Applied Linear Algebra in Data Analysis: Course Notes

Sivakumar Balasubramanian
CMC Vellore

Update on September 23, 2024

Contents

CONTENTS

List of Figures

Part I

Linear Algebra

Chapter 1

Vectors

1.1 n-Vectors

A collection of an ordered list of n numbers is called an n-vector. We will use bold lower case alphabets to represent such vectors, and we will represent these as a column of numbers, which is referred to as a *column vector*. We will look at *row vectors* at a later stage. Consider the following example:

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

The elements of the n-vector $x_1, x_2, \ldots, x_n$ are called the *components* of the vector $\mathbf{x}$; x_i is the i^{th} component of the vector $\mathbf{x}$. If these components are all real numbers, the set of all such n-vectors is the set $\mathbb{R}^n$.

Where do we come across such n-vectors? In many places, such as in physics, engineering, economics, medicine, etc. Any application where we deal with multiple pieces of information that can be represented as a list of numbers can be represented as an n-vector. When we deal with systems with multiple inputs, multiple output, or multiple states, we can represent these as n-vectors. We talk about the state of a system in a later chapter.

1.1.1 Some common n-vectors

We will often come across some special n-vectors in this document course and in many applications. We will define some of these vectors here.

- **Zero vector:** The n-vector whose components are all zeros is called the *zero vector*. $\mathbf{0} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$
- **One vector:** The n-vector whose components are all ones is called the *one vector*. $\mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$
- **Unit vectors:** The n-vectors whose components are all zeros except for one component which is 1. These are called the *standard basis vectors* and are denoted by $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$. The n-vector $\mathbf{e}_i$ has all components as zeros except for the i^{th} component which is 1. For example, the unit vectors in $\mathbb{R}^2$ are:

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

We now look at some examples of n-vectors that we come aross in applications.

Example 1.1. Basic clinical information during a hospital visit. When a patient visits a hostpial, several clinically relevant variables are captured, for instance:

Index	Variable	Units
1	Sex	None (0: Male, 1: Female)
2	Age	Years
3	Height	cm
4	Weight	kg
5	Heart rate	count
6	Systolic pressure	mm of Hg
7	Diastolic pressure	mm of Hg
8	Temperature	celcius

The following are some examples of n-vectors generated from three different patients visiting the hospital.

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 67 \\ 152 \\ 56 \\ 132 \\ 102 \\ 37.1 \end{bmatrix} \quad \mathbf{x}_2 = \begin{bmatrix} 0 \\ 36 \\ 172 \\ 97 \\ 156 \\ 97 \\ 36.5 \end{bmatrix} \quad \mathbf{x}_3 = \begin{bmatrix} 1 \\ 22 \\ 162 \\ 56 \\ 121 \\ 78 \\ 38.2 \end{bmatrix}$$

Example 1.2. Time series data. We often collect data over time, often at regular intervals. For example, consider the example of an attending nurse taking the temperature of a patient admitted to the hospital for an infectious disease. The nurse records the temperature of the patient every hour, without fail for the 48 hours the patient spent in the hosptial. This temperature record will have a total of 49 measurements, which can conveniently think of as a n-vector, in this case a 49-vector. Instead of writing down the entire 49-vector, we depict it as a time series plot.

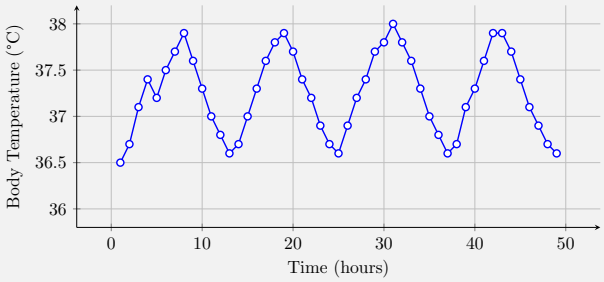


Figure 1.1: Body temperature recorded at multiple time points.

1.2 Visualizing n-vectors

The n-vectors can be visualized as points in n-dimensional space. For example, A 1-vector or just single real number or a *scalar* can be thought of as a point on the real line. The 1-vector $x = 2.45$ is shown in Figure 1.2 is the black point. But we will find it useful to visualize a 1-vector as an arrow starting at the origin and ending at the point on the real line. The arrow is shown in blue in Figure 1.2.

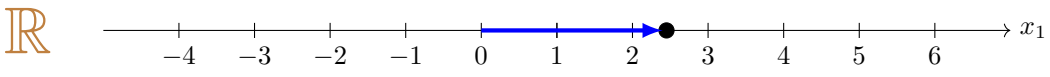
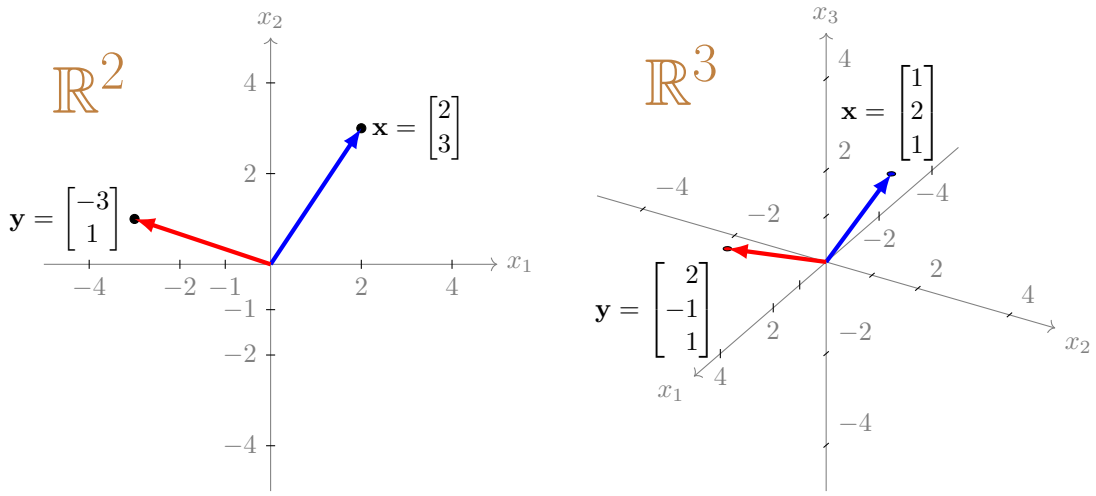


Figure 1.2: The real line $\mathbb{R}$ contains the 1-vectors.



(a) The $\mathbb{R}^2$ plane contains the 2-vectors. **(b)** $\mathbb{R}^3$ contains the 3-vectors.

Figure 1.3: The $\mathbb{R}^2$ and $\mathbb{R}^3$ sets.

The elements of $\mathbb{R}^2$ are points on the plane, and we can visualize them as points in the plane. The 2-vectors $\mathbf{x} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$ and $\mathbf{x} = \begin{bmatrix} -3 \\ 1 \end{bmatrix}$ are shown in Figure 1.3a. A similar visualization is shown for $\mathbb{R}^3$ (Figure 1.3b), and for $\mathbb{R}^4$ and beyond you simply pretend that you can visualize things in your head like your instructor does.

1.3 Operations on n-vectors

There are many operations we can perform on n-vectors, but we will only focus on two operations for this course:

- **Scalar multiplication:** Given a scalar $c \in \mathbb{R}$ and an n-vector $\mathbf{x}$. The scalar multiplication operation produces another n-vector $c\mathbf{x}$ whose components are $cx_1, cx_2, \ldots, cx_n$.

$$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \longrightarrow 2\mathbf{x} = \begin{bmatrix} 2\,(1) \\ 2\,(2) \end{bmatrix} = \begin{bmatrix} 2 \\ 8.2 \end{bmatrix}$$

  The geometric interpretation scalar multiplication is shown in Figure 1.4. Scalar multiplication stretches or shrinks the vector without rotating it. When the scalar is positive the direction of the scaled vector is the same as the original vector, and when the scalar is negative the direction is opposite. When the scalar is zero, the scaled vector is the zero vector $\mathbf{0}$.

- **Vector Addition:** Given two n-vectors $\mathbf{x}$ and $\mathbf{y}$, the vector addition operation, represented by $\mathbf{x} + \mathbf{y}$, producs another n-vector whose components are $x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n$.

$$\mathbf{x} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \longrightarrow \mathbf{x} + \mathbf{y} = \begin{bmatrix} 1+2 \\ 3+1 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$$

  The geometric interpretation the vector addition operation is shown in Figure 1.5. Geometrically, the vector addition operation follows the parallelogram law of addition, where the resulting vector $\mathbf{x}+\mathbf{y}$ is a diagonal of the parallelogram formed by the two vectors $\mathbf{x}$ and $\mathbf{y}$. Another way to think about this, is that you first move along $\mathbf{x}$ to its end point, and starting from there then move along $\mathbf{y}$ to its end point or vice versa.

  You can add more than two vectors to obtain a new vector, like below:

$$\mathbf{w} = \mathbf{x} + \mathbf{y} + \mathbf{z}$$

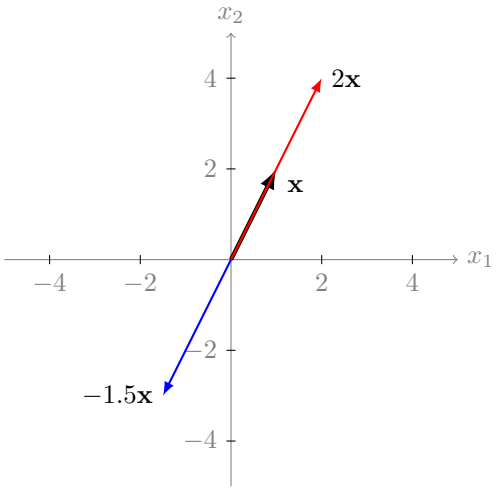


Figure 1.4: Scalar multiplication of a vector.

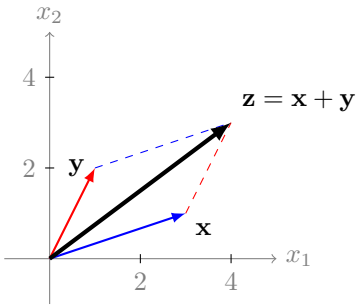


Figure 1.5: Vector addition.

Geometrically, we can first apply the parallelogram law to $\mathbf{x}$ and $\mathbf{y}$, and then apply the parallelogram law to $\mathbf{x} + \mathbf{y}$ and $\mathbf{z}$ to get $\mathbf{w}$.

1.4 Vector spaces

Vector spaces are *sets* with some special properites. More specifically, a vector space is a set V of elements called *vectors* that are closed under two operations called *addition* and *scalar multiplication*. This simply means that if you perfom these operations using elements from the set V, the result is also an element of the set V. A vector space must satisfy the following properties:

- **Closure under addition:** For any two vectors $\mathbf{x}, \mathbf{y} \in V$, the sum $\mathbf{x} + \mathbf{y} \in V$.
- **Closure under scalar multiplication:** For any scalar $c \in \mathbb{R}$ and any vector $\mathbf{x} \in V$, the product $c\mathbf{x} \in V$.
- **Additive identity:** There exists a vector $\mathbf{0} \in V$ such that for any vector $\mathbf{x} \in V$, $\mathbf{x} + \mathbf{0} = \mathbf{x}$.
- **Additive inverse:** For any vector $\mathbf{x} \in V$, there exists a vector $-\mathbf{x} \in V$ such that $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$.
- **Commutativity of addition:** For any two vectors $\mathbf{x}, \mathbf{y} \in V$, $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$.
- **Associativity of addition:** For any three vectors $\mathbf{x}, \mathbf{y}, \mathbf{z} \in V$, $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$.
- **Distributive property:** For any scalar $c \in \mathbb{R}$ and any two vectors $\mathbf{x}, \mathbf{y} \in V$, $c(\mathbf{x} + \mathbf{y}) = c\mathbf{x} + c\mathbf{y}$.
- **Distributive property:** For any two scalars $c, d \in \mathbb{R}$ and any vector $\mathbf{x} \in V$, $(c + d)\mathbf{x} = c\mathbf{x} + d\mathbf{x}$.
- **Associativity of scalar multiplication:** For any two scalars $c, d \in \mathbb{R}$ and any vector $\mathbf{x} \in V$, $(cd)\mathbf{x} = c(d\mathbf{x})$.

- **Multiplicative identity:** For any vector $\mathbf{x} \in V$, $1\mathbf{x} = \mathbf{x}$.

These properties are satisfied by the set $\mathbb{R}^n$ of n-vectors, and hence $\mathbb{R}^n$ is a vector space. Geometrically, the concept of a vector space corresponds to flat spaces that contain the origin. This will become more clear when we talk about subspaces. Notice that definition of the vector space given above does not make any specific mention of n-vectors. The definition is general and can be applied to any set of elements that satisfy the properties listed above. The following are some examples of vector spaces with the addition and scalar multiplication operations defined on them.

Example 1.3. Set of $m \times n$ matrices. The set M of all $m \times n$ matrices of real numbers is a vector space.

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{11} & a_{12} & \cdots & a_{1n} \end{bmatrix}, \ a_{ij} \in \mathbb{R}$$

We define scalar multiplication and addition of matrices as follows:

$$c\mathbf{A} = \begin{bmatrix} ca_{11} & ca_{12} & \cdots & ca_{1n} \\ ca_{11} & ca_{12} & \cdots & ca_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ ca_{11} & ca_{12} & \cdots & ca_{1n} \end{bmatrix}, \ c \in \mathbb{R}$$

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \cdots & a_{1n} + b_{1n} \\ a_{11} + b_{11} & a_{12} + b_{12} & \cdots & a_{1n} + b_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{11} + b_{11} & a_{12} + b_{12} & \cdots & a_{1n} + b_{1n} \end{bmatrix}, \ \mathbf{A}, \mathbf{B} \in M$$

Since each element of $c\mathbf{A}$ and $\mathbf{A} + \mathbf{B}$ is a real number, M is a vector space.

Example 1.4. Set of polynomials of order $\leq n$. Now we look at strange example of a vector space. The set P_n of all polynomials of degree at most n with real coefficients, defined over an interval $[a, b]$.

$$p(x) = \sum_{k=0}^{n-1} a_k x^k, \ x \in [a, b], \ a_k \in \mathbb{R}$$

The set P_n contains all polynomials of the form shown above. We define scalar multiplication and addition of polynomials as follows:

$$cp(x) = c \sum_{k=0}^{n-1} a_k x^k = \sum_{k=0}^{n-1} c a_k x^k, \ p(x) \in P$$

$$p(x) + q(x) = \sum_{k=0}^{n-1} a_k x^k + \sum_{k=0}^{n-1} b_k x^k = \sum_{k=0}^{n-1} (a_k + b_k) x^k, \ p(x), q(x) \in P_n$$

The set P_n is a vector space because the sum and product of any two polynomials from P_n is also a polynomial of degree at most n with real coefficients.

Example 1.5. Set of continuous functions. The set $C[0, 1]$ of all continuous functions $f(x)$ over the time interval $x \in [0, 1]$ is a vector space. We define scalar multiplication and addition of functions as follows:

$$cf(x) = cf(x), \ f(x) \in C(0, 1)$$

$$f(x)+g(x)=f(x)+g(x),\ \ f(x),g(x)\in C(0,1)$$

The set $C(0,1)$ is a vector space because the sum and product of any two continuous functions from $C(0,1)$ is also a continuous function.

1.5 Subspaces – "Little" Vector Spaces

These are little vector spaces in the sense that they are subsets of a larger vector space that are themselves vector spaces. More formally, a subspace U of a vector space V is a subset of V that is itself a vector space. The subspace U of the vector space V must satisfy the following properties:

- **Closure under addition:** For any two vectors $\mathbf{x},\mathbf{y}\in U$, the sum $\mathbf{x}+\mathbf{y}\in U$.
- **Closure under scalar multiplication:** For any scalar $c\in\mathbb{R}$ and any vector $\mathbf{x}\in U$, the product $c\mathbf{x}\in U$.

One immediate consequence of the above definition is that the zero element of the vector space V must be present in every subspace of V. If the zero element is not in a subset, then it cannot be a subspace. Geometrically subspace are flat structures (or surfaces or manifolds) in $\mathbb{R}^n$ (or the parent vector space) that contain the origin, and extends infinitely. Let's look at some examples of subspaces of $\mathbb{R}^2$ and $\mathbb{R}^3$, which are easier to visualize.

Example 1.6. A straight line through the origin. We know that $\mathbb{R}^2$ is a vector space. Now consider the set of all points in $\mathbb{R}^2$ that lie on a straight line passing through the origin, defined as follows:

$$S=\left\{\mathbf{x}\,:\,x=\begin{bmatrix}x_1\\x_2\end{bmatrix}\in\mathbb{R}^2,\,x_1=m\cdot x_2,\,m\in\mathbb{R}\right\}$$

How do we verfiy this is a subspace of $\mathbb{R}^2$? The definition above shows that any x in S comes from $\mathbb{R}^2$, which means its a subset of $\mathbb{R}^2$. Figure 1.6a shows the set S for $m=2$. How do we verify if S is a subspace of $\mathbb{R}^2$? We need to now verify that S satisfies the properties of a vector space.

1. First, let's check if S contains the zero vector. If it does not contain the zero vector, then it cannot be a subspace. The elements from S are of the form $\begin{bmatrix}x\\mx\end{bmatrix}$, thus if we choose $x=0$, then we get $\begin{bmatrix}0\\0\end{bmatrix}\in S$. So, S contains the zero vector. This means that S can be a subspace space of $\mathbb{R}^2$.

2. Let's verify vector scaling. Scaling the element $\begin{bmatrix}x\\mx\end{bmatrix}\in S$ by a scalar c we get,

$$c\begin{bmatrix}x\\mx\end{bmatrix}=\begin{bmatrix}cx\\cmx\end{bmatrix}=\begin{bmatrix}cx\\m(cx)\end{bmatrix}=\begin{bmatrix}y\\my\end{bmatrix},\quad\text{where } y=cx\in\mathbb{R}$$

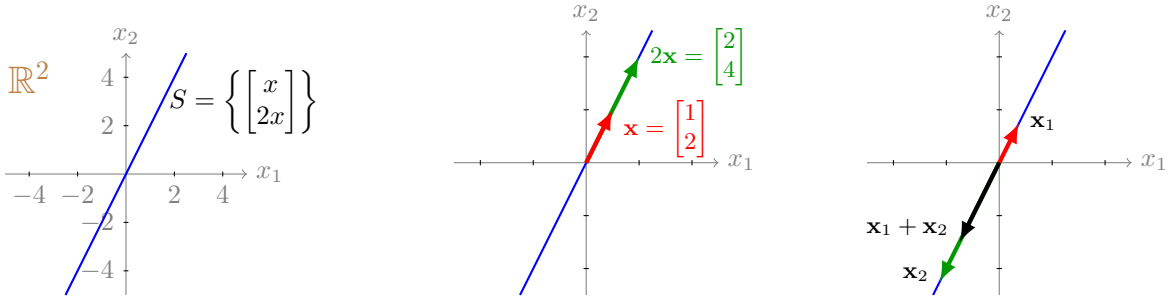
   This means that $c\begin{bmatrix}x\\mx\end{bmatrix}$ belongs to S, this the set S is closed under scalar multiplication. This still means that S can be a subspace of $\mathbb{R}^2$.

3. Let's verify vector addition. Adding two elements $\begin{bmatrix}x_1\\mx_1\end{bmatrix},\begin{bmatrix}x_2\\mx_2\end{bmatrix}\in S$ we get,

$$\begin{bmatrix}x_1\\mx_1\end{bmatrix}+\begin{bmatrix}x_2\\mx_2\end{bmatrix}=\begin{bmatrix}x_1+x_2\\mx_1+mx_2\end{bmatrix}=\begin{bmatrix}y_1\\my_1\end{bmatrix},\quad\text{where } y_1=x_1+x_2\in\mathbb{R}$$

   This means that $\begin{bmatrix}y_1\\my_1\end{bmatrix}$ belongs to S, this the set S is closed under vector addition. This means that S is a subspace of $\mathbb{R}^2$.

Since, the subset S is closed under both vector addition and scalar multiplication, it is a subspace of $\mathbb{R}^2$.



(a) A subspace of $\mathbb{R}^2$. **(b)** Vector scaling. **(c)** Vector addition.

Figure 1.6: Example of a subspace of $\mathbb{R}^2$. (a) Shows the set of all points in $\mathbb{R}^2$ corresponding to the subset $S = \left\{ \begin{bmatrix} x \\ 2x \end{bmatrix} \right\} \subset \mathbb{R}^2$. (b) Shows that the set S is closed under scalar multiplication. Take any vecotr from the line, and scale it and it remains on that blue line. (c) Shows that S is closed under vector addition. If we take any two vectors from the blue line and add them, the resulting vector remains in the blue line.

Example 1.7. A straight line not through the origin. Consider the set of all points in $\mathbb{R}^2$ of the following form:

$$S = \left\{ \mathbf{x} \,:\, x = \begin{bmatrix} x \\ mx + c \end{bmatrix} \in \mathbb{R}^2,\, m, c \in \mathbb{R} \right\}$$

This is shown in the Figure 1.7a.

How do we verfiy this is a subspace of $\mathbb{R}^2$? The definition above shows that any x in S comes from $\mathbb{R}^2$, which means its a subset of $\mathbb{R}^2$. Figure 1.7a shows the set S for $m = -\frac{1}{2}$ and $c = 1$. How do we verify if S is a subspace of $\mathbb{R}^2$? We need to now verify that S satisfies the properties of a vector space.

1. First, let's check if S contains the zero vector. If it does not contain the zero vector, then it cannot be a subspace. The elements from S are of the form $\begin{bmatrix} x \\ mx + c \end{bmatrix}$, thus if we choose $x = 0$, then we get $\begin{bmatrix} 0 \\ c \end{bmatrix} \in S$. So, S does not contain the zero vector, which implies that S is not a subspace space of $\mathbb{R}^2$. We need not check the other two conditions; but we will test them just to see which of these two fails.

2. Scaling the element $\begin{bmatrix} x \\ mx + c \end{bmatrix} \in S$ by a scalar d we get,

$$d \begin{bmatrix} x \\ mx + c \end{bmatrix} = \begin{bmatrix} dx \\ dmx + dc \end{bmatrix} = \begin{bmatrix} dx \\ m\,(dx) + dc \end{bmatrix} \neq \begin{bmatrix} y \\ my + c \end{bmatrix}, \quad \text{where } y = dx \in \mathbb{R}$$

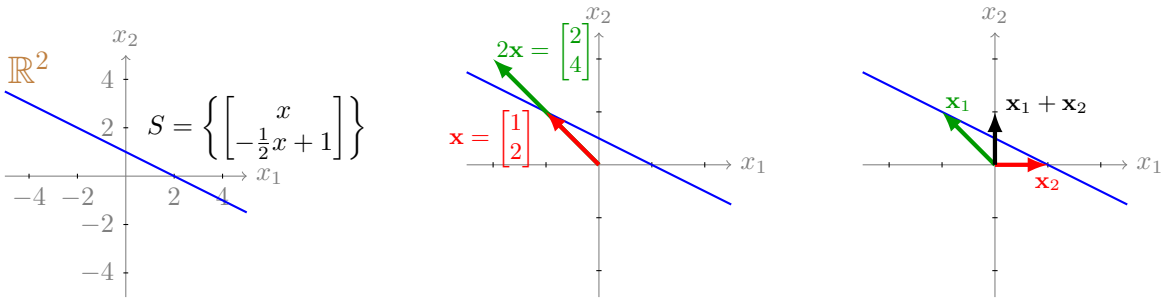
   This means that $d \begin{bmatrix} x \\ mx + c \end{bmatrix} \notin S$. Thus, the set S is not closed under scalar multiplication. Another confirmation that it is not a subspace. This can be seen in Figure 1.7b, which shows that when we choose an elment from $\mathbf{x}$ (red arrow) from S (blue line), the scaled version of this vector leaves the set S, i.e., the tip of the green arrow does not stay on the blue line.

3. Let's verify vector addition. Adding two elements $\begin{bmatrix} x_1 \\ mx_1 + c \end{bmatrix}, \begin{bmatrix} x_2 \\ mx_2 + c \end{bmatrix} \in S$ we get,

$$\begin{bmatrix} x_1 \\ mx_1 + c \end{bmatrix} + \begin{bmatrix} x_2 \\ mx_2 + c \end{bmatrix} = \begin{bmatrix} x_1 + x_2 \\ mx_1 + mx_2 + 2c \end{bmatrix} \neq \begin{bmatrix} y_1 \\ my_1 + c \end{bmatrix}, \quad \text{where } y_1 = x_1 + x_2 \in \mathbb{R}$$

This means that $\begin{bmatrix} x_1 \\ mx_1 + c \end{bmatrix} + \begin{bmatrix} x_2 \\ mx_2 + c \end{bmatrix} \notin S$. Thus the set S is not closed under vector addition. We see this geometrically in Figure 1.7c, where the sum of two vectors in S does not stay in the set S. Even though the tips of the green and red arrow are on the blue line, the tip of the black arrow is not on the blue line.

Since, the subset S is not closed under both vector addition and scalar multiplication, it is not a subspace of $\mathbb{R}^2$.



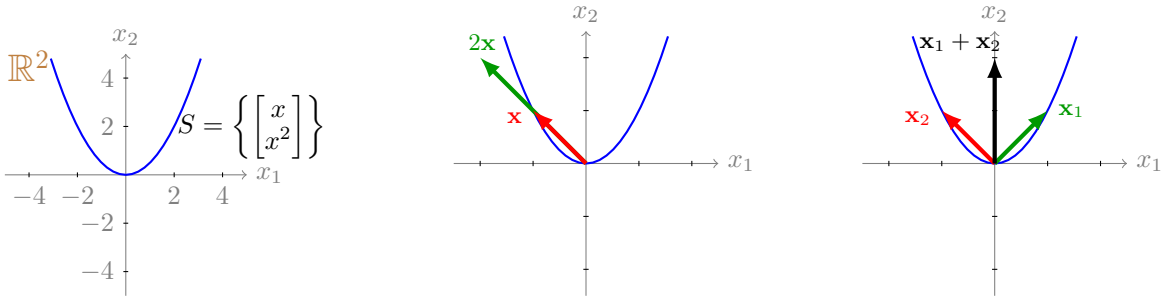
(a) Not a subspace of $\mathbb{R}^2$. **(b)** Vector scaling. **(c)** Vector addition.

Figure 1.7: Example of a subspace of $\mathbb{R}$. (a) Shows the set of all points in $\mathbb{R}^2$ corresponding to the subset $S = \left\{ \begin{bmatrix} x \\ 2x \end{bmatrix} \right\} \subset \mathbb{R}^2$. (b) Shows that the set S is closed under scalar multiplication. Take any vecotr from the line, and scale it and it remains on that blue line. (c) Shows that S is closed under vector addition. If we take any two vectors from the blue line and add them, the resulting vector remains in the blue line.

Example 1.8. A parabola through the origin. Consider the set of all points in $\mathbb{R}^2$ of the following form:

$$S = \left\{ \mathbf{x} : x = \begin{bmatrix} x \\ \frac{1}{2}x^2 \end{bmatrix} \in \mathbb{R}^2,\ m, c \in \mathbb{R} \right\}$$

This is not a subspace of $\mathbb{R}^2$. This is geometrically depicted in Figure 1.8a, Figure 1.8b and Figure 1.8c. You are encouraged to verify this algebraically by checking the properties of a vector space.



(a) Not a subspace of $\mathbb{R}^2$. **(b)** Vector scaling. **(c)** Vector addition.

Figure 1.8: Example of a subspace of $\mathbb{R}$. (a) Shows the set of all points in $\mathbb{R}^2$ corresponding to the subset $S = \left\{ \begin{bmatrix} x \\ 2x \end{bmatrix} \right\} \subset \mathbb{R}^2$. (b) Shows that the set S is closed under scalar multiplication. Take any vecotr from the line, and scale it and it remains on that blue line. (c) Shows that S is closed under vector addition. If we take any two vectors from the blue line and add them, the resulting vector remains in the blue line.

1.6 Linear combination

Linear combination is an *algebraic operation* performed on a set of vectors. We can combine the two fundamental operations on vectors into a single operation called the *linear combination* of a set of vectors. Given a set of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m \in \mathbb{R}^n$ and scalars $c_1, c_2, \ldots, c_n \in \mathbb{R}$, the linear combination of the vectors is given by:

$$\mathbf{v} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_m \in \mathbb{R}^n \tag{1.1}$$

Notice that the linear cominations of single vector $\mathbf{v}_1$ are simply different scaled versions of the vector $c_1\mathbf{v}_1$. Linear cominations are the bread-and-butter of linear algebra and we will encounter them again and again. An informal way to think of a linear combination of a set of vectors as process of mixing the set of vectors together with the corresponding scalar c_i determining the amount of a vector in the mixture. There are other types of combinations of vectors, which we will not discuss further in this book.

- **Affine combination:** $\mathbf{v} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_m, \quad \sum_{i=1}^{m} c_i = 1$
- **Convex combinations:** $\mathbf{v} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_m, \quad c_i \geq 0, \ \sum_{i=1}^{m} c_i = 1$
- **Conic combinations:** $\mathbf{v} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_m, \quad c_i \geq 0$

1.7 Linear independence of a set of vectors

- Linear independence is a property of a set of vectors.
- A set is either linear independent or its not.
- No element of a linearly independent set can be represented as linear combination of the other elements in the set.
- Linearly independent set does not have any redundancy.

Linear independece is a *property* of a set of vectors; a set of vector is either linearly independent or linearly dependent. The concept of linear independence is easy to understand but the algebraic condition for independence can seem a bit unintuitive. A set of vectors is said to be linearly independent if no vector in the set can be expressed as a linear combination of the other vectors in the set. This means that there is some unique information contained in each element of the set, which cannot be obtained from the other elements of the set, i.e., there is no redundancy, so to speak.

More formally, a set of vectors $V = \{\mathbf{v}_i\}_{i=1}^{m}$ is said to be linearly independent if and only if the only way to produce the zero vector $\mathbf{0}$ through the linear combination of the set V is by setting all the scalars to zero, i.e.,

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_m = \mathbf{0} \quad \text{if and only if} \quad c_1 = c_2 = \cdots = c_m = 0 \tag{1.2}$$

To understand this better, let's assume that the set V is linear dependent and let's assume that the vector $\mathbf{v}_m$ can be represented as the linear combination of the vectors $\mathbf{v}_1, \mathbf{v}_2, \cdots \mathbf{v}_{m-1}$. This means that there exist a set of scalar $\alpha_i, \ 1 \leq i \leq m-1$, such that

$$\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_{m-1}\mathbf{v}_{m-1} = \mathbf{v}_m$$

We can rewrite this as the following,

$$\begin{aligned} &\alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_{m-1}\mathbf{v}_{m-1} - \mathbf{v}_m = \mathbf{0} \\ \implies &c_m\alpha_1\mathbf{v}_1 + c_m\alpha_2\mathbf{v}_2 + \cdots + c_m\alpha_{m-1}\mathbf{v}_{m-1} - c_m\mathbf{v}_m = \mathbf{0} \end{aligned}$$

This implies that there exist a set of scalar $c_i = \alpha_i, \ 1 \leq i \leq m-1$, and $c_m = -1$ such that $\sum_{i=1}^{n} c_i\mathbf{v}_i = \mathbf{0}$, where not all c_i are zero. So when a set is linearly dependent, then there are scalars c_i, not all zero, such that the linear combination of the vectors from V with these scalars produces the zero vector.

Now, let's assume that the set V is linearly independent, that is no vector in the set V can be expressed as a linear combination of other vectors in that set. And let's assume that there are scalars c_i, not all zero, such that the linear combination of the vectors from V with these scalars produces the zero vector, i.e.,

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_m\mathbf{v}_m = \mathbf{0}$$
$$\implies \frac{c_1}{c_m}\mathbf{v}_1 + \frac{c_2}{c_m}\mathbf{v}_2 + \cdots + \frac{c_{m-1}}{c_m}\mathbf{v}_{m-1} = \mathbf{v}_m,\ c_m \neq 0$$

But this a contradiction because we have just expressed $\mathbf{v}_m$ is expressed as a linear combination of the vectors $\mathbf{v}_1, \mathbf{v}_2, \cdots \mathbf{v}_{m-1}$.

Example 1.9. Consider the set of vectors $\{\mathbf{v}_1, \mathbf{v}_2\}$, such that $\mathbf{v}_1 = \begin{bmatrix}1\\0\end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix}0\\1\end{bmatrix}$ in $\mathbb{R}^2$. This set is linearly independent. Let's verify this algebraically. Let's assume that there exist scalars c_1, c_2 such that $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 = \mathbf{0}$. This implies that, $c_1\begin{bmatrix}1\\0\end{bmatrix} + c_2\begin{bmatrix}0\\1\end{bmatrix} = \begin{bmatrix}c_1\\c_2\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix} \implies c_1 = c_2 = 0$. Thus the set $\{\mathbf{v}_1, \mathbf{v}_2\}$ is linearly independent.

Example 1.10. Consider the set of vectors $\{\mathbf{v}_1, \mathbf{v}_2\}$, such that $\mathbf{v}_1 = \begin{bmatrix}1\\0\end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix}0\\1\end{bmatrix}, , \mathbf{v}_3 = \begin{bmatrix}23\\-5\end{bmatrix}$ in $\mathbb{R}^2$. This set is not linearly independent, i.e., its linear dependent. Let's assume that there exist scalars c_1, c_2, c_3 such that $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 = \mathbf{0}$. This implies that, $c_1\begin{bmatrix}1\\0\end{bmatrix} + c_2\begin{bmatrix}0\\1\end{bmatrix} + c_3\begin{bmatrix}23\\-5\end{bmatrix} = \begin{bmatrix}c_1 + 23c_3\\c_2 - 5c_3\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix} \implies c_1 = -23c_3$ and $c_2 = 5c_3$. If we choose c_3 to be a non-zero value, we have a set of non-zero scalars such that linear combination of the set $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ produces the zero vector. Thus, this set is linearly dependent.

Example 1.11. Consider the set of vectors $\{\mathbf{v}_1\}$, $\mathbf{v}_1 = \begin{bmatrix}-1\\1\\1\end{bmatrix}$ in $\mathbb{R}^3$. This set is linearly independent. Let's assume that there exist scalars c_1 such that $c_1\mathbf{v}_1 = \mathbf{0}$. This implies that, $\begin{bmatrix}-c_1\\c_1\\c_1\end{bmatrix} = \begin{bmatrix}0\\0\\0\end{bmatrix} \implies c_1 = 0$. Thus, the set $\{\mathbf{v}_1\}$ is linearly independent.

Example 1.12. Consider the set $\{\mathbf{0}\}$ in $\mathbb{R}^3$. This set is linearly dependent. Here, $c_1\mathbf{0} = \begin{bmatrix}0c_1\\0c_1\\0c_1\end{bmatrix} = \mathbf{0}$. Any non-zero c_1 will produce the zero vector. Thus, the set $\{\mathbf{v}_1\}$ is linearly dependent. In fact, any set that contains the zero vector is linearly dependent. (Why? Can you show this algebraically?)

1.8 Span of a set of vectors

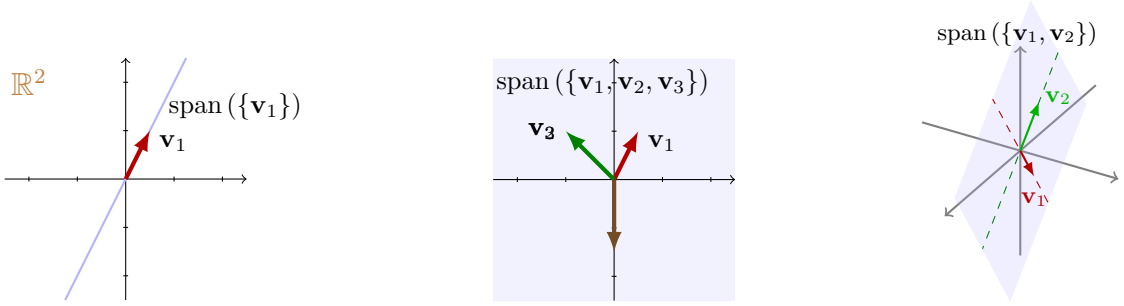
- The span of a set of vectors V is another set.
- It is generated through the linear combination of the elements of V.
- The span of a set of vector V is a subsapce of the original vectors space the elements of V are from.

So, linear combinations of a set of vectors $V = \{\mathbf{v}_i\}_{i=1}^m$ ($\mathbf{v}_i \in \mathbb{R}^n$) is a way of generating new vectors not in that set. All we need to do is choose a random set of real numbers $\{c_i\}_{i=1}^m$, and "mix" the vectors $\mathbf{v}_i$ from the set using these as weights. Clearly there are infinite number of vectors we could generate through this process, and we can put them all together in a set. And this set has a name – the *span* of the set V. The span of a set of vectors $V = \{\mathbf{v}_i\}_{i=1}^m$ is denoted by span (V) and is defined as:

$$\text{span}\,(V) = \{c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_m \ :\ c_i \in \mathbb{R}\} \subseteq \mathbb{R}^n \tag{1.3}$$

Its clear that this will be a subset of $\mathbb{R}^n$, but it turns out that it is also a subspace of $\mathbb{R}^n$. Why? Can you verify this fact algebraically? (*Hint*: Just follow the steps in Examples 1.6-1.8).

Geometrically, this means that the span (V) will be a flat surface in $\mathbb{R}^n$. Which means that the linear combination operation generates vectors that lie on a flat surface spanned by the vectors employed in the linear combination.



(a) Not a subspace of $\mathbb{R}^2$. **(b)** Vector scaling. **(c)** Vector addition.

Figure 1.9: Span of a set of vectors in $\mathbb{R}^2$ and $\mathbb{R}^3$.

1.9 How big is a vector?

The size of a vector is an extension of the idea of the magnitude of a real number. The magnitude of a real number $a \in \mathbb{R}$ tells us how big the number is irrespective of its sign:

$$|a| = \begin{cases} a, & a \geq 0 \\ -a, & a < 0 \end{cases} \tag{1.4}$$

The "magnitude" or size of an element of a vector space (such as $\mathbb{R}^n$) is called the *norm* of the vector. The norm is a generalization of the magnitude of a real number to a vector. The norm of a vector is a function that maps a vector to a non-negative real number, and satisfies the following properties:

- **Non-negativity:** For any vector $\mathbf{x} \in \mathbb{R}^n$, $\|\mathbf{x}\| \geq 0$.
- **Definiteness:** The norm of a vector is zero if and only if the vector is the zero vector, i.e., $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$.
- **Homogeneity:** Scaling a vector by a scalar c, scales the norm of the vector by $|c|$. For any vector $\mathbf{x} \in \mathbb{R}^n$ and any scalar $c \in \mathbb{R}$, $\|c\mathbf{x}\| = |c|\|\mathbf{x}\|$.
- **Triangle inequality:** For any vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$.

According to this definition, the magnitude of real numbers (Eq. 1.4) is a norm of the vector space $\mathbb{R}$. The most common norm of a vector is the *Euclidean norm* or the *2-norm* of a vector. The Euclidean norm of a vector $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ is defined as:

$$\|\mathbf{x}\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \tag{1.5}$$

We are well-versed with this as the length of a vector in $\mathbb{R}^2$ and $\mathbb{R}^3$. The properties of non-negativity, definiteness, and homogeneity are easy to verify. The triangle inequality is a bit more involved.

The subscript 2 in Eq. 1.5 is used to indicate that it is the 2-norm, which is a special case of a general class of norms in $\mathbb{R}^n$ – the *p-norm*. The *p-norm* is defined as the following:

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}, \quad p \in \mathbb{Z}, \ p \geq 1 \tag{1.6}$$

Apart from the *2-norm*, the *1-norm* and the ∞-*norm* are also commonly used norms, which are defined as the following:

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|, \qquad \|\mathbf{x}\|_\infty = \max_i |x_i| \tag{1.7}$$

The *1-norm* is the sum of the absolute value of the elements of the vector, and the ∞-*norm* is the maximum of the absolute value of the elements of the vector. The *1-norm* is also called the *Manhattan norm* or the *Taxicab norm* because it measures the distance between two points in a city if you can only travel along the grid of streets.

Example 1.13. Let's calculate the 1-norm, 2-norm, and ∞-norm of the some vectors:

1. $\mathbf{x}_1 = \begin{bmatrix} 1 \\ -1 \\ 3 \end{bmatrix} \longrightarrow \|\mathbf{x}_1\|_1 = 5, \quad \|\mathbf{x}_1\|_2 = \sqrt{1+1+9} = \sqrt{11}, \quad \|\mathbf{x}_1\|_\infty = 3.$
2. $\mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \longrightarrow \|\mathbf{x}_1\|_1 = 1, \quad \|\mathbf{x}_1\|_2 = 1, \quad \|\mathbf{x}_1\|_\infty = 1.$ All the p-norms of the unit vectors are 1. No wonder we call them "unit" vectors.
3. $\|\mathbf{0}\|_1 = \|\mathbf{0}\|_2 = \|\mathbf{0}\|_\infty = 0.$ All p-norms will produce 0, otherwise it is not a norm (remember the definiteness property?)

Problem 1.1. Why does the ∞-norm measure have this weird looking definition compared to the other p-norms?

Solution. Consider the vector $\mathbf{x} \in \mathbb{R}^n$, and $x_{max} = \max_{0 \le i \le n} |x_i|$; let's also assume that the j^{th} element of $\mathbf{x}$ has the maximum absolute value, i.e. $x_{max} = |x_j|$. The p-norm is defined as the following:

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} = x_{max} \left(1 + \sum_{\substack{1 \le i \le n \\ i \ne j}} \left| \frac{x_i}{x_j} \right|^p \right)^{1/p} = x_{max} \, (N)^{1/p}$$

where, N is a real number between 1 and n, because $|\frac{x_i}{x_j}| \le 1$ (why?). Now, if we increase the value of p to infinity, then the term $\lim_{p \to infty} (N)^{1/p} = 1$. Thus, we have $\|\mathbf{x}\|_\infty = \lim_{p \to \infty} \|\mathbf{x}\|_p = x_{max} = \max_i |x_i|$. □

1.9.1 Geometry of the p-norms

In the case of real numbers, the set of all numbers with a magnitude of 1 is the set $\{-1, 1\}$. We can plot these points in the real line as below.

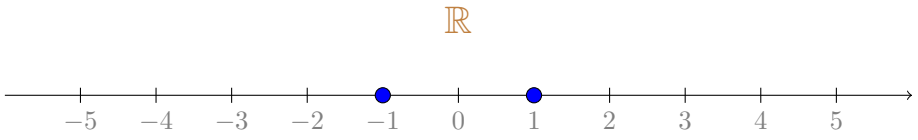
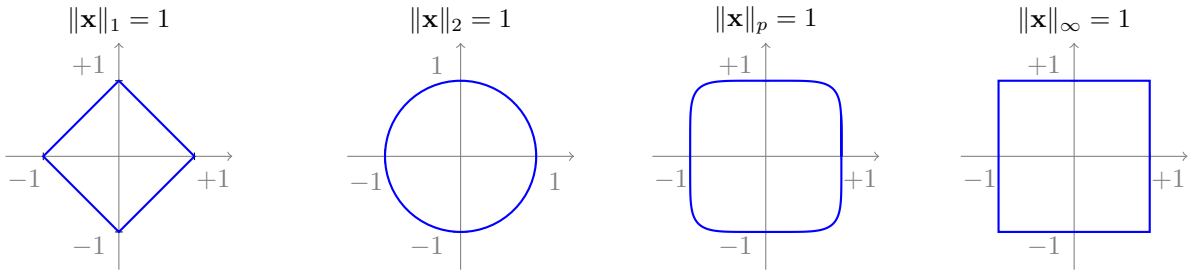


Figure 1.10: The set of all real numbers with magnitude 1. This set contains two numbers $\{-1, 1\}$.

In $\mathbb{R}^2$, the set of all vectors from $\mathbb{R}^2$ with a *2-norm* of 1 is the unit circle. The following figure shows the set of all points in $\mathbb{R}^2$ with unit 1, 2, p, and ∞ norm.

Problem 1.2. Can you explain why the different norms have these shapes?

Problem 1.3. Can you write a Python program to generate the above plots for different values of $p =$



(a) 1-norm **(b)** 2-norm **(c)** p-norm **(d)** ∞-norm

Figure 1.11: Locus of all points with unit 1, 2, p, and ∞ norms in $\mathbb{R}^2$.

$1, 2, 3, 10$ and ∞?

Problem 1.4. Can describe what these 1, 2, p and ∞ norms will look like in $\mathbb{R}^3$?

1.10 How similar are two vectors?

The idea of how similar two or more vectors are is an important topic in data analysis, in particular in classification problems in machine learning. Vectors that are "similar" somehow belong to the same "category" or "class", while vectors that are "dissimilar" belong to different categories or classes. There are various ways to measure the similarity between two vectors. We will look at two methods in this section where similarity is mesasured by computing the distance between two vectors or by computing the angle between two vectors.

1.10.1 Distance between two vectors

The logic here is that similar vectors correspond to points that are close together, while disimilar vectors are father away. We can make use of the norm to compute the distance between two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Since the the difference between these two vectors $\mathbf{x} - \mathbf{y}$ is also another vector, we can compute the distance between vectors $\mathbf{x}$ and $\mathbf{y}$ as the norm of the vector $\mathbf{x} - \mathbf{y}$ (Figure 1.12a).

$$\text{Distance between } \mathbf{x} \text{ and } \mathbf{y} = d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_p$$

We could use any of the p-norms to compute this or come-up with a new norm depending on the application we are dealing with. Take look at the clusters of points shown in Figure 1.12b, we would agree that the different colored points each form a cluster, since the points of the same color are closer to each other than points from another color.

Example 1.14. Test scores in ALADA. The ALADA course has three segments: linear algebra, optimization, and probability/statistics. Let's assume that the final exam contains three sections with a maximum of 25 marks students. The scores from these three segments are stored in a 3-vector of the form $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \in \mathbb{R}^3$, where x_1, x_2, x_3 are the marks obtained for linear algebra, optimization, and probability/statistics section, respectively. Consider the scores from the 6 students that took the course:

$$\mathbf{x}_1 = \begin{bmatrix} 2 \\ 5 \\ 1 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 3 \\ 3 \\ 4 \end{bmatrix}, \quad \mathbf{x}_3 = \begin{bmatrix} 8 \\ 20 \\ 22 \end{bmatrix}, \quad \mathbf{x}_4 = \begin{bmatrix} 9 \\ 20 \\ 21 \end{bmatrix}, \quad \mathbf{x}_5 = \begin{bmatrix} 24 \\ 24 \\ 23 \end{bmatrix}, \quad \mathbf{x}_6 = \begin{bmatrix} 24 \\ 23 \\ 22 \end{bmatrix}$$

The distance between the scores of these students tells us something about the ability of the students in the course. Let $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$ be the Euclidean distance between the scores of student i and j; notice that $d_{ij} = d_{ji}$. The distance between the different scores is given by the following table.

		$\mathbf{x}_i$					
		1	2	3	4	5	6
$\mathbf{x}_j$	1	0.0	3.7	26.5	26.0	36.5	35.3
	2	θ_{12}	0.0	25.3	24.8	35.3	34.1
	3	θ_{13}	θ_{23}	0.0	1.4	16.5	16.3
	4	θ_{14}	θ_{24}	θ_{34}	0.0	15.7	15.3
	5	θ_{15}	θ_{25}	θ_{35}	θ_{45}	0.0	1.4
	6	θ_{16}	θ_{26}	θ_{36}	θ_{46}	θ_{56}	0.0

The following observations can be made from the table:

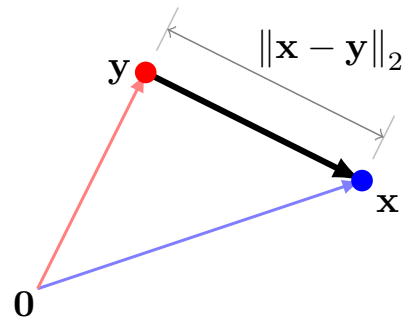
- Students 1 and 2 have similar scores, compared to the other students.
- Students 3 and 4 have very similar scores, compared to the other students.
- Students 5 and 6 have very similar scores, compared to the other students.
- Students 1 and 2, are closer to 3 and 4 than to 5 and 6.

Notice that you could have also used the other norms to creata table similar to the above one. Its left as an exercise for you generate a similar table using the 1-norm and the ∞-norm to define the distance between the scores of the students.

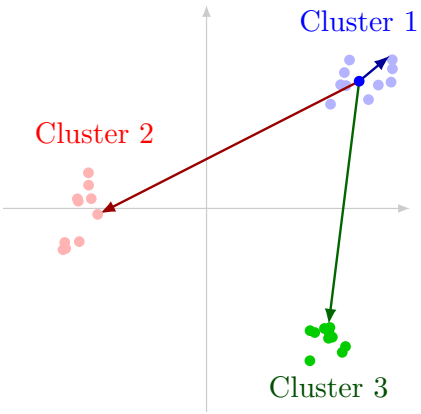
Using norms instead of norms of differences. Another way to understand the score vectors is to directly compute the norms of $\mathbf{x}_i$ and see what information they convey about the students' performance in the ALADA final exam.

$\|\mathbf{x}_1\|_2$	$\|\mathbf{x}_2\|_2$	$\|\mathbf{x}_3\|_2$	$\|\mathbf{x}_4\|_2$	$\|\mathbf{x}_5\|_2$	$\|\mathbf{x}_6\|_2$
5.5	5.8	30.8	30.4	41.0	39.9

The size of the score vectors tells us that students 1 and 2 performed worst among the six students, while students 5 and 6 performed the best; the performance of students 3 and 4 was somehwere in the middle. Would we have reached similar conclusions if we had used the 1-norm or the ∞-norm to compute the norms of the score vectors?



(a) Distance between two vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^2$. This figures depicts the 2-norm, but any p-norm or valid norm function could be used to quantify the distance between two vectors or points.



(b) Distance between two vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^2$. This figures depicts the 2-norm, but any p-norm or valid norm function could be used to quantify the distance between two vectors or points. could be used to quantify the distance between

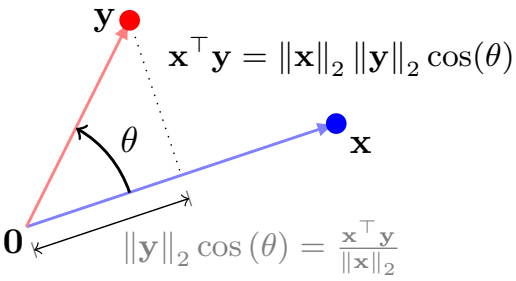
Figure 1.12

1.10.2 Angle between two vectors

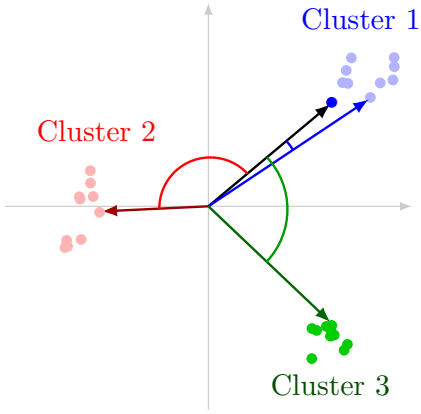
This approach is based on the idea that the direction of the vector repesenting a point contains information about the point. Thus, vectors that point in a similar direction could be considered similar. But how do we measure the angle between two vectors in $\mathbb{R}^n$? This is where the concept of the *standard inner product* (or the dot product from vectors from high school math and physics). The standard inner product of two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ is defined as:

$$\mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{n} x_i y_i$$

The superscript '$\top$' represents the transpose operation. We will not worry about what it means until the next chapter. The standard inner product takes in two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and returns a scalar value $\mathbb{R}$; it can be both positive and negative. We compute it by simply taking the two vectors $\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$, and multiply the two of them element-wise $x_i y_i$, $1 \leq i \leq n$ and add the n products together $\sum_{i=1}^{n} x_i y_i$ to obtain the inner product.



(a) Angle between two vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^2$. The standard inner product provides a measure of the cosine of the angle between the two vectoes.



(b) The relative angle between the points of the same colors is smaller than that of points of different colors.

Figure 1.13

The standard inner product of two vectors $\mathbf{x}$ and $\mathbf{y}$ is related to the cosine of the angle θ between the two vectors, and the 2-norm of the two vectors.

$$\mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{n} x_i y_i = \|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2 \cdot \cos(\theta)$$

The angle θ between the two vectors $\mathbf{x}$ and $\mathbf{y}$ can be computed as:

$$\theta = \cos^{-1}\left(\frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2}\right), \ \|\mathbf{x}\|_2 \neq 0, \|\mathbf{y}\|_2 \neq 0$$

If the 2-norms of the $\mathbf{x}$ and $\mathbf{y}$, then $\mathbf{x}^\top \mathbf{y}$ is simply the cosine of the angle between the vectors.

Example 1.15. Let's look at the data from Example 1.14, but this time using the angle between the vectors to understand the performance of the students in the ALADA final exam. The angle between the scores (in degrees) of the students is given by the following table. Let $\theta_{ij} = \cos^{-1}\left(\frac{\mathbf{x}_i^\top \mathbf{x}_j}{\|\mathbf{x}_i\|_2 \|\mathbf{x}_j\|_2}\right)$ be the angle between the scores of student i and j; notice that $\theta_{ij} = \theta_{ji}$.

		$\mathbf{x}_i$					
		1	2	3	4	5	6
$\mathbf{x}_j$	1	0.0	38.5	35.1	33.3	31.7	32.1
	2	θ_{12}	0.0	16.7	15.0	9.2	9.9
	3	θ_{13}	θ_{23}	0.0	2.5	21.1	22.2
	4	θ_{14}	θ_{24}	θ_{34}	0.0	18.7	19.9
	5	θ_{15}	θ_{25}	θ_{35}	θ_{45}	0.0	1.2
	6	θ_{16}	θ_{26}	θ_{36}	θ_{46}	θ_{56}	0.0

It looks like the angles do not do a good job of capturing the similarity between the scores of the students like the distance between the scores, in particular θ_{12} is quite large, while θ_{34} and θ_{56} are quite small. Why do you think this is so?

1.11 Standard and other inner products

$\mathbf{x}^\top\mathbf{y}$ is the standard inner product, which of course means there are non-standard inner products. But before we look at generalizing the concept of an inner product, let's look at some properties of the standard inner product.

- **Connection to the 2-norm**. The standard inner product of a vector $\mathbf{x}$ with itself is the square of the 2-norm of the vector, i.e., $\mathbf{x}^\top\mathbf{x} = \|\mathbf{x}\|_2^2 = \sum_{i=1}^n x_i^2$.
- **Cauchy-Bunyakovski-Schwartz Inequality**:

$$\mathbf{x}^\top\mathbf{y} \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \tag{1.8}$$

The concept of an inner product is a general one. An inner product $\langle\cdot,\cdot\rangle$ is a function that maps two vectors from $\mathbb{R}^n$ to a scalar value, and satisfies the following properties:

- **Positive definiteness:** $\langle\mathbf{x},\mathbf{x}\rangle \geq 0$, and $\langle\mathbf{x},\mathbf{x}\rangle = 0$ if and only if $\mathbf{x} = \mathbf{0}$.
- **Symmetry:** For any vectors $\mathbf{x},\mathbf{y}$, $\langle\mathbf{x},\mathbf{y}\rangle = \langle\mathbf{y},\mathbf{x}\rangle$.
- **Linearity:** For any vectors $\mathbf{x},\mathbf{y},\mathbf{z}$, and any scalars $\alpha,\beta \in \mathbb{R}$, $\langle\alpha\mathbf{x}+\beta\mathbf{y},\mathbf{z}\rangle = \alpha\langle\mathbf{x},\mathbf{y}\rangle + \beta\langle\mathbf{y},\mathbf{z}\rangle$.

We will come across other inner products in due course, but we will stick to the standard inner product for most problem in $\mathbb{R}^n$ in this course. A class of inner products in $\mathbb{R}^n$ is of the form $\langle\mathbf{x},\mathbf{y}\rangle = \mathbf{x}^\top\mathbf{A}\mathbf{y}$, where $\mathbf{A}$ is a $n \times n$ *positive definite* matrix. The standard inner product is a special case of this class of inner products where $\mathbf{A} = \mathbf{I}$, the identity matrix. That may sound like too much jargon for now, but we will come to these concepts in the upcoming chapters.

Problem 1.5. Consider the vector space $\mathbb{R}^n$. Is the the following a valid inner product of $\mathbb{R}^n$?

$$\langle\mathbf{x},\mathbf{y}\rangle = \sum_{i=1}^n w_1 x_i y_i, \;\; w_i \in \mathbb{R}, \;\; w_i > 0$$

Solution. To find out if this is a valid inner product, we need to verify by if it satisfies the properties of an inner product.

- **Positive definiteness**: The first property is positive definiteness, which is satisfied because $w_i > 0$.

$$\langle\mathbf{x},\mathbf{x}\rangle = \sum_{i=1}^n w_i x_i^2 \geq 0, \;\; \text{since, } w_i > 0 \text{ and } x_i^2 \geq 0$$

  Notice, that if any of the w_i is zero or negative, then $\langle\mathbf{x},\mathbf{y}\rangle$ will not be a valid inner product (why?).
- **Symmetry**: From the commutativity and associativity of multiplication of real numbers, we have $\langle\mathbf{x},\mathbf{y}\rangle = \langle\mathbf{y},\mathbf{x}\rangle$.
- **Linearity**: This is also satisfied. You should verify this yourself.

The given function is a valid inner product of $\mathbb{R}^n$. □

Problem 1.6. Consider the vector space $\mathbb{R}^n$. Is the the following a valid inner product of $\mathbb{R}^n$?

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} w_i x_i y_i, \;\; w_i = \begin{cases} +1, & i \text{ is odd} \\ -1, & i \text{ is even} \end{cases}$$

Solution. To find out if this is a valid inner product, we need to verify by if it satisfies the properties of an inner product.

- **Positive definiteness**: This is not satisfied. Can you provide an example where positive definiteness fails?

The given function is a *not* valid inner product of $\mathbb{R}^n$. □

1.12 Orthogonality of vectors

The concept of orthogonality is a generalization of the concept of perpendicularity in $\mathbb{R}^2$ and $\mathbb{R}^3$. Two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ are said to be orthogonal if their standard inner product is zero, i.e., $\mathbf{x}^\top \mathbf{y} = 0$.

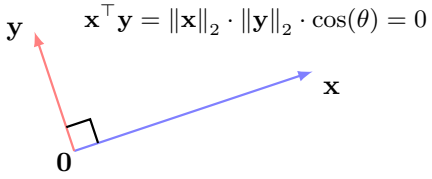


Figure 1.14: Orthogonal vectors in $\mathbb{R}^2$.

Geometrically, when two vectors $\mathbf{x}$ and $\mathbf{y}$ are orthogonal, then when we move along the direction of one vector, we are not moving along the direction of the other vector. If directions of vector convey some information about something, then vectors that are orthogonal to each other are vectors that convey mutually exclusive information,i.e. the two vectors share nothing in common. The concept of orthogonality is a very important concept in linear algebra and we will encounter it again and again.

Note that this definition of orthogonality also implied that $\mathbf{0}$ is othorognal to all vectors!

Problem 1.7. Explain why the following statement about the unit vectors of $\mathbb{R}^n$ is true.

$$\mathbf{e}_i^\top \mathbf{e}_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

All the unit vectors of $\mathbb{R}^n$ are orthogonal to each other, and have unit length.

Problem 1.8. Show if two vectors $\mathbf{x}$ and $\mathbf{y}$ are orthogonal to each other, then the following is true.

$$\|\mathbf{x} - \mathbf{y}\|_2^2 = \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2$$

1.13 Basis of a vector space

A set of vectors $V = \{\mathbf{v}_1, \mathbf{v}_2, \cdots \mathbf{v}_k\}$, $\mathbf{v}_i \in \mathbb{R}^n$ is called a basis for $\mathbb{R}^n$ if it is a linearly independent set and if it spans $\mathbb{R}^n$. This effectively means the following two things. If V is a basis for a vector space, then,

1. " spans $\mathbb{R}^n$ " $\longrightarrow$ It can be used to generate every element of $\mathbb{R}^n$ through a linear comibination operation.
2. " linearly independent " $\longrightarrow$ There is a unique linear combination of the elements of V that produces every element of $\mathbb{R}^n$.

A basis is the smallest possible set for generating a vector space.

Note that although the above definition and description is done usign $\mathbb{R}^n$, the concept of a basis applies to any vector space and all its subspaces.

Example 1.16. Consider the set of vectors $V = \{\mathbf{e}_1\} \subset \mathbb{R}^2$. This set V forms a basis for the subspace $S_1 = \left\{ \begin{bmatrix} \alpha \\ 0 \end{bmatrix} \ : \ \alpha \in \mathbb{R} \right\}$. This is because, the span $(V) = S_1$ (verify this) and the set V is linearly independent, because $\beta \mathbf{e}_1 = \mathbf{0}$ only if $\beta = 0$.

Example 1.17. Consider the set of vectors $V = \{\mathbf{e}_1, 3\mathbf{e}_1\} \subset \mathbb{R}^2$. This set V does not form a basis for the subspace $S_1 = \left\{ \begin{bmatrix} \alpha \\ 0 \end{bmatrix} \ : \ \alpha \in \mathbb{R} \right\}$. The span $(V) = S_1$ (verify this). But the set V is linearly dependent, because $-3\mathbf{e}_1 + 1\,(3\mathbf{e}_1) = \mathbf{0}$, thus a non-zero set of coefficients result in the zero vector.

Example 1.18. Consider the set of vectors $V = \{\mathbf{e}_1, \mathbf{e}_2\} \subset \mathbb{R}^2$. This set V is a basis for $\mathbb{R}^2$. The span $(V) = \mathbb{R}^2$ (verify this). And this set V is linearly independent, because $\beta_1 \mathbf{e}_1 + \beta_2 \mathbf{e}_2 = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$, thus the only was we get the zero vector through the linear combination is if $\beta_1 = \beta_2 = 0$.

Example 1.19. Consider the set of vectors $V = \left\{ \mathbf{e}_1, \mathbf{e}_2, \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right\} \subset \mathbb{R}^2$. This set V is a basis for $\mathbb{R}^2$. The span $(V) = \mathbb{R}^2$ (verify this). And this set V is linearly dependent, because $\mathbf{e}_1 + \mathbf{e}_2 - 1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \mathbf{0}$; a non-zero linear comination produces the zero vector. Thus, V is not a basis for $\mathbb{R}^2$.

How many different basis does a vector space have? For instance, how many different basis does $\mathbb{R}^2$ have? For a set to be a basis, all we need to ensure is that the set spans $\mathbb{R}^2$ and is linearly independent. Thus, there are infinitely many basis for $\mathbb{R}^2$. For instance, the sets $\{\alpha_1 \mathbf{e}_1, \alpha_2 \mathbf{e}_2\}$ with $0 \neq \alpha_1, \alpha_2 \in \mathbb{R}$ are all basis for $\mathbb{R}^2$. Since there are inifinite number of choices for α_1, α_2, we have an infite number of basis for $\mathbb{R}^2$. The same argument applies to $\mathbb{R}^n$.

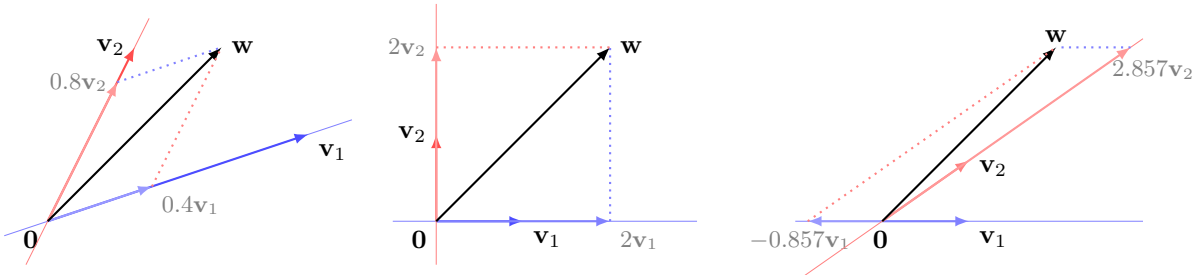


Figure 1.15: Representation of $\mathbf{w}$ in three different basis of $\mathbb{R}^2$. (a) and (c) are some arbitrary basis, while (b) is an orthonormal basis.

1.13.1 Orthonormal basis

Among the infinite number of basis for a vector space, there is a special class of basis called the *orthonormal basis*. Let $V = \{\mathbf{v}_1, \mathbf{v}_2, \cdots \mathbf{v}_n\}$, then for an orthonormal basis, the following properties hold:

$$\mathbf{v}_i^\top \mathbf{v}_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} \tag{1.9}$$

This means that all basis vector have unit 2-norm, and are orthogonal to each other. We will come across orthonormal basis often in this course, and for a good reason. Orthonormal basis is easy to work with and its easy to compute the representation of a vector in an orthonormal basis. Let V be an orthonormal basis for

$\mathbb{R}^n$, and let $\mathbf{w} \in \mathbb{R}^n$ with the following representation,

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i \mathbf{v}_i$$

Then, the coefficients α_i can be computed as the following:

$$\mathbf{v}_j^\top \mathbf{w} = \sum_{i=1}^{n} \alpha_i \mathbf{v}_j^\top \mathbf{v}_i = \alpha_j, \text{ because } \mathbf{v}_i^\top \mathbf{v}_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

A special orthonormal basis for the $\mathbb{R}^n$ is the *standard basis* $\{\mathbf{e}_1, \mathbf{e}_2 \cdots \mathbf{e}_n\}$.

1.14 Dimension of a vector space

Although there can be infinite number of basis for a vector space – they all have the same number of elements or basis vectors. This number is called the *dimension* of the vector space that they span. This is also sometimes called the *degrees of freedom* of the vector space. We represent the dimension of a vector space V as $\dim(V)$, e.g. $\dim(\mathbb{R}^n) = n$.

The dimension of a vector space also tells us that the maximum number of linearly independent vectors we can choose from that vector space. For instance, in $\mathbb{R}^n$ we can only choose n vectors that can form a linearly independent set. If we already have a linearly independent set with n elements, then adding even one more vector (any vector) to the set will make it linearly dependent. Proper subspaces of a vector space will have dimensions less than the vector space itself.

Example 1.20. The dimension of $\mathbb{R}^n$ is n. The following are subspaces of $\mathbb{R}^n$ and their dimensions. Consider the set $V = \{\mathbf{v}_1\}$ with $\mathbf{v}_1 \neq \mathbf{0}$.

- $\dim(\text{span}(V)) = 1$
- Now, let's add another vector $\mathbf{v}_2$ to V to get $V_1 = \{\mathbf{v}_1, \mathbf{v}_2\}$ which is still linearly independent, then

  $$\dim(\text{span}(V_1)) = 2$$

- Let's now add the vector $\mathbf{v}_1 - \mathbf{v}_2$ to V_1 to get $V_2 = \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_1 - \mathbf{v}_2\}$, then

  $$\dim(\text{span}(V_2)) = 2$$

  Why?

- Consider the $V_k = \{\mathbf{v}_1, \mathbf{v}_2, \cdots \mathbf{v}_k\}$ which is linearly independent.

  $$\dim(\text{span}(V_k)) = ?$$

  Can you find out the answer and explain?

1.15 Linear functions

We will conclude this chapter with a brief introduction to linear functions. A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be linear if it satisfies the following property $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\forall \alpha, \beta \in \mathbb{R}$:

$$f(\alpha \mathbf{x} + \beta \mathbf{y}) = \alpha f(\mathbf{x}) + \beta f(\mathbf{y}) \tag{1.10}$$

This means that $f(\mathbf{0}) = 0$ for all linear functions. if a function does not satisfy this property, then it is not linear.

The standard inner product $\mathbf{w}^\top \mathbf{x}$ with a fixed vector $\mathbf{w}$ is a linear function of $\mathbf{x}$,

$$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$$

An interesting fact about linear functions is that every possible $f : \mathbb{R}^n \to \mathbb{R}$ can be presented as a inner product operation with a fixed vector $\mathbf{w} \in \mathbb{R}^n$. This means the following: if f is a linear function $\mathbb{R}^n$ to $\mathbb{R}$, then there exists a vector $\mathbf{w} \in \mathbb{R}^n$ such that $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}, \forall \mathbf{x} \in \mathbb{R}^n$. This might seem like a strange fact at first. Let's now look at how we could find the vector $\mathbf{w}$ if the function f is linear. This is very simple. We first compute the value of the function for the n unit vectors of $\mathbb{R}^n$, i.e. $f(\mathbf{e}_1), f(\mathbf{e}_2), \cdots f(\mathbf{e}_n)$. Then, the vector $\mathbf{w}$ is simply the vector of these values, i.e. $\mathbf{w} = \begin{bmatrix} f(\mathbf{e}_1) \\ f(\mathbf{e}_2) \\ \vdots \\ f(\mathbf{e}_n) \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$. For any given vector $\mathbf{x}$, let the representation of $\mathbf{x}$ in the standard basis be $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$; this simply means that $\mathbf{x} = \sum_{i=1}^n x_i \mathbf{e}_i$.

$$f(\mathbf{x}) = f\left(\sum_{i=1}^n x_i \mathbf{e}_i\right) = \sum_{i=1}^n x_i f(\mathbf{e}_i) = \sum_{i=1}^n x_i w_i = \mathbf{w}^\top \mathbf{x}$$

Here is another interesting consequence of linearity. If we know the value of a linear function for a set of vectors $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots \mathbf{x}_k\}$, then know the output of the function for the set span (X). Can you think of why it is so?

1.16 Applications

The concepts convered in this chapter are enough to understand the following important and useful applications:

- k-nearest neighbors (k-NN) classification/regression algorithm
- k-means clustering algorithm

We will now look two applications in data analysis based on the concepts we have discussed in this chapter. We will present the bare-bones version of two commonly used algorithms in data analysis – the k-nearest neighbors (k-NN) classification algorithm and the k-means clustering algorithm. Numerous variations and improvements of these algorithms exist, which are beyond the scope of this course.

1.16.1 k-nearest neighbors (k-NN) classification and regression algorithms

The k-NN algorithm is a simple and intuitive algorithm for classification and regression problems. But before that, what are classification and regression problems? Both of them are problems where one is interested in finding a function or a map from a set of features (or inputs) to a target value. The features are often the form of n-vectors, and the target value is a scalar value. Classificaiton and regression problems differ in the nature of the target value. In a classification problem, the target value is a class label (from set of finite values), while in a regression problem, the target value is a real-valued number (from an interval on the real line).

Examples of classification problems

Example 1.21. Disease diagnosis. We are often interested in knowing whether or not a patient presenting with a set of symptoms at the hospital has a particular disease, based on clinical symptoms, and clinical lab tests. This is a typical example of a classification problem encountered in medicine. Given a set of features of a patient, the goal of the classifier is to determine whether the patient has a particular

disease or not.

- **Inputs**: Set of demographic data, clinical tests, imaging data, etc.
- **Output**: Disease label (e.g., positive, negative, 0: no disease/1: disease, etc.)

Example 1.22. Treatment prognosis: Let's assume there is a treatment that can be used for curing a particular disease. This treatment works well on a group of patients who recovery full after the treatment, while some only recover partially, and the rest do not recover at all. Given a patient with this disease, the goal of the treatment prognosis classifer is to predict the effectiveness of the treatment for patient. This classifier would take the features of the patient as input and predict the effectiveness of the treatment as one of three possible labels – *full recovery*, *partial recover* or *no recovery*.

- **Inputs**: Set of demographic data, clinical tests, imaging data, severity of the disease, etc.
- **Output**: Recovery label (e.g., full recovery, partial recovery, or no recovery)

Example 1.23. Spam email detection: This is a classifier that we see in action on a daily basis. Our email managers/service provides automatically send certain emails to the spam folder to weed out the usless emails from the useful ones. Given an email, the goal of such a classifier is to determine whether the email is spam or not.

- **Inputs**: Features extracted from the email content, sender, subject, etc.
- **Output**: Spam label (e.g., spam, not spam)

Example 1.24. Handwritten digit recognition: Given an image of a handwritten digit, the classifier here needs to determine the corresponding digit of the image.

- **Inputs**: Image features extracted from the image of a handwritten digit.
- **Output**: Digit label (e.g., 0, 1, 2, ..., 9)

Examples of regression problems

Example 1.25. Growth prediction: We are interested in knowing the effect iof the addition of protien supplements to the daily diet of childen over a period of three months. The goal here is to develop a model that can predict the increase in height of the children after six months given the certain level of protien supplement in their diet.

- **Inputs**: Demographic details and amount of protien supplement in the diet.
- **Output**: Increase in child's height after six months.

Example 1.26. Clinical score estimation: There are clinical tests that are considered gold standard for know the health status of a patient. These gold standard tests are often time consuming, expensive, and difficult to administer on a regular basis. Thus, we are often interestd in estimating the outcome of this gold standard test using other clinically relevant variables that are easily measured.

- **Inputs**: A set of clinically relevant parameters that are easy to measure.
- **Output**: Estimate of the value of the gold standard test.

k-NN classification algorithm

The k-NN classifier is a very simple and intuitive algorithm. Let's assume that we have a dataset with N data points/samples $(\mathbf{x}_i, y_i)\,,\ 1 \leq i \leq N$, where $\mathbf{x}_i$ is the vector of features and y_i is the know label for the i^{th} sample. The feature vector is an elements from $\mathbb{R}^n$, i.e. $\mathbf{x}_i \in \mathbb{R}^n$. The labels taken on values from a finite set with L distinct labels, $y_i \in \{1, 2, \ldots, L\}$. We would like to use this available dataset, which will

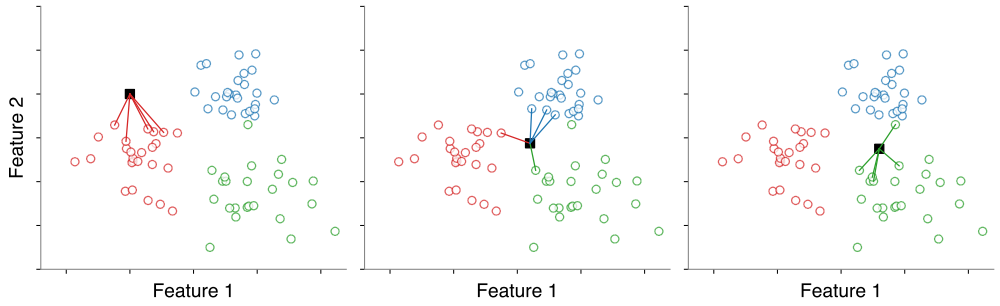


Figure 1.16: Demonstration of the k-NN classification algorithm. There are three classes or labels, which are shown in different colors. Three test points (black filled sqaure) are considered in the three plots shown in this figure. For each point the $k = 5$ nearest neighbours are depicted through lines joining the test point with the nearest neighbours. The colors of the line also indicate the class of that neighbour.

often be called the *training dataset*, to learn to correctly classify new samples that we may encounter in the future, where we have the feature vector $\mathbf{x}_{new}$ but we do not know the corresponding label y_{new}; we want a prediction of the label from the k-NN algorithm.

Although, the k-NN algorithm is called a *supervised learning* algorithm, there is really no learning in this algorithm. The algorithm keeps the entire training dataset in its memory, and assign the label to a new feature vector to be the label of the most similar feature vectors from the training dataset. The similarity can be measured through a multitude of ways, but the most common way is to use the Euclidean distance between the feature vectors. The only (hyper)parameter required for the k-NN algorithm is the value of k, which is the number of nearest neighbors to consider. The label of the new feature vector is assigned by taking a vote from the k-nearest neighbors. The label with the highest number of votes is assigned to the new feature vector. The outline of the k-NN algorithm is given below in Algorithm 1.1.

Algorithm 1.1: k-Nearest Neighbors (k-NN) Algorithm

Data: Training data $(\mathbf{x}_i, y_i)\ \ 1 \leq i \leq N$; test point $\mathbf{x}_{new}$; number of neighbors k.
Result: Predicted label for test point $\mathbf{x}_{new}$.

foreach $\mathbf{x}_j$ *in* $(\mathbf{x}_i)_{i=1}^{N}$ **do**
  Compute the distance $d(\mathbf{x}_j, \mathbf{x}_{new})$ between $\mathbf{x}_j$ and $\mathbf{x}_{new}$;
end
Sort the training instances/samples $(\mathbf{x}_i, y_i)$ by distance in ascending order;
Select the k closest samples to $\mathbf{x}_{new}$; let $\mathcal{K}$ be the set of indices between 1 to N corresponding to these k closest samples;
let $Y_{\mathcal{K}}$ be the set of labels of these k closest samples;
foreach *label in* $Y_{\mathcal{K}}$ **do**
  Compute the frequency of each label;
end
Return the label with the highest frequency;

This is depicted in Figure 1.16, which shows the example of a 3-class classification problem. The three classes are shown in different colors. The $k = 5$ nearest neighbors of three test points are shown. The label of the test point is assigned based on the majority vote of the k nearest neighbors; whenever there is a tie, a tie-breaking rule is applied to choose the label of the test point. Although, simple and intuitive, the k-NN is a computationally expensive algorithm, especially when the number of samples in the training dataset is large. The advantage, however, is that there is no training process required and there is only only hyperparameter to choose.

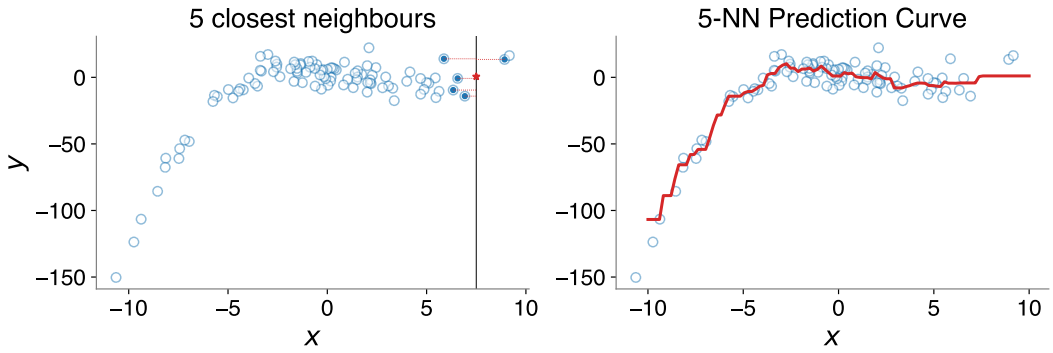


Figure 1.17: Demonstration of the k-NN regression algorithm. In this example, $\mathbf{x} \in \mathbb{R}$. The left plot demonstrates the algorithm, where the vertical black line is the $\mathbf{x}_{new}$, the filled blue circles are the 5 closest neighbours. The red star along the black line is the predicted value for $\mathbf{x}_{new}$. The right plot shows the 5-NN prediction curve for the given data in red.

k-NN regression algorithm

The k-NN regression algorithm uses the same principle as the k-NN classification algorothm. The only difference is that the output of the algorithm is the average of the y_i of the k nearest neighbors. This algorithm is detailed in Algorithm 1.2.

Algorithm 1.2: k-Nearest Neighbors (k-NN) Regression Algorithm

Data: Training data $\{(\mathbf{x}i, y_i)\}\, i = 1^N$; test point $\mathbf{x}new$; number of neighbors k.
Result: Predicted value for test point $\mathbf{x}new$.

1 **foreach** $\mathbf{x}_j$ *in* $\{\mathbf{x}_i\}_{i=1}^N$ **do**
2 | Compute the distance $d(\mathbf{x}_j, \mathbf{x}_{new})$ between $\mathbf{x}_j$ and $\mathbf{x}_{new}$;
3 **end**
4 Sort the training instances/samples $(\mathbf{x}_i, y_i)$ by distance in ascending order;
5 Select the k closest samples to $\mathbf{x}_{new}$; let $\mathcal{K}$ be the set of indices between 1 to N corresponding to these k closest samples;
6 let $Y_{\mathcal{K}}$ be the set of values of these k closest samples;
7 Compute the average of the values in $Y_{\mathcal{K}}$;
8 Return the computed average as the predicted value;

1.16.2 k-mean clustering algorithm

The k-mean is a popular clustering algorithm, which unlike the k-NN algorithms, is an *unsupervised learning* algorithm. In a *supervised learning* algorithm, we have a dataset that has an output label or numerical value of interest, which can be used to learn the association between the given features and the output label/numerical value. However, we will often come across datasets where there is no such pre-existing output label or numerical value or it is unknown. In such cases, we are simply interested in interesting patterns (clusters or groups) in the data; we need to mathematically define what we mean by "interesting" to find such patterns. The k-means algorithm lumps data points into k clusters or groups, where the elements of the cluster/group are considered to be similar; remember the discussion on measuring similarity between vectors in Section 1.10.

Let's assume that our dataset consists of N samples, each of which is a feature vector, $\mathbf{x}_i \in \mathbb{R}^n$, $1 \leq i \leq N$. We want to group the data points into k clusters, with $k < N$. Given the data points $\mathbf{x}_1 \ldots \mathbf{x}_N$, the k-means algorithm produces two outputs:

1. **Cluster means**: A set of k points $\mathbf{m}_j \in \mathbb{R}^n$, $1 \leq j \leq k$ that are supposed to be representatives of the k clusters identified. You can think of $\mathbf{m}_j$ to be a typical member of the j^{th} cluster.

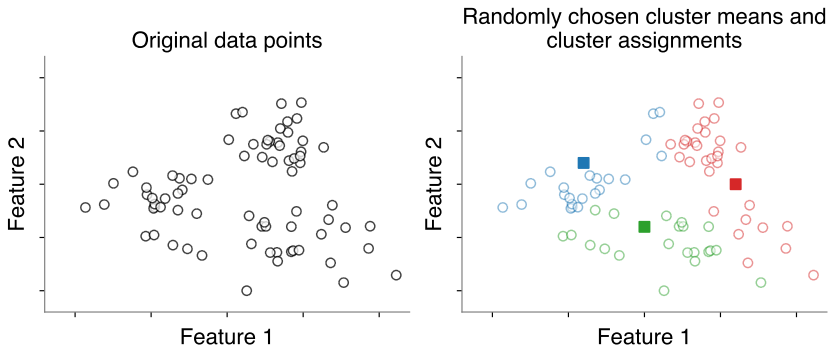


Figure 1.18: The clustering problem tackled by the k-means algorithm. The lft plot shows

2. **Cluster assignment**: Each data point $\mathbf{x}_i$ is given a cluster assignment j, such that $\mathbf{x}_i$ is closest to $\mathbf{m}_j$. This is an N-tuple of the form $(c_i)_{i=1}^N$, with $1 \leq c_i \leq k$. If $c_i = j$, then the data point $\mathbf{x}_i$ belongs to the j^{th} cluster.

This is demonstrated in Figure 1.18, where the left plot shows the dataset with N samples, each belonging to $\mathbb{R}^2$. The right plot shows three randomly chosen cluster means $\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_3$ and the cluster assignment of the data points to the closest $\mathbf{m}_j$. The three cluster means are shown in different colors and the corresponding data points in those clusters are shown in the same color (but a lighter shade). The goal for the k-means algorithm is to find the optimal cluster means and the clustern assignment such that the spread of points within each cluster is minimized across all clusters. We can measure this spread as the following,

$$J_{clust} = \frac{1}{N} \sum_{j=1}^{k} \sum_{i \in C_j} \|\mathbf{x}_i - \mathbf{m}_j\|_2^2 \tag{1.11}$$

where, $C_j = \{i : 1 \leq i \leq N, c_i = j\}$ is the set of indices of the data points that belong to the cluster j.

Minimizing J_{clust} for a given dataset is a computationally intensive problem. Note that the cluster means and the cluster assingments both depend on each other, and optimally choosing both of them to minimize J_{clust} is not easy.

The k-means simplifies the computational problem because, choosing the optimal cluster means for a fixed cluster assigment, and choosing the optimal cluster assingment for fixed cluster means is easy to do. The k-mean algorithm optmizes the cluster means and cluster assingments, while fixing the other, and iterates these steps until the cluster means and cluster assingments converge. This algorithm is detailed in Algorithm 1.3.

Algorithm 1.3: k-means Clustering Algorithm

Data: Dataset $\mathbf{x}_1 \dots \mathbf{x}_N$; number of cluster k tp be identified.
Result: A set of k cluster means $\mathbf{m}_1, \dots, \mathbf{m}_k$ and a N-tuple of cluster assingments $(c_i)_{i=1}^N$.

1 Choose a random set of k cluster means $\mathbf{m}_1, \dots, \mathbf{m}_k$;
2 **repeat**
3 **Update cluster assignment**: For the current means find the best cluster assingment. $c_i = j$, such that $\|\mathbf{x}_i - \mathbf{m}_j\|_2 < \|\mathbf{x}_i - \mathbf{m}_l\|_2, \forall\ 1 \leq l \leq k, l \neq j$;
4 **Update cluster means**: For the new cluster assignment, find the best cluster means. $\mathbf{m}_j = \frac{1}{|C_j|} \sum_{i \in C_j}$, where C_j is the set of indices of data points for the j^{th} cluster, and $|\cdot|$ is a function that returns the number of elements in a set;
5 **until** *until convergence*;
6 Return the cluster means $\mathbf{m}_1 \dots \mathbf{m}_k$ and the cluster assignment $(c_i)_{i=1}^N$;

This is demonstrated in Figure 1.18. The top left plot shows the dataset with N samples without any cluster assignment; all data points are colored black. The top right plot shows the first step in the iteration where the cluster means were chosen randomly; these are shown in different colored filled triangles. The

corresponding optimal cluster assingment of points that are closest to each mean are also depicted in this plot.

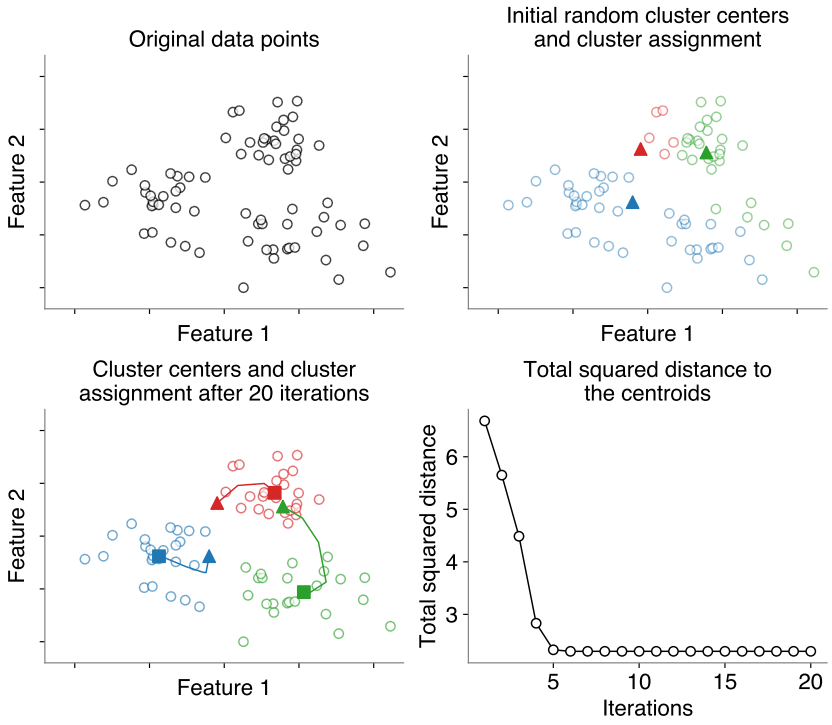


Figure 1.19: The clustering problem tackled by the k-means algorithm. The left plot shows

The bottom left plot shows trajectory of the cluster means as the k-means algorithm iterates repeating the process of updating the cluster means and cluster assingments. The two steps in of the k-means algorithm are guaranteed to reduce J_{clust}, and thus with each iteration the value of J_{clust} for the current cluster means and cluster assingments will reduce. This is depicted in the bottom right plot, which shows the trend of J_{clust} as a function of the iteration number.

1.17 Exercise

1. Is this set of vectors $\left\{ \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right\}$ independent? Explain your answer.

2. Show that the set $\{\mathbf{0}\}$, $\mathbf{0} \in \mathbb{R}^n$ a subspace of $\mathbb{R}^n$? This is called the trivial subspace of $\mathbb{R}^n$. What is the dimension of this subspace?

3. Find the 1, 2 and ∞ norms of the following vectors from $\mathbb{R}^3$:

   (a) $\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$ (b) $\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ (c) $\mathbf{e}_3$ (d) $\begin{bmatrix} a \\ b \\ c \end{bmatrix}$ (e) $\mathbf{e}_1 - \mathbf{e}_2 + \mathbf{e}_3$

4. If $S_1, S_2 \subseteq V$ are subspaces of a vectors space V then, is $S_1 \cap S_2$ a subspace? Is $S_1 \cup S_2$ a subspace? Explain your answers.

5. Consider a basis $B = \{\mathbf{b}_i\}_{i=1}^n$ of $\mathbb{R}^n$. Let the vectors $\mathbf{x}$ and $\mathbf{x}_b$ be the representations in the standard

and B basis respectively.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \sum_{i=1}^{n} x_i \mathbf{e}_i \quad \text{and} \quad \mathbf{x}_b = \begin{bmatrix} x_{b1} \\ x_{b2} \\ \vdots \\ x_{bn} \end{bmatrix} = \sum_{i=1}^{n} x_{bi} \mathbf{b}_i$$

Evaluate the $\|\mathbf{x}\|_2^2$ and $\|\mathbf{x}_b\|_2^2$. Determine what happens to $\|\mathbf{x}_b\|_2^2$ under the following conditions on the basis vectors:

(a) $\|\mathbf{b}_i\| = 1, \forall i$

(b) $\left\|\mathbf{b}_i^\top \mathbf{b}_j\right\| = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$

6. Prove the following for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

(a) **Triangle Inequality**:

$$\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$$

(b) **Backward Triangle Inequality**:

$$\|\mathbf{x} - \mathbf{y}\| \geq |\|\mathbf{x}\| - \|\mathbf{y}\||$$

(c) **Parallelogram Identity**:

$$\frac{1}{2}\left(\|\mathbf{x} + \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2\right) = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$$

7. Consider a set of vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. When is $\|\mathbf{x} - \mathbf{y}\| = \|\mathbf{x} + \mathbf{y}\|$? What can you say about the geometry of the vectors $\mathbf{x}$, $\mathbf{y}$, $\mathbf{x} - \mathbf{y}$ and $\mathbf{x} + \mathbf{y}$?

8. Consider a vector $\mathbf{v} = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}^\top$. Express the following in-terms of inner product between a constant vector $\mathbf{u}$ and the given vector $\mathbf{v}$, and in each case specify the vector $\mathbf{u}$.

(a) $\sum_{i=1}^{n} v_i$

(b) $\frac{1}{n} \sum_{i=1}^{n} v_i$

(c) $\frac{1}{5} \sum_{i=3}^{5} v_i$

9. Which of the following are linear functions of $\{x_1, x_2, \ldots, x_n\}$?

(a) $\min_i \{x_i\}_{i=1}^n$

(b) $\left(\sum_{i=1}^{n} x_i^2\right)^{1/2}$

(c) x_6

10. Consider a linear function $f : \mathbb{R}^n \to \mathbb{R}$. Prove that every linear function of this form can be represented in the following form.

$$y = f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} = \sum_{i=1}^{n} w_i x_i, \qquad \mathbf{x}, \mathbf{w} \in \mathbb{R}^n$$

11. An *affine* function f is defined as the sum of a linear function and a constant. It can in general be represented in the form,

$$y = f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + \beta, \qquad \mathbf{x}, \mathbf{w} \in \mathbb{R}^n, \beta \in \mathbb{R}$$

Prove that affine functions are not linear. Prove that any affine function can be represented in the form $\mathbf{w}^\top \mathbf{x} + \beta$.

12. **[Programming]** Let's build a simple classifier using the concepts you've learned in this chapter. We will pretend that we are doing this for classifying or detecting the presence or absence of a disease called - *vector space sickness*. We wish to diagnoze using two clinical tests - (1) *Subspace assay* and (2) *Basis balance scale*. Both these serious clinical tests generate numerical outcomes that can takeon any real number value.

    The department has been conducting a large scale clinical study for the last 5-6 years collecting data from participants, from different background, with and without vector space sickness by administering the subspace assay and the basis balance test. The data from this study is stored as a CSV file with four columns: (a) `subjectno` – subject numbers, (b) `x1` – value of the subspace assay test, (c) `x2` – value of the basis balance scale, and (d) `vss` – presence (1) or absence (0) of the vector space sickness condition. Each row of this CSV file corresponds to a individual subject who participated in the study. Your goal here is to look at the data from this experiment and propose a classifier to determine if a person has vector space sickness if we are given their scores on the subspace assay and the basis balance scale clinical tests. A group of Master's students participated in the study. Unfortunately, almost half of these students were diagnosed with vector space sickness. This data is stored in `expt1.csv`. Read this data, make a scatter plot in 2D (`x1` versus `x2`) with the data points for participants with vector space sickness in blue and the one without in red. Look at the data, and propose how you could use the measurements `x1` and `x2` to distinguish between participants from the two groups. Implement the classifier you've proposed and find out how well it performs in correctly classifying the two groups.

13. **[Programming]** Consider a set of measurements made from adult male subjects, where their height, weight and BMI (body mass index) were recorded and stored as vectors of length three; the first element is the height in cm, second is the weight in Kg, and the last is the BMI. Consider the following four subjects,

$$\mathbf{s}_1 = \begin{bmatrix} 167 \\ 102 \\ 36.6 \end{bmatrix}; \ \mathbf{s}_2 = \begin{bmatrix} 180 \\ 87 \\ 26.9 \end{bmatrix} \ \mathbf{s}_3 = \begin{bmatrix} 177 \\ 78 \\ 24.9 \end{bmatrix}; \ \mathbf{s}_4 = \begin{bmatrix} 152 \\ 76 \\ 32.9 \end{bmatrix}$$

    You can use the distance between these vectors $\|\mathbf{s}_i - \mathbf{s}_j\|_2$ as a measure of the similiarity between the four subjects. Generate a 4×4 table comparing the distance of each subject with respect to another subject; the diagonal elements of this table will be zero, and it will be symmetric about the main diagonal.

    (a) Based on this table, how do the different subjects compare to each other?

    (b) How do the similarities change if the height had been measured in m instead of cm? Can you explain this difference?

    (c) Consider the weighted norm presented in one of the earlier problems.

$$\|x\|_{\mathbf{w}} = \left(w_1 x_1^2 + w_2 x_2^2 + \cdots + w_n x_n^2\right)^{\frac{1}{2}}$$

    Will this fix the problem? What would be a good choice for $\mathbf{w}$ to address the problems with comparing distance between vectors due to unit change?

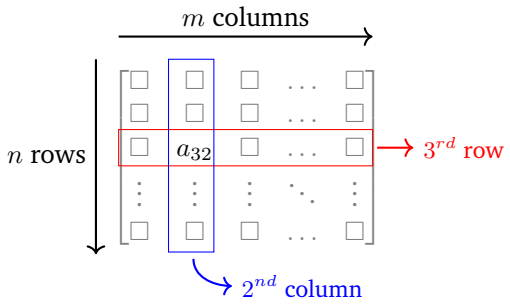
    (d) Can the angle between two vectors be used as a measure of similarity between vectors? Does this suffer from the problem of the 2-norm $\|\cdot\|_2$?

Chapter 2

Matrices

2.1 Matrcies

Matrices are rectangular arrangement of numbers, with a finite number of rows and columns. Matrices are represented as rectangular grid of numbers with n rows and m columns, with the grid contained between two square brackets, as shown below:



The matrix is referred to as an $n \times m$ matrix, where n is the number of rows and m is the number of columns. The numbers in the matrix are called the *elements* of the matrix. The element in the i^{th} row and j^{th} column of the matrix is denoted by a_{ij}. The elements of a matrix with the same row and column index are called the *diagonal elements* of the matrix; these are elements of the form a_{ii}.

We will denote matrices using bold capital letters, and its elements by the corresponding lowercase letter with the row and column indices as subscripts; the row index will be written first, followed by the column index. For example, in the picture above a_{32} is the element corresponding to the element in the 3^{rd} row and 2^{nd} column. This must make you wonder if the n-vectors we talked about earlier are just $n \times 1$ matrices. You are right! We can interpret these as $n \times 1$ matrices. In fact, these single column matrices are also referred to as *column vectors* or a *column matrix*. Now hold on, does that mean that we can also have *row vectors* or *row matrices*? Yes, we do have row vectors, which are $1 \times m$ matrices. We will talk about these in a later section in this chapter. Note that we can also have a 1×1 matrix!

Depending on the number of rows and columns, we group matrices in three categories based on their shape:

- **Square matrix:** A matrix is said to be square if it has the same number of rows and columns, $n = m$. A square matrix is denoted by $n \times n$.
- **Wide/Fat matrix:** A matrix with more columns than rows, $n < m$.
- **Tall/Skinny matrix:** A matrix with more rows than columns, $n > m$.

We will refer to $n \times m$ as the *shape of a matrix* $\mathbf{A}$, where n, m are the number of rows and columns of $\mathbf{A}$, respectively. The set of all $n \times m$ matrices is denoted by the set $\mathbb{R}^{n \times m}$, where $\mathbb{R}$ is the set of real numbers. Later on we will come across matrices where the elements are complex numbers and these would be elements from the set $\mathbb{C}^{n \times m}$.

Block Matrices and Submatrices: We will often also come across matrices where the elements themselves are matrices. These are called *block matrices*. The following is an example of a block matrix $\mathbf{M}$:

$$\mathbf{M} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 \\ \mathbf{A}_3 & \mathbf{A}_4 \end{bmatrix} \tag{2.1}$$

Here, $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, \mathbf{A}_4$ are themselves matrices, which are refered to as the submatrices of $\mathbf{M}$. We cannot have any arbitrary matrices as submatrices of a block matrix. The submatrices must satisfy some constraints.

- The submatrices in a column must have the same number of columns, but have arbitrary number of rows.
- The submatrices in a row must have the same number of rows, but have arbitrary number of columns.

Let the shape of the submatrix $\mathbf{A}_i$ in Eq. 2.1 be $n_i \times m_i$. Then, $n_1 = n_2$, $n_3 = n_4$, $m_1 = m_3$, and $m_2 = m_4$. The shape of the block matrix $\mathbf{M}$ is $(n_1 + n_3) \times (m_1 + m_2)$.

Matrices also are a conveninent way of represent a set of indexed column n-vectors, $\mathbf{x}_1, \mathbf{x}_2, \dots \mathbf{x}_m$. We can treat these columns vectors are $n \times 1$ matrices and form a block matrix as shown below:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_m \end{bmatrix} \tag{2.2}$$

Can I call this matrix a block row matrix? What is the shape of this matrix?

2.1.1 Some special matrices

We will now define some special matrices that we will come across in this course.

- **Zero matrix:** The matrix whose elements are all zeros is called the *zero matrix*. These are often represented by $\mathbf{0}_{n \times m}$ – the matrix of shape $n \times m$ with all elements as zeros.
- **Identity matrix:** The square matrix whose diagonal elements are all ones and all other elements are zeros is called the *identity matrix*. The identity matrix of shape $n \times n$ is denoted by $\mathbf{I}_n$.

  $$\mathbf{I}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \mathbf{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

  Notice that we can also represent identify matrices as a block row matrix of the following form,

  $$\mathbf{I}_n = \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \dots & \mathbf{e}_n \end{bmatrix}$$

  where, $\mathbf{e}_1, \mathbf{e}_2, \dots \mathbf{e}_n$ are the n-unit vectors of $\mathbb{R}^n$.
- **Diagonal matrix:** A square matrix whose non-diagonal elements are all zeros is called a *diagonal matrix*. The diagonal matrix of shape $n \times n$ with diagonal elements $d_1, d_2, \dots d_n$ is denoted by $\mathbf{D}$.

  $$\begin{bmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & d_n \end{bmatrix} = \mathbf{diag}\,(d_1, d_2, \dots d_n)$$

- **Upper triangular matrix:** A square matrix whose elements below the diagonal are all zeros is called an *upper triangular matrix*.
- **Lower triangular matrix:** A square matrix whose elements above the diagonal are all zeros is called a *lower triangular matrix*.
- **Sparse matrix:** A matrix with a large number of zeros is called a *sparse matrix*. We will not come across sparse matrices in this course.

2.1.2 Why do I need to know about matrices?

Matrices are a fundamental concept in linear algebra, and are used in many applications. There are two ways to interpret a matrix – the rectangular arrangment of numbers:

Data representation: Matrices are a convenient way to represent data. Any application where there is a set of parameters are measured multiple times - across space, time, individuals, etc. These are usually represented as a table of entires. A table of entries can be thought of as a matrix. For example,

- The temperature recorded at different locations at different times can be represented as a matrix; the different locations could correspond to the columns and different measurement timepoints could correspond rows.
- The clinical information of patients visiting a hospital can be represented as a matrix. The different patient details could the columns and the rows could correspond to different patients.
- A grayscale image is a matrix, with each element representing the the intensity of a pixel located at a partiular horizontal and vertical position.
- and so on ...

Linear transformations: Matrices are used to represent linear transformations. A linear transformation is a function that maps a vector to another vector, which means these can be use to manipulate vectors. We will have a detailed discussion on this section xxx.

2.2 Matrix operations

We will now define some important matrix operations involving matrices. These are the operations that we will use in this course.

2.2.1 Matrix transpose

This may sound like a strange operation at first, but it turns out be an important operation. The transpose of a matrix is obtained by interchanging the rows and columns of the matrix. The transpose of a matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$ is denoted by $\mathbf{A}^\top$. This matrix is a member of the set $\mathbb{R}^{m\times n}$. The element in the i^{th} row and j^{th} column of the matrix $\mathbf{A}$ is the j^{th} row and i^{th} column of the matrix $\mathbf{A}^\top$. The transpose of a matrix is defined as:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{bmatrix} \longrightarrow \mathbf{A}^\top = \begin{bmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ a_{12} & a_{22} & \dots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1m} & a_{2m} & \dots & a_{nm} \end{bmatrix}$$

Problem 2.1. What is the transpose of the block matrix $\mathbf{M} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 \\ \mathbf{A}_3 & \mathbf{A}_4 \end{bmatrix}$?.

2.2.2 Matrix scalar multiplication

We can also multiply a matrix by a scalar. Given a scalar $c \in \mathbb{R}$ and a matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$, the scalar multiplication operation produces another matrix $c\mathbf{A}$ whose elements are $ca_{11}, ca_{12}, \dots, ca_{nm}$. The scalar multiplication operation is defined as:

$$c\mathbf{A} = c \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{bmatrix} = \begin{bmatrix} ca_{11} & ca_{12} & \dots & ca_{1m} \\ ca_{21} & ca_{22} & \dots & ca_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ ca_{n1} & ca_{n2} & \dots & ca_{nm} \end{bmatrix} \in \mathbb{R}^{n\times m} \tag{2.3}$$

2.2.3 Matrix addition

We can define a matrix addition operation between two matrices. We can add two matrices only if they have the same shape, i.e. they have the same number of rows and same number of columns. Consider two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times m}$. The addition of these matrices is defined as follows:

$$\begin{aligned}\mathbf{A} + \mathbf{B} &= \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \dots & b_{nm} \end{bmatrix} \\ &= \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \dots & a_{1m} + b_{1m} \\ a_{21} + b_{21} & a_{22} + b_{22} & \dots & a_{2m} + b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} + b_{n1} & a_{n2} + b_{n2} & \dots & a_{nm} + b_{nm} \end{bmatrix} \in \mathbb{R}^{n \times m}\end{aligned} \tag{2.4}$$

Its very simple. You simply add the individual elements of the two matrices to get the elements of the resulting matrix. The resulting matrix is also a member of the set $\mathbb{R}^{n \times m}$. Note that this is consistent with the definition of the vector addition operation defined in the previous chapter.

Properties of matrix addition

Here are some of the properties of matrix addition:

- **Commutative property:** $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$.
- **Associative property:** $\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C}$.
- **Distributive property:** $c\,(\mathbf{A} + \mathbf{B}) = c\mathbf{A} + c\mathbf{B}$.
- **Additiob with the zero matrix:** $\mathbf{A} + \mathbf{0}_{n \times m} = \mathbf{A}$.
- **Tranpose of the sum:** $(\mathbf{A} + \mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top$.

We leave it as a exercise for you to prove these properties using the properties of real number addition.

Problem 2.2. The set of matrices $\mathbf{R}^{n \times m}$ for a vector space! The set of $n \times m$ matrices with real numbers is the set $\mathbb{R}^{n \times m}$. We have defined scalar matrix multiplication and matrix addition operations on the elements of this set. Show that the set $\mathbb{R}^{n \times m}$ is a vector space.
Hint: You just need to show the set of closed under scalar multiplication and vector addition. You can also verify that the additional properties are also satisfied by this set.

2.2.4 Matrix multiplication

This is the most important operation involving matrices. Understanding matrix multiplication is vital to understanding the rest of the course material. So the importance of this section cannot be overstated. Unlike matrix scalar multiplication and matrix addition, the concept of matrix multiplication will seem a bit strange at first. But we will see later that the definition of matrix multiplication is a natural when matrices are viewed as reprensenting linear transformations.

Consder two matrices $\mathbf{A} \in \mathbb{R}^{n \times m}$ and $\mathbf{B} \in \mathbb{R}^{p \times q}$. A matric multiplication operation $\mathbf{AB}$ is defined if and only if the number of columns of $\mathbf{A}$ is equal to the number of rows of $\mathbf{B}$, i.e. $m = p$. The result of the matrix multiplication operation is a matrix $\mathbf{C} \in \mathbb{R}^{n \times q}$, where $\mathbf{C} = \mathbf{AB}$. The elements of the resulting matrix $\mathbf{C}$ are defined as:

$$c_{ij} = \sum_{k=1}^{m} a_{ik} b_{kj}, \ \forall i \in \{1, \dots n\}, \ j \in \{1, \dots q\} \tag{2.5}$$

This for sure looks confusing and makes little sense. It turns out the matrix multiplication operation is a natural operation representing composition of linear transformation, and all matrices represent linear transformations.

Hadamard product: Element-wise matrix multiplication

On a separate note, you are probably asking yourself, why not just define matrix multiplication like we defined matrix addition. Just multiply the individual elements together. the element-wise multilication is also a useful operation and is supported by scientific computing programs like MATLAB, Python, etc. This operation is called the *Hadamard product* and is denoted by $\circ$. The Hadamard product of two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times m}$ is defined as:

$$\mathbf{A}\circ\mathbf{B} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{bmatrix} \circ \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \dots & b_{nm} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} & a_{12}b_{12} & \dots & a_{1m}b_{1m} \\ a_{21}b_{21} & a_{22}b_{22} & \dots & a_{2m}b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}b_{n1} & a_{n2}b_{n2} & \dots & a_{nm}b_{nm} \end{bmatrix} \in \mathbb{R}^{n\times m} \tag{2.6}$$

This element-wide multiplication operation is useful when matrices are used to represent data. For example, we might use the Hadamard product when wish to apply a mask to a image represented by the matrix.

We will not discuss the Hadamard product in this course, but it is good to know that this operation exists. We will not discuss the Hadamard product in this course, but it is good to know that this operation exists.

The definition of matrix multiplication in Eq. **??** is hardly illuminating. Its hard to see what is going on. Before we dive a littel deeper into matrix multiplication, we will first define some coventions we will follow in the course:

- All n-vectors are will be column vectors. Row vectors will be represented as the transpose of a column vector.
- For a matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$, the columns of the matrix will be represented as $\mathbf{a}_1, \mathbf{a}_2, \dots \mathbf{a}_m$, where each vector is from $\mathbb{R}^n$. Then, we can write the matrix as the following,

  $$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \dots & \mathbf{a}_m \end{bmatrix} \tag{2.7}$$

  We represent the rows of the matrix $\mathbf{A}$ as $\tilde{\mathbf{a}}_1^\top, \tilde{\mathbf{a}}_2^\top, \dots \tilde{\mathbf{a}}_n^\top$, where each vector $\tilde{\mathbf{a}}_i$ is a column vector from $\mathbb{R}^m$. We will reserve the tilde notation for the rows of a matrix. Then, we can write the matrix as the following,

  $$\mathbf{A} = \begin{bmatrix} \tilde{\mathbf{a}}_1 & \tilde{\mathbf{a}}_2 & \dots & \tilde{\mathbf{a}}_n \end{bmatrix}^\top = \begin{bmatrix} \tilde{\mathbf{a}}_1^\top \\ \tilde{\mathbf{a}}_2^\top \\ \vdots \\ \tilde{\mathbf{a}}_n^\top \end{bmatrix} \tag{2.8}$$

Problem 2.3. Elements of a row and column of a matrix $\mathbf{A}$ Consider a matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$ with elements a_{ij}. Can you write down the elements of the k^{th} row and the l^{th} row of the matrix $\mathbf{A}$?.

We will first start with a simplified versions of the matric multiplication problem before tackling the most general problem. It is left as an exercise for you to verify that the following four simplified version comply with Eq. 2.5.

- **Inner product:** We have already seen this in the previous chapter. Consider two column vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. We can this of these two as $n \times 1$ matrices. The inner product is a product between a row matrix and a column matrix. It is defined as follows,

  $$\mathbf{x}^\top\mathbf{y} = \sum_{i=1}^{n} x_i y_i$$

  Note that the matrices $\mathbf{x}^\top \in \mathbb{R}^{1\times n}$ and $\mathbf{y} \in \mathbb{R}^{n\times 1}$. Thus, matrix multiplication is defined.

- **Post-multiplying a matrix $\mathbf{A}$ by a column vector $\mathbf{x}$:** Consider a matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$ and a column vector $\mathbf{x} \in \mathbb{R}^m$. The product $\mathbf{Ax}$ produces a column vector $\mathbf{y} \in \mathbb{R}^n$.

$$\mathbf{y} = \mathbf{Ax} = \sum_{i=1}^{n} x_i \mathbf{a}_i = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_m\mathbf{a}_m \tag{2.9}$$

The above equation has a nice interpreation. The vector $\mathbf{y}$ is the linear combination of the columns of $\mathbf{A}$, where the mixture for the linear combination come from the elements of the vector $\mathbf{x}$. This means that post-multiplying a matrix $\mathbf{A}$ by a column vector is a process of generating a vector from the span of the set formed by the columns of the matrix $\mathbf{A}$ (recall span of a set of vector from Section 1.8). We will now show how Eq. 2.9 is the consequence of Eq. 2.5. From, Eq. 2.5, we have:

$$\mathbf{y} = \mathbf{Ax} = \begin{bmatrix} \sum_{i=1}^m a_{1i}x_i \\ \sum_{i=1}^m a_{2i}x_i \\ \vdots \\ \sum_{i=1}^m a_{ni}x_i \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1m}x_m \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2m}x_m \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nm}x_m \end{bmatrix} = \begin{bmatrix} a_{11}x_1 \\ a_{21}x_1 \\ \vdots \\ a_{n1}x_1 \end{bmatrix} + \begin{bmatrix} a_{12}x_2 \\ a_{22}x_2 \\ \vdots \\ a_{n2}x_2 \end{bmatrix} + \cdots + \begin{bmatrix} a_{1m}x_m \\ a_{2m}x_m \\ \vdots \\ a_{nm}x_m \end{bmatrix}$$

$$= x_1 \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{bmatrix} + x_2 \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{n2} \end{bmatrix} + \cdots + x_m \begin{bmatrix} a_{1m} \\ a_{2m} \\ \vdots \\ a_{nm} \end{bmatrix} = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_m\mathbf{a}_m$$

- **Pre-multiplying a matrix $\mathbf{A}$ by a row vector $\mathbf{x}^\top$:** Consider a matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$ and a row vector $\mathbf{x}^\top \in \mathbb{R}^{1\times n}$. The product $\mathbf{x}^\top\mathbf{A}$ produces a row vector $\mathbf{y}^\top$, with $\mathbf{y} \in \mathbb{R}^{1\times m}$.

$$\mathbf{y}^\top = \mathbf{x}^\top\mathbf{A} = \sum_{i=1}^{m} x_i \tilde{\mathbf{a}}_i^\top = x_1\tilde{\mathbf{a}}_1^\top + x_2\tilde{\mathbf{a}}_2^\top + \cdots + x_n\tilde{\mathbf{a}}_n^\top \tag{2.10}$$

The row vector $\mathbf{y}^\top$ is a linear combination of the rows of the matrix $\mathbf{A}$, where the miscture fot the linear combination comes from the elements of the row vector $\mathbf{x}^\top$. Pre-multiplying a matrix by a row vector is a process of generating a vector from the span of the set formed by the rows of the matrix $\mathbf{A}$. We will leave it as a exercise for you to show how how Eq. 2.10 is the consequence of Eq. 2.5.

- **Outer product:** The outer product, like the inner product, is a product between a row matrix and a column matrix, except in the reverse order. Consider two column vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. The outer product is defined as follows,

$$\mathbf{xy}^\top = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix} = \begin{bmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_n \\ x_2y_1 & x_2y_2 & \cdots & x_2y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_ny_1 & x_ny_2 & \cdots & x_ny_n \end{bmatrix} \in \mathbb{R}^{n\times n} \tag{2.11}$$

While the inner product produces a scalar, the outer product produces a matrix. The outer product is a very useful operation in many applications, as we will see in the multiple occasions in this course.

Problem 2.4. Outer product to produce a rectangular matrix. In the Eq. 2.11, the outer product operation produces a square matrix. Can the outer product operation ever produce a rectangular matrix of shape $n \times m$. If yes, how? If no, explain why not.

We now look at some numerical examples to get a handle on the ideas discussed above.

Example 2.1. Inner product Let $\mathbf{x} = \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix}$. The inner product of these two vectors or

column matrices is as follows:

$$\mathbf{x}^\top \mathbf{y} = \begin{bmatrix} 1 & 3 & 4 \end{bmatrix} \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix} = 1 \cdot (-1) + 3 \cdot 2 + 4 \cdot 1 = -1 + 6 + 4 = 9$$

Example 2.2. Post-mulitplying a matrix by a column vector Let $\mathbf{x} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$ and $\mathbf{A} = \begin{bmatrix} 2 & -1 \\ 1 & 1 \end{bmatrix}$. The product $\mathbf{Ax}$ is given as follows:

$$\mathbf{y} = \mathbf{Ax} = \begin{bmatrix} 2 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 2 \cdot 1 + -1 \cdot 3 \\ 1 \cdot 1 + 1 \cdot 3 \end{bmatrix} = \begin{bmatrix} -1 \\ 4 \end{bmatrix} = 1 \begin{bmatrix} 2 \\ 1 \end{bmatrix} + 3 \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

Example 2.3. Pre-mulitplying a matrix by a row vector Let $\mathbf{x}^\top = \begin{bmatrix} 1 & 3 \end{bmatrix}$ and $\mathbf{A} = \begin{bmatrix} 2 & -1 \\ 1 & 1 \end{bmatrix}$. The product $\mathbf{Ax}$ is given as follows:

$$\mathbf{y}^\top = \mathbf{x}^\top \mathbf{A} = \begin{bmatrix} 1 & 3 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 \cdot 2 + 3 \cdot 1 & 1 \cdot (-1) + 3 \cdot 1 \end{bmatrix} = \begin{bmatrix} 5 & 2 \end{bmatrix} = 1 \begin{bmatrix} 2 & -1 \end{bmatrix} + 3 \begin{bmatrix} 1 & 1 \end{bmatrix}$$

Example 2.4. Outer product Let $\mathbf{x} = \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix}$. The outer product of these two vectors or column matrices is as follows:

$$\mathbf{x}\mathbf{y}^\top = \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix} \begin{bmatrix} -1 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 1 \cdot (-1) & 1 \cdot 2 & 1 \cdot 1 \\ 3 \cdot (-1) & 3 \cdot 2 & 3 \cdot 1 \\ 4 \cdot (-1) & 4 \cdot 2 & 4 \cdot 1 \end{bmatrix} = \begin{bmatrix} -1 & 2 & 1 \\ -3 & 6 & 3 \\ -4 & 8 & 4 \end{bmatrix}$$

Now we are ready to handle the general matrix multiplication problem. Consider two matrices $\mathbf{A} \in \mathbb{R}^{n \times p}$ and $\mathbf{B} \in \mathbb{R}^{p \times m}$. These two matrices can be multiplied to get a matrix $\mathbf{C} = \mathbf{AB} \in \mathbb{R}^{n \times m}$. We can write the matrices $\mathbf{A}$ and $\mathbf{B}$ as follows:

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \dots & \mathbf{a}_p \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{a}}_1^\top \\ \tilde{\mathbf{a}}_2^\top \\ \vdots \\ \tilde{\mathbf{a}}_n^\top \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} \mathbf{b}_1 & \mathbf{b}_2 & \dots & \mathbf{b}_m \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{b}}_1^\top \\ \tilde{\mathbf{b}}_2^\top \\ \vdots \\ \tilde{\mathbf{b}}_p^\top \end{bmatrix} \tag{2.12}$$

There four ways to interpret the matrix multiplication operation: inner-product interpretation, column interpretation, row interpretation, and the outer-product interpretation.

Inner product interpretation

The ij^{th} element of $\mathbf{C} = \mathbf{AB}$ is given by the inner product of the i^{th} row of $\mathbf{A}$ and the j^{th} column of $\mathbf{B}$.

$$c_{ij} = \tilde{\mathbf{a}}_i^\top \mathbf{b}_j = \sum_{k=1}^{p} a_{ik} b_{kj} \implies \mathbf{C} = \begin{bmatrix} \tilde{\mathbf{a}}_1^\top \mathbf{b}_1 & \tilde{\mathbf{a}}_1^\top \mathbf{b}_2 & \dots & \tilde{\mathbf{a}}_1^\top \mathbf{b}_m \\ \tilde{\mathbf{a}}_2^\top \mathbf{b}_1 & \tilde{\mathbf{a}}_2^\top \mathbf{b}_2 & \dots & \tilde{\mathbf{a}}_2^\top \mathbf{b}_m \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\mathbf{a}}_n^\top \mathbf{b}_1 & \tilde{\mathbf{a}}_n^\top \mathbf{b}_2 & \dots & \tilde{\mathbf{a}}_n^\top \mathbf{b}_m \end{bmatrix} \tag{2.13}$$

In fact, there is beautiful way of looking at this. If we express $\mathbf{A}$ as a block column and $\mathbf{B}$ as a block row, then we have an outer-product between the block column and the block row.

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} \tilde{\mathbf{a}}_1^\top \\ \tilde{\mathbf{a}}_2^\top \\ \vdots \\ \tilde{\mathbf{a}}_n^\top \end{bmatrix} \begin{bmatrix} \mathbf{b}_1 & \mathbf{b}_2 & \dots & \mathbf{b}_m \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{a}}_1^\top \mathbf{b}_1 & \tilde{\mathbf{a}}_1^\top \mathbf{b}_2 & \dots & \tilde{\mathbf{a}}_1^\top \mathbf{b}_m \\ \tilde{\mathbf{a}}_2^\top \mathbf{b}_1 & \tilde{\mathbf{a}}_2^\top \mathbf{b}_2 & \dots & \tilde{\mathbf{a}}_2^\top \mathbf{b}_m \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\mathbf{a}}_n^\top \mathbf{b}_1 & \tilde{\mathbf{a}}_n^\top \mathbf{b}_2 & \dots & \tilde{\mathbf{a}}_n^\top \mathbf{b}_m \end{bmatrix} \tag{2.14}$$

Note that each entry of this block outer-product matrix is an inner product of of two vectors from $\mathbb{R}^p$.

Column interpretation

The columns of the matrix $\mathbf{C} = \mathbf{AB}$ are the linear combinations of the columns of the matrix $\mathbf{A}$, where the mixture for the linear combination come from the columns of the matrix $\mathbf{B}$.

$$\mathbf{c}_i = \mathbf{A}\mathbf{b}_i \implies \begin{bmatrix} \mathbf{c}_1 & \mathbf{c}_2 & \cdots & \mathbf{c}_m \end{bmatrix} = \mathbf{A} \begin{bmatrix} \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_m \end{bmatrix} = \begin{bmatrix} \mathbf{A}\mathbf{b}_1 & \mathbf{A}\mathbf{b}_2 & \cdots & \mathbf{A}\mathbf{b}_m \end{bmatrix} \tag{2.15}$$

Row interpretation

The rows of the matrix $\mathbf{C} = \mathbf{AB}$ are the linear combinations of the rows of the matrix $\mathbf{B}$, where the mixture for the linear combination come from the rows of the matrix $\mathbf{A}$.

$$\tilde{\mathbf{c}}_i^\top = \mathbf{a}_i^\top \mathbf{B} \implies \begin{bmatrix} \tilde{\mathbf{c}}_1^\top \\ \tilde{\mathbf{c}}_2^\top \\ \vdots \\ \tilde{\mathbf{c}}_n^\top \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1^\top \\ \mathbf{a}_2^\top \\ \vdots \\ \mathbf{a}_n^\top \end{bmatrix} \mathbf{B} = \begin{bmatrix} \mathbf{a}_1^\top \mathbf{B} \\ \mathbf{a}_2^\top \mathbf{B} \\ \vdots \\ \mathbf{a}_n^\top \mathbf{B} \end{bmatrix} \tag{2.16}$$

Outer product interpretation

The outer-product view is complementary to the inner-product view. If we express the matrices $\mathbf{A}$ and $\mathbf{B}$ as block column and block row matrix, then the matrix $\mathbf{C} = \mathbf{AB}$ is the outer-product of the block column and the block row.

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \dots & \mathbf{a}_p \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{b}}_1^\top \\ \tilde{\mathbf{b}}_2^\top \\ \dots \\ \tilde{\mathbf{b}}_p^\top \end{bmatrix} = \mathbf{a}_1 \tilde{\mathbf{b}}_1^\top + \mathbf{a}_2 \tilde{\mathbf{b}}_2^\top + \cdots + \mathbf{a}_p \tilde{\mathbf{b}}_p^\top \tag{2.17}$$

In this view, we see that matrix mulitplication can be viewed the inner-product of a block row matrix with a block column matrix, while the individual elements of the inner-product sum are the outer-products.

Each of these four views will be useful under different circumstances to help interpret the meaning of the matrix mulitplication operation. We will now look at some numerical examples for the four views to get comfortable with them.

Example 2.5. Four interpretations of matrix multiplication Let $\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}$. The matrix multiplication $\mathbf{C} = \mathbf{AB}$ is given as follows:

- **Inner product interpretation:**

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 5 \\ 7 \end{bmatrix} & \begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 6 \\ 8 \end{bmatrix} \\ \begin{bmatrix} 3 & 4 \end{bmatrix} \begin{bmatrix} 5 \\ 7 \end{bmatrix} & \begin{bmatrix} 3 & 4 \end{bmatrix} \begin{bmatrix} 6 \\ 8 \end{bmatrix} \end{bmatrix} = \begin{bmatrix} 1 \cdot 5 + 2 \cdot 7 & 1 \cdot 6 + 2 \cdot 8 \\ 3 \cdot 5 + 4 \cdot 7 & 3 \cdot 6 + 4 \cdot 8 \end{bmatrix} = \begin{bmatrix} 19 & 22 \\ 43 & 50 \end{bmatrix}$$

- **Column interpretation:**

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = \begin{bmatrix} 5\begin{bmatrix} 1 \\ 3 \end{bmatrix} + 7\begin{bmatrix} 2 \\ 4 \end{bmatrix} & 6\begin{bmatrix} 1 \\ 3 \end{bmatrix} + 8\begin{bmatrix} 2 \\ 4 \end{bmatrix} \end{bmatrix} = \begin{bmatrix} 19 & 22 \\ 43 & 50 \end{bmatrix}$$

- **Row interpretation**

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = \begin{bmatrix} 1\begin{bmatrix} 5 & 6 \end{bmatrix} + 2\begin{bmatrix} 7 & 8 \end{bmatrix} \\ 3\begin{bmatrix} 5 & 6 \end{bmatrix} + 4\begin{bmatrix} 7 & 8 \end{bmatrix} \end{bmatrix} = \begin{bmatrix} 19 & 22 \\ 43 & 50 \end{bmatrix}$$

- **Outer product interpretation**

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \begin{bmatrix} 5 & 6 \end{bmatrix} + \begin{bmatrix} 2 \\ 4 \end{bmatrix} \begin{bmatrix} 7 & 8 \end{bmatrix} = \begin{bmatrix} 5 & 6 \\ 15 & 18 \end{bmatrix} + \begin{bmatrix} 14 & 16 \\ 28 & 32 \end{bmatrix} = \begin{bmatrix} 19 & 22 \\ 43 & 50 \end{bmatrix}$$

Properties of matrix multiplication

We will not demonstrate any of these properties, but we will list them here for your reference. It is left as an exercise for you to demonstrate these properties. Here are some of the properties of matrix multiplication:

- **Non-commutative property:** $\mathbf{AB} \neq \mathbf{BA}$. You could simply use the matrices in Example 2.5 to verify that this is true. Can give an example of two matrices where their product does commute?.
- **Dsitributive**: $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$ and $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$.
- **Associative**: $\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$
- **Transpose of the product:** $(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top$.
- **Identity matrix:** $\mathbf{I}_n \mathbf{A} = \mathbf{A}\mathbf{I}_m = \mathbf{A}$
- **Scalar multiplication:** $(\alpha \mathbf{A})\mathbf{B} = \alpha(\mathbf{AB}) = \mathbf{A}(\alpha \mathbf{B})$

2.3 Rank of a matrix

The rank of a matrix is a very important concept in linear algebra. There are different ways of defining it, but we will use a geometric definition that is easy to wrap our head around.

The *rank of a matrix* $\mathbf{A} \in \mathbb{R}^{n \times m}$ is the dimension of the subspace spanned by the set of columns of $\mathbf{A}$ or the set of rows of $\mathbf{A}$.

$$\begin{aligned} \text{rank}(\mathbf{A}) &= \dim(\text{span}(\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m\})) \longrightarrow \text{Column rank} \\ &= \dim(\text{span}(\{\tilde{\mathbf{a}}_1, \tilde{\mathbf{a}}_2, \ldots, \tilde{\mathbf{a}}_n\})) \longrightarrow \text{Row rank} \\ &\leq \min(n, m) \end{aligned} \tag{2.18}$$

This is also equivalent to saying it is the smallest number of columns or rows of $\mathbf{A}$ that form a linearly independent set; basically, the number of linearly independent columns or rows of the matrix $\mathbf{A}$. A matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ is said to be a **full rank matrix** if $\text{rank}(\mathbf{A}) = \min(n, m)$. A matrix is said to be **rank deficient** if $\text{rank}(\mathbf{A}) < \min(n, m)$.

2.4 Matrix inverse

We will talk about matrix inverses in more detail in Chapter xxx, but we will conclude this chapter with a provide brief discussion of this concept. Consider a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ that is full rank (if so, what is the rank of $\mathbf{A}$?). A matrix $\mathbf{B} \in \mathbb{R}^{n \times n}$ is called the *inverse* of the matrix $\mathbf{A}$ if and only if $\mathbf{AB} = \mathbf{BA} = \mathbf{I}_n$. This matrix $\mathbf{B}$ is denoted by $\mathbf{A}^{-1}$.

All full rank square matrices have a unique inverse. These matrices are said to be *invertible* or *non-singular*, which rank deficient square matrices are said to *non-invertible* or *singular*. For a non-singular matrix $\mathbf{A}$, $\mathbf{A}$ and $\mathbf{A}^{-1}$ are inverses of each other. Because both pre- and post-multiplying $\mathbf{A}^{-1}$ with $\mathbf{A}$ gives us the identity matrix, we say $\mathbf{A}^{-1}$ is both the left and right inverse of $\mathbf{A}$. We will see in Chapter xxx that full rectangular matrices can either have non-unique left or right inverses, and not both.

The inverse of matrix follows the following two properites, which you are en encouraged to demonstrate:

- $\left(\mathbf{A}^{-1}\right)^{-1} = \mathbf{A}$
- $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$
- $\left(\mathbf{A}^{-1}\right)^{\top} = \left(\mathbf{A}^{\top}\right)^{-1}$

2.5 Exercise

1. Elements of the matrix $\mathbf{C} \in \mathbb{R}^{m\times n}$ obtained as the product of two matrices $\mathbf{A} \in \mathbb{R}^{m\times p}$ and $\mathbf{B} \in \mathbb{R}^{p\times n}$ is given by,
$$c_{ij} = \sum_{k=1}^{p} a_{ik}b_{kj}$$
We had discussed four different ways to think of matrix multiplication. By algebraically manipulating the previous equation, arrive at these four views (inner product view, column view, row view and outer product view)?
2. Show that $(\mathbf{AB})^{\top} = \mathbf{B}^{\top}\mathbf{A}^{\top}$.
3. Derive the inverse of the matrix $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$.
4. Prove that the rank of an outer product $\mathbf{xy}^{\top}$ is 1, where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\mathbf{x}, \mathbf{y} \neq \mathbf{0}$. **[Marks: 1]**
5. For a $n \times n$ square matrix $\mathbf{A}$, prove that if $\mathbf{AX} = \mathbf{I}$, then $\mathbf{XA} = \mathbf{I}$ and $\mathbf{X} = \mathbf{A}^{-1}$.
6. Prove the following for the non-singular square matrices $\mathbf{A}$ and $\mathbf{B}$:
   (a) $\mathbf{AB}$ is non-singular.
   (b) $\left(\mathbf{A}^{-1}\right)^{-1} = \mathbf{A}$.
   (c) $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$
   (d) $\left(\mathbf{A}^{\top}\right)^{-1} = \left(\mathbf{A}^{-1}\right)^{\top}$
7. **Computational cost of different operations.** What is the computational cost of the following matrix operations? Computational cost refers to the number of arithmetic operations required to carry out a particular matrix operation. Computational cost is a measure of the efficiency of an algorithm. For example, consider the operation of vector addition, $\mathbf{a} + \mathbf{b}$, where $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$. This requires n addition/-subtraction operations and zero multiplication/division operations.
   (a) Matrix multiplication: $\mathbf{AB}$, where $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$
   (b) Inner product: $\mathbf{u}^{\top}\mathbf{v}$

   Report the counts for the addition/subtraction and multiplication/division operations separately.
8. Consider the following matrix,
$$\mathbf{A} = \begin{bmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{bmatrix} \begin{bmatrix} 0.1 & 0 \\ 0 & 0.9 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{3}}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{\sqrt{3}}{2} \end{bmatrix}$$
Find out the expression for $\mathbf{A}_n = \mathbf{A}^n$. What is $\mathbf{A}_\infty = \lim_{n\to\infty} \mathbf{A}^n$?

9. Prove that a matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$ can always be written as a sum of a symmetric matrix $\mathbf{S}$ and a skew-symmetric matrix $\mathbf{A}$.

$$\mathbf{M} = \mathbf{S} + \mathbf{A}, \;\; \mathbf{S}^\top = \mathbf{S} \text{ and } \mathbf{A}^\top = -\mathbf{A}$$

10. The trace of a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is defined as, $trace\,(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$. Prove the following,

    (a) $trace\,(\mathbf{A})$ is a linear function of $\mathbf{A}$.

    (b) $trace\,(\mathbf{AB}) = trace\,(\mathbf{BA})$

    (c) $trace\left(\mathbf{A}^\top \mathbf{A}\right) = 0 \implies \mathbf{A} = 0$

11. Consider upper triangular and lower triangular matrices $\mathbf{U}$ and $\mathbf{L}$, respectively.

    (a) Is the product of two upper triangular matrices $\mathbf{U}_1\mathbf{U}_2$ upper triangular?

    (b) Is the product of two lower triangular matrices $\mathbf{L}_1\mathbf{L}_2$ upper triangular?

    (c) What is the $trace\,(\mathbf{LU})$?

12. Consider the following upper-triangular matrix,

$$\mathbf{U} = \begin{bmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{1n} \\ 0 & u_{22} & u_{23} & \cdots & u_{2n} \\ 0 & 0 & u_{33} & \cdots & u_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & u_{nn} \end{bmatrix}$$

    where, $u_{ii} \neq 0, \;\; 1 \leq i \leq n$. Do the columns of this matrix form a linearly independent set? Explain your answer.

13. Verify that $\mathbf{A}$ and $\mathbf{B}$ are inverses of each other,

    (a) $\mathbf{A} = \mathbf{I} - \mathbf{uv}^\top$ and $\mathbf{B} = \mathbf{I} + \mathbf{uv}^\top / \left(1 - \mathbf{v}^\top \mathbf{u}\right)$

    (b) $\mathbf{A} = \mathbf{C} - \mathbf{uv}^\top$ and $\mathbf{B} = \mathbf{C}^{-1} + \mathbf{C}^{-1}\mathbf{uv}^\top\mathbf{C}^{-1} / \left(1 - \mathbf{v}^\top \mathbf{C}^{-1}\mathbf{u}\right)$

    (c) $\mathbf{A} = \mathbf{I} - \mathbf{UV}$ and $\mathbf{B} = \mathbf{I}_n + \mathbf{U}\left(\mathbf{I}_m - \mathbf{VU}\right)^{-1}\mathbf{V}$

    (d) $\mathbf{A} = \mathbf{C} - \mathbf{UD}^{-1}\mathbf{V}$ and $\mathbf{B} = \mathbf{C}^{-1} + \mathbf{C}^{-1}\mathbf{U}\left(\mathbf{D} - \mathbf{VC}^{-1}\mathbf{U}\right)^{-1}\mathbf{VC}^{-1}$

    where, $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{n \times n}$, $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, $\mathbf{U} \in \mathbb{R}^{n \times m}$, $\mathbf{V} \in \mathbb{R}^{m \times n}$ and $\mathbf{D} \in \mathbb{R}^{m \times m}$.

Chapter 3

Linear Transformations

"Though a bit of an exaggeration, it can be said that a mathematical problem can be solved only if it can be reduced to a calculation in linear algebra. And a calculation in linear algebra will reduce ultimately to the solving of a system of linear equations, which in turn comes down to the manipulation of matrices."

Thomas A Garrity in *All the Mathematics You Missed.*

I concede that the first two chapters were a bit dry. Hopefully, you will find topics from now on a bit more interesting. Linear algebra is the study of linear transformations. We already saw an example of a linear transformation in Section 1.15. We will look at linear transformations in in their general form and see how matrices can be used to represent and understand them.

3.1 What is a linear transformation?

A *linear transformation* or *linear map* T is a function between two vector spaces that satisfies the homgeneity (scaling) and additivity properties. In this course, we will particularly be interested in linear transformation from $\mathbb{R}^m$ to $\mathbb{R}^n$, i.e. $T : \mathbb{R}^m \to \mathbb{R}^n$. These satisfy the following two properties:

1. **Additivity:** For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$, $T(\mathbf{x} + \mathbf{y}) = T(\mathbf{x}) + T(\mathbf{y})$.
2. **Homogeneity:** For all $\mathbf{x} \in \mathbb{R}^m$ and $c \in \mathbb{R}$, $T(c\mathbf{x}) = cT(\mathbf{x})$.

Note that linearity property allow us to move the transformation operation into the paranthesis for the addition or scalar multiplication of the arguments. The above properties also imply the following:

- **Superposition:** For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$ and $c, d \in \mathbb{R}$, $T(c\mathbf{x} + d\mathbf{y}) = T(c\mathbf{x}) + T(d\mathbf{y}) = cT(\mathbf{x}) + dT(\mathbf{y})$.

  Linear combination of two vectors $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^m$ in the transformations arugment results in the same linear combination of the individual transformed vectors $T(\mathbf{x}_1), T(\mathbf{x}_2) \in \mathbb{R}^n$.
- **Zero input:** $T(\mathbf{0}_m) = \mathbf{0}_n$, where $\mathbf{0}_m \in \mathbb{R}^m$ and $\mathbf{0}_n \in \mathbb{R}^n$.

  The zero vector from $\mathbb{R}^m$ maps to the zero vector from $\mathbb{R}^m$ under any linear transformation.

The linear functions we considered in Section 1.15 are special cases of linear transformations, where $n = 1$. We will now look at some examples of transformations that are linear and some that are not.

Example 3.1. Consider the transformation $T : \mathbb{R}^2 \to \mathbb{R}^2$ defined by $T\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} 2x_1 \\ 3x_2 \end{bmatrix}$. This transformation is linear because it satisfies the properties of additivity and homogeneity. For example, $T\left(\alpha \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \beta \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}\right) = T\left(\begin{bmatrix} \alpha x_1 + \beta y_1 \\ \alpha x_2 + \beta y_2 \end{bmatrix}\right) = \begin{bmatrix} 2(\alpha x_1 + \beta y_1) \\ 3(\alpha x_2 + \beta y_2) \end{bmatrix} = \alpha \begin{bmatrix} 2x_1 \\ 3x_2 \end{bmatrix} + \beta \begin{bmatrix} 2y_1 \\ 3y_2 \end{bmatrix} = \alpha T\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) + \beta T\left(\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}\right)$.

Example 3.2. Consider the transformation $T : \mathbb{R}^2 \to \mathbb{R}^2$ defined by $T\left(\begin{bmatrix} x_1 \\ y \end{bmatrix}\right) = \begin{bmatrix} x^2 \\ y^2 \end{bmatrix}$. This transformation is not linear because it does not satisfy the property of homogeneity. For example, $T\left(2\begin{bmatrix} 1 \\ 2 \end{bmatrix}\right) = T\left(\begin{bmatrix} 2 \\ 4 \end{bmatrix}\right) = \begin{bmatrix} 4 \\ 16 \end{bmatrix} \neq 2\begin{bmatrix} 1 \\ 4 \end{bmatrix} = 2T\left(\begin{bmatrix} 1 \\ 2 \end{bmatrix}\right)$.

Now you can try your hand at the following exercises. Note that to show $T : \mathbb{R}^m \to \mathbb{R}^n$ is not a linear transformation, you only need to show an example that does not satisfy additivity and homogeneity. However, to show that a transformation is linear, you need to demonstrate that T satisfies both properties for all possible vectors in $\mathbb{R}^m$. The proof of linearity can be a bit more involved than showing non-linearity.

Problem 3.1. Which of the following transformations are linear?

1. $T : \mathbb{R}^2 \to \mathbb{R}^2$, $T\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} x_1 + x_2 \\ x_1 - x_2 \end{bmatrix}$
2. $T : \mathbb{R}^2 \to \mathbb{R}^3$, $T\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \begin{bmatrix} x_1 + 1 \\ x_2 + 2 \\ 3x_1 + 2x_2 - 1 \end{bmatrix}$
3. $T : \mathbb{R}^3 \to \mathbb{R}^2$, $T\left(\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}\right) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

3.2 Matrices represent linear transformations

In Section 1.15, we saw that any linear function $f : \mathbb{R}^m \to \mathbb{R}$ can be represented by a fixed $\mathbf{w} \in \mathbb{R}^m$, and $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$. We can extend this idea to linear transformations from $\mathbb{R}^m$ to $\mathbb{R}^n$. Any linear transformation $T : \mathbb{R}^m \to \mathbb{R}^n$ can be represented by a fixed $n \times m$ matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$, such that $T(\mathbf{x}) = \mathbf{A}\mathbf{x}$; $\mathbf{A}$ is the *matrix representation* of the linear transformation T.

$$\begin{aligned} \mathbf{y} = T(\mathbf{x}) = \mathbf{A}\mathbf{x} = & \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \\ \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = & \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1m}x_m \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2m}x_m \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nm}x_m \end{bmatrix} \end{aligned} \tag{3.1}$$

This shows that all linear transformation are essentially a set of simultaneous linear equations, where $y_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{im}x_m$, $1 \leq i \leq n$.

Let's assume that I give you a Python/Julia function, which takes in vectors from $\mathbb{R}^m$, does some computation, and returns vectors from $\mathbb{R}^n$. I also tell to you that this function is linear. You do not want to use this function. If you knew the matrix $\mathbf{A}$ corresponding to this linear transformation, then you can compute the function yourself. But, how can you find the matrix corresponding this function? Remember how how we identified the entries of $\mathbf{w} \in \mathbb{R}^m$ (Section 1.15) in the case of linear functions? We had used the unit vectors of $\mathbb{R}^m$ to get the w_is. Turns out, we do the exact same thing, and the elements of $\mathbf{A}$ will be revealed to us by the output of the linear transformation to the m unit vectors of $\mathbb{R}^m$. If $\mathbf{a}_i$ is the i^{th} column of $\mathbf{A}$, then $\mathbf{a}_i = T(\mathbf{e}_i)$, $\mathbf{e}_i \in \mathbb{R}^m$. Can you explain why?

3.3 Matrix multiplication and linear transformations

Consider the following two transformations T_A and T_B:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = T_A(\mathbf{x}) = \mathbf{A}\mathbf{x} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{21}x_1 + a_{22}x_2 \end{bmatrix}$$
$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = T_B(\mathbf{u}) = \mathbf{B}\mathbf{u} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} b_{11}u_1 + b_{12}u_2 \\ b_{21}u_1 + b_{22}u_2 \end{bmatrix}$$

Let $\mathbf{u} \in \mathbb{R}^2$ and $\mathbf{v} = T_C(\mathbf{u}) = T_A \circ T_B(\mathbf{u}) = T_A(T_B(\mathbf{u}))$. We can write this as:

$$\mathbf{v} = T_C(\mathbf{u}) = T_A(T_B(\mathbf{u})) = T_A(\mathbf{B}\mathbf{u}) = \mathbf{A}\mathbf{B}\mathbf{u}$$

This implies that the matrix $\mathbf{AB}$ corresponds to the transformation T_C. This is an important result. Matrix multiplication can be seen as the process of composing two linear transformations, i.e. applying two linear transoformation one after the other on a vector $\mathbf{x}$ (from right to left).

$$\begin{aligned}\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = T_A(T_B(\mathbf{u})) = T_A\left(\begin{bmatrix} b_{11}u_1 + b_{21}u_2 \\ b_{21}u_1 + b_{22}u_2 \end{bmatrix}\right) &= \begin{bmatrix} a_{11}b_{11}u_1 + a_{11}b_{21}u_2 + a_{12}b_{21}u_1 + a_{12}b_{22}u_2 \\ a_{21}b_{11}u_1 + a_{21}b_{21}u_2 + a_{22}b_{21}u_1 + a_{22}b_{22}u_2 \end{bmatrix} \\ = \begin{bmatrix} (a_{11}b_{11} + a_{12}b_{21})u_1 + (a_{11}b_{21} + a_{12}b_{22})u_2 \\ (a_{21}b_{11} + a_{22}b_{21})u_1 + (a_{21}b_{21} + a_{22}b_{22})u_2 \end{bmatrix} &= \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{11}b_{21} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} + a_{21}b_{21} + a_{22}b_{22} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}\end{aligned}$$

The above equation shows how the strange definition of matrix multiplication comes out beautifully from the composition of linear transformations. Although, we have taken a simple example of 2×2 matrices, the idea extends to matrices of any size, as long as the dimensions are compatible for multiplication.

Problem 3.2. Consider the matrix multiplication, $\mathbf{C} = \mathbf{AB}$ with $\mathbf{A} \in \mathbb{R}^{n \times p}$ and $\mathbf{B} \in \mathbb{R}^{p \times m}$. We know that this operation can be seen as the composition of two linear transformations. Using the idea of composition of linear transformations, show that the the i^{th} column of the matrix $\mathbf{C}$ is given by the linear combination of the columns of $\mathbf{A}$ with the coefficients from the i^{th} column of $\mathbf{B}$.

3.4 System of linear equations

An important use of matrices and linear transformations is in solving systems of linear equations. Consider the following system of linear equations which we have seen from our high school years,

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1m}x_m &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2m}x_m &= b_2 \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nm}x_m &= b_n \end{aligned} \tag{3.2}$$

Here, the coefficients a_{ij} and b_i are known, and we are interested in solving for the unknowns x_j. We know how to solve these equations and also know the geometric associated with these equations. Each row represents a hyperplane in $\mathbb{R}^m$ and the solution, if it exists, is the set of all points in the intersection of the n planes in $\mathbb{R}^m$.

We can write the above system of linear equations in the matrix form as the following,

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1m}x_m &= b_1 \\ &\vdots \\ a_{n1}x_1 + \cdots + a_{nm}x_m &= b_n \end{aligned} \qquad \longrightarrow \mathbf{A}\mathbf{x} = \mathbf{b}, \quad \mathbf{A} \in \mathbb{R}^{\mathbf{n} \times \mathbf{m}}, \mathbf{x} \in \mathbb{R}^m, \mathbf{b} \in \mathbb{R}^n \tag{3.3}$$

3.4.1 Geometry interpretation of linear equations

The matrix representation offers a new geometrical view of the system of linear equations in Eq 3.2. The above geometrical interpreation of viewing each equation as a hyperplane in $\mathbb{R}^m$ is referred to as the *row view*. The other geomtric view is the *column view*, which can be easily seen from the following way of writing Eq. 3.2.

$$\mathbf{Ax} = \mathbf{b} \longrightarrow x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_m\mathbf{a}_m = \mathbf{b} \tag{3.4}$$

The vector $\mathbf{b}$ is a linear combination of the columns of $\mathbf{A}$, where the m columns $\mathbf{a}_i$ and $\mathbf{b}$ are know, and the unknows x_i are the mixture of the linear combination that we are interested in determining. The row and column view of a system of two linear equations with two unknowns are depicted in Figure 3.1 and Figure 3.2, respectively.

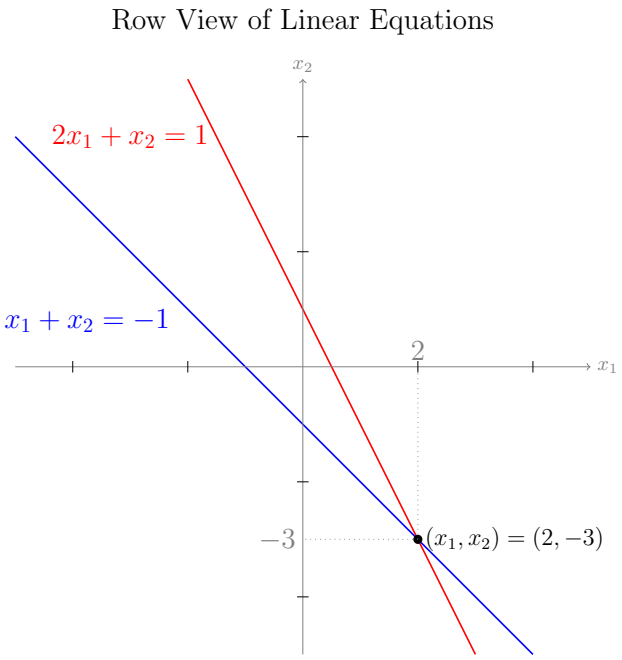


Figure 3.1: The row view of a system of two linear equations with two unknowns. The individual equations and lines are depicted in red and blue color in the plot. The intersection of these two lines is the solution to the problem, which is depicted by the black circle at $(2, -3)$.

The row view for a system of n linear equations with m unknowns is is a set of n is depicted in $\mathbb{R}^m$, while the column view is depicted in $\mathbb{R}^n$. We will almost exclusively only make use of the column view for understanding linear equations and their solutions, as this view provides a more intuitive understanding than the row view. The row view is useful for when dealing with two unknowns, but becomes complicated for system of equations with more than two unknowns.

We will now look at two types of applications where we come across linear equations: control problems and estimation problems. The equations have the same form in both cases, its our interpretation of ther different terms that changes.

3.4.2 Linear equations in control problems

Control problems are ones where the unknown $\mathbf{x}$ is the input to a system whose output is $\mathbf{b}$, and the matrix $\mathbf{A}$ is the system matrix mapping the inputs to the outputs. In control problems of the form $\mathbf{Ax} = \mathbf{b}$, the system matrix is known, and we wish to produce a desired output $\mathbf{b}$. We would like to find out the input $\mathbf{x}$ that produces the output for this system. The goal of solving $\mathbf{Ax} = \mathbf{b}$ in control problems, are to then apply these inputs to the real system to have it produce the output we desire. In such control problems, we ideally would like to have more inputs than outputs, which gives us more flexibility in controlling the system's output. This corresponds to system matrix is fat in this case ($n > m$); when A is full rank, there are infinitely many inputs that produce the same output.

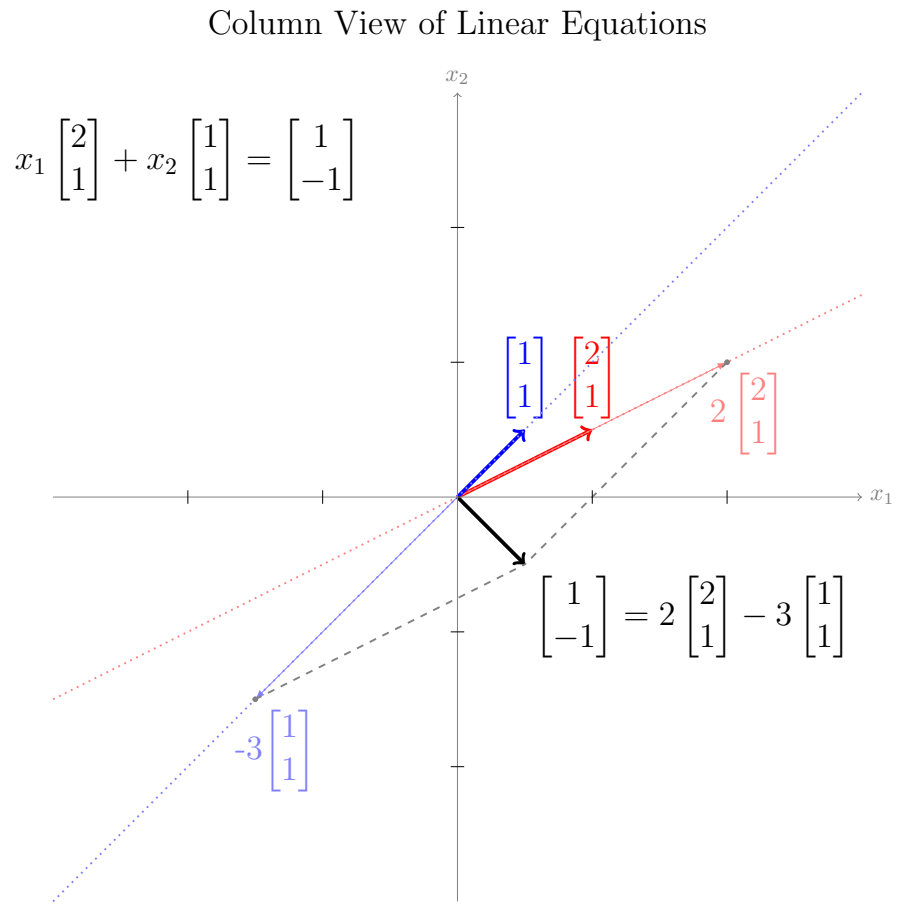


Figure 3.2: The set of all real numbers with magnitude 1. This set contains two numbers $\{-1, 1\}$.

Examples of control problems

Example 3.3. Radiotherapy. Radiotheray or radiation therapy is a medical procedure where high-energy radiation is used to target and destroy cancer cells. Radiation therapy involves a radiation planning step, where the appropriate radiation dose is determined for delivery the desired dose to the cancers cells. The dose plannign step ca be formulated as a $\mathbf{Ax} = \mathbf{b}$ problem, where $\mathbf{A}$ is the radiation dose distribution matrix, $\mathbf{b}$ is the vector of desired radiation dose to the cancer cells, and $\mathbf{x}$ is the vector of unknown radiation dose to be delivered to the patient. We note that this descritpion is a simplification of the actual radiation therapy process. We will look at a more realistic description in Chapter 7 on Linear Programming.

Example 3.4. Robotics: In robotics, we can control the motion and interaction of a robot by controlling the input to the actuators of the robot. Consider a robot with a serial, open kinematic chain consisting of m joints (with one actuator per joint) which control the 3D position of its end-effector. The relationship between the joint angles and the end-effector position is non-linear. However, for a given joint configuration (i.e., a fixed set of values for the joint angles), the relationship between the joint velocities $\boldsymbol{\omega} \in \mathbb{R}^m$ and the end-point velocity $\mathbf{v} \in \mathbb{R}^3$ is linear, which are related through the robot's *Jacobian* matrix $\mathbf{J} \in \mathbb{R}^{3\times m}$.

$$\mathbf{v} = \mathbf{J}\boldsymbol{\omega}$$

Given the desired end-effector velocity $\mathbf{v} \in \mathbb{R}^3$ for a given joint configuration (i.e., know Jacobian matrix $\mathbf{J} \in \mathbb{R}^{3\times m}$), the control problem is to determine the unknown joint velocities $\boldsymbol{\omega} \in \mathbb{R}^m$ that will produce the desired end-effector velocity.

3.4.3 Linear equations in estimation problems

Estimation problems are ones where the unknown $\mathbf{x}$ is the set of parameter associated with a system that cannot be directly observed or measured, adn $\mathbf{b}$ is the of measureable quantities of the system, and the matrix $\mathbf{A}$ is the system matrix that determines how the parameters parameters influnce the system's measurable quantities. In estimation problems we are interested in solving $\mathbf{Ax} = \mathbf{b}$, because want to the know the internal parmeters $\mathbf{x}$ of the system that are responsible for generating the observed quantities $\mathbf{b}$. Note that unlike the control problems, we cannot modify that system parameters $\mathbf{x}$; in these problem solving for $\mathbf{x}$ is a way of knowing something about the system that we cannot directly observe. In these problems, the system matrix and measurements are known, and we ideally would like to have more measurements than the number of parameters associated with the system, as this allow redundancy in the measurements to account for noise in our measurements and obtain more accurate estimates of the system parameters. This corresponds to system matrix is tall in this case ($n > m$); when A is full rank, then the outputs are forces to be constrained to lie in a smaller subspace of dimension m in $\mathbb{R}^n$; these constrains are what help us beat noise when we have multiple measurements. We will discuss this in more detail in Chapter 6 on Statistical Estimation.

Examples of estimation problems

Example 3.5. Computed tomography (CT). CT is a medical imagining modality of creating cross-sectional images of the body using x-rays. X-ray passed through a cross-section of the body, and received by detectors on the other side. The internal distribution of the absoptivity of the body tissue is estimated from measurements made by shooting x-ray through the body from different angles. This can be fomulated as a $\mathbf{Ax} = \mathbf{b}$ problem, which is explained in a bit more details in the Exercises section of this chapter.

Example 3.6. System identification of linear dynamical systems: The process of estimating the parameters of a linear dynamical system from records of its input and output data is referred to as *system identification*. This discussed in detail in the Applications section (Section 3.8) of this chapter.

3.5 Solutions of linear equations

We will not look at the procedure for solving a system of linear equations, but rather understand the different possible solution(s) to a system of linear equations, and how the properties of the matrix $\mathbf{A}$ and the vector $\mathbf{b}$ determine the solution(s) of $\mathbf{Ax} = \mathbf{b}$. A system of linear equations can be one of the following three types: (a) unique solution, (b) infinitely many solution, and (c) no solution. We will now look at the conditions under which each of these types of solutions occur.

3.5.1 Unique solution

A unique solution is possible under two conditions:

- **Square, full rank** $\mathbf{A} \in \mathbb{R}^{n \times n}$: When $\mathbf{A}$ is square and full rank, then the span of the columns of $\mathbf{A}$ is $\mathbb{R}^n$ and are linearly independent. This implies that the columns of $\mathbf{A}$ form a basis for $\mathbb{R}^n$. Thus, any vector in $\mathbb{R}^n$ can be represents as a linear combination of the columns of $\mathbf{A}$ and in a unique manner. Thus, there is a unique solution to the system of linear equations.

- **Tall, full rank** $\mathbf{A} \in \mathbb{R}^{n \times m}$ and $\mathbf{b} \in$ **span** $\{\mathbf{a}_1, \mathbf{a}_2, \cdots \mathbf{a}_m\}$: When a tall matrix $\mathbf{A}$ is full rank, then rank $(\mathbf{A}) = m$. The columns of $\mathbf{A}$ are linearly independent and span a subspace $\mathcal{S}$ of $\mathbb{R}^n$ of dimension m. The columns of $\mathbf{A}$ also form basis for the subspace $\mathcal{S}$. Thus, if $\mathbf{b}$ is in $\mathcal{S}$, then a unique linear combination of the columns of $\mathbf{A}$ would produce $\mathbf{b}$. But what happens when $\mathbf{b} \notin \mathcal{S}$, then what?

The geometry of the unique solution is depicted in Figure xxx.

3.5.2 Infinitely many solutions

3.5.3 No solution

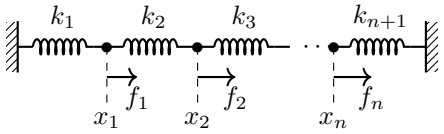
3.6 Revisiting linear independence

3.7 Four fundamental subspaces of A

3.8 Applications

3.9 Exercise

1. Derive force and displacement relationship for a series of $n+1$ springs (with spring constants k_i) connected in a line. There are n nodes, with f_i and x_i representing the force applied and resulting displacement at the i^{th} node.



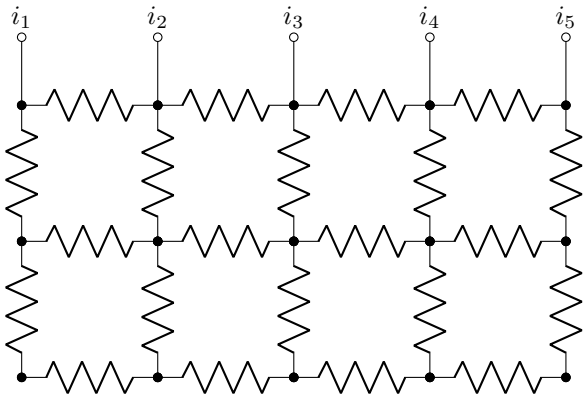
   (a) Represent the relationship in the following form,

$$\mathbf{f} = \mathbf{K}\mathbf{x};\ \ \mathbf{f} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix};\ \ \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

   (b) What kind of a pattern does $\mathbf{K}$ have?

   (c) **[Programming]** Consider a specific case where $n = 4$ and $k = 1.5N.m^{-1}$. What should be forces applied at the four nodes in order to displace the spring $\mathbf{x} = \begin{bmatrix} 0.5 \\ -0.5 \\ 0 \\ 0 \end{bmatrix} m$.

2. Consider the following electrical circuit with rectangular grid of resistors R. The input to this grid is a set of current injected at the top node as shown in the figure, such that $\sum_{k=1}^{5} i_k = 0$.



   Express the relationship between the voltages at the different nodes (represented by • in the figure) and the net current flowing in/out of the node in the following form, $\mathbf{G}\mathbf{v} = \mathbf{i}$. Where, $\mathbf{G}$ is the conductance matrix, $\mathbf{v}$ is the vector of node voltages, and $\mathbf{i}$ is the vector representing the net current flow in/out of the different node.

3. **Two point boundary problem.** $\mathbf{Ax} = \mathbf{b}$ is often encountered in many practical applications. One such application is the numerical solution of differential equations of the following form,

$$\sum_{i=0}^{M} a_i(x)\, y^{(i)}(x) = f(x)$$

where, $x \in [a, b]$ and $y(a) = \alpha, y(b) = \beta$.

Numerical methods are often employed for obtaining an approximate estimate of $y(x)$ at discrete points in the interval $[a, b]$. The interval is divided into subintervals of width Δx. The derivate of $y(x)$ at the different nodes (points between two subintervals) can be approximated as the following,

$$y'(x_i) = \frac{y(x_i + \Delta x) - y(x_i - \Delta x)}{\Delta x}$$
$$y''(x_i) = \frac{y(x_i + \Delta x) + 2y(x_i) - y(x_i - \Delta x)}{\Delta x^2}$$

where, $x_i = a + i\Delta x,\ 0 \leq i \leq N + 1$, and $b - a = (N + 1)\Delta x$. Addition and subtracting the above two equations and neglecting terms involving higher orders of Δx, we get the following approximations for the derivatives of $y(x)$ at x_i.

Replacing the derivatives of $y(x)$ by the above approximations and evaluating the equation at the different nodes x_is, we arrive a set of N linear equations with N unknowns $y(x_1), y(x_2), \ldots y(x_N)$.

Using this approach, compute an approximate solution for $y(x)$ for the following differential equations over the interval $x \in [0, 1]$.

(a) $y''(x) = -x$

(b) $y''(x) + y'(x) = x$

[Programming] Solve these equations for different values of Δx, and compare the resulting approximate solution for $y(x)$ with the exact solution. Present your results as a plot the solution $y(x_i)$ versus x_i.

Comment on the dependence of the solution (x) on Δx. What is the best value for Δx to use in solving these equations?

4. **Ill-conditioned systems.** A system $\mathbf{Ax} = \mathbf{b}$ is said to be ill-conditioned when small changes in the components of $\mathbf{A}$ or $\mathbf{b}$ can produce large changes in the solution $\mathbf{x}$. Consider the following system,

$$x - y = 100$$
$$10 + (9 + \Delta)\, y = 0$$

[Programming] Find the solutions of the system for different values of $\Delta = -2, -1, 0, 1, 2$. How do the solutions change with Δ.

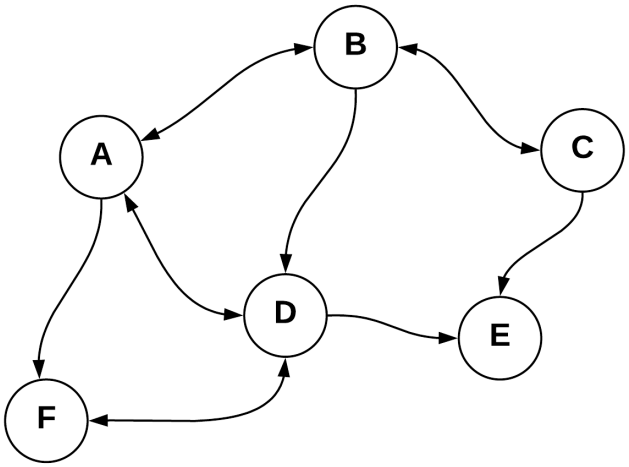
Now consider the following system,

$$x - y = 100$$
$$10 - (9 + \Delta)\, y = 0$$

[Programming] Find the solutions of the system for different values of $\Delta = -2, -1, 0, 1, 2$. How do the solutions change with Δ.

The second system is an example of an ill-conditioned system. What can you say about the geometries of these two systems?

5. **Connectivity matrices.** Another common application of matrices is in graph theory. A graph is a set of vertices or nodes connected by edges, as show in the following figure. A-F are the nodes of the graph, and the lines with the arrows are the edges that convey information about the connections or relationships between the nodes.



The above graph can be thought as a representation of different places in a city (represented by the nodes), and the lines with the arrows represent the roads connecting these different places. A line with two arrows allow two-way traffic, while line with single arrow only allow one way traffic. The connectivity between the different places can be summarized though the connectivity matrix $\mathbf{C} \in \mathbb{R}^{n \times n}$, where n is the number of nodes in the graph. The elements of this connectivity matrix represents whether or not there is a direct path between two places.

$$c_{ij} = \begin{cases} 1 & \text{there is a direct road between places } i \,\&\, j. \\ 0 & \text{otherwise.} \end{cases}$$

The diagonal element of $\mathbf{C}$ are zero, $c_{ii} = 0$.

Write down the connectivity matrix $\mathbf{C}$ for the graph shown above. How can we use the matrix $\mathbf{C}$ to answer the following questions? Explain exact matrix operation you would perform to answer these questions (Hint: Consider higher power of $\mathbf{C}$).

(a) Is there a path between two places i and j that goes via one other place? For example, we can go from A to D via B.

(b) How many paths are there between places i and j that goes via three other places?

6. **Computed Tomography** (CT) is a medical imaging technique that is used to reconstruct the internal structure of an object from a set of X-ray measurements. The object is placed between an X-ray source and a detector. The X-ray source and the detector are rotated around the object, and the X-ray measurements are recorded at different angles. The X-ray measurements are then used to reconstruct the internal structure of the object.

   The x-ray attentuation equation is given by,

$$I_o = I_i \exp\left(-\mu l\right)$$

   where, I_i is the intensity of the x-ray entering an object with fixed attentuation coefficient μ, I_o is the intensity of the x-ray existing the object, and l is the path length of the x-ray in the object.

   In general, the attenuation coefficient is a function of the position within the object, $\mu = \mu\left(x, y\right)$. The goal of CT is to reconstruct the spatial map of the attenuation coefficient $\mu\left(x, y\right)$.

   Let L represent the line segment of the x-ray within the object as shown in Figure 3.3, and the attentuation outside the object is assumed to be zero. If I_s is the intensity of the x-ray leaving the source, the intensity of the x-ray reaching the detector I_d is given by,

$$I_d = I_s \exp\left(-\int_L \mu\left(x, y\right) dl\right)$$

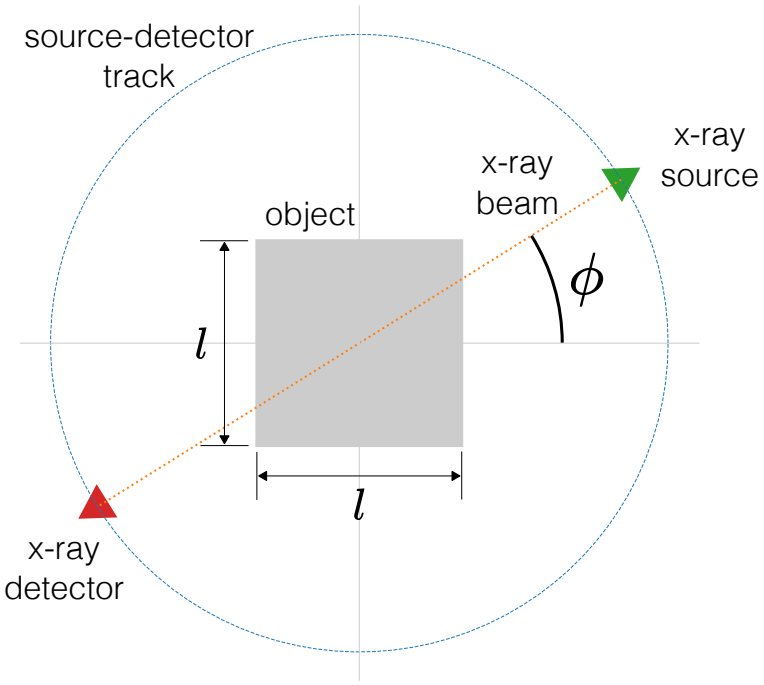


Figure 3.3: A simplified CT set-up with a single X-ray source and a single detector, that are located diametrically opposite to each other and can rotate to any scan angle ϕ. The object is placed between the X-ray source and the detector, which is depicted by the gray square of side l. The x-ray originates from the green triangle (x-ray source), passes through the object, and is detected by the red triangle (detector). The x-ray undergoes attenuation as it passes through the object. Different points in the objects are most likely to have different attentuation coefficients, and the goal of CT is toe reconstruct of spatial map of the attentuation within the object, which provides a measure of the internal structure of the object.

where, dl is the differential length along the line segment L, which is a function of x, y. The integral in the above equation is the line integral of the attenuation coefficient $\mu(x, y)$ along the line segment L.

We wish to solve this problem using a computer by posing it as a set of linear equations relating the attenuation coefficient $\mu(x, y)$ to the x-ray intensity measurements I_d. For simplicity, we will assume $I_s = 1$. First, we simplify the above integration equation by taking the log on both sides, which results in,

$$\ln I_d = -\int_L \mu(x, y)\, dl$$

Explain the why above step is valid. Do we need to worry about the case where $I_d = 0$?

Discretize the object into $n \times n$ grid of pixels, and assume that attentuation coefficient within each pixel to be constant. For any given scan angle ϕ you need to find out the pixels the line segment L passes through, along with the path length of the x-ray line segment within each pixel. Derive the discrete expression relating the x-ray intensity measurements I_d to the attenuation coefficient $\mu(x, y)$ for a given scan angle ϕ, assuming a discretized object $n \times n$. Write down the above expression in the following form

$$\tilde{\mathbf{a}}(\phi)^\top \mathbf{x} = y(\phi)$$

where, $\mathbf{x} \in \mathbb{R}^{n^2}$ is the vector of unknown attenuation coefficients of the pixels in the image, $\tilde{\mathbf{a}}(\phi) \in \mathbb{R}^{n^2}$ is the vector relating the attentuation coefficients to the detector measurement for the given source-detector angle ϕ and $y(\phi) \in \mathbb{R}$ is the log of the x-ray intensity measured by the detector.

If we make a set of such measurements for m different angles $\phi_1, \phi_2, \cdots \phi_m$, we can write the above equation in the following matrix form,

$$\mathbf{A}\mathbf{x} = \mathbf{y}$$

$$\mathbf{A} = \begin{bmatrix} \tilde{\mathbf{a}}_1 & \tilde{\mathbf{a}}_2 & \cdots & \tilde{\mathbf{a}}_m \end{bmatrix}^\top \in \mathbb{R}^{m \times n^2} \quad \mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_m \end{bmatrix}^\top \in \mathbb{R}^m$$

where, $\tilde{\mathbf{a}}_i$ is the vector relating the attentuation coefficients to the detector measurement for the source-detector angle ϕ_i and y_i is the log of the x-ray intensity measured by the detector for the angle ϕ_i.

Forward problem for CT. Once you've derived the above matrix linear equation, we can use it to simulate a CT scan by computing $\mathbf{y}$ for a given $\mathbf{x}$ and scan angle ϕ, we can compute intensity of the x-ray that will be measured by the detector. Consider the following three objects that we wish to scan using out CT scanner. Given the relatively simple geometry of the spatial distribution of the attenuation coefficients $\mu(x, y)$, you can use the $\mathbf{Ax} = \mathbf{y}$ equation to solve the forward CT problem, i.e., compute the detector output for different scan angles for the given object. For the three objects below, come up with a pixelation scheme for the objects (n number of pixels and pixel dimensions) and come up with the matrix $\mathbf{A}$ for the three objects. Use this matrix A and the known pixel attentualtion coefficients $\mathbf{x}$ to compute the detector outputs for 360 scan angle $\phi = 0\,\text{deg}, 1\,\text{deg}, 2\,\text{deg}, \ldots 359\,\text{deg}$. Assume $l = 1unit$.

[Programming] Write a python program to compute and plot the detector output for the different scan angles.

Suggestion: *For each object write a function that will take in the scan angle, the object side length l, and return the row vector $\tilde{\mathbf{a}}(\phi)^\top$. Use this function to compute the matrix $\mathbf{A}$ for the three objects.*

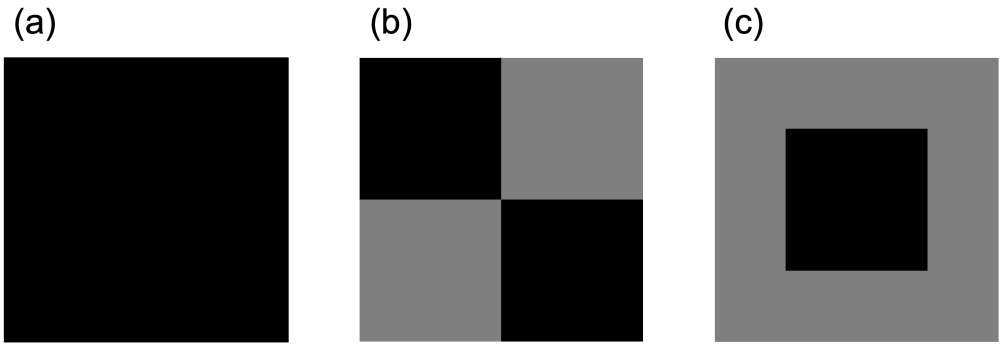


Figure 3.4: Three objects that are to be scanned using the CT scanner described in the figure above. The black regions in this image represent the pixels with attenuation coefficient $\mu = 1$, and the gray regions represent the pixels with attenuation coefficient $\mu = 0.5$.

Chapter 4

Orthogonality

"Every pair of perpendicular vectors are orthogonal, but not every pair of orthogonal vectors is perpendicular."

Ben Grossmann in StackExchange.

4.1 Exercise

1. Consider an orthonormal set of vectors,

$$V = \{\mathbf{v}_1, \mathbf{v}_2, \dots \mathbf{v}_r\} \quad \mathbf{v}_i \in \mathbb{R}^n \quad \forall i \in \{1, 2, \dots r\}$$

   If there is a vector $\mathbf{w} \in \mathbb{R}^n$ such that $\mathbf{v}_i^T \mathbf{w} = 0 \ \forall i \in \{1, 2, \dots r\}$. Prove that $\mathbf{w} \notin span\,(V)$.

2. Prove that the rank of an orthogonal projection matrix $\mathbf{P}_S = \mathbf{U}\mathbf{U}^T$ onto a subspace $\mathcal{S}$ is equal to the dim $\mathcal{S}$, where the columns of $\mathbf{U}$ form an orthonormal basis of $\mathcal{S}$.

3. For a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, prove that $C\,(\mathbf{A}) \perp N\left(\mathbf{A}^T\right)$ and $C\left(\mathbf{A}^T\right) \perp N\,(\mathbf{A})$.

4. If the columns of a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ are orthonormal, prove that $\mathbf{A}^{-1} = \mathbf{A}^T$. What is $\mathbf{A}^T\mathbf{A}$ when $\mathbf{A}$ is rectangular $(\mathbf{A} \in \mathbb{R}^{m \times n})$ with orthonormal columns?

5. What will happen when the Gram-Schmidt procedure is applied to: (a) orthonormal set of vectors; and (b) orthogonal set of vectors? If the set of vectors are columns of a matrix $\mathbf{A}$, then what are the corresponding $\mathbf{Q}$ and $\mathbf{R}$ matrices for the orthonormal and orthogonal cases?

6. If the columns of $\mathbf{A} \in \mathbb{R}^{m \times n}$ represent a basis for the subspace $\mathcal{S} \subset \mathbb{R}^m$. Find the orthogonal projection matrix $\mathbf{P}_{\mathcal{S}}$ onto the subspace $\mathcal{S}$. Hint: Gram-Schmidt orthogonalization.

7. Consider two orthogornal matrices $\mathbf{Q}_1$ and $\mathbf{Q}_2$. Is the $\mathbf{Q}_2^T\mathbf{Q}_1$ an orthogonal matrix? If yes, prove that it is so, else provide a counter-example showing $\mathbf{Q}_2^T\mathbf{Q}_1$ is not orthogonal.

8. Consider a 1 dimensional subspace spanned by the vector $\mathbf{u} \in \mathbb{R}^n$. What kind of a geometric operation does the matrix $\mathbf{I} - 2\frac{\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}}$ represent?

9. Prove that when a triangular matrix is orthogonal, it is diagonal.

Chapter 5

Matrix Inverses

"Every pair of perpendicular vectors are orthogonal, but not every pair of orthogonal vectors is perpendicular."

Ben Grossmann in StackExchange.

5.1 Exercise

1. When does the following diagnoal matrix have an inverse?

$$\mathbf{D} = \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{bmatrix}$$

   Write down an expression for $\mathbf{D}^{-1}$.

2. Consider a 2×2 block matrix, $\mathbf{A} = \begin{bmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{D} & \mathbf{E} \end{bmatrix}$, where $\mathbf{A} \in \mathbb{R}^{m \times m}$. Find an expression for the inverse $\mathbf{A}^{-1}$ interms of the block components and their inverses of $\mathbf{A}$. Hint: Consider $\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{P} & \mathbf{Q} \\ \mathbf{R} & \mathbf{S} \end{bmatrix}$, and solve $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$.

3. Consider a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ with linearly independent columns. Prove that the Gram matrix $\mathbf{A}^T\mathbf{A}$ is invertible.

4. Consider the scalar equation, $ax = ay$. Here we can cancel a from the equation when $a \neq 0$. When can we carry out similar cancellations for matrcies?

   (a) $\mathbf{AX} = \mathbf{AY}$. Prove that here $\mathbf{X} = \mathbf{Y}$ only when $\mathbf{A}$ is left invertible.

   (b) $\mathbf{XA} = \mathbf{YA}$. Prove that here $\mathbf{X} = \mathbf{Y}$ only when $\mathbf{A}$ is right invertible.

5. Consider two non-singular matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$. Explain whether or not the following matrices are invertible. If they are, then provide an expression for it inverse.

   (a) $\mathbf{C} = \mathbf{A} + \mathbf{B}$

   (b) $\mathbf{C} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{bmatrix}$

   (c) $\mathbf{C} = \begin{bmatrix} \mathbf{A} & \mathbf{A} + \mathbf{B} \\ \mathbf{0} & \mathbf{B} \end{bmatrix}$

   (d) $\mathbf{C} = \mathbf{ABA}$

6. For a square matrix $\mathbf{A}$ with non-signular $\mathbf{I} - \mathbf{A}$, prove that,

$$\mathbf{A}\left(\mathbf{I} - \mathbf{A}\right)^{-1} = \left(\mathbf{I} - \mathbf{A}\right)^{-1}\mathbf{A}$$

7. Consider the non-singular matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{A} + \mathbf{B}$. Prove that,

$$\mathbf{A}\left(\mathbf{A} + \mathbf{B}\right)^{-1}\mathbf{B} = \mathbf{B}\left(\mathbf{A} + \mathbf{B}\right)^{-1}\mathbf{A} = \left(\mathbf{A}^{-1} + \mathbf{B}^{-1}\right)^{-1}$$

Part II

Optimization

Part III

Probability and Statistics

Chapter 6

Statistical Estimation

"Every pair of perpendicular vectors are orthogonal, but not every pair of orthogonal vectors is perpendicular."

Ben Grossmann in StackExchange.

Part IV

Linear Programming

Chapter 7

Linear Programs

"Every pair of perpendicular vectors are orthogonal, but not every pair of orthogonal vectors is perpendicular."

Ben Grossmann in StackExchange.