1}{4} & \frac{1}{4} \end{array}\right]$$

Row Operation: $R_3 + R_2 \to R_2$

$$\Rightarrow \left[\begin{array}{ccc|ccc} 0 & 1 & 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 & \frac{9}{4} & \frac{3}{4} & \frac{7}{4} \\ 1 & 0 & 0 & \frac{3}{4} & \frac{1}{4} & \frac{1}{4} \end{array}\right]$$

Row Operation: $R_3 \leftrightarrow R_2$

$$\Rightarrow \left[\begin{array}{ccc|ccc} 0 & 1 & 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 0 & \frac{3}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 1 & \frac{9}{4} & \frac{3}{4} & \frac{7}{4} \end{array}\right]$$

Row Operation: $R_1 \leftrightarrow R_2$

$$\Rightarrow \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & \frac{3}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 1 & 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 & \frac{9}{4} & \frac{3}{4} & \frac{7}{4} \end{array}\right]$$

Thus,

$$\boldsymbol{A^{-1}} = \begin{bmatrix} \frac{3}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{9}{4} & \frac{3}{4} & \frac{7}{4} \end{bmatrix} = \frac{1}{4}\begin{bmatrix} 3 & 1 & 1 \\ 2 & 2 & 2 \\ 9 & 3 & 7 \end{bmatrix}$$

Option 2: The next way involves using Minors, Cofactors, and Adjugate. First, calculate the matrix of minors: for each element of the matrix ignore the values of the current row and column and calculate the determinant of the remaining values. Second, apply a "checkerboard" of minuses to the "Matrix of Minors". In other words, we need to change the sign of alternate cells (+, -, +, etc.). Next, transpose all elements of the previous matrix (swap their positions over the diagonal) (This matrix will be called the Adjugate). Finally, find the determinant of the original matrix and multiply 1/determinant to the Adjugate matrix.

Example. Let's use the same $\boldsymbol{A}$ matrix in the previous example. Now let's calculate the determinant, which will be helpful for later before calculating the matrix of minors.

$$\det(A) = \begin{vmatrix} 2 & -1 & 0 \\ 1 & 3 & -1 \\ -3 & 0 & 1 \end{vmatrix} = 2\begin{vmatrix} 3 & -1 \\ 0 & 1 \end{vmatrix} - (-1)\begin{vmatrix} 1 & -1 \\ -3 & 1 \end{vmatrix} + 0\begin{vmatrix} 1 & 3 \\ -3 & 0 \end{vmatrix}$$

$$= 2((3 \times 1) - (-1 \times 0)) + (1)((1 \times 1) - (-1 \times -3)) + 0$$

$$= 2(3) + (1 - 3)$$
$$= 6 - 2$$
$$= 4$$

Now, we know the determinant let's solve for the matrix of minors.

Matrix of Minors =

$$\begin{bmatrix} ((3 \times 1) - (-1 \times 0)) & ((1 \times 1) - (-1 \times -3)) & ((1 \times 0) - (3 \times -3)) \\ ((-1 \times 1) - (0 \times 0)) & ((2 \times 1) - (-3 \times 0)) & ((2 \times 0) - (-1 \times -3)) \\ ((-1 \times -1) - (0 \times 3)) & ((2 \times -1) - (0 \times -3)) & ((2 \times 3) - (-1 \times 1)) \end{bmatrix} = \begin{bmatrix} 3 & -2 & 9 \\ -1 & 2 & -3 \\ 1 & -2 & 7 \end{bmatrix}$$

Matrix of Cofactors =

$$\begin{bmatrix} + & - & + \\ - & + & - \\ + & - & + \end{bmatrix} \Rightarrow \begin{bmatrix} +(3) & -(-2) & +(9) \\ -(-1) & +(2) & -(-3) \\ +(1) & -(-2) & +(7) \end{bmatrix} = \begin{bmatrix} 3 & 2 & 9 \\ 1 & 2 & 3 \\ 1 & 2 & 7 \end{bmatrix}$$

Adjugate Matrix is the transpose of the matrix of cofactors:

$$\begin{bmatrix} 3 & 2 & 9 \\ 1 & 2 & 3 \\ 1 & 2 & 7 \end{bmatrix}^{\top} = \begin{bmatrix} 3 & 1 & 1 \\ 2 & 2 & 2 \\ 9 & 3 & 7 \end{bmatrix}$$

Finally, to get $\boldsymbol{A^{-1}}$, we multiply $\frac{1}{\det(A)}$ to the Adjugate matrix:

$$\frac{1}{\det(A)} \begin{bmatrix} 3 & 1 & 1 \\ 2 & 2 & 2 \\ 9 & 3 & 7 \end{bmatrix} = \frac{1}{4} \begin{bmatrix} 3 & 1 & 1 \\ 2 & 2 & 2 \\ 9 & 3 & 7 \end{bmatrix} = \boldsymbol{A^{-1}}$$

□

Exercise 6. What is the determinant of a square matrix? How is it calculated? What is the connection of determinant to eigenvalues?

Proof. The determinant of a square matrix $\boldsymbol{A}$, denoted as $\det(A)$, is a function mapping matrices to real scalars. The absolute value of the determinant can be thought of as a measure of how much multiplication by the matrix expands or contracts space. If the determinant is 0, then space is contracted completely along at least one dimension, causing it to lose all of its volume. If the determinant is 1, then the transformation preserves volume. The determinant is equal to the product of all the eigenvalues of the matrix.

Example. You can calculate determinants either via Row or Column. Just remember the signs from the matrix of cofactors!

Using the matrix $\boldsymbol{A} = \begin{bmatrix} 2 & -1 & 0 \\ 1 & 3 & -1 \\ -3 & 0 & 1 \end{bmatrix}$, as our example let's start by solving for the $\det(A)$ via the first row.

$$\det(A) = \begin{vmatrix} 2 & -1 & 0 \\ 1 & 3 & -1 \\ -3 & 0 & 1 \end{vmatrix} = 2\begin{vmatrix} 3 & -1 \\ 0 & 1 \end{vmatrix} - (-1)\begin{vmatrix} 1 & -1 \\ -3 & 1 \end{vmatrix} + 0\begin{vmatrix} 1 & 3 \\ -3 & 0 \end{vmatrix}$$

$$= 2((3 \times 1) - (-1 \times 0)) + (1)((1 \times 1) - (-1 \times -3)) + 0$$

$$= 2(3) + (1 - 3)$$

$$= 6 - 2$$

$$= 4$$

Now, let's solve for $\det(A)$ via the second column:

$$\det(A) = \begin{vmatrix} 2 & -1 & 0 \\ 1 & 3 & -1 \\ -3 & 0 & 1 \end{vmatrix} = -(-1)\begin{vmatrix} 1 & -1 \\ -3 & 1 \end{vmatrix} + (3)\begin{vmatrix} 2 & 0 \\ -3 & 1 \end{vmatrix} - 0$$

$$= ((1 \times 1) - (-1 \times -3)) + 3((2 \times 1) - (0 \times -3))$$

$$= (1 - 3) + 3(2)$$

$$= 6 - 2$$

$$= 4$$

□

Exercise 7. Discuss span and linear dependence.

Proof. Using $\boldsymbol{Ax} = \boldsymbol{b}$ as a reference, we can look at the columns of $\boldsymbol{A}$. Think of the columns of $\boldsymbol{A}$ as specifying different directions we can travel from the origin (point specified to be 0 of all vectors), and determine how many ways there are to reach $\boldsymbol{b}$. Therefore, we can think of $\boldsymbol{x}$ specifying how far we should travel in each of these directions like this:

$$\boldsymbol{Ax} = \sum_i \boldsymbol{x}_i \boldsymbol{A}_{:,i}.$$

This is a linear combination in which each vector is multiplied by a scalar coefficient and adding the results. The span of a set of vectors is the set of all points obtainable by linear combination of the original vectors. (sounds sort of like a range?)

Example. Now consider $\boldsymbol{A} = \begin{bmatrix} 2 & 2 \\ 1 & 1 \end{bmatrix}$. Both columns ($\begin{bmatrix} 2 \\ 1 \end{bmatrix}$, $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$) are identical, so this matrix has the same column space or more formally put: $\text{Span}(\boldsymbol{A}) = \text{Span}(\begin{bmatrix} 2 & 2 \\ 1 & 1 \end{bmatrix}) = \text{Span}(\begin{bmatrix} 2 \\ 1 \end{bmatrix})$. The second column is a replica of the first column, so the span of $\boldsymbol{A}$ is 2×1. This type of redundancy is called linear dependence. A set of vectors is linearly independent if no vector in the set if a linear combination of the other vectors. Such as:

$$\boldsymbol{M} = \begin{bmatrix} 2 & 4 \\ 1 & 3 \end{bmatrix}$$

□

Exercise 8. What is $\boldsymbol{Ax} = \boldsymbol{b}$? When does $\boldsymbol{Ax} = \boldsymbol{b}$ has a unique solution?

Proof. $\boldsymbol{Ax} = \boldsymbol{b}$ is a system of linear equations where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is a known matrix, $\boldsymbol{b} \in \mathbb{R}^m$ is a known vector, and $\boldsymbol{x} \in \mathbb{R}^n$ is a vector of unknown variables we would like to solve for. Observe matrix $\boldsymbol{A}^{m \times n}$. $\boldsymbol{Ax} = \boldsymbol{b}$ will have a unique solution if each column is linearly independent and $\boldsymbol{b}$ lies in the column space of $\boldsymbol{A}$.

Example. Say $\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & -1 & 1 \\ 3 & 0 & -1 \end{bmatrix}$ and $\boldsymbol{b} = \begin{bmatrix} 9 \\ 8 \\ 3 \end{bmatrix}$, then the augmented matrix $\boldsymbol{A}|\boldsymbol{b}$ in row-reduced echelon form is:

$$\left[\begin{array}{ccc|c} 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 3 \end{array}\right]$$

Each of the columns is linearly independent and $\boldsymbol{b}$ lies in the column space, so $\boldsymbol{Ax} = \boldsymbol{b}$ has a unique solution in which $\boldsymbol{x} = \begin{bmatrix} 2 \\ -1 \\ 3 \end{bmatrix}$.

□

Exercise 9. In $\boldsymbol{Ax} = \boldsymbol{b}$, what happens when $\boldsymbol{A}$ is fat or tall?

Proof. Now depending on the **rank**—the number of first nonzero entry of each row (**pivots**) in a reduced row-echelon form matrix—affects the number of solutions for either a tall or fat matrix.

Note. A matrix is in reduced row-echelon form *if (1) it is in row-echelon form, (2) all of the pivots are equal to 1, and (3) all entries in the pivot columns, except for the pivots themselves, are equal to zero.*

$$A_{rref} = \begin{bmatrix} 1 & 0 & 4 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Here the $\text{rank}[A_{rref}] = 3$.

Now the general rule of thumb is:

1) $\boldsymbol{Ax} = \boldsymbol{b}$ *is inconsistent (i.e. no solution exists) if and only if* $\text{rank}[\boldsymbol{A}] < \text{rank}[\boldsymbol{A}\,|\,\boldsymbol{b}]$.

2) $\boldsymbol{Ax} = \boldsymbol{b}$ *has a unique solution if and only if* $\text{rank}[\boldsymbol{A}] = \text{rank}[\boldsymbol{A}\,|\,\boldsymbol{b}] = n$.

3) $\boldsymbol{Ax} = \boldsymbol{b}$ *has infinitely many solutions if and only if* $\text{rank}[\boldsymbol{A}] = \text{rank}[\boldsymbol{A}\,|\,\boldsymbol{b}] < n$.

Since $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, if $\boldsymbol{A}$ is a tall matrix, then the numbers of rows $m >$ the number of columns n. If $\text{rank}(\boldsymbol{A}) < \text{rank}[\boldsymbol{A}\,|\,\boldsymbol{b}]$, then there is no solution—no linear combination will reach the desired $\boldsymbol{b}$—for all $\boldsymbol{b}$. If $\text{rank}(\boldsymbol{A}) = n$, then there is ONLY one unique solution for every $\boldsymbol{b}$. In laymen's terms, if $\text{rank}(\boldsymbol{A}) = n$, then there is one solution that is unique; otherwise, there does not exist a solution. Note if $\boldsymbol{b} = 0$, there there is unique (very obvious) solution if $\text{rank}(\boldsymbol{A}) = n$—hint: it is 0—but if $\text{rank}(\boldsymbol{A}) < n$, there are infinitely many solutions.

TL;DR. For most tall matrices, there are usually no solutions.

However, if $\boldsymbol{A}$ is a fat matrix then the number of columns $n >$ the number of rows m. If $\text{rank}[\boldsymbol{A}] < m$, then either there does not exist a solution or there are infinitely many solutions. However, if $\text{rank}[\boldsymbol{A}] = m$, there exists at least one solution for all $\boldsymbol{b}$—which suggests that there can be more than one solution (thereby infinitely many solutions). Note, if $\boldsymbol{b} = 0$ then there are infinite solutions.

TL;DR. For most fat matrices, there are usually infinitely many solutions that exist—there are an infinite number of linear combinations to reach the desired $\boldsymbol{b}$.

Example. Let's look for a solution to $\boldsymbol{Ax} = \boldsymbol{b}$, where

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 3 & 0 \\ 2 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \ \boldsymbol{x} = \begin{bmatrix} w \\ x \\ y \\ z \end{bmatrix}, \ \boldsymbol{b} = \begin{bmatrix} 9 \\ 8 \\ b_3 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 & 2 & 3 & 0 \\ 2 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} w \\ x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 9 \\ 8 \\ b_3 \end{bmatrix}.$$

The augmented matrix $\boldsymbol{A}|\boldsymbol{b}$ in row-reduced echelon form:

$$\left[\begin{array}{cccc|c} 1 & 0 & -1 & 0 & -5 \\ 0 & 1 & 1 & 0 & 2 \\ 0 & 0 & 0 & 0 & b_3 \end{array}\right] \Rightarrow \begin{bmatrix} w \\ x \\ y \\ z \end{bmatrix} = \begin{bmatrix} y-5 \\ 2-y \\ y \\ z \end{bmatrix} = \begin{bmatrix} -5+y \\ 2-y \\ y \\ z \end{bmatrix} = \begin{bmatrix} -5 \\ 2 \\ 0 \\ 0 \end{bmatrix} + y \begin{bmatrix} 1 \\ -1 \\ 1 \\ 0 \end{bmatrix} + z \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

Both y, z can be any $\mathbb{R}$. Now, let's pay attention to the constant b_3. If $b_3 = 0$, then $\text{rank}[\boldsymbol{A}] = \text{rank}[\boldsymbol{A}\,|\,\boldsymbol{b}]$, which $< n$— it is also $< m$. So, there are an infinite number of linear combinations to reached the desired $\boldsymbol{b}$. On the other hand, if $b_3 \neq 0$, then $\text{rank}[\boldsymbol{A}] < \text{rank}[\boldsymbol{A}\,|\,\boldsymbol{b}]$, so there does not exist a solution. Therefore, matrix $\boldsymbol{A}$ is a fat matrix.

Note, if $\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 3 & 0 \\ 2 & -1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \Rightarrow \left[\begin{array}{cccc|c} 1 & 0 & -1 & 0 & -5 \\ 0 & 1 & 1 & 0 & 2 \\ 0 & 0 & 0 & 1 & b_3 \end{array}\right] \Rightarrow z = b_3.$

So $b_3 \in \mathbb{R} \rightarrow z \in \mathbb{R}$, which means that any value of z can satisfy this equation. Hence, there are an infinite number of solutions.

Example. Let's look another example and find a solution to $\boldsymbol{Ax} = \boldsymbol{b}$, where

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & -1 & 1 \\ 3 & 0 & -1 \\ 0 & 0 & 0 \end{bmatrix}, \ \boldsymbol{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \ \boldsymbol{b} = \begin{bmatrix} 9 \\ 8 \\ 3 \\ b_4 \end{bmatrix}.$$

The augmented matrix $\boldsymbol{A}|\boldsymbol{b}$ in row-reduced echelon form:

$$\left[\begin{array}{ccc|c} 1 & 0 & 0 & 2 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & b_4 \end{array}\right] \Rightarrow \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

If the values for $b_4 = 0$ then $\text{rank}[\boldsymbol{A}] = \text{rank}[\boldsymbol{A}\,|\,\boldsymbol{b}]$, which $= n$. Thus, there is a unique solution. However, $b_4 \neq 0$, then $\text{rank}[\boldsymbol{A}] < \text{rank}[\boldsymbol{A}\,|\,\boldsymbol{b}]$, which means there does not exist a solution. Therefore, at most this system has one solution. Thus, $\boldsymbol{A}$ is a tall matrix.

□

Exercise 10. When does an inverse of A exist?

Proof. Inverse of $\boldsymbol{A}$ (or $\boldsymbol{A^{-1}}$) exists if in $\boldsymbol{Ax} = \boldsymbol{b}$, there has to be n linearly independent columns and at most one solution for every value of b. This will make the matrix a square matrix (number of m rows = number of n columns) and nonsingular (columns are linearly independent). Also, another check is that the determinant cannot be equal to 0. □

Exercise 11. What is a norm? What is L^1, L^2 and L^∞ norm?

Proof. A norm is a function that measures the size of vectors. Simply speaking, the norm of a $\boldsymbol{x}$ measures the distance from the origin to the point x.

L^2 norm (pronounced L-two) is known as the Euclidean Distance: $||\boldsymbol{x}||_2 = \sqrt{\sum_i |\boldsymbol{x}_i|^2}$. It is so commonly used in machine learning that the 2 subscript is dropped out, so it's just $||\boldsymbol{x}||$. Remember, it's the Euclidean distance from the origin to the point x.

L^1 norm (pronounced L-one) is commonly referred to as Manhattan distance:

$$||\boldsymbol{x}||_1 = \sum_i |\boldsymbol{x}_i|.$$

L^∞ norm (also known as max norm) is the absolute value of the element with the largest magnitude in the vector: $||\boldsymbol{x}||_\infty = \max_i |x_i|$.

Example. Say $\boldsymbol{x} = \begin{bmatrix} -6 \\ 3 \\ 4 \end{bmatrix}$ then

L^2 norm =

$$||\boldsymbol{x}||_2 = \sqrt{(-6)^2 + (3)^2 + (4)^2} = \sqrt{36 + 9 + 16} = \sqrt{61} \approx 7.810.$$

L^1 norm =

$$||\boldsymbol{x}||_1 = |-6| + |3| + |4| = 13.$$

L^∞ norm =

$$\max_i |\boldsymbol{x}| = \max_{i=3} |x_3| = 6$$

□

Exercise 12. What are the conditions a norm has to satisfy?

Proof. More formally put, a norm is any function which satisfies the following properties:

1) $f(\mathbf{x}) = 0 \Rightarrow \mathbf{x} = \mathbf{0}$
2) $f(\mathbf{x}+\mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y})$ (the triangle inequality)
3) $\forall \alpha \in \mathbb{R}, f(\alpha \mathbf{x}) = |\alpha| f(\mathbf{x})$

□

Exercise 13. Why is squared of L^2 norm preferred in Machine Learning than just L^2 norm?

Proof. The squared L^2 norm is mathematically and computationally convenient to work with than the L^2 norm itself. For example, the derivatives of the squared L^2 norm with respect to each element of $\boldsymbol{x}$ depend only on the corresponding element of $\boldsymbol{x}$, while all the derivatives of the L^2 norm may depend on the entire vector.

Example. From our definition,

$$L^2 \text{ norm} = ||\boldsymbol{x}||_2 = \sqrt{\sum_i |\boldsymbol{x}_i|^2}\ , \quad \text{so squared of } L^2 \text{ norm} = ||\boldsymbol{x}||_2^2 = \sum_i |\boldsymbol{x}_i|^2.$$

Now, if we take the derivative of the squared L^2 norm:

$$\frac{\partial ||\boldsymbol{x}||_2^2}{\partial x_j} = \frac{\partial}{\partial x_j}\Big[\sum_i |\boldsymbol{x}_i|^2\Big] = \sum_i \frac{\partial}{\partial x_j}|\boldsymbol{x_i}|^2 = 0 + ... + 0 + 2|x_j| + 0 + ... + 0 = 2|x_j|$$

when $i = j$.

Now, if we take the derivative of the L^2 norm:

$$\frac{\partial ||\boldsymbol{x}||_2}{\partial x_j} = \frac{\partial}{\partial x_j}\Big[\sqrt{\sum_i |\boldsymbol{x}_i|^2}\Big] = \frac{1}{2}\Big(\sum_i |\boldsymbol{x_i}|^2\Big)^{-\frac{1}{2}} \cdot \sum_i \frac{\partial}{\partial x_j}|\boldsymbol{x_i}|^2 = \frac{|x_j|}{\sqrt{\sum_i |\boldsymbol{x_i}|^2}}.$$

Also when $i = j$ for the numerator.

Therefore, it is easier to rely on the squared L^2 because you do not need the whole vector to compute its gradient. You just need the specific element.

□

Exercise 14. When L^1 norm is preferred over L^2 norm?

Proof. However, in many contexts, the squared L^2 norm may be undesirable because it increases very slowly near the origin. And in several machine learning applications, it is important to discriminate between elements that are exactly 0 and elements that are small but nonzero. Hence, we turn to the L^1 norm which grows at the same rate in all locations. Every time an element of x moves away from 0 by ϵ, the L^1 norm increases by ϵ. □

Exercise 15. Can the number of nonzero elements in a vector be defined as L^0 norm? If not, why?

Proof. No, this is not correct. The number of nonzero entries in a vector is not a norm because scaling the vector by α (property 3) does not change the number of nonzero entries. The L^1 norm is often used as a substitute for the number of nonzero entries.

□

Exercise 16. What is Frobenius norm?

Proof. The Frobenius norm measures the size of a matrix:

$$||\boldsymbol{A}||_F = \sqrt{\sum_{i,j} \boldsymbol{A}_{i,j}^2},$$

which is analogous to L^2 norm of a vector.

Example. Say $\boldsymbol{A} = \begin{bmatrix} 2 & -1 & 5 \\ 0 & 2 & 1 \\ 3 & 1 & 1 \end{bmatrix}$, so $||\boldsymbol{A}||_F = \sqrt{\sum_{i=1}^{3}\sum_{j=1}^{3} \boldsymbol{A}_{i,j}^2}$

$$= \sqrt{(2)^2 + (-1)^2 + (5)^2 + (0)^2 + (2)^2 + (1)^2 + (3)^2 + (1)^2 + (1)^2}$$

$$= \sqrt{4 + 1 + 25 + 4 + 1 + 9 + 1 + 1}$$

$$= \sqrt{2 \cdot 4 + 4 \cdot 1 + 25 + 9}$$

$$= \sqrt{3 \cdot 4 + 34}$$

$$= \sqrt{46}$$

□

Exercise 17. What is a diagonal matrix?

Proof. Diagonal matrices consist mostly of zeros and have nonzero entries only along the main diagonal. Formally, a matrix $\boldsymbol{D}$ is diagonal if and only if $D_{i,j} = 0$ for all $i \neq j$.

Notation: diag($\mathbf{v}$) denotes a square diagonal matrix whose diagonal entries are given by the entries of the vector v.

Example. Say $\mathbf{v} = \begin{bmatrix} 6 \\ 7 \\ 19 \end{bmatrix}$, so $\text{diag}(\mathbf{v}) = \begin{bmatrix} 6 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & 19 \end{bmatrix}$. The identity matrix ($I$) is also a diagonal matrix.

□

Exercise 18. Why is multiplication by diagonal matrix computationally cheap? How is the multiplication different for square vs. non-square diagonal matrix?

Proof. Multiplication by diagonal matrix is computationally cheap because to compute diag($\mathbf{v}$)$\mathbf{x}$, we only need to scale each element x_i by v_i. Simply put: $\text{diag}(\mathbf{v})\mathbf{x} = \mathbf{v} \odot \mathbf{x}$.

Example. $\text{diag}(\mathbf{v}) = \begin{bmatrix} 6 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & 19 \end{bmatrix}$ and $\boldsymbol{x} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$, so

$$\text{diag}(\mathbf{v})\mathbf{x} = \begin{bmatrix} 6 \cdot 1 & 0 \cdot 2 & 0 \cdot 3 \\ 0 \cdot 4 & 7 \cdot 5 & 0 \cdot 6 \\ 0 \cdot 7 & 0 \cdot 8 & 19 \cdot 9 \end{bmatrix} = \begin{bmatrix} 6 & 0 & 0 \\ 0 & 35 & 0 \\ 0 & 0 & 171 \end{bmatrix}$$

However, not diagonal matrices need be square. It is possible to construct a rectangular diagonal matrix. Non-square diagonal matrices do not have inverses but it is still possible to multiply by them cheaply. For a non-square diagonal matrix $\boldsymbol{D}$, the product $\boldsymbol{Dx}$ will involve scaling each element of $\boldsymbol{x}$, and either concatenating some zeros to the result if $\boldsymbol{D}$ is taller than it is fat, or discarding some of the last elements of the vector if $\boldsymbol{D}$ is fatter than it is tall.

Example. $\boldsymbol{D}$ is taller than it is fat. $\boldsymbol{D} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & -3 \\ 0 & 0 & 0 \end{bmatrix}$ and $\boldsymbol{x} = \begin{bmatrix} 3 & 2 & 1 \\ 6 & 5 & 4 \\ 9 & 8 & 7 \end{bmatrix}$, so

$$\boldsymbol{Dx} = \begin{bmatrix} 1 \cdot 3 & 1 \cdot 2 & 1 \cdot 1 \\ 4 \cdot 6 & 4 \cdot 5 & 4 \cdot 4 \\ -3 \cdot 9 & -3 \cdot 8 & -3 \cdot 7 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 3 & 2 & 1 \\ 24 & 20 & 16 \\ -27 & -24 & -21 \\ 0 & 0 & 0 \end{bmatrix}.$$

Now, let's concatenate another column of 0's to make this a square matrix:

$$\boldsymbol{Dx} + \boldsymbol{0} = \begin{bmatrix} 3 & 2 & 1 & 0 \\ 24 & 20 & 16 & 0 \\ -27 & -24 & -21 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Example. $\boldsymbol{D}$ is fat. $\boldsymbol{D} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & -3 & 0 \end{bmatrix}$ and $\boldsymbol{x} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 9 & 8 & 7 & 6 \\ 5 & 4 & 3 & 2 \end{bmatrix}$, so

$$\boldsymbol{Dx} = \begin{bmatrix} 1 \cdot 1 & 1 \cdot 2 & 1 \cdot 3 & 1 \cdot 4 \\ 4 \cdot 5 & 4 \cdot 6 & 4 \cdot 7 & 4 \cdot 8 \\ -3 \cdot 9 & -3 \cdot 8 & -3 \cdot 7 & -3 \cdot 6 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 20 & 24 & 28 & 32 \\ -27 & -24 & -21 & -18 \end{bmatrix}.$$

Now, let's remove the last column to turn this into a square matrix:

$$\rightarrow \begin{bmatrix} 1 & 2 & 3 \\ 20 & 24 & 28 \\ -27 & -24 & -21 \end{bmatrix}$$

□

Exercise 19. At what conditions does the inverse of a diagonal matrix exist?

Proof. The inverse of a diagonal matrix exists only if the matrix is square and every diagonal entry is nonzero, and in that case: $\text{diag}(\mathbf{v})^{-1} = \text{diag}([\frac{1}{\mathbf{v}_1}, ..., \frac{1}{\mathbf{v}_n}])^\top$

Example. Say $\text{diag}(\mathbf{v}) = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 7 \end{bmatrix}$, so $\text{diag}(\mathbf{v})^{-1} = \begin{bmatrix} \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{5} & 0 \\ 0 & 0 & \frac{1}{7} \end{bmatrix}$.

To prove that $\text{diag}(\mathbf{v})^{-1}$ is in fact the inverse: $\text{diag}(\mathbf{v})^{-1}\text{diag}(\mathbf{v}) =$

$$\begin{bmatrix} 3 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 7 \end{bmatrix} \begin{bmatrix} \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{5} & 0 \\ 0 & 0 & \frac{1}{7} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \boldsymbol{I_3}$$

Note, that the diagonal matrices are symmetric.

□

Exercise 20. What is a symmetric matrix?

Proof. A symmetric matrix is any matrix that is equal to its own transpose: $\boldsymbol{A} = \boldsymbol{A}^\top$.

Example. $\boldsymbol{A} = \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix}$, so $\boldsymbol{A}^\top = \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix}$. Thus, $\boldsymbol{A} = \boldsymbol{A}^\top$.

□

Exercise 21. What is a unit vector?

Proof. A unit vector is a vector with unit norm: $||\boldsymbol{x}||_2 = 1$ □

Exercise 22. When are two vectors x and y orthogonal?

Proof. A vector $\boldsymbol{x}$ and a vector $\boldsymbol{y}$ are orthogonal to each other if $\boldsymbol{x}^\top \boldsymbol{y} = 0$. If both vectors have nonzero norm, this means that they are at a 90° angle to each other.

Note. You can represent the dot product of two vectors in terms of norms, which helps prove when two vectors are orthogonal:

$$\boldsymbol{x}^\top \boldsymbol{y} = ||\boldsymbol{x}||_2 ||\boldsymbol{y}||_2 \cos\theta.$$

Hence, this equation holds true when $\theta = \frac{\pi}{2}$ or 90°, which means that vectors $\boldsymbol{x}, \boldsymbol{y}$ must be perpendicular to each other to be orthogonal.

Note. The dot product of two vectors is also called the **inner product**, which is denoted as:

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \boldsymbol{x}^\top \boldsymbol{y} = \sum_{i=1}^{m} x_i y_i. \tag{1}$$

Note the output is a scalar not a matrix.

□

Exercise 23. At $\mathbb{R}^n$, what is the maximum possible number of orthogonal vectors with non-zero norm?

Proof. In $\mathbb{R}^n$, at most n vectors may be mutually orthogonal with nonzero norm. □

Exercise 24. When are two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ orthonormal?

Proof. If two vectors are not only orthogonal but also have unit norm, then they are orthonormal. □

Exercise 25. What is an orthogonal matrix? Why is computationally preferred?

Proof. An orthogonal matrix is a square matrix whose rows are mutually orthonormal and whose columns are mutually orthonormal: $\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{I} \rightarrow \boldsymbol{A^{-1}} = \boldsymbol{A}^\top$.

Orthogonal matrices are of interest because their inverse is very cheap to compute.

Example. Orthogonal matrix $\boldsymbol{A} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$

□

Exercise 26. What is eigendecomposition, eigenvectors and eigenvalues?

Proof. Eigendecompostion decomposes a matrix into a set of eigenvectors and eigenvalues (kind of like how we can break down a number into its prime factors).

An eigenvector of a square matrix $\boldsymbol{A}$ is a nonzero vector $\mathbf{v}$ such that multiplication by $\boldsymbol{A}$ alters only on the scale of $\mathbf{v}$: $\boldsymbol{Av} = \lambda \boldsymbol{v}$ (the scalar *lambda* is known as the eigenvalue corresponding to the eigenvector). □

Exercise 27. How to find eigenvalues of a matrix?

Proof. To find the eigenvalues of a matrix, we find the values of λ which satisfy the characteristic equation of the matrix $\boldsymbol{A}$: $\det(\boldsymbol{A} - I\lambda) = 0$. Then, plug the solved lambda values in to the matrix $\boldsymbol{A}$ to get the augmented matrix. Then row-reduce to find the eigenvectors.

Note: matrix $\boldsymbol{A}$ must be a square matrix.

Example. Let's find the eigenvalues and then the eigenvectors of matrix $\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}$.

$$\det(\boldsymbol{A} - I_2\lambda) = 0 \rightarrow \det(\begin{bmatrix} -\lambda & 1 \\ -2 & -3-\lambda \end{bmatrix}) = 0$$

$$\rightarrow \begin{vmatrix} -\lambda & 1 \\ -2 & -3-\lambda \end{vmatrix} = 0$$

$$(-\lambda)(-3-\lambda) - (-2)(1) = 0$$

$$(\lambda)(3+\lambda) + 2 = 0$$

$$3\lambda + \lambda^2 + 2 = 0$$

$$(\lambda+1)(\lambda+2) = 0$$

$$\lambda = -1, -2$$

These are our eigenvalues. Now, let's find their corresponding eigenvectors.

Using the equation $\boldsymbol{Av} = \lambda \boldsymbol{v}$, isolate $\boldsymbol{v}$ (our eigenvector): $(\boldsymbol{A} - \boldsymbol{I_2}\lambda)\boldsymbol{v} = 0$. Plug in the first λ value of -1 into the equation $\rightarrow (\boldsymbol{A} - I_2(-1))\boldsymbol{v} = (\boldsymbol{A} + I_2)\boldsymbol{v} = 0$

$$= \begin{bmatrix} 0+1 & 1 \\ -2 & -3+1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -2 & -2 \end{bmatrix}.$$

Row-reducing this matrix:

Row Operation: $2R_1 + R_2 \rightarrow R_2$

$$\Rightarrow \left[\begin{array}{cc|c} 1 & 1 & 0 \\ 0 & 0 & 0 \end{array}\right].$$

Now,

$$\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \boldsymbol{v} = 0$$

$$\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0 \Rightarrow x + y = 0 \Rightarrow x = -y,$$

so $\boldsymbol{v}_1 = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -y \\ y \end{bmatrix} = y \begin{bmatrix} -1 \\ 1 \end{bmatrix} \Rightarrow k \begin{bmatrix} -1 \\ 1 \end{bmatrix}$. To normalize vector, set $k = \frac{1}{\sqrt{(-1)^2+(1)^2}} = \frac{1}{\sqrt{2}}$.

Plug in the second λ value of -2 into the equation $\rightarrow (\boldsymbol{A} - I_2(-2))\boldsymbol{v} = (\boldsymbol{A} + 2I_2)\boldsymbol{v} = 0$

$$= \begin{bmatrix} 0+2 & 1 \\ -2 & -3+2 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ -2 & -1 \end{bmatrix}.$$

Row-reducing this matrix:

Row Operation: $R_1 + R_2 \rightarrow R_2$

$$\Rightarrow \left[\begin{array}{cc|c} 2 & 1 & 0 \\ 0 & 0 & 0 \end{array}\right].$$

Now,

$$\begin{bmatrix} 2 & 1 \\ 0 & 0 \end{bmatrix} \boldsymbol{v} = 0$$

$$\begin{bmatrix} 2 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0 \Rightarrow 2x + y = 0 \Rightarrow x = -\frac{y}{2},$$

so $\boldsymbol{v}_2 = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -\frac{y}{2} \\ y \end{bmatrix} = y \begin{bmatrix} -\frac{1}{2} \\ 1 \end{bmatrix} \Rightarrow k \begin{bmatrix} -\frac{1}{2} \\ 1 \end{bmatrix}$. To normalize vector, set $k = \frac{1}{\sqrt{\left(-\frac{1}{2}\right)^2+(1)^2}} = \frac{2}{\sqrt{5}}$

□

Exercise 28. Write the eigendecomposition formula for a matrix. If the matrix is real symmetric, how will this change?

Proof. The eigendecomposition of $\boldsymbol{A}$ is given by:

$$\boldsymbol{A} = \boldsymbol{V}\text{diag}(\boldsymbol{\lambda})\boldsymbol{V}^{-1}, \tag{2}$$

where matrix $\boldsymbol{V}$ is all the eigenvectors with one eigenvector per column: $\boldsymbol{V} = [\boldsymbol{v}^{(1)}, ..., \boldsymbol{v}^{(n)}]$ and likewise concatenating all the eigenvalues to form a vector $\boldsymbol{\lambda} = [\lambda_1, ..., \lambda_n]^\top$

Example. Using our previous example from above, find the eigendecomposition of $\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}$.

Now, as calculated from above, both the eigenvalues ($\lambda = -1, -2$) and the eigenvectors ($\boldsymbol{v}_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \boldsymbol{v}_2 = \begin{bmatrix} -\frac{1}{2} \\ 1 \end{bmatrix}$) are known. So to create matrix $\boldsymbol{V}$,stack the eigenvector corresponding to the largest eigenvalue first followed by the second.

$$\boldsymbol{V} = \begin{bmatrix} -1 & -\frac{1}{2} \\ 1 & 1 \end{bmatrix}.$$

So,

$$\det(\boldsymbol{V}) = \begin{vmatrix} -1 & -\frac{1}{2} \\ 1 & 1 \end{vmatrix} = -1(1) - (-\frac{1}{2})(1) = -1 + \frac{1}{2} = -\frac{1}{2}.$$

Because the determinant is nonzero, $\boldsymbol{V}^{-1}$ does exist. $\boldsymbol{V}^{-1} = \frac{1}{\det(\boldsymbol{V})} \begin{bmatrix} 1 & \frac{1}{2} \\ -1 & -1 \end{bmatrix} = \begin{bmatrix} -2 & -1 \\ 2 & 2 \end{bmatrix}$ and $\text{diag}(\boldsymbol{\lambda}) = \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix}$.

Plugging this all into the equation:

$$\boldsymbol{A} = \boldsymbol{V}\text{diag}(\boldsymbol{\lambda})\boldsymbol{V}^{-1} = \begin{bmatrix} -1 & -\frac{1}{2} \\ 1 & 1 \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix} \begin{bmatrix} -2 & -1 \\ 2 & 2 \end{bmatrix} = \begin{bmatrix} -1(-1)+0 & 0+(-2)(-\frac{1}{2}) \\ -1(1)+0 & 0-2 \end{bmatrix} \begin{bmatrix} -2 & -1 \\ 2 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 1 \\ -1 & -2 \end{bmatrix} \begin{bmatrix} -2 & -1 \\ 2 & 2 \end{bmatrix} = \begin{bmatrix} -2(1)+2(1) & -1(1)+2(1) \\ -2(-1)+(-2)(2) & (-1)(-1)+2(-2) \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}.$$

Note: there are infinite eigenvectors corresponding with a specific eigenvalues. While this may change the determinant, the important point here is that the ratio of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ remains the same.

For real symmetric matrices,

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top, \tag{3}$$

where $\boldsymbol{Q}$ is an orthogonal matrix composed of eigenvectors of $\boldsymbol{A}$, and Λ is a diagonal matrix. The eigenvalue $\Lambda_{i,j}$ is associated with the eigenvector of column i of $\boldsymbol{Q}_{:,i}$. Because $\boldsymbol{Q}$ is an orthogonal matrix think of $\boldsymbol{A}$ as scaling space by λ_i in the direction $\mathbf{v}^{(i)}$.

Example. Matrix $\boldsymbol{A} = \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix}$ is real symmetric. So, let's find the eigendecomposition.

First, solve the characteristic equation: $\det(\boldsymbol{A} - I_2\lambda) = 0$.

$$\rightarrow \det(\begin{bmatrix} 3-\lambda & 2 \\ 2 & 3-\lambda \end{bmatrix}) = 0 \rightarrow \begin{vmatrix} 3-\lambda & 2 \\ 2 & 3-\lambda \end{vmatrix} = 0$$

Difference of Squares: $\rightarrow (3-\lambda)^2 - (2)^2 = 0$

$$(3-\lambda+2)(3-\lambda-2) = 0$$

$$(5-\lambda)(1-\lambda) = 0$$

$$\lambda = 5, 1$$

These are our eigenvalues. Let's find their corresponding eigenvectors.

$\lambda = 5 : \boldsymbol{A} - \boldsymbol{I}_2(5)$

$$= \begin{bmatrix} 3-5 & 2 \\ 2 & 3-5 \end{bmatrix} = \begin{bmatrix} -2 & 2 \\ 2 & -2 \end{bmatrix}$$

Row-reducing this matrix:
Row Operation: $R_1 + R_2 \to R_2$

$$\Rightarrow \left[\begin{array}{cc|c} -2 & 2 & 0 \\ 0 & 0 & 0 \end{array}\right]$$

Row Operation: $-\frac{1}{2}R_1 \to R_1$

$$\Rightarrow \left[\begin{array}{cc|c} 1 & -1 & 0 \\ 0 & 0 & 0 \end{array}\right]$$

Now, remember the equation $(\boldsymbol{A} - \boldsymbol{I}_2\lambda)\boldsymbol{v} = 0$? Time to use it again:

$$\begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} \mathbf{v} = 0 \to \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0 \to x = y.$$

Thus, $\boldsymbol{v}_1 = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} y \\ y \end{bmatrix} = y \begin{bmatrix} 1 \\ 1 \end{bmatrix} = k \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. To normalize vector, let $k = \frac{1}{\sqrt{2}}$
$\lambda = 1 : \boldsymbol{A} - \boldsymbol{I}_2(1)$

$$= \begin{bmatrix} 3-1 & 2 \\ 2 & 3-1 \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$$

Row-reducing this matrix:
Row Operation: $R_1 - R_2 \to R_2$

$$\Rightarrow \left[\begin{array}{cc|c} 2 & 2 & 0 \\ 0 & 0 & 0 \end{array}\right]$$

Row Operation: $\frac{1}{2}R_1 \to R_1$

$$\Rightarrow \left[\begin{array}{cc|c} 1 & 1 & 0 \\ 0 & 0 & 0 \end{array}\right]$$

$$\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \mathbf{v} = 0 \to \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0 \to x = -y.$$

Therefore, $\boldsymbol{v}_2 = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -y \\ y \end{bmatrix} = y \begin{bmatrix} -1 \\ 1 \end{bmatrix} = k \begin{bmatrix} -1 \\ 1 \end{bmatrix}$. To normalize vector, let $k = \frac{1}{\sqrt{2}}$.
Now, stacking these eigenvectors together (from highest eigenvalue to lowest),

$$\boldsymbol{Q} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$$

$$\boldsymbol{Q}^\top = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$$

So,

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$$

$$\Rightarrow \boldsymbol{A} = \left(\frac{1}{\sqrt{2}}\right)^2 \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 5 & -1 \\ 5 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 6 & 4 \\ 4 & 6 \end{bmatrix}$$
$$= \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix}$$

□

Exercise 29. Is the Eigendecomposition guaranteed to be unique? If not, then how do we represent it?

Proof. While any real symmetric matrix $\boldsymbol{A}$ is guaranteed to have an eigendecomposition, the eigendecomposition may not be unique. If any two or more eigenvectors share the same eigenvalue, then any set of orthogonal vectors lying in their span are also eigenvectors with that eigenvalue, and we could equivalently choose a $\boldsymbol{Q}$ using those eigenvectors instead. By convention, we usually sort the entries of $\boldsymbol{\Lambda}$ in descending order. Under this convention, the eigendecomposition is unique only if all of the eigenvalues are unique. □

Exercise 30. What are positive definite, negative definite, positive semi definite and negative semi definite matrices?

Proof. A matrix whose eigenvalues are all positive is called positive definite. A matrix whose eigenvalues are all positive or zero-valued is called positive semidefinite. Likewise, if all eigenvalues are negative, the matrix is negative definite, and if all eigenvalues are negative or zero-valued, it is negative semidefinite.

Positive semidefinite matrices are interesting because they guarantee that $\forall \boldsymbol{x}, \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \geq 0$.

Positive definite matrices additionally guarantee that $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = 0, \rightarrow \boldsymbol{x} = 0$.

Note. So why do we care?

Imagine, there is a vector $\boldsymbol{z}$, which will have a certain direction. When we multiply matrix $\boldsymbol{M}$ with $\boldsymbol{z}$, $\boldsymbol{z}$ no longer points in the same direction. Therefore, the direction of $\boldsymbol{z}$ is transformed by $\boldsymbol{M}$. Would it not be nice in an abstract sense to be able to multiply some matrices multiple times and they will not change the sign of the vectors? If you multiply positive numbers to other positive numbers, it does not change its sign. So let's apply the same logic to Linear Algebra using the eigenvalues.

Example. Say, we have a matrix $\boldsymbol{A}$ and eigenvalues λ. Now, each eigenvector of $\boldsymbol{A}$ can be represented by vector $\boldsymbol{x}$:

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x} \rightarrow \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}^\top \lambda \boldsymbol{x}.$$

Now using norms instead of dot products,

$$\boldsymbol{x}^\top \lambda \boldsymbol{x} = ||\boldsymbol{x}||_2 ||\lambda \boldsymbol{x}||_2 \cos\theta = \lambda ||\boldsymbol{x}||_2^2 \cos(0) = \lambda ||\boldsymbol{x}||_2^2.$$

And,

$$\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = ||\boldsymbol{x}||_2 ||\boldsymbol{A}\boldsymbol{x}||_2 \cos\theta$$

Plugging this all back in, we have:

$$\Rightarrow ||\boldsymbol{x}||_2 ||\boldsymbol{A}\boldsymbol{x}||_2 \cos\theta = \lambda ||\boldsymbol{x}||_2^2$$

$$\Rightarrow \lambda = (||\boldsymbol{x}||_2^2)^{-1}||\boldsymbol{x}||_2||\boldsymbol{A}\boldsymbol{x}||_2 \cos\theta.$$

On an intuitive note, if $\boldsymbol{A}$ is a positive definite matrix, the new direction will always point in the "same general" direction (here "same general" means $\theta < \frac{\pi}{2}$). $\boldsymbol{A}$ is "semi" definite, if $\theta \leq \frac{\pi}{2}$. However, it will reverse the original direction if $\theta \geq \frac{\pi}{2}$.

Now how does this relate to eigenvalues? Well, if we assume $||\boldsymbol{x}||_2||\boldsymbol{A}\boldsymbol{x}||_2 \cos\theta > 0$ and $||\boldsymbol{x}||_2^2 > 0$, then eigenvalues ($\lambda$) must be greater than 0! So, a positive definite matrix must have positive eigenvalues.

If $\boldsymbol{A} = \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix}$, then our eigenvalues ($\lambda$) = 5,1. So, matrix $\boldsymbol{A}$ is a positive definite matrix.

If $\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}$, then since our eigenvalues (λ) = -1,-2. So, therefore, matrix $\boldsymbol{A}$ is a negative definite matrix.

Note. Positive definite matrices are extremely useful tools. For example, if the Hessian of a function is a positive definite matrix then the function is convex. Otherwise if the Hessian is a negative definite matrix, then the function is non-convex.

□

Exercise 31. What is Singular Value Decomposition? Why do we use it? Why not just use ED?

Proof. The **singular value decomposition** (SVD) provides another way to factorize a matrix, into singular vectors and singular values. Like eigendecompostion (ED), the SVD allows one to discover some of the same kind of information. However, the SVD is generally more applicable. Every real matrix has a singular value decomposition, but the same is not true of the eigenvalue decomposition. For example, if a matrix is not square, the eigendecomposition is not defined, and we must use a singular value decomposition instead. □

Exercise 32. Given a matrix $\boldsymbol{A}$, how will you calculate its Singular Value Decomposition?

Proof. Suppose that $\boldsymbol{A}$ is an $m \times n$ matrix. Then $\boldsymbol{U}$ is defined to be an $m \times m$ matrix, $\boldsymbol{D}$ to be an $m \times n$ matrix, and $\boldsymbol{V}$ to be an $n \times n$ matrix, then the Singular Value Decomposition (SVD):

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top \tag{4}$$

Note, matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are both defined to be orthogonal matrices and matrix $\boldsymbol{D}$ is defined to be a diagonal matrix (note it does not have to be a square).

Example. Say matrix $\boldsymbol{A} = \begin{bmatrix} 2 & 2 \\ 1 & 1 \end{bmatrix}$, so $\boldsymbol{A}^\top = \begin{bmatrix} 2 & 1 \\ 2 & 1 \end{bmatrix}$.

$$\boldsymbol{A}^\top\boldsymbol{A} = \begin{bmatrix} 2 & 1 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} 2 & 2 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 5 & 5 \\ 5 & 5 \end{bmatrix}, \ \boldsymbol{A}\boldsymbol{A}^\top = \begin{bmatrix} 2 & 2 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 2 & 1 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 8 & 4 \\ 4 & 2 \end{bmatrix}.$$

So,

$$\det(\boldsymbol{A}^\top\boldsymbol{A} - \boldsymbol{I}_2\lambda) = \begin{vmatrix} 5-\lambda & 5 \\ 5 & 5-\lambda \end{vmatrix} = 0 \rightarrow (5-\lambda)^2 - 25 = 0 \rightarrow \lambda^2 - 10\lambda = 0$$
$$\rightarrow \lambda = 0, 10$$

$$\det(\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{I}_2\lambda) = \begin{vmatrix} 8-\lambda & 4 \\ 4 & 2-\lambda \end{vmatrix} = 0 \to (8-\lambda)(2-\lambda) - 16 = 0 \to \lambda^2 - 10\lambda = 0$$

$$\to \lambda = 0, 10$$

These are our eigenvalues. Let's find their corresponding eigenvectors.

$\lambda = 10$: $\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{I}_2(10)$

$$\to \left[\begin{array}{cc|c} 8-10 & 4 & 0 \\ 4 & 2-10 & 0 \end{array}\right] = \left[\begin{array}{cc|c} -2 & 4 & 0 \\ 4 & -8 & 0 \end{array}\right]$$

Row Operation: $2R_1 + R_2 \to R_2$

$$\to \left[\begin{array}{cc|c} -2 & 4 & 0 \\ 0 & 0 & 0 \end{array}\right]$$

Row Operation: $-\frac{1}{2}R_1 \to R_1$

$$\to \left[\begin{array}{cc|c} 1 & -2 & 0 \\ 0 & 0 & 0 \end{array}\right] \Rightarrow \begin{bmatrix} 1 & -2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \Rightarrow x = 2y.$$

So, $\boldsymbol{v}_1 = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2y \\ y \end{bmatrix} = y \begin{bmatrix} 2 \\ 1 \end{bmatrix} = k \begin{bmatrix} 2 \\ 1 \end{bmatrix}$. To normalize vector, set $k = \frac{1}{\sqrt{5}}$.

$\lambda = 0$: $\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{I}_2(0)$

$$\to \left[\begin{array}{cc|c} 8-0 & 4 & 0 \\ 4 & 2-0 & 0 \end{array}\right] = \left[\begin{array}{cc|c} 8 & 4 & 0 \\ 4 & 2 & 0 \end{array}\right]$$

Row Operation: $\frac{1}{2}R_1 - R_2 \to R_2$

$$\to \left[\begin{array}{cc|c} 8 & 4 & 0 \\ 0 & 0 & 0 \end{array}\right]$$

Row Operation: $\frac{1}{4}R_1 \to R_1$

$$\to \left[\begin{array}{cc|c} 2 & 1 & 0 \\ 0 & 0 & 0 \end{array}\right] \Rightarrow \begin{bmatrix} 2 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \Rightarrow x = -\frac{y}{2}.$$

So, $\boldsymbol{v}_2 = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -\frac{y}{2} \\ y \end{bmatrix} = y \begin{bmatrix} -\frac{1}{2} \\ 1 \end{bmatrix} = k \begin{bmatrix} -\frac{1}{2} \\ 1 \end{bmatrix}$. To normalize, set $k = \frac{2}{\sqrt{5}}$.

$\lambda = 10$: $\boldsymbol{A}^\top\boldsymbol{A} - \boldsymbol{I}_2(10)$

$$\to \left[\begin{array}{cc|c} 5-10 & 5 & 0 \\ 5 & 5-10 & 0 \end{array}\right] = \left[\begin{array}{cc|c} -5 & 5 & 0 \\ 5 & -5 & 0 \end{array}\right]$$

Row Operation: $R_1 + R_2 \to R_2$

$$\to \left[\begin{array}{cc|c} -5 & 5 & 0 \\ 0 & 0 & 0 \end{array}\right]$$

Row Operation: $-\frac{1}{5}R_1 \to R_1$

$$\rightarrow \left[\begin{array}{cc|c} 1 & -1 & 0 \\ 0 & 0 & 0 \end{array}\right] \Rightarrow \begin{bmatrix} 1 & -1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \Rightarrow x = y.$$

So, $\boldsymbol{v}_1 = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} y \\ y \end{bmatrix} = y \begin{bmatrix} 1 \\ 1 \end{bmatrix} = k \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. To normalize, set $k = \frac{1}{\sqrt{2}}$.

$\lambda = 0$: $\boldsymbol{A}^\top \boldsymbol{A} - \boldsymbol{I}_2(0)$

$$\rightarrow \left[\begin{array}{cc|c} 5-0 & 5 & 0 \\ 5 & 5-0 & 0 \end{array}\right] = \left[\begin{array}{cc|c} 5 & 5 & 0 \\ 5 & 5 & 0 \end{array}\right]$$

Row Operation: $R_1 - R_2 \rightarrow R_2$

$$\rightarrow \left[\begin{array}{cc|c} 5 & 5 & 0 \\ 0 & 0 & 0 \end{array}\right]$$

Row Operation: $\frac{1}{5} R_1 \rightarrow R_1$

$$\rightarrow \left[\begin{array}{cc|c} 1 & 1 & 0 \\ 0 & 0 & 0 \end{array}\right] \Rightarrow \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \Rightarrow x = -y.$$

So, $\boldsymbol{v}_2 = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -y \\ y \end{bmatrix} = y \begin{bmatrix} -1 \\ 1 \end{bmatrix} = k \begin{bmatrix} -1 \\ 1 \end{bmatrix}$. To normalize, set $k = \frac{1}{\sqrt{2}}$.

Now, since $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \rightarrow \boldsymbol{A}^\top = (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top)^\top = \boldsymbol{V}\boldsymbol{\Sigma}^\top\boldsymbol{U}^\top$

$$\Rightarrow \boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Sigma}^\top\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top = \boldsymbol{V}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}\boldsymbol{V}^\top = \boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^\top.$$

$$\Rightarrow \boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{\Sigma}^\top\boldsymbol{U}^\top = \boldsymbol{U}\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}\boldsymbol{U}^\top = \boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^\top.$$

Now, since we calculated 2 pairs of eigenvalues for $\boldsymbol{A}\boldsymbol{A}^\top$ and $\boldsymbol{A}^\top\boldsymbol{A}$, we know which respective matrices we were building. $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top$ or $\boldsymbol{\Sigma}^\top\boldsymbol{\Sigma}$ is just the diagonal matrix of eigenvalues for $\boldsymbol{A}^\top\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^\top$

$$\Rightarrow \begin{bmatrix} 10 & 0 \\ 0 & 0 \end{bmatrix} \Rightarrow \boldsymbol{\Sigma} = \begin{bmatrix} \sqrt{10} & 0 \\ 0 & \sqrt{0} \end{bmatrix}$$

.

Therefore, $\boldsymbol{U} = \frac{1}{\sqrt{5}} \begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix}$ and $\boldsymbol{V} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \rightarrow \boldsymbol{V}^\top = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$.

$$\Rightarrow \frac{1}{\sqrt{10}} \begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} \sqrt{10} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} = \boldsymbol{A}$$

Note. Now, what does this show intuitively? More importantly, what does this mean? Think of the $\boldsymbol{A}$ as a type of linear transformation. Therefore, SVD breaks down the transformation into three simple steps:

1) The initial rotation (or reflection) $\boldsymbol{V}^\top$.

2) The $\boldsymbol{\Sigma}$ matrix has to do with rescaling vertically or horizontally (really in any dimension depending on the matrix).

3) The final rotation (or reflection) $\boldsymbol{U}$.

By looking at the SVD, we can tell which matrix has a significant impact on transforming the matrix and how much as well.

If we apply that intuition to this problem, we explain break up the transformation:
$\boldsymbol{V}^\top = \frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ -1 & 1\end{bmatrix} = \frac{1}{\sqrt{2}}\Big(\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} + \begin{bmatrix}0 & 1\\ -1 & 0\end{bmatrix}\Big) = \frac{1}{\sqrt{2}}(\boldsymbol{I}_2 + \boldsymbol{R}(\theta = 270°))$.

So $\boldsymbol{V}^\top$ boils down to summing the initial vector with the rotation of the initial vector by 270°.

$\boldsymbol{\Sigma} = \begin{bmatrix}\sqrt{10} & 0\\ 0 & 0\end{bmatrix} \to \Sigma$ basically scales the x coordinate by $\sqrt{10}$.

$\boldsymbol{U} = \frac{1}{\sqrt{5}}\begin{bmatrix}2 & 1\\ -1 & 2\end{bmatrix} = \frac{1}{\sqrt{5}}\Big(\begin{bmatrix}2 & 0\\ 0 & 2\end{bmatrix} + \begin{bmatrix}0 & -1\\ 1 & 0\end{bmatrix}\Big) = \frac{1}{\sqrt{5}}(2\boldsymbol{I}_2 + \boldsymbol{R}(\theta = 90°))$.

$\boldsymbol{U}$ boils down to summing two times the initial vector with the rotation of the initial vector by 90°.

□

Exercise 33. What are singular values, left singulars, and right singulars?

Proof. The elements along the diagonal of $\boldsymbol{D}$ are known as the **singular values** of the matrix $\boldsymbol{A}$. The columns of $\boldsymbol{U}$ are known as the **left-singular vectors**. The columns of $\boldsymbol{V}$ are known as as the **right-singular vectors**. □

Exercise 34. What is the connection of Singular Value Decomposition of $\boldsymbol{A}$ with functions of $\boldsymbol{A}$?

Proof. We can actually interpret the singular value decomposition of A in terms of the eigendecomposition of functions of $\boldsymbol{A}$. The left-singular vectors of $\boldsymbol{A}$ are the eigenvectors of $\boldsymbol{A}\boldsymbol{A}^\top$. The right-singular vectors of $\boldsymbol{A}$ are the eigenvectors of $\boldsymbol{A}^\top\boldsymbol{A}$. The non-zero singular values of $\boldsymbol{A}$ are the square roots of the eigenvalues of $\boldsymbol{A}^\top\boldsymbol{A}$. The same is true for $\boldsymbol{A}\boldsymbol{A}^\top$. □

Exercise 35. Why are singular values always non-negative?

Proof. Let's assume matrix $\boldsymbol{A}$ has real entries—otherwise consider $\boldsymbol{A}^H\boldsymbol{A}$—then $\boldsymbol{A}^\top\boldsymbol{A}$ is positive semidefinite because it is an inner product. One of the properties of an inner product is that for any vector $\boldsymbol{u}$, $\langle \boldsymbol{u}, \boldsymbol{u}\rangle \geq 0$ or $\langle \boldsymbol{u}, \boldsymbol{u}\rangle = 0$ if and only if $\boldsymbol{u} = \boldsymbol{0}$. Now, since $\boldsymbol{A}^\top\boldsymbol{A}$ is matrix (and we want a vector), if $\boldsymbol{v}$ is a nonzero vector in $\mathbb{R}^n$, then we have

$$\langle \boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{v}, \boldsymbol{v}\rangle := (\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{v})^\top\boldsymbol{v} = \boldsymbol{v}^\top\boldsymbol{A}^\top\boldsymbol{A}\boldsymbol{v}$$

$$= \langle \boldsymbol{A}\boldsymbol{v}, \boldsymbol{A}\boldsymbol{v}\rangle \geq 0, \forall \boldsymbol{v}.$$

Therefore, the eigenvalues of $\boldsymbol{A}^\top\boldsymbol{A}$ are non-negative, and the singular values of $\boldsymbol{A}$ are the square roots of the eigenvalues. □

Exercise 36. What is the Moore Penrose pseudo inverse and how to calculate it?

Proof. The Moore Penrose pseudo inverse definition is:

$$\boldsymbol{A}^+ = \lim_{\alpha\to}(\boldsymbol{A}^\top\boldsymbol{A} + \alpha\boldsymbol{I})^{-1}\boldsymbol{A}^\top. \tag{5}$$

Practical algorithms for computing the pseudo inverse are not based on this definition, but rather the formula:

$$\boldsymbol{A}^{+} = \boldsymbol{V}\boldsymbol{D}^{+}\boldsymbol{U}^{\top} \tag{6}$$

where $\boldsymbol{U}$, $\boldsymbol{D}$, and $\boldsymbol{V}$ are the singular value decomposition of $\boldsymbol{A}$. The pseudo inverse $\boldsymbol{D}^{+}$ of a diagonal matrix $\boldsymbol{D}$ is obtained by taking the reciprocal of its non-zero elements then taking the transpose of the resulting matrix.

Example. Say $\boldsymbol{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \sqrt{3} & 0 \end{bmatrix}$, so $\rightarrow \boldsymbol{A}^{\top} = \begin{bmatrix} 1 & 0 & \sqrt{3} \\ 0 & 1 & 0 \end{bmatrix}$.

$$\boldsymbol{A}^{\top}\boldsymbol{A} = \begin{bmatrix} 1 & 0 & \sqrt{3} \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \sqrt{3} & 0 \end{bmatrix} = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}, \; \boldsymbol{A}\boldsymbol{A}^{\top} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \sqrt{3} & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & \sqrt{3} \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & \sqrt{3} \\ 0 & 1 & 0 \\ \sqrt{3} & 0 & 3 \end{bmatrix}.$$

So,

$$\det(\boldsymbol{A}^{\top}\boldsymbol{A} - \boldsymbol{I}_2\lambda) = \begin{vmatrix} 4-\lambda & 0 \\ 0 & 1-\lambda \end{vmatrix} = 0 \rightarrow (4-\lambda)(1-\lambda) - 0 = 0 \rightarrow \lambda = 1, 4$$

$$\det(\boldsymbol{A}\boldsymbol{A}^{\top} - \boldsymbol{I}_2\lambda) = \begin{vmatrix} 1-\lambda & 0 & \sqrt{3} \\ 0 & 1-\lambda & 0 \\ \sqrt{3} & 0 & 3-\lambda \end{vmatrix} = 0 \rightarrow (1-\lambda)[(1-\lambda)(3-\lambda)-0]-0+\sqrt{3}[0-\sqrt{3}(1-\lambda)] = 0$$

$$(1-\lambda)[(1-\lambda)(3-\lambda)] - 3(1-\lambda) = 0$$

$$(1-\lambda)[(1-\lambda)(3-\lambda) - 3] = 0$$

$$(1-\lambda)[3-\lambda-3\lambda+\lambda^2-3] = 0$$

$$(1-\lambda)[-4\lambda+\lambda^2] = 0$$

$$(1-\lambda)(-\lambda)(4\lambda) = 0$$

$$\rightarrow \lambda = 0, 1, 4$$

These are our eigenvalues. Let's find their corresponding eigenvectors.
$\lambda = 4$: $\boldsymbol{A}\boldsymbol{A}^{\top} - \boldsymbol{I}_3(4)$

$$\rightarrow \left[\begin{array}{ccc|c} (1-4) & 0 & \sqrt{3} & 0 \\ 0 & (1-4) & 0 & 0 \\ \sqrt{3} & 0 & (3-4) & 0 \end{array}\right] = \left[\begin{array}{ccc|c} -3 & 0 & \sqrt{3} & 0 \\ 0 & -3 & 0 & 0 \\ \sqrt{3} & 0 & -1 & 0 \end{array}\right]$$

Row Operation: $-\sqrt{3}R_3 - R_1 \rightarrow R_1, -\frac{1}{3}R_2 \rightarrow R_2$

$$\Rightarrow \left[\begin{array}{ccc|c} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \sqrt{3} & 0 & -1 & 0 \end{array}\right]$$

Row Operation: $R_3 \leftrightarrow R_1$

$$\Rightarrow \left[\begin{array}{ccc|c} \sqrt{3} & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] \Rightarrow \begin{bmatrix} \sqrt{3} & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = 0$$

$$\Rightarrow x\sqrt{3} - z = 0 \Rightarrow z = x\sqrt{3}$$

$$\Rightarrow y = 0.$$

Thus, $\boldsymbol{v}_1 = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x \\ 0 \\ x\sqrt{3} \end{bmatrix} = x \begin{bmatrix} 1 \\ 0 \\ \sqrt{3} \end{bmatrix} = k \begin{bmatrix} 1 \\ 0 \\ \sqrt{3} \end{bmatrix}$. To normalize, let $k = \frac{1}{2}$.

$\lambda = 1$: $\boldsymbol{A}\boldsymbol{A}^\top - \boldsymbol{I}_3(1)$

$$\rightarrow \left[\begin{array}{ccc|c} (1-1) & 0 & \sqrt{3} & 0 \\ 0 & (1-1) & 0 & 0 \\ \sqrt{3} & 0 & (3-1) & 0 \end{array}\right] = \left[\begin{array}{ccc|c} 0 & 0 & \sqrt{3} & 0 \\ 0 & 0 & 0 & 0 \\ \sqrt{3} & 0 & 2 & 0 \end{array}\right]$$

Row Operation: $\frac{1}{\sqrt{3}} R_1 \rightarrow R_1$

$$\Rightarrow \left[\begin{array}{ccc|c} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ \sqrt{3} & 0 & 2 & 0 \end{array}\right]$$

Row Operation: $R_3 \leftrightarrow R_1$

$$\Rightarrow \left[\begin{array}{ccc|c} \sqrt{3} & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{array}\right]$$

Row Operation: $2R_3 - R_1$

$$\Rightarrow \left[\begin{array}{ccc|c} \sqrt{3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{array}\right] \Rightarrow \begin{bmatrix} \sqrt{3} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = 0$$

$$\Rightarrow x\sqrt{3} = 0 \Rightarrow x = 0$$

$$\Rightarrow z = 0.$$

Thus, $\boldsymbol{v}_2 = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ y \\ 0 \end{bmatrix} = y \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = k \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$.

$\lambda = 0$: $\boldsymbol{A}\boldsymbol{A}^\top$

$$\rightarrow \left[\begin{array}{ccc|c} (1-0) & 0 & \sqrt{3} & 0 \\ 0 & (1-0) & 0 & 0 \\ \sqrt{3} & 0 & (3-0) & 0 \end{array}\right] = \left[\begin{array}{ccc|c} 1 & 0 & \sqrt{3} & 0 \\ 0 & 1 & 0 & 0 \\ \sqrt{3} & 0 & 3 & 0 \end{array}\right]$$

Row Operation: $\sqrt{3}R_1 - R_3 \to R_3$

$$\Rightarrow \left[\begin{array}{ccc|c} 1 & 0 & \sqrt{3} & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] \Rightarrow \begin{bmatrix} 1 & 0 & \sqrt{3} \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = 0$$

$$\Rightarrow x + z\sqrt{3} = 0 \Rightarrow x = -z\sqrt{3}$$

$$\Rightarrow y = 0.$$

Thus, $\boldsymbol{v}_3 = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} -z\sqrt{3} \\ 0 \\ z \end{bmatrix} = z \begin{bmatrix} -\sqrt{3} \\ 0 \\ 1 \end{bmatrix} = k \begin{bmatrix} -\sqrt{3} \\ 0 \\ 1 \end{bmatrix}$. To normalize, $k = \frac{1}{2}$.

$\lambda = 4$: $\boldsymbol{A}^\top \boldsymbol{A} - \boldsymbol{I}_2(4)$

$$\to \left[\begin{array}{cc|c} (4-4) & 0 & 0 \\ 0 & (1-4) & 0 \end{array}\right] = \left[\begin{array}{cc|c} 0 & 0 & 0 \\ 0 & -3 & 0 \end{array}\right] \Rightarrow \begin{bmatrix} 0 & 0 \\ 0 & -3 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0 \to y = 0.$$

Thus, $\boldsymbol{v}_1 = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x \\ 0 \end{bmatrix} = x \begin{bmatrix} 1 \\ 0 \end{bmatrix} = k \begin{bmatrix} 1 \\ 0 \end{bmatrix}$.

$\lambda = 1$: $\boldsymbol{A}^\top \boldsymbol{A} - \boldsymbol{I}_2(1)$

$$\to \left[\begin{array}{cc|c} (4-1) & 0 & 0 \\ 0 & (1-1) & 0 \end{array}\right] = \left[\begin{array}{cc|c} 3 & 0 & 0 \\ 0 & 0 & 0 \end{array}\right] \Rightarrow \begin{bmatrix} 3 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0 \to x = 0.$$

Thus, $\boldsymbol{v}_2 = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} = y \begin{bmatrix} 0 \\ 1 \end{bmatrix} = k \begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

Now, we calculated 2 eigenvalues for $\boldsymbol{A}^\top \boldsymbol{A}$ and 3 eigenvalues for $\boldsymbol{A}\boldsymbol{A}^\top$. However, our $\boldsymbol{\Sigma}$ is not a square matrix, so it cannot be symmetric. Note, that since this is pseudoinverse problem, we will now denote $\boldsymbol{D}$ to be $\boldsymbol{\Sigma}$. Because we have an extra eigenvalue (0), we will add to it to the subsequent row.

Therefore our $\boldsymbol{D}$ matrix is: $\begin{bmatrix} \sqrt{4} & 0 \\ 0 & \sqrt{1} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \to \boldsymbol{D}^+ = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{1} \\ 0 & 0 \end{bmatrix}^\top = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$.

Therefore, $\boldsymbol{U} = \begin{bmatrix} \frac{1}{2} & 0 & -\frac{\sqrt{3}}{2} \\ 0 & 1 & 0 \\ \frac{\sqrt{3}}{2} & 0 & \frac{1}{2} \end{bmatrix} \to \boldsymbol{U}^\top = \begin{bmatrix} \frac{1}{2} & 0 & \frac{\sqrt{3}}{2} \\ 0 & 1 & 0 \\ -\frac{\sqrt{3}}{2} & 0 & \frac{1}{2} \end{bmatrix}$ and $\boldsymbol{V} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \boldsymbol{V}^\top$.

Thus, $\boldsymbol{A}^\top = \boldsymbol{V}\boldsymbol{D}^+\boldsymbol{U}^\top$

$$\Rightarrow \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 & \frac{\sqrt{3}}{2} \\ 0 & 1 & 0 \\ -\frac{\sqrt{3}}{2} & 0 & \frac{1}{2} \end{bmatrix} = \frac{1}{4} \begin{bmatrix} 1 & 0 & \sqrt{3} \\ 0 & 4 & 0 \end{bmatrix}.$$

□

Exercise 37. If we do Moore Penrose pseudo inverse on $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, what solution is provided is $\boldsymbol{A}$ is fat? Moreover, what solution is provided if $\boldsymbol{A}$ is tall?

Proof. When $\boldsymbol{A}$ has more columns than rows (fat matrix), then solving a linear equation using the pseudoinverse provides one of the many possible solutions. Specifically, it provides the solution $\boldsymbol{x} = \boldsymbol{A}^+\boldsymbol{b}$ with minimal Euclidean norm $||\boldsymbol{x}||_2$ among all possible solutions.

When $\boldsymbol{A}$ has more rows than columns (tall matrix), it is possible for there to be no solution. In this case, using the pseudoinverse gives us the $\boldsymbol{x}$ for which $\boldsymbol{A}\boldsymbol{x}$ is as close as possible to $\boldsymbol{b}$ in terms of Euclidean norm $||\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}||_2$. □

Exercise 38. Which matrices can be decomposed by ED?

Proof. Only square matrices have eigenvalue decomposition. □

Exercise 39. Which matrices can be decomposed by SVD?

Proof. Every real matrix has a singular value decomposition square. However, matrix $\boldsymbol{V}$ can either be conjugate transpose or normal transpose depending on whether matrix $\boldsymbol{A}$ is complex or real. □

Exercise 40. What is the trace of a matrix?

$$\mathrm{Tr}(\boldsymbol{A}) = \sum_i \boldsymbol{A}_{i,i} \tag{7}$$

Exercise 41. How to write Frobenius norm of a matrix $\boldsymbol{A}$ in terms of trace?

$$||A||_F = \sqrt{\mathrm{Tr}(\boldsymbol{A}\boldsymbol{A}^\top)} \tag{8}$$

Exercise 42. Why is trace of a multiplication of matrices invariant to cyclic permutations?

Proof. Because the trace of a matrix is only concerned with the diagonal entries, regardless of the cyclic permutations (even if the two matrices are different in shape), the sum of the diagonals will stay the same. □

Exercise 43. What is the trace of a scalar?

$$a = \mathrm{Tr}(a) \tag{9}$$

2 Extra Stuff

Exercise 44. What is a Hermitian Matrix?

Proof. A **Hermitian matrix** is a complex square matrix that is equal to its own conjugate transpose—more specifically the element of the i-th row and the j-th column is equal to the complex conjugate of the element in the j-th row and the i-column.

$$\boldsymbol{A} = \boldsymbol{A}^H = \overline{\boldsymbol{A}^\top}, \tag{10}$$

where $\boldsymbol{A}^H$ denotes the **conjugate transpose**, which is done by first taking the transpose of $\boldsymbol{A}$ and then taking the complex conjugate of each entry in $\boldsymbol{A}^\top$ (denoted as $\overline{\boldsymbol{A}^\top}$).

Example. Say $\boldsymbol{A} = \begin{bmatrix} 2 & 2+i & 4 \\ 2-i & 3 & i \\ 4 & -i & 1 \end{bmatrix} \to \boldsymbol{A}^\top = \begin{bmatrix} 2 & 2-i & 4 \\ 2+i & 3 & -i \\ 4 & i & 1 \end{bmatrix} \to \overline{\boldsymbol{A}^\top} = \begin{bmatrix} 2 & 2+i & 4 \\ 2-i & 3 & i \\ 4 & -i & 1 \end{bmatrix}$.

$$\overline{\boldsymbol{A}^\top} = \boldsymbol{A}^H = \boldsymbol{A}.$$

Note. Hermitian matrices can be thought of as a complex extension of real symmetric matrices because they always have real eigenvalues. Sometimes in SVD, $\boldsymbol{U}, \boldsymbol{V}$ matrices can be complex— so $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^H \to \boldsymbol{A}^H = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^H$.

More properties of Hermitian matrices can be found here. □

Exercise 45. What is an outer product?

Proof. The outer product $\boldsymbol{u} \otimes \boldsymbol{v}$ is equivalent to a matrix multiplication $\boldsymbol{u}$, provided that u is represent $\boldsymbol{u}$ is represented as a $m \times 1$ column vector and $\boldsymbol{v}$ as a $n \times 1$ column vector (which makes $\boldsymbol{v}^\top$ a row vector). For instance, if $m = 4$ and $n = 3$, then

$$\boldsymbol{u} \otimes \boldsymbol{v} = \boldsymbol{u}\boldsymbol{v}^\top = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix} = \begin{bmatrix} u_1v_1 & u_1v_2 & u_1v_3 \\ u_2v_1 & u_2v_2 & u_2v_3 \\ u_3v_1 & u_3v_2 & u_3v_3 \\ u_4v_1 & u_4v_2 & u_4v_3 \end{bmatrix}. \tag{11}$$

Note. This returns a matrix, not a scalar.

□

Exercise 46. What is the Rotation Matrix?

Proof.

$$\boldsymbol{R}(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{12}$$

□

Exercise 47. What is a projection? What is a projection matrix? Why do we care?

Proof. If we have a vector $\boldsymbol{b}$ and a line determined by a vector $\boldsymbol{a}$, how do we find the point on the line that is closest to $\boldsymbol{b}$?

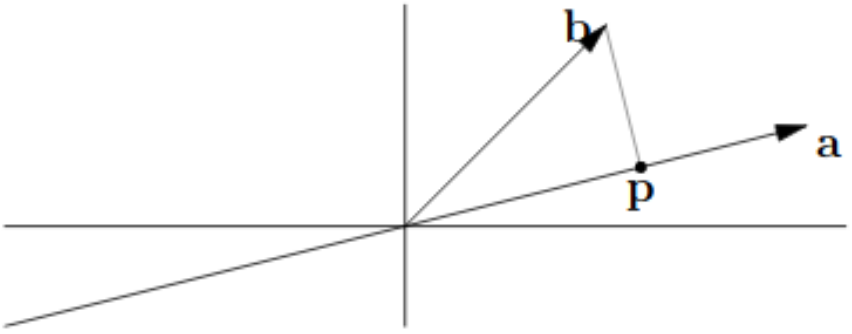

Figure 1: The point closest to **b** on the line determined by **a**.

As we see from Figure 1, the closest point $\boldsymbol{p}$ is at the intersection formed by a line through $\boldsymbol{b}$ that is orthogonal to $\boldsymbol{a}$. If we think of $\boldsymbol{p}$ as an approximation of $\boldsymbol{b}$, then the length of $\boldsymbol{e} = \boldsymbol{b} - \boldsymbol{p}$ is the error in that approximation. We could try to find $\boldsymbol{p}$ using trigonometry or calculus, but it's easier to use linear algebra. Since $\boldsymbol{p}$ lies on the line through $\boldsymbol{a}$, we know $\boldsymbol{p} = \boldsymbol{a}x$ for some number x. We also know that $\boldsymbol{a}$ is perpendicular to $\boldsymbol{e} = \boldsymbol{b} - \boldsymbol{a}x$, so

$$\boldsymbol{a} \cdot \boldsymbol{e} = 0 \rightarrow \boldsymbol{a}^\top(\boldsymbol{b} - \boldsymbol{a}x) = 0$$
$$\boldsymbol{a}^\top \boldsymbol{b} - \boldsymbol{a}^\top \boldsymbol{a}x = 0$$
$$x = \frac{\boldsymbol{a}^\top \boldsymbol{b}}{\boldsymbol{a}^\top \boldsymbol{a}} = \frac{\langle \boldsymbol{a}, \boldsymbol{b} \rangle}{||\boldsymbol{a}||^2},$$

and $\boldsymbol{p} = \boldsymbol{a}x = \boldsymbol{a}\frac{\boldsymbol{a}^\top \boldsymbol{b}}{\boldsymbol{a}^\top \boldsymbol{a}}$. Note, doubling $\boldsymbol{b}$ doubles $\boldsymbol{p}$. Doubling $\boldsymbol{a}$ does not affect $\boldsymbol{p}$. $\boldsymbol{p}$ is called the **projection**.

We'd like to write this projection in terms of a **projection matrix** $P : \boldsymbol{p} = P\boldsymbol{b}$

$$\boldsymbol{a}\frac{\boldsymbol{a}^\top \boldsymbol{b}}{\boldsymbol{a}^\top \boldsymbol{a}} = P\boldsymbol{b} \rightarrow \left[\frac{\boldsymbol{a}\boldsymbol{a}^\top}{||\boldsymbol{a}||^2}\right]\boldsymbol{b} \rightarrow P = \frac{\boldsymbol{a}\boldsymbol{a}^\top}{||\boldsymbol{a}||^2}.$$

Note that $\boldsymbol{a}\boldsymbol{a}^\top$ is a three by three matrix, not a number; matrix multiplication is not commutative. The column space of P is spanned by $\boldsymbol{a}$ because for any $\boldsymbol{b}$, $P\boldsymbol{b}$ lies on the line determined by $\boldsymbol{a}$. The rank of P is 1. P is symmetric. $P^2\boldsymbol{b} = P\boldsymbol{b}$ because the projection of a vector already on the line through $\boldsymbol{a}$ is just that vector. In general, projection matrices have the properties:

$P^\top = P$ and $P^2 = P$

As we know, the equation $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ may have no solution. The vector $\boldsymbol{A}\boldsymbol{x}$ is always in the column space of $\boldsymbol{A}$, and $\boldsymbol{b}$ is unlikely to be in the column space. So, we project $\boldsymbol{b}$ onto a vector $\boldsymbol{p}$ in the column space of $\boldsymbol{A}$ and solve $\boldsymbol{A}\hat{\boldsymbol{x}} = \boldsymbol{p}$.

More info can be found here. □