Notes on Core Topics in Mathematical Economics

Matthew Lee Chen*

University of Cambridge

Disclaimer: These notes were written as preparatory material for the mathematics and statistics exam in the second year of my undergraduate degree in economics at Cambridge University (academic year 2019-2020). These are not intended to substitute for the course materials, but rather to supplement them with additional explanations. I take full responsibility for all errors and will be more than happy to promptly make corrections if brought to my attention. Please feel free to email to alert me to any errors or improvements at: mlc82 [at] cam [dot] ac [dot] uk.

Summary: This set of notes reviews core concepts in linear algebra, static optimisation and dynamic optimisation at an advanced level for economics undergraduate students. It starts with an overview of principles in linear algebra followed by an introduction to topics in real analysis and static optimisation. It finishes with a review of differential equations and dynamic optimisation. Without prior background in linear algebra, the notes should be read chronologically since the content from Section 1 is used in Section 2. However, Section 3 is reasonably standalone, and prior knowledge of linear algebra should render Section 2 very accessible.

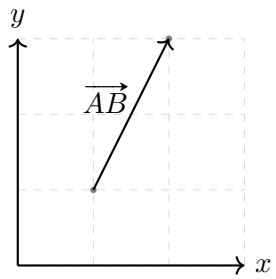
*These notes are based on lecture notes, problem sets and written solutions by Mikhail Safronov, as well as Mas-Colell, Whinston & Green (1995), Sydsæter & Hammond (2005), Sydsæter (1981), Simon & Blume (1994) and Carter (2001). All errors are solely my own.

Contents

1 Linear Algebra

1.1 Vectors and Matrices

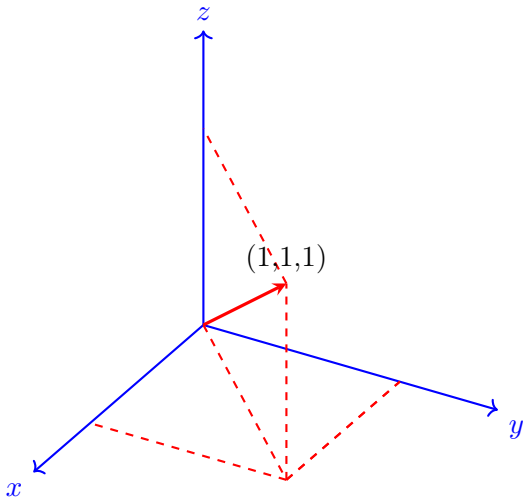
A **vector** can be represented in $\mathbb{R}^2$ as an arrow between two points. For instance, if we connect $A = (1, 1)$ and $B = (2, 3)$ with an arrow, then $\overrightarrow{AB}$ is a vector that lives in two-dimensional space $\mathbb{R}^2$ with coordinates $(2 - 1, 3 - 1) = (1, 2)$.



It is conventional to represent vectors as **column vectors**, as shown below

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$$

Such vectors live in space $\mathbb{R}^n$, meaning that if we were to draw the vector, we would need $\mathbb{R}^n$. For instance, if we wanted to draw the vector $\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$, we would need space $\mathbb{R}^3$. Often, we represent this as an arrow from the **zero vector** $\mathbf{0} = (0, 0, 0)$ to the coordinates $(1, 1, 1)$ in (x, y, z)-space.



Hence, a vector can be represented as a column of real numbers. In the above case, we have a $n \times 1$ column vector, with n denoting the number of rows and 1 the number of columns. If we have an array of real numbers consisting of multiple columns, we have an $n \times k$ **matrix**

$$X = \begin{pmatrix} x_{11} & \dots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nk} \end{pmatrix} \in \mathbb{R}^{n \times k}$$

We denote X as $n \times k$ to mean it has n rows and k columns. We often write $X = (x_{ij})$, meaning X consists of elements x_{ij} in row i and column j. For instance, x_{23} means the element of X in the 2nd row and 3rd column. The notation (x_{ij}) thus identifies the unique position of elements in X.

Definition 1. *The **transpose** of a matrix $X = (x_{ij})$ is the matrix $X^T = (x_{ji})$.*

Thus the matrix X has transpose X^T given by

$$X^T = \begin{pmatrix} x_{11} & \dots & x_{n1} \\ \vdots & \ddots & \vdots \\ x_{1k} & \dots & x_{nk} \end{pmatrix} \in \mathbb{R}^{n \times k}$$

The transpose operation simply holds the **main diagonal** (the set of elements of a matrix where $i = j$) fixed and then swaps all elements above the main diagonal with the corresponding elements below the main diagonal.

Transposing vectors works in exactly the same way. For instance, we have $(1 \;\; 2)^T = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$. The main diagonal consists of the element 1. So we flip the element 2, which is above the main diagonal, to the position below the main diagonal.

We often write a matrix as a set of its columns or rows. For instance, matrix X written as a set of columns is $(x_1, ..., x_k)$, with each x_i for $i = 1, ..., k$ being a column vector in $\mathbb{R}^n$.

Alternatively, as a set of rows, it can be written as $\begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}$ with each x_i for $i = 1, ..., n$ being a row vector in $\mathbb{R}^k$. We transpose row vectors to maintain the convention of writing vectors as columns.

Matrices can be added element-by-element provided they are **conformable**. That is, they have the required dimensions. If we add two matrices X and Y, we require both matrices to have the same dimensions. That is, for a $n \times k$ matrix X, Y must also be $n \times k$. If both matrices are conformable, adding them yields

$$X + Y = \begin{pmatrix} x_{11} & \dots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nk} \end{pmatrix} + \begin{pmatrix} y_{11} & \dots & y_{1k} \\ \vdots & \ddots & \vdots \\ y_{n1} & \dots & y_{nk} \end{pmatrix} = \begin{pmatrix} x_{11} + y_{11} & \dots & x_{1k} + y_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} + y_{n1} & \dots & x_{nk} + y_{nk} \end{pmatrix}$$

If we have matrix $X = (x_{ij})$, we can also multiply every element by any **scalar** (i.e. any single real number) $\alpha \in \mathbb{R}$ with the resulting matrix $\alpha X = (\alpha x_{ij})$ having the same dimensions as X.

If we are trying to multiply matrices X and Y, we require the number of columns of X be equal to the number of rows of Y for the operation to be conformable. That is, suppose that X is an $n \times k$ matrix and Y is $k \times m$, then

$$XY = \begin{pmatrix} \sum_{j=1}^{k} x_{1j} y_{j1} & \dots & \sum_{j=1}^{k} x_{1j} y_{jm} \\ \vdots & \ddots & \vdots \\ \sum_{j=1}^{k} x_{nj} y_{j1} & \dots & \sum_{j=1}^{k} x_{nj} y_{jm} \end{pmatrix}$$

Hence, the procedure of **matrix multiplication** simply means that each entry of the new matrix is given by the sum of the product of the corresponding entries of that row of X and column of Y. For instance, position $(xy)_{23}$ in the new matrix XY consists of the sum along the entries of the 2nd row of X multiplied by the entry along the 3rd column of Y i.e. $\sum_{j=1}^{k} x_{2j} y_{j3}$.

Note that if XY is conformable, in general, YX is not conformable. As such, $XY \neq YX$ in general. Clearly the number of columns in Y is not generally equal to the number of rows in X given Y is $k \times m$ and X is $n \times k$. Typically if YX is conformable, its dimensions will differ from XY. We can see that the resulting matrix XY from the multiplication will have dimensions $n \times m$ as we have added the product of elements in each column of X with the elements in each row of Y.

We can think of vectors as a particular type of matrix with a single column, allowing us to apply matrix addition and multiplication to vectors. Whether we write a vector horizontally or as a column vector does not affect the information it encodes.

For example, $x = (x_1 \;\; x_2)$ conveys the same information as $x^T = (x_1 \;\; x_2)^T = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$. Now, if we seek to compute Ax^T where $A = \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix}$, we have $Ax^T = \begin{pmatrix} a_1x_1 + a_2x_2 \\ a_3x_1 + a_4x_2 \end{pmatrix}$.

However, if we seek to compute Ax, we encounter non-conformability as A is a 2×2 matrix but x is a 1×2 vector. Thus, writing vectors as columns helps us deal with non-conformability issues which may arise.

From our discussion of matrix addition and multiplication, we can observe that for conformable matrices X, Y and Z and scalar α, it is true that $X + Y = Y + X$ and hence by multiplicity with a scalar, $\alpha(X + Y) = \alpha X + \alpha Y$. It also follows readily that $(X + Y) + Z = X + (Y + Z)$.

Some less trivial properties include the fact that $(XY)Z = X(YZ)$[1] and $X(Y+Z) = XY+XZ$[2].

Definition 2. *A **square matrix** is an $n \times n$ matrix.*

Definition 3. *A **symmetric matrix** X is a square matrix in which $X = X^T$.*

An example of a symmetric matrix is

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{12} & x_{22} & x_{23} \\ x_{13} & x_{23} & x_{33} \end{pmatrix}$$

Two explicit examples are the **identity matrix** and **zero matrix**, displayed respectively as

$$I_n = \begin{pmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{pmatrix} \qquad 0_n = \begin{pmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{pmatrix}$$

The identity matrix I_n is such that $XI_n = I_nX = X$ for all conformable X and the zero matrix 0_n is such that $X0_n = 0_nX = 0_n$.

[1] We can show this by observing that $XY = (xy)_{ik}$ can be indexed by $\sum_l x_{il}y_{lk}$. Hence, $(XY)Z = \sum_k (xy)_{ik}z_{kj} = \sum_k \sum_l x_{il}y_{lk}z_{kj}$. Similarly, each element of $YZ = (yz)_{lj}$ can be indexed by $\sum_l y_{lk}z_{kj}$. Hence, $X(YZ) = \sum_l x_{il}(yz)_{lj} = \sum_l \sum_k x_{il}y_{lk}z_{kj}$. Sums can be interchanged, giving the result.

[2] We may write $X(Y + Z)$ as $\sum_k x_{ik}(y_{kj} + z_{kj}) = \sum_k x_{ik}y_{kj} + \sum_k x_{ik}z_{kj} = XY + YZ$.

Definition 4. *A diagonal matrix D is a symmetric matrix with $(d_{ij}) = 0,\ \forall i \neq j$.*

$$X = \begin{pmatrix} d_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & d_n \end{pmatrix} = diag(d_1, ..., d_n)$$

That is, the collection of entries of a matrix except those running from the top left to bottom right are all zero. Note that $DX \neq XD$ in general for any conformable matrix X.

An exception is when $D = I_n$ since the identity matrix is a special diagonal matrix.

Definition 5. $\forall \boldsymbol{x} = (x_1, ..., x_n),\ \boldsymbol{y} = (y_1, ..., y_n) \in \mathbb{R}^n$, *the* ***scalar product***[3] *of* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *is given by*

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^T \boldsymbol{y} = \sum_{i=1}^{n} x_i y_i = \boldsymbol{y}^T \boldsymbol{x} = \langle \boldsymbol{y}, \boldsymbol{x} \rangle$$

As an example, consider the vectors $\mathbf{x} = (a, b)^T$ and $\mathbf{y} = (b, -a)^T$. The scalar product of $\mathbf{x}$ and $\mathbf{y}$ is $\mathbf{x}^T\mathbf{y} = ab - ba = 0$ and hence these vectors $\mathbf{x}$ and $\mathbf{y}$ are orthogonal.

From the definition of scalar products, we can establish that for scalars α and β, $\langle \alpha\mathbf{x}, \beta\mathbf{y} \rangle = \alpha\beta\langle \mathbf{x}, \mathbf{y} \rangle$.

In addition, for vectors $\mathbf{x}, \mathbf{y}$ and $\mathbf{z}$, the fact that $\langle \mathbf{x}, \alpha\mathbf{y} + \beta\mathbf{z} \rangle = \alpha\langle \mathbf{x}, \mathbf{y} \rangle + \beta\langle \mathbf{x}, \mathbf{z} \rangle$ follows readily from the definition of scalar products and the properties of matrix multiplication.

However, note that $\langle \langle \mathbf{x}, \mathbf{y} \rangle, \mathbf{z} \rangle$ and $\langle \mathbf{x}, \langle \mathbf{y}, \mathbf{z} \rangle \rangle$ are undefined as $\langle \mathbf{x}, \mathbf{y} \rangle$ and $\langle \mathbf{y}, \mathbf{z} \rangle$ are scalars. Therefore, we cannot perform the scalar product between a scalar and a vector.

Also, if $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle$, then it is not true to say that $\mathbf{y} = \mathbf{z}$. Instead, $\langle \mathbf{x}, \mathbf{y} - \mathbf{z} \rangle = 0$, meaning that $\mathbf{x}$ is orthogonal to **y-z**. Hence, **y-z** need not be zero, and so $\mathbf{y}$ need not be equal to $\mathbf{z}$.

Definition 6. *The* ***norm*** *of a vector is given by*

$$||\boldsymbol{x}|| = \langle \boldsymbol{x}, \boldsymbol{x} \rangle^{1/2} = (\boldsymbol{x}^T \boldsymbol{x})^{1/2} = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2}$$

[3] This is also called the **dot product** or **inner product**. Differing notation is also used sometimes, in particular note that $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x} \cdot \mathbf{y}$.

Intuitively, the norm[4] captures the distance of the vector relative to the zero-vector $\mathbf{0} = (0, ..., 0)$.

Definition 7. *Vectors* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *are* ***orthogonal****, denoted* $\boldsymbol{x} \perp \boldsymbol{y}$*, if and only if* $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{0}$*.*

Definition 8. *A set of vectors* $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_k\}$ *with* $\boldsymbol{x}_i \in \mathbb{R}^n$ *for all* $i = 1, ..., k$ *is* ***linearly dependent*** *if and only if there exists scalars* $\alpha_1, ..., \alpha_k$ *not all zero such that* $\alpha_1 \boldsymbol{x}_1 + ... + \alpha_k \boldsymbol{x}_k = \boldsymbol{0}$*. If no such scalars exist, the set of vectors is* ***linearly independent****.*

Linearly dependent vectors are such that at least one vector can be represented as a linear combination of the other vectors[5].

Supposing that $\alpha_i \neq 0$, we therefore have that

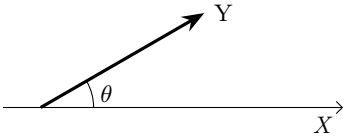
$$\mathbf{x}_i = \frac{-1}{\alpha_i} \sum_{j \neq i} \alpha_j \mathbf{x}_j$$

For example, vectors $\mathbf{x}_1 = (1, 1)^T$ and $\mathbf{x}_2 = (-3, -3)^T$ are linearly dependent because we can write $3\mathbf{x}_1 + \mathbf{x}_2 = \mathbf{0}$. We can also express $\mathbf{x}_1$ as a linear combination of $\mathbf{x}_2$ by saying $\mathbf{x}_1 = -\frac{1}{3}\mathbf{x}_2$.

Proposition 1. *The zero-vector is linearly dependent but any non-zero single vector is linearly independent.*

Proof. With one vector, linear dependence requires $\alpha\mathbf{x} = 0$ for $\alpha \neq 0$. Clearly, this is only possible when $\mathbf{x} = 0$. Linear independence requires $\alpha\mathbf{x} = \mathbf{0}$ with $\alpha = 0$ which holds for any $\mathbf{x} \neq \mathbf{0}$. □

[4]There is a geometric interpretation of scalar products and norms where $\langle \mathbf{x}, \mathbf{y} \rangle = ||\mathbf{x}||||\mathbf{y}||\cos\theta$ such that θ is the angle between $\mathbf{x}$ and $\mathbf{y}$. When the vectors are orthogonal, then $\cos 90 = 0$ (or $\cos \pi/2 = 0$) and hence $\langle \mathbf{x}, \mathbf{y} \rangle = 0$. Intuitively, the scalar product tells us the magnitude of the projection of one vector onto another. Imagine pulling an object located at the base of Y in direction Y. The horizontal projection of this force is given by $Y\cos\theta$. The magnitude of this projection is the force multiplied by the distance it travels along X i.e. $||X||||Y||\cos\theta$.



[5]It is a common misconception that *any* vector in a linearly dependent set $\{\mathbf{x}_1, ..., \mathbf{x}_k\}$ can be represented as a linear combination of the others. Suppose that $\{\mathbf{x}_1, \mathbf{x}_2\}$ are linearly independent, meaning that neither vector can be proportional to the other. Now, add $\mathbf{x}_3 = 2\mathbf{x}_2$ to the set. Clearly, $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ is now linearly dependent but $\mathbf{x}_1$ can still not be expressed as a linear combination of $\mathbf{x}_2$. Importantly, it takes just one redundancy of linear independence to obtain linear dependence, and hence to assess linear dependence, we look to see if we can find just one vector that is representable as a linear combination of others. If we find one, the set is linearly dependent.

Proposition 2. *Any set of non-zero orthogonal vectors is linearly independent.*

Proof. Let $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ be a set of non-zero orthogonal vectors. The scalar product with any vector in the set gives $0 = \langle \mathbf{x}_i, \alpha_1 \mathbf{x}_1 + ... + \alpha_n \mathbf{x}_n \rangle = \alpha_1 \langle \mathbf{x}_i, \mathbf{x}_1 \rangle + ... + \alpha_i \langle \mathbf{x}_i, \mathbf{x}_i \rangle + ... + \langle \mathbf{x}_i, \mathbf{x}_n \rangle$, but clearly $\langle \mathbf{x}_i, \mathbf{x}_j \rangle = 0, \ \ \forall j \neq i$. We are left with $0 = \langle \mathbf{x}_i, \mathbf{x}_i \rangle = \alpha_i ||\mathbf{x}_i||^2$. As $\mathbf{x}_i \neq 0$, $\alpha_i = 0$. This holds for all $i = 1, ..., n$ and hence $\alpha_1 = ... = \alpha_n = 0$, so $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ is linearly independent. □

Definition 9. *A **basis** of $\mathbb{R}^n$ is a set of linearly independent vectors $\boldsymbol{x}_1, ..., \boldsymbol{x}_n \in \mathbb{R}^n$ such that $\forall \boldsymbol{y} \in \mathbb{R}^n$, $\exists \alpha_1, ..., \alpha_n \in \mathbb{R}$ such that $\boldsymbol{y} = \alpha_1 \boldsymbol{x}_1 + ... + \alpha_n \boldsymbol{x}_n$. The vectors $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are called **basis vectors** and coefficients $\alpha_1, ..., \alpha_n$ are called **coordinates** of $\boldsymbol{y}$.*

The basis is not unique. If we think about vectors $\mathbf{x}_1 = (1, 0)^T$ and $\mathbf{x}_2 = (0, 1)^T$, these form a basis in $\mathbb{R}^n$. However, the vectors $\mathbf{x}_1 = (1, 0)^T$ and $\mathbf{x}_3 = (1, 1)^T$ also form a basis.

The vector $\mathbf{y} = (2, 3)^T$ has coordinates $(2, 3)$ in the basis $\{\mathbf{x}_1, \mathbf{x}_2\}$ but coordinates $(-1, 3)$ in the basis $\{\mathbf{x}_1, \mathbf{x}_3\}$.

Definition 10. *An **orthonormal basis** is a basis whose basis vectors $e_1, ..., e_n \in \mathbb{R}^n$ have a norm of 1 each, and $\langle e_i, e_j \rangle = e_i^T e_j = 0$ for all $i \neq j$.*

A typical example of an orthonormal basis is

$$e_1 = (1, 0, ..., 0), e_2 = (0, 1, ..., 0), ..., e_n = (0, ..., 1)$$

Another example of an orthonormal basis is

$$\hat{e}_1 = (\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, 0, ..., 0), \hat{e}_2 = (-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}, 0, ..., 0), ..., \hat{e}_n = (0, 0, ..., 1)$$

If we assume that vector $\mathbf{x}$ has coordinates $(\alpha_1, \alpha_2, ..., \alpha_n)$ in an orthonormal basis $(e_1, e_2, ..., e_n)$, the norm of $\mathbf{x}$ is given as

$$||\mathbf{x}|| = (\mathbf{x}^T \mathbf{x})^{1/2} = \Big[(\alpha_1 e_1 +, ..., +\alpha_n e_n)^T (\alpha_1 e_1 +, ..., +\alpha_n e_n) \Big]^{1/2} = \Big(\sum_{i=1}^{n} \alpha_i^2 \Big)^{1/2}$$

To clarify the above expression, consider the vector $\mathbf{x} = (x_1, x_2)^T = \alpha_1 e_1 + \alpha_2 e_2$.

The norm can be expressed as $||\mathbf{x}|| = (\alpha_1^2 ||e_1||^2 + 2\alpha_1 \alpha_2 \langle e_1, e_2 \rangle + \alpha_2^2 ||e_2||^2)^{1/2}$.

Since the norm of all vectors in the orthonomal basis is 1, e_1^2 and e_2^2 are both 1, and the dot product $\langle e_1, e_2 \rangle = 0$ since vectors are orthogonal. Hence, we are reduced to $(\alpha_1^2 + \alpha_2^2)^{1/2}$.

1.2 Matrix Inversion, Determinant and Rank

Definition 11. *An **inverse matrix** X^{-1} is such that $XX^{-1} = X^{-1}X = I$ for all conformable matrices X.*

Definition 12. *A matrix X for which an inverse matrix X^{-1} exists is called **non-singular** or **invertible**. If no such X^{-1} exists, the matrix X is called **singular** or **non-invertible**.*

Suppose we have a matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$. If we seek to find its inverse, we are looking for a matrix such that $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

It turns out that we can write such a matrix as

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \begin{pmatrix} \frac{d}{ad-bc} & \frac{-b}{ad-bc} \\ \frac{-c}{ad-bc} & \frac{a}{ad-bc} \end{pmatrix}$$

A different case might be when we have a diagonal matrix $\begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix}$. Clearly, we can multiply this matrix by $\begin{pmatrix} \frac{1}{d_1} & 0 \\ 0 & \frac{1}{d_2} \end{pmatrix}$ and we yield the 2×2 identity matrix. More generally, we say that a diagonal matrix $D = diag(d_1, ..., d_n)$ has an inverse given by

$$\begin{pmatrix} \frac{1}{d_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{1}{d_n} \end{pmatrix}$$

Proposition 3. *For all non-singular square matrices, a unique inverse exists.*

Proof. We proceed by contradiction. It is established that $XX^{-1} = X^{-1}X = I$ for a conformable X and inverse X^{-1}. Assume there also exists a $Y \neq X^{-1}$ such that $XY = YX = I$. This implies that $XX^{-1} = XY$, and hence $X(X^{-1} - Y) = 0$. Premultiplying by X^{-1} gives $X^{-1} - Y = 0$, and so $X^{-1} = Y$ which is a contradiction. Thus the inverse is necessarily unique. □

Proposition 4. *For non-singular square matrices X and Y, $(XY)^{-1} = Y^{-1}X^{-1}$.*

Proof. For conformable matrices, $(XY)(XY)^{-1} = I$ and $XYY^{-1}X^{-1} = XX^{-1} = I$. Similarly, $(XY)^{-1}(XY) = I$ and $Y^{-1}X^{-1}XY = Y^{-1}Y = I$. □

By definition of transposes, $(X^T)^T = X$. That is, if we flip the elements of a matrix off its main diagonal and repeat again, we will yield the same matrix. Furthermore, by looking at the definition of transpose and element-by-element matrix addition, we can deduce that $(X + Y)^T = X^T + Y^T$.

Proposition 5. *For conformable matrices X and Y, $(XY)^T = Y^T X^T$.*

Proof. [6] Observe that $(xy)^T_{ij} = (xy)_{ji} = \sum_k x_{jk}y_{ki}$ and $(y^T x^T)_{ij} = \sum_k y^T_{ik}x^T_{kj} = \sum_k y_{ki}x_{jk}$. □

Proposition 6. $(X^T)^{-1} = (X^{-1})^T$

Proof. Premultiply by X^T to give the identity matrix on the left-hand side; then, apply Proposition 3 to the right-hand side, yielding $X^T(X^{-1})^T = (X^{-1}X)^T = I$. □

Definition 13. *The **determinant** of a square matrix X is a number defined as*

$$det(X) = \sum_{j=1}^{n}(-1)^{i+j}x_{ij}det(X_{ij}) \quad \textit{for any } i = \sum_{i=1}^{n}(-1)^{i+j}x_{ij}det(X_{ij}) \quad \textit{for any } \ j$$

*where X_{ij} is the $(n-1)\times(n-1)$ **cofactor matrix** with the ith row and jth column removed so that, for example*

$$X_{11} = \begin{pmatrix} x_{22} & \dots & x_{2n} \\ \vdots & \ddots & \vdots \\ x_{n2} & \dots & x_{nn} \end{pmatrix}$$

We say that a matrix is non-invertible or singular if and only if its determinant is zero. Therefore, for an inverse of a square matrix to exist, its determinant must be non-zero.

For 2×2 matrices such as $\begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix}$, we can apply Definition 13 to obtain a quick and easy formula for the determinant.

[6] This proof can be extended to $(X_1X_2...X_n)^T = X_n^T X_{n-1}^T ... X_1^T$ by induction. We have already proved the basis case $n = 2$. Assume it is true for $n = k$, and hence for $n = k+1$, we have $(A_1A_2...A_kA_{k+1})^T = (A_1A_2...A_k)A_{k+1})^T = A_{k+1}^T(A_1A_2...A_k)^T = A_{k+1}^T A_k^T ... A_1^T$. Hence, we have proven that the proposition holds for all cases $n \geq 2$.

First, we can use either summation in Definition 13. We will sum over all the j's (i.e. the columns) for the first row (i.e. $i = 1$). Now we proceed through x_{1j} though only sum across the columns: $x_{11}x_{22} - x_{12}x_{21}$. We could have picked any row when summing over j or instead summed over i and picked any column and we would have got the same answer.

It is important to note that each of the cofactor matrices in the above calculation is 1×1. The determinant of a 1×1 matrix is the sole entry. The reason for this can be seen through the formula. There is only one entry with positive sign and the cofactor matrix has no entries (sometimes called the **empty matrix**). The empty matrix is defined to have determinant 1, giving the required result.

For 3×3 matrices such as $\begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{pmatrix}$, we can repeat and obtain a formula for the determinant: $x_{11}x_{22}x_{33} + x_{12}x_{23}x_{31} + x_{13}x_{21}x_{32} - x_{13}x_{22}x_{31} - x_{12}x_{21}x_{33} - x_{11}x_{23}x_{32}$.

By applying the formula, we can obtain a few facts about determinants and their properties. First, observe that $det(I_n) = 1$ and more generally for diagonal matrices, $det(D) = \prod_i d_i$. Transposing a matrix does not change its determinant, which should follow from the fact that we can sum over any row and column in Definition 13. That is, $det(X^T) = det(X)$.

Two trickier properties involve matrix multiplication and inverse matrices. First, $det(XY) = det(X)det(Y)$ for conformable matrices, which can be verified with any two conformable matrices. A corollary of this property is that $det(X^{-1}) = \dfrac{1}{det(X)}$. This can be derived by observing that $1 = det(XX^{-1}) = det(X)det(X^{-1})$.

Definition 14. *An **upper triangular matrix** is a square matrix such that all elements below its main diagonal are zero. A **lower triangular matrix** is a square matrix such that all elements above its main diagonal are zero.*

It can be shown from Definition 13 that the determinant of any triangular matrix (upper or lower) is the product of elements on its main diagonal.

When we calculate determinants of general matrices, there exist some helpful rules which are summarised below.

1. Multiplying any row or column by a scalar α multiplies the determinant of the matrix by α.
2. A matrix with any row or column of all zeros has zero determinant.

3. Adding one row or column to another does not change the determinant.

4. Swapping any two rows or columns multiplies the determinant by -1.

These rules above help us compute determinants of matrices when we perform **row reduction** - that is, we perform elementary operations on rows and columns of the matrix. We can use row reduction to convert any non-singular square matrix to a diagonal matrix, a process called **diagonalisation**. As an example, consider the following matrix

$$X = \begin{pmatrix} 2 & 4 & 6 \\ 1 & 3 & 4 \\ 1 & 4 & 6 \end{pmatrix}$$

First, let's divide the first row by 2, which halves the determinant.

$$\begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 4 \\ 1 & 4 & 6 \end{pmatrix}$$

Subtracting the first row from the second, and also from the third doesn't change the determinant.

$$\begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 1 \\ 0 & 2 & 3 \end{pmatrix}$$

Now subtract the third row from the first, again not changing the determinant.

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 2 & 3 \end{pmatrix}$$

Subtract the second row twice from the third, not changing the determinant.

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

Subtract the third row from the second, not changing the determinant.

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

The matrix above is the identity matrix and has determinant 1. The only elementary row operation we conducted which changed the determinant was to divide the first row by 2, which halved the determinant. Hence, the determinant of the original matrix X must be 2.

Proposition 7. *For a set of linearly independent vectors $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$, if we add two vectors together, the vectors will still be linearly independent.*

Proof. Consider $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$. Without loss of generality, we can check for $\{\mathbf{x}_1, \mathbf{x}_1 + \mathbf{x}_2, \mathbf{x}_3\}$. By linear independence, $\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \alpha_3\mathbf{x}_3 = 0 \implies \alpha_1 = \alpha_2 = \alpha_3 = 0$. Hence, for the new set to be linearly independent, we require $\delta_1\mathbf{x}_1 + \delta_2(\mathbf{x}_1 + \mathbf{x}_2) + \delta_3\mathbf{x}_3 = 0 \implies \delta_1 = \delta_2 = \delta_3 = 0$, which is clearly true since some manipulation gives us $(\delta_1 + \delta_2)\mathbf{x}_1 + \delta_2\mathbf{x}_2 + \delta_3\mathbf{x}_3 = 0 \implies \delta_1 + \delta_2 = \delta_2 = \delta_3 = 0$. □

Proposition 8. *For a set of linearly independent vectors $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$, multiplication of one or more vectors by non-zero scalars preserves linear independence.*

Proof. Analogous to the proof of Proposition 7. □

Definition 15. *The **column rank** of an $n \times K$ matrix X is the number of linearly independent columns. The **row rank** is the number of linearly independent rows. The column rank and row rank are always equal, and this number is called the **rank** of X. For X to be **full rank**, it must be that $rank(X) = \min(n, K)$.*

In the context of row reduction, we can provide an alternative definition: the rank of matrix X is the size of its largest non-singular square **minor** or **submatrix**.

In other words, remove one or more rows or columns of X to obtain a square submatrix.

Then, the size of the largest square submatrix with non-zero determinant (i.e. no columns or rows of all-zero entries) is the rank of X.

Interpreting the rank as the size of the largest non-singular square submatrix allows us to say that if by row reduction, we obtain one row of all-zero entries, the determinant of that matrix is zero, so we need to proceed to a smaller submatrix to find the rank.

Hence, a square matrix is non-singular if and only if all rows (and columns) are linearly independent.

As an example, suppose we seek to find the rank of matrix X below.

$$X = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 4 & 4 \\ 3 & 6 & 5 & 4 \end{pmatrix}$$

The first row has non-zero elements, and the first entry is 1.

Performing row reduction by subtracting the first row multiplied by 2 from the second row and subtracting the first row multiplied by 3 from the first row gives us

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & -2 & -4 \\ 0 & 0 & -4 & -8 \end{pmatrix}$$

Now, since all entries in the first column below 1 are zero, 1 cannot possibly be written as a linear combination of two zeros, and thus the first row must be linearly independent.

The second row has non-zero element -2. Multiplying the second row by $-\frac{1}{2}$ gives

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & -4 & -8 \end{pmatrix}$$

Now, we can make all entries below 1 zero, so add the second row multiplied by 4 to the third row, giving

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Now, observe that the third row of the original matrix was linearly dependent since through row operations, we transformed it to all zero entries. However, this is not possible with the second row. Thus, we have 2 linearly independent rows (and columns) so the original matrix has rank 2.

It is worth noting that if we had a different set of row operations, we may have produced a different set of linearly independent rows. Therefore, it must be the case that the total number of linearly independent rows (and columns) does not depend on the exact procedure of row reduction.

In general, if one procedure found rows $\mathbf{x}_1$ and $\mathbf{x}_2$ as linearly independent, while another found $\mathbf{x}'_1$, $\mathbf{x}'_2$ and $\mathbf{x}'_3$ as linearly independent, then it must be true that

$$\mathbf{x}'_1 = \alpha_{11}\mathbf{x}_1 + \alpha_{12}\mathbf{x}_2$$

$$\mathbf{x}'_2 = \alpha_{21}\mathbf{x}_1 + \alpha_{22}\mathbf{x}_2$$

$$\mathbf{x}'_3 = \alpha_{31}\mathbf{x}_1 + \alpha_{32}\mathbf{x}_3$$

Alternatively, consider the matrix form

$$\begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \\ \alpha_{31} & \alpha_{32} \end{pmatrix}$$

Proceeding by row reduction for the above matrix, one of the rows must be a linear combination of the other two.

That is, one row must be linearly dependent and reducible to all-zero entries.

Hence, only two of the rows are linearly independent so we again get the result of the original matrix having rank 2.

Let's return to the following reduced matrix

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Observe that the elements in the first and second row and the first and third column together form a square non-singular upper-triangular matrix $\begin{pmatrix} 1 & 3 \\ 0 & 1 \end{pmatrix}$.

This matrix is the largest square non-singular submatrix and has size 2, which is the same as the number of linearly independent rows.

Alternatively, we can look at the original matrix

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 4 & 4 \\ 3 & 6 & 5 & 4 \end{pmatrix}$$

Observe here that the submatrix $\begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}$ is also non-singular.

Our discussion of upper triangular submatrices allows us to argue the proposition suggested in Definition 15.

Proposition 9. *The column rank and row rank are equal for every matrix.*

Proof. Let the row rank of a matrix be given by n and the size of the largest nonsingular square submatrix m.

Row reduction gives us a non-singular upper-triangular submatrix, and hence the size of the largest non-singular square submatrix cannot be smaller than the number of linearly independent rows - that is, $m \geq n$.

Furthermore, a non-singular submatrix cannot have linearly dependent rows so $n \geq m$.

Hence, $n = m$ must hold.

Performing similar operations with columns allows us to say that the number k of linearly independent columns equals m which in turn equals n. Hence, row rank always equals column rank. □

Proposition 10. *Let V be an $n \times k$ matrix with $k \leq n$ and rank k. Then, VV^T is of rank k (and is singular when $k < n$), while the $k \times k$ matrix V^TV is of rank k and hence non-singular.*

Consider V as an $n \times k$ matrix, meaning V^T is $k \times n$. Hence, VV^T must be $n \times n$. Then, if VV^T has rank $k < n$, then VV^T is singular. Only when $k = n$ is VV^T non-singular. Similarly, V^TV must be $k \times k$ so with rank k, it must always be non-singular.

For example, let's define

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 1 & 1 \end{pmatrix}$$

Notice that A is a 2×3 matrix with rank 2 (as the last column is simply the sum of the first two so is linearly dependent). Now, we observe that

$$AA^T = \begin{pmatrix} 14 & 5 \\ 5 & 2 \end{pmatrix}$$

$$A^TA = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 5 & 7 \\ 3 & 7 & 10 \end{pmatrix}$$

Observe that AA^T has rank 2 since the rows and columns are disproportional to one another. Also A^TA has rank 2 since the last column equals the sum of the first two columns.

However, we can say much less if we have general matrices X $(n \times k)$ and Y $(k \times m)$.

Proposition 11. $rank(XY) \leq \min\{rank(X), rank(Y)\}$

Let's now define two unrelated matrices A and B. Observe that

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix} \quad B = \begin{pmatrix} 1 & 1 \\ -1 & -1 \\ 1 & -1 \\ -1 & 1 \end{pmatrix}$$

$$AB = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \quad BA = \begin{pmatrix} 2 & 2 & 1 & 1 \\ -2 & -2 & -1 & -1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & -1 & -1 \end{pmatrix}$$

Matrices A and B have rank 2 but AB clearly has rank 0 but BA has rank 2.

Proposition 12. *$rank(ABC) = rank(B)$ for conformable A, B and C and non-singular A and C.*

Proof. By Proposition 11, $rank((AB)C) \leq rank(AB) \leq rank(B)$. Let $D \equiv ABC$, then $B = A^{-1}DC^{-1}$ so $rank(B) = rank((A^{-1}D)C^{-1}) \leq rank(A^{-1}D) \leq rank(D)$. Combining the two gives the result. □

1.3 Change of Basis

Consider $\mathbf{x} \in \mathbb{R}^n$ with coordinates $(x_1, ..., x_n)$ in basis $(e_1, ..., e_n)$. If we want to change to a new basis $(e'_1, ..., e'_n)$, we need to find the coordinates $(x'_1, ...x'_n)$ of $\mathbf{x}$ in the new basis.

We can express each new basis vector e'_i in terms of old basis vectors as follows

$$e'_1 = q_{11}e_1 + q_{12}e_2 + ... + q_{1n}e_n$$

$$\vdots$$

$$e'_n = q_{n1}e_1 + q_{n2}e_2 + ... + q_{nn}e_n$$

where, for instance, q_{11} represents the coordinates of vector e'_1 in the old basis.

Writing the above system of equations in matrix form gives us

$$(e'_1, e'_2, ..., e'_n) = (e_1, e_2, ..., e_n) \cdot Q$$

where the columns of the right-hand side matrix correspond to coordinates of $(e'_1, ..., e'_n)$ in the old basis, and that

$$Q = \begin{pmatrix} q_{11} & q_{21} & \cdots & q_{n1} \\ q_{12} & q_{22} & \cdots & q_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ q_{1n} & q_{2n} & \cdots & q_{nn} \end{pmatrix}$$

Q must be invertible for $(e'_1, ..., e'_n)$ to be a basis.

In the old basis, vector $\mathbf{x}$ is given by

$$\mathbf{x} = (e_1, e_2, ..., e_n) \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Using the matrix form of the system of equations for the new basis vectors, we can deduce that

$$(e'_1, e'_2, ..., e'_n) \cdot Q^{-1} = (e_1, e_2, ..., e_n)$$

$$\mathbf{x} = (e'_1, e'_2, ..., e'_n) \cdot Q^{-1} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Thus vector $\mathbf{x} = x'_1 e'_1 +, ..., + x'_n e'_n$ with new coordinates

$$\begin{pmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_n \end{pmatrix} = Q^{-1} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Estimating the inverse of the $n \times n$ matrix above can be quite difficult. However, things become easier when we are transferring between two orthonormal bases, which we define $(e_1, ..., e_n)$ and $(e'_1, ..., e'_n)$, and so under this condition, Q is an **orthonormal matrix**.

Definition 16. *A square matrix Q is orthonormal if and only if $QQ^T = I$. Equivalently $Q^T = Q^{-1}$.*

Definition 16 implies that

$$q_i^T q_j = \begin{cases} 1 & if \ \ i = j \\ 0 & if \ \ i \neq j \end{cases}$$

That is, an orthonormal matrix has its columns equal the coordinates of basis vectors of one orthonormal basis in another. Then q_i and q_j correspond to different basis vectors e'_i and e'_j.

Proposition 13. *The determinant of an orthonormal matrix is either 1 or -1.*

Proof. $1 = det(I_n) = det(QQ^T) = det(Q)det(Q^T) = (det(Q))^2$. □

However, it is important to note that if the determinant of a matrix is 1 or -1, it doesn't necessarily mean that the matrix is orthonormal.

To consider an example of an orthonormal matrix, think of the matrix below. When we premultiply any vector by the matrix below, it rotates vectors counterclockwise through an angle θ of the origin.

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

We can verify that it is orthonormal using the definition of orthonormal matrices and the trigonometric identity that $\cos^2\theta + \sin^2\theta \equiv 1$.

In the event that we have two orthonormal bases, the formula for changing coordinates is

$$\begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix} = Q^T \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

As an explicit example, consider the orthonormal basis $e_1 = (1,0)^T, e_2 = (0,1)^T$ in $\mathbb{R}^2$ and a new orthonormal basis $e_1' = (\frac{\sqrt{3}}{2}, \frac{1}{2})^T, e_2' = (-\frac{1}{2}, \frac{\sqrt{3}}{2})^T$. Then, we have

$$Q = \begin{pmatrix} \frac{\sqrt{3}}{2} & -\frac{1}{2} \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{pmatrix}$$

Let's say that in the old basis, vector $\mathbf{x}$ had coordinates $(2,1)$. New coordinates are given by

$$\begin{pmatrix} x_1' \\ x_2' \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{3}}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{\sqrt{3}}{2} \end{pmatrix} \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} \sqrt{3} + \frac{1}{2} \\ \frac{\sqrt{3}}{2} - 1 \end{pmatrix}$$

This gives us the vector $\mathbf{x}$ after change of basis consisting of $\mathbf{x} = (\sqrt{3} + \frac{1}{2})e_1' + (\frac{\sqrt{3}}{2} - 1)e_2'$.

New basis vectors are represented via old ones as

$$e_1' = (e_1, e_2)\begin{pmatrix} \frac{\sqrt{3}}{2} \\ \frac{1}{2} \end{pmatrix} \qquad e_2' = (e_1, e_2)\begin{pmatrix} -\frac{1}{2} \\ \frac{\sqrt{3}}{2} \end{pmatrix}$$

Thus we can represent $\mathbf{x}$ via old basis vectors as follows

$$\mathbf{x} = (\sqrt{3} + \frac{1}{2})(e_1, e_2)\begin{pmatrix} \frac{\sqrt{3}}{2} \\ \frac{1}{2} \end{pmatrix} + (\frac{\sqrt{3}}{2} - 1)(e_1, e_2)\begin{pmatrix} -\frac{1}{2} \\ \frac{\sqrt{3}}{2} \end{pmatrix} = 2e_1 + e_2$$

1.4 Eigenvalues and Eigenvectors

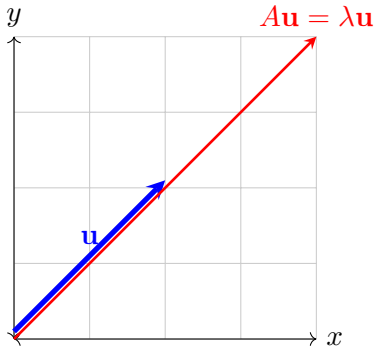
Consider a square matrix A. Given a vector $\mathbf{u} \in \mathbb{R}^n$, $A\mathbf{u}$ is also a vector in $\mathbb{R}^n$.

We can think of the matrix A as a linear mapping or **transformation** from $\mathbb{R}^n \to \mathbb{R}^n$.

That is, premultiplying by A has not changed the dimension of the vector. Rather, it has changed its direction and magnitude in the same space. Often, we are interested in describing what has happened under such a transformation.

To do so, we can look at vectors $\mathbf{u} \in \mathbb{R}^n$ such that after the linear transformation, the new vector $A\mathbf{u}$ is parallel to the original vector. For instance, if we have $A\mathbf{u} = 2\mathbf{u}$, we can say that the linear transformation stretched the space $\mathbb{R}^n$ in the direction of $\mathbf{u}$ by scale factor 2.

More generally, we can say that $A\mathbf{u} = \lambda\mathbf{u}$ where the transformation stretches the space by scale factor λ, as shown below.



Definition 17. *For a real symmetric $n \times n$ matrix A, a vector $\boldsymbol{u} \in \mathbb{R}^n \backslash \boldsymbol{0}$ and scalar $\lambda \in \mathbb{R}$ are an* ***eigenvector*** *and* ***eigenvalue*** *respectively if $A\boldsymbol{u} = \lambda\boldsymbol{u}$.*

We rule out vectors $\mathbf{u} = 0$ because they trivially satisfy every equation.

Proposition 14. *If there exists a $\lambda = 0$ with $\boldsymbol{u} \neq 0$, it means that $A\boldsymbol{u} = 0$, and so A must be singular (i.e. A cannot have all linearly independent rows).*

For an identity matrix $A = I_n$, we have $\mathbf{u} = \lambda\mathbf{u}$, which is true only when $\lambda = 1$. The eigenvectors are all vectors $\mathbf{u} \neq 0$.

For general diagonal matrices, we have the eigenvalues $\lambda = d_1, ..., d_n$ and corresponding eigenvectors $\mathbf{u} = (1, 0, ..., 0)^T, (0, 1, ..., 0)^T, ..., (0, ..., 0, 1)^T$.

We can compute the eigenvalues and eigenvectors of a matrix by writing that $(A - \lambda I_n)\mathbf{u} = \mathbf{0}$. If we have $\mathbf{u} \neq \mathbf{0}$, then $A - \lambda I_n$ must be singular (as the row vectors must be linearly dependent) - otherwise, there is a contradiction.

We therefore state that λ is an eigenvalue if and only if $det(A - \lambda I_n) = 0$ - this expression is called the **characteristic polynomial**. If we have real symmetric matrices, then the eigenvalues are all real-valued.[7]

As an example of solving for eigenvalues and eigenvectors, consider the matrix $A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$. The characteristic polynomial suggests that we need the determinant of $A - \lambda I = \begin{pmatrix} 2-\lambda & -1 \\ -1 & 2-\lambda \end{pmatrix}$ to be zero. Hence, we need $3 - 4\lambda + \lambda^2 = 0$, which factorises to $(\lambda - 3)(\lambda - 1) = 0$, giving two distinct eigenvalues $\lambda_1 = 3$ and $\lambda_2 = 1$.

Now, we need to find the corresponding eigenvectors. First, consider $\lambda_1 = 3$. This means that $A - \lambda I = \begin{pmatrix} -1 & -1 \\ -1 & -1 \end{pmatrix}$. Hence, we solve $\begin{pmatrix} -1 & -1 \\ -1 & -1 \end{pmatrix} \mathbf{u} = 0$. Clearly we need a $\mathbf{u} \in \mathbb{R}^2 \backslash (0, 0)^T$. One example is $\mathbf{u}_1 = (-1 \quad 1)^T$. Notice that every vector proportional to $\mathbf{u}_1$ (i.e. $\gamma\mathbf{u}_1$ for all $\gamma \in \mathbb{R}\backslash 0$) is also an eigenvector that corresponds to λ_1.

---

[7] The proof is involved and relies on knowledge of complex numbers. Start by supposing that $\lambda \in \mathbb{C}$ is an eigenvalue of a symmetric matrix $A \in \mathbb{R}^{n \times n}$ and $\mathbf{u} \in \mathbb{C}^n$ is a corresponding eigenvector such that $A\mathbf{u} = \lambda\mathbf{u}$. Take complex conjugates, which gives us $A\bar{\mathbf{u}} = \bar{\lambda}\bar{\mathbf{u}}$ as A is real so is unaffected.

Now observe that premultiplying $A\mathbf{u} = \lambda\mathbf{u}$ by $\bar{\mathbf{u}}^T$ suggests $\bar{\mathbf{u}}^T A\mathbf{u} = \bar{\mathbf{u}}^T\lambda\mathbf{u} \implies (A\bar{\mathbf{u}})^T u = \lambda\bar{\mathbf{u}}^T\mathbf{u}$ by the rules of transposes i.e. $(A\bar{\mathbf{u}})^T = \bar{\mathbf{u}}^T A^T$ and the definition of a symmetric matrix i.e. $A = A^T$. Using the conjugate equation $A\bar{\mathbf{u}} = \bar{\lambda}\bar{\mathbf{u}}$, we have $\lambda\bar{\mathbf{u}}^T\mathbf{u} = (\bar{\lambda}\bar{\mathbf{u}})^T\mathbf{u} = \bar{\lambda}\bar{\mathbf{u}}^T\mathbf{u}$. Some manipulation gives $(\lambda - \bar{\lambda})\bar{\mathbf{u}}^T\mathbf{u} = \mathbf{0}$.

Notice that $\bar{\mathbf{u}}^T\mathbf{u} = \sum_i \bar{u}_i^T u_i > 0$ as for any complex number, we have $z\bar{z} = (a + bi)(a - bi) = a^2 + b^2 \geq 0$ and $\mathbf{u} \neq \mathbf{0}$. Hence, it must be the case that $\lambda = \bar{\lambda}$. Thinking about $\lambda = a + bi$ and $\bar{\lambda} = a - bi$, for $a + bi = a - bi$, the imaginary component must be zero so $b = 0$. Hence, $\lambda \in \mathbb{R}$, completing the proof.

It is worth noting that for real asymmetric matrices (i.e. for which $A \neq A^T$), the eigenvalues may be complex as when we take the conjugate equation, $\bar{A} \neq A$.

We can repeat the exercise for $\lambda_2 = 1$ and find that $\mathbf{u}_2 = (1 \;\; 1)^T$ or any vector proportional to $(1 \;\; 1)^T$.

Proposition 15. *If $A\boldsymbol{u} = \lambda\boldsymbol{u}$, then for any scalar k, $A(k\boldsymbol{u}) = \lambda(k\boldsymbol{u})$.*

Hence, $k\mathbf{u}$ is a corresponding eigenvector of eigenvalue λ. Therefore, we conventionally normalise eigenvector $\mathbf{u}$ to norm 1 such that $||\mathbf{u}|| = \mathbf{u}^T\mathbf{u} = 1$.

Proposition 16. *Suppose A has eigenvalues λ_i. Then, the eigenvalues of $A + cI_n$ are $\lambda_i + c$ for $c \in \mathbb{R}$.*

Proof. $A\mathbf{u} + cI_n\mathbf{u} = \lambda\mathbf{u} + cI_n\mathbf{u} \implies (A + cI)\mathbf{u} = (\lambda + c)\mathbf{u}$. □

Proposition 17. *An eigenvector $\boldsymbol{u}$ of A corresponding to λ is also an eigenvector of A^k corresponding to λ^k.*

Proof. By induction, the basis case is $A^2\mathbf{u} = A(A\mathbf{u}) = \lambda A\mathbf{u} = \lambda^2\mathbf{u}$. Assume it to be true for $n = k$ i.e. $A^k\mathbf{u} = \lambda^k\mathbf{u}$ and proceed for the case $n = k + 1$. Observe that $\lambda^{k+1}\mathbf{u} = \lambda\lambda^k\mathbf{u} = \lambda A^k\mathbf{u} = A^k A\mathbf{u} = A^{k+1}\mathbf{u}$. Hence, it holds for all integers $n \geq 2$. □

When we worked through an example of finding eigenvalues and eigenvectors earlier, we found that $\mathbf{u}_1 = (-1 \;\; 1)^T$ and $\mathbf{u}_2 = (1 \;\; 1)^T$. Notice that $\langle \mathbf{u}_1, \mathbf{u}_2 \rangle = \mathbf{u}_1^T\mathbf{u}_2 = 0$ i.e. the eigenvectors are orthogonal. This is not coincidental but rather a general feature.

Proposition 18. *For symmetric matrix A, eigenvectors $\boldsymbol{u}$ and $\boldsymbol{v}$ corresponding to distinct eigenvalues λ and μ are **orthogonal**.*

Proof. Suppose $A\mathbf{u} = \lambda\mathbf{u}$ and $A\mathbf{v} = \mu\mathbf{v}$ where $\mu \neq \lambda$. Then $\mathbf{v}^T A\mathbf{u} = \lambda\mathbf{v}^T\mathbf{u}$ and $\mathbf{u}^T A\mathbf{v} = \mu\mathbf{u}^T\mathbf{v}$. Since $(\mathbf{v}^T(A\mathbf{u}))^T = (A\mathbf{u})^T\mathbf{v} = \mathbf{u}^T A\mathbf{v}$ (A is symmetric) and $\mathbf{v}^T A\mathbf{u}$ is a number, $\mathbf{v}^T A\mathbf{u} = \mathbf{u}^T A\mathbf{v} \implies \mathbf{v}^T A\mathbf{u} - \mathbf{u}^T A\mathbf{v} = 0 \implies \lambda\mathbf{v}^T\mathbf{u} - \mu\mathbf{u}^T\mathbf{v} = (\lambda - \mu)\langle \mathbf{v}, \mathbf{u} \rangle = 0$. Therefore, $\langle \mathbf{v}, \mathbf{u} \rangle = 0$. □

By Proposition 2, we can extend Proposition 18 to say that eigenvectors $\mathbf{u}$ and $\mathbf{v}$ are necessarily linearly independent.

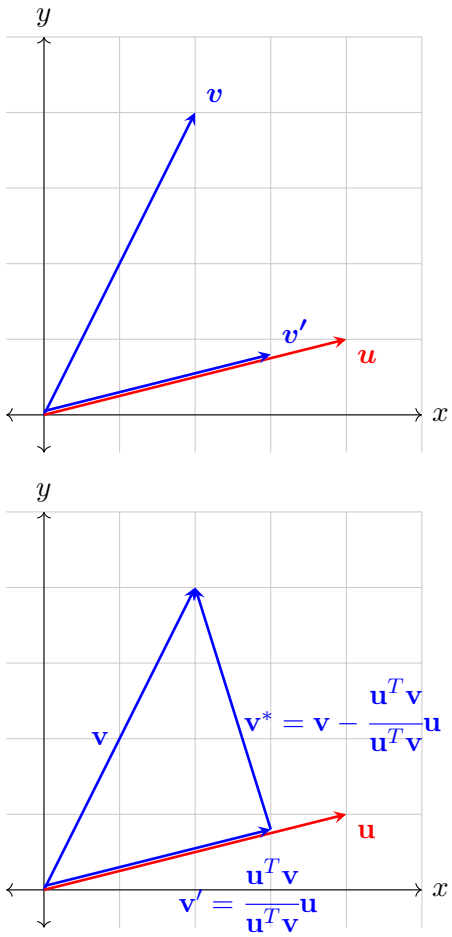
However, Proposition 18 does not extend to asymmetric matrices. Rather, for asymmetric matrices, eigenvectors are linearly independent but not necessarily orthogonal.

Proposition 19. *If* $\boldsymbol{u}$ *and* $\boldsymbol{v}$ *are distinct proportional eigenvectors corresponding to eigenvalue* λ*, then so is* $\alpha\boldsymbol{u} + \beta\boldsymbol{v}$ *for scalars* $\alpha, \beta \in \mathbb{R}\backslash 0$.

If $\boldsymbol{u}$ *and* $\boldsymbol{v}$ *are distinct eigenvectors corresponding to* λ *but disproportional to one another, then so are* $\boldsymbol{u}$ *and* $\boldsymbol{v}^*$ *where* $\boldsymbol{v}^* = \boldsymbol{v} - \dfrac{\boldsymbol{u}^T\boldsymbol{v}}{\boldsymbol{u}^T\boldsymbol{u}}\boldsymbol{u}$.

In the case where $\mathbf{u}$ and $\mathbf{v}$ are proportional, this follows because $A(\alpha\mathbf{u} + \beta\mathbf{v}) = \alpha A\mathbf{u} + \beta A\mathbf{v} = \alpha\lambda\mathbf{u} + \beta\lambda\mathbf{v} = \lambda(\alpha\mathbf{u} + \beta\mathbf{v})$, so $\alpha\mathbf{u} + \beta\mathbf{v}$ is an eigenvector of A corresponding to λ.

In the disproportionate eigenvectors case, we are projecting $\mathbf{v}$ onto $\mathbf{u}$ and taking the orthogonal component $\mathbf{v}^*$. This can be visualised in the diagrams below where $\mathbf{u}$ and $\mathbf{v}$ are not proportional but the projection $\mathbf{v}' = \dfrac{\mathbf{u}^T\mathbf{v}}{\mathbf{u}^T\mathbf{v}}\mathbf{u}$ is proportional to $\mathbf{u}$. Taking $\mathbf{v}^*$ gives a vector orthogonal to $\mathbf{u}$.



One method of compactly writing the eigenvalues and eigenvectors of a matrix is as follows. First, a real symmetric $n \times n$ matrix A has n eigenvalues $\lambda_1, ..., \lambda_n$ (could all be distinct or not) and associated eigenvectors $\mathbf{u}_1, ..., \mathbf{u}_n \in \mathbb{R}^n$ so that $A\mathbf{u}_i = \lambda_i\mathbf{u}_i$ for $i = 1, ..., n$.

This can be compactly written as

$$A(\mathbf{u}_1 \ \ldots \ \mathbf{u}_n) = (\mathbf{u}_1 \ \ldots \ \mathbf{u}_n) \begin{pmatrix} \lambda_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \lambda_n \end{pmatrix}$$

Or even more compactly

$$AU = U\Lambda$$

where U is an $n \times n$ matrix with the columns being the eigenvectors of A, and Λ is the diagonal matrix containing the eigenvalues on the main diagonal.

If U is invertible, we can write $A = U\Lambda U^{-1}$ and $U^{-1}AU = \Lambda$. Note that the set of eigenvectors is linearly independent if and only if U is non-singular (i.e. U is full rank), and hence constitutes a basis for $\mathbb{R}^n$ if and only if U is non-singular.

We can think about whether or not U is invertible. Let's consider two cases.

First, suppose the eigenvalues are all distinct. Then, we have already proven that the eigenvectors are all orthogonal, and hence linearly independent. Thus, U is full rank, and equivalently non-singular. As such, the eigenvectors constitute a basis. Since they are all orthogonal and we can normalise them to norm 1, they constitute an orthonormal basis. As such, $U^{-1} = U^T$ i.e. U is an orthonormal matrix.

Now, consider the case when the eigenvalues are not all distinct. In this case, for symmetric matrices, it can be shown that there exists a set of orthogonal eigenvectors that forms an orthonormal basis.

Consider the 2×2 identity matrix as an example: the characteristic polynomial suggests that $(1 - \lambda)^2 = 0$, which has eigenvalues $\lambda_1 = 1$ and $\lambda_2 = 1$, hence repeated. The number of corresponding eigenvectors is infinite: the identity matrix has every non-zero vector as an eigenvector. However, eigenvectors will clearly be repeated, so matrix U will not be full rank and hence singular. Nonetheless, we are able to find a subset of the eigenvectors of A that is linearly independent.

However, we may not be able to find such a basis for asymmetric matrices.

Our discussions above can be summarised succinctly as follows.

Proposition 20. *Any real symmetric matrix A possesses an* ***eigendecomposition****. That is, A can be written as $A = U\Lambda U^{-1} = U\Lambda U^T = \sum_{i=1}^{n} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T$ where $\boldsymbol{u}_i \in \mathbb{R}^n$, $\lambda \in \mathbb{R}$, $i = 1, ..., n$.*

We can apply Proposition 20 to a couple of examples.

First, consider the real symmetric matrix $\begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ with three distinct eigenvalues (for $\rho \neq 0$) $\lambda_1 = \rho + 1$, $\lambda_2 = 1$ and $\lambda_3 = 1 - \rho$. Associated with these eigenvalues are the distinct eigenvectors $\mathbf{v}_1 = (1 \;\; 1 \;\; 0)^T$, $\mathbf{v}_2 = (0 \;\; 0 \;\; 1)^T$ and $\mathbf{v}_3 = (-1 \;\; 1 \;\; 0)^T$.

The eigenvectors $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$ are mutually orthogonal and recall that any vector that is proportional to each of $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$ is also an eigenvector of that corresponding eigenvalue.

Thus, we can normalise (by dividing by $\sqrt{2}$) to obtain an orthonormal basis in $\mathbb{R}^3$ represented by matrix $U = \begin{pmatrix} \frac{\sqrt{2}}{2} & 0 & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & 0 & \frac{\sqrt{2}}{2} \\ 0 & 1 & 0 \end{pmatrix}$. Now, to check that U is non-singular, observe that $det(U) \neq 0$.

Another example is the real symmetric matrix $\begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}$ with two distinct eigenvalues $\lambda_1 = 1 - \rho$ (repeated twice) and $\lambda_2 = 1 + 2\rho$, and distinct eigenvectors $\mathbf{u}_1 = (-1 \; 1 \; 0)^T$ and $\mathbf{u}_2 = (-1 \; 0 \; 1)^T$ both corresponding to λ_1 and $\mathbf{u}_3 = (1 \;\; 1 \;\; 1)^T$ corresponding to λ_2.

Observe, however, that while $\langle \mathbf{u}_1, \mathbf{u}_3 \rangle = \langle \mathbf{u}_2, \mathbf{u}_3 \rangle = 0$, $\langle \mathbf{u}_1, \mathbf{u}_2 \rangle = 1 \neq 0$. However, we can project $\mathbf{u}_2$ onto $\mathbf{u}_1$ to obtain an orthogonal component $\mathbf{u}_2^*$ that is mutually orthogonal to $\mathbf{u}_1$ and $\mathbf{u}_3$. We find this component by computing $\mathbf{u}_2^* = \mathbf{u}_2 - \frac{\mathbf{u}_2^T \mathbf{u}_1}{\mathbf{u}_2^T \mathbf{u}_2} \mathbf{u}_1 = \begin{pmatrix} -\frac{1}{2} \\ -\frac{1}{2} \\ 1 \end{pmatrix}$ and scaling to norm 1, giving $U = (\mathbf{u}_1 \;\; \mathbf{u}_2^* \;\; \mathbf{u}_3)^T$ as an orthonormal basis in $\mathbb{R}^3$.

Note that this is not unique: we could have projected $\mathbf{u}_1$ onto $\mathbf{u}_2$ and obtained orthogonal component $\mathbf{u}_1^*$ which would have allowed for an orthonormal basis represented by $U = (\mathbf{u}_1^* \; \mathbf{u}_2 \; \mathbf{u}_3)^T$.

When we calculated the determinant of a matrix, we performed row reduction to diagonalise the matrix. However, another way to diagonalise a matrix is through the eigendecomposition as $U^{-1}XU = \Lambda$ or $U^T XU = \Lambda$ for orthonormal U.

We can use eigendecompositions to help us solve matrix equations. We are essentially using the eigendecomposition to diagonalise the matrix into an easier form to work with. For instance, consider $2X^2 - 3X + I_n = 0_n$.

Let's look at the eigendecompositions of all matrices via the same set of eigenvectors in U: for X, we write $U^T XU = diag(\lambda_1, ..., \lambda_n)$, for X^2 (recalling that for real symmetric matrix X, $X^2\mathbf{u} = \lambda^2\mathbf{u}$) it can be written $U^T X^2 U = diag(\lambda_1^2, ..., \lambda_n^2)$, and for I_n, we write $U^T I_n U = diag(1, ..., 1)$ where $\lambda_1, ..., \lambda_n$ are the eigenvalues of X and U is the orthonormal matrix of eigenvectors generated by normalising all eigenvectors to norm 1.

Now, write the equation as $U^T[2X^2 - 3X + I_n]U = diag(2\lambda_1^2 - 3\lambda_1 + 1, ..., 2\lambda_n^2 - 3\lambda_n + 1) = 0_n$. In other words, any λ_i must satisfy $2\lambda_i^2 - 3\lambda_i + 1 = 0$. Hence, the only possibilities are $\lambda_i = 1$ or $\lambda_i = \frac{1}{2}$.

A geometric interpretation of this result might be that if we consider any n orthogonal vectors, and shrink some of them by half, the related matrix of such a linear transformation must satisfy the matrix quadratic equation.

Definition 18. *A real symmetric $n \times n$ matrix A is*

Positive definite *if and only if $\boldsymbol{x}^T A\boldsymbol{x} = \sum_{i=1}^n \sum_{j=1}^n a_{ij}x_i x_j > 0$ for all $\boldsymbol{x} \in \mathbb{R}^n \backslash \boldsymbol{0}$.*

Negative definite *if and only if $\boldsymbol{x}^T A\boldsymbol{x} < 0$ for all $\boldsymbol{x} \in \mathbb{R}^n \backslash \boldsymbol{0}$.*

Positive semi-definite *if and only if $\boldsymbol{x}^T A\boldsymbol{x} \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$.*

Negative semi-definite *if and only if $\boldsymbol{x}^T A\boldsymbol{x} \leq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$.*

If A is neither positive semi-definite nor negative semi-definite, then it is ***indefinite****.*

Using the eigendecomposition of matrix A, we can provide alternative definitions of definiteness and semi-definiteness.

We have that $A = U\Lambda U^T$ and hence $\mathbf{x}^T U\Lambda U^T\mathbf{x} = \mathbf{y}^T\Lambda\mathbf{y} = \sum_{i=1}^n \lambda_i \mathbf{y}_i^2$ where $\mathbf{y} = U^T\mathbf{x}$ (recall the main diagonal entries of a diagonal matrix constitute the eigenvalues).

For A to be positive definite, we require $\sum_{i=1}^n \lambda_i \mathbf{y}_i^2 > 0$. As $\mathbf{y}_i^2 \geq 0$ for all $i = 1, ..., n$, we require $\lambda_i > 0$ for all $i = 1, ..., n$. For negative definiteness, we need $\lambda_i < 0$. For positive semi-definiteness, we need $\lambda_i \geq 0$ and for negative semi-definiteness, $\lambda_i \leq 0$.

Definition 19. *The **trace** of a square matrix A to be $tr(A) = \sum_{i=1}^{n} a_{ii}$. That is, the trace is the sum of entries along a matrix's main diagonal.*

Proposition 21. *For conformable A and B, $tr(AB) = tr(BA)$. This can be extended to $tr(ABC) = tr(BCA) = tr(CAB)$ with conformable C.*

Proof. By definition of the trace, we can label A and B without loss of generality to give $tr(AB) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}b_{ji} = \sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}a_{ji} = tr(BA)$. The extension comes from setting $X = AB$ and $Y = C$, and then $X = BC$ and $Y = A$ and observing $tr(XY) = tr(YX)$. □

Proposition 22. *For a real symmetric $n \times n$ matrix A, $tr(A) = \sum_{i=1}^{n} \lambda_i$.*

Proof. Observe that $tr(A) = tr(U\Lambda U^T) = tr(U^T U\Lambda) = tr(\Lambda) = \sum_{i=1}^{n} \lambda_i$. □

Proposition 23. *For a real symmetric $n \times n$ matrix A, $det(A) = \prod_{i=1}^{n} \lambda_i$.*

Proof. Observe that $det(A) = det(U\Lambda U^T) = det(U^T U\Lambda) = det(\Lambda) = \prod_{i=1}^{n} \lambda_i$. □

Propositions 10 and 11 give us a quick way of determining the magnitude of eigenvalues. For instance, we may seek to know whether some eigenvalues are greater than 1 or smaller than -1. One way is to check if $|det(A)|=| \prod_{i=1}^{n} \lambda_i |> 1$. Alternatively, if $| tr(A) |> n$, then the absolute value of at least one eigenvalue exceeds 1.

Proposition 11 leads us to a key result about eigenvalues and matrix invertibility. Recall that if $det(A) = \prod_{i=1}^{n} \lambda_i$, then if the determinant is equal to zero, at least one eigenvalue must have been zero to generate this. Similarly, if at least one eigenvalue is zero, the determinant must be equal to zero, giving a non-singular matrix.

Proposition 24. *A square matrix is nonsingular if and only if it has all eigenvalues non-zero. That is, positive and negative definite matrices are necessarily invertible but matrices which are not positive or negative definite, but positive or negative semi-definite are singular.*

It is worth qualifying that Proposition 12 is not asserting that all semi-definite matrices are singular. That is untrue. But rather, if a semi-definite matrix is to be non-singular, it must also be definite. The reason follows from the definition of a definite matrix. We cannot have eigenvalues equal to zero if we want a matrix to be non-singular.

Proposition 25. *For all positive definite matrices A, A^{-1} is also positive definite. Similarly for negative definite matrices, A^{-1} is negative definite.*

Proof. Consider $A\mathbf{u} = \lambda\mathbf{u}$, which means that A has eigenvalue λ. If A is definite, it has an inverse so premultiplying by A^{-1} and rearranging gives $A^{-1}\mathbf{u} = \frac{1}{\lambda}\mathbf{u}$. If A is positive definite, every $\lambda > 0$ so every $\frac{1}{\lambda} > 0$. If A is negative definite, every $\lambda < 0$ so every $\frac{1}{\lambda} < 0$. □

Proposition 26. *When X is positive semi-definite or negative semi-definite, X^TX and XX^T are necessarily positive semi-definite. If and only if X^TX and XX^T are positive definite, then X must be either positive or negative definite.*

Proof. Consider $\mathbf{z}^T(X^TX)\mathbf{z} = (X\mathbf{z})^T(X\mathbf{z}) = ||X\mathbf{z}||^2 \geq 0$ and hence X^TX is positive semi-definite. For X^TX to be positive definite, we must guarantee that X is full rank (i.e. non-singular), which implies that X is positive definite or negative definite. The same argument applies to XX^T. □

Proposition 27. *For a real symmetric matrix A, its rank is equal to the number of non-zero eigenvalues (counting repetitions).*

Proof. Observe that $rank(A) = rank(U\Lambda U^T) = rank(\Lambda)$, which is exactly the number of non-zero eigenvalues of $diag(\lambda_1, ..., \lambda_n)$ as the rank is the size of the largest square non-singular submatrix, and invertibility requires linear independence of all rows and columns. □

Our discussion about eigenvalues and eigenvectors has focused on real symmetric matrices. However, when we have asymmetric matrices, there may exist complex eigenvalues. Consider $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. Premultiplying a vector by this matrix rotates it by 90 degrees counterclockwise. Hence, all non-zero vectors will change direction. The characteristic polynomial suggests that $\lambda^2 + 1 = 0$ which gives solutions $\lambda = \pm i$ where $i = \sqrt{-1}$.

With asymmetric matrices, eigenvectors need not constitute a basis even though the eigenvalues are real. For instance, consider $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, which has one unique eigenvalue $\lambda = 1$ but the set of corresponding eigenvectors is $(1, 0)^T$ and all other proportional vectors. However, the set of linearly independent eigenvectors is only $(1, 0)^T$ which is a vector in $\mathbb{R}^2$ but the matrix is in space $\mathbb{R}^{2\times 2} = \mathbb{R}^4$, so we need another eigenvector to form a basis that spans the matrix.

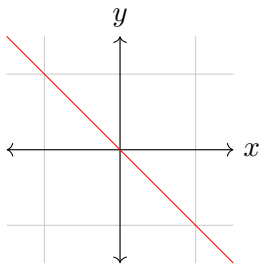
1.5 Linear Subspaces

Think about $\mathbb{R}^3$ characterised by vectors $\mathbf{x} = (x_1 \quad x_2 \quad x_3)^T$. Let's suppose that we are trying to characterise all vectors in $\mathbb{R}^3$ that satisfy $x_3 = 0$ (i.e. the set $\{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_3 = 0\}$). We can describe this set of vectors as a two-dimensional **linear subspace** in $\mathbb{R}^3$.

Definition 20. *A set $\mathbb{L} \subseteq \mathbb{R}^n$ is a **linear subspace** if $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{L}$ and any $\alpha, \beta \in \mathbb{R}$, we have $\alpha\boldsymbol{x} + \beta\boldsymbol{y} \in \mathbb{L}$.*

Definition 19 implies that every linear subspace contains the zero vector $\mathbf{0}$. Note that subspaces can be trivial, as we see with $\mathbb{L} = \{\mathbf{0} \in \mathbb{R}^n\}$. We may describe the set $\mathbb{L} = \{\mathbf{x} \in \mathbb{R}^n : x_k = x_{k+1} = ... = x_n = 0\}$ as a **proper subspace** (i.e. $\mathbb{L} \subsetneq \mathbb{R}^n$). To verify this, let $\mathbf{x}, \mathbf{y} \in \mathbb{L}$ and then notice that $\alpha\mathbf{x}_i + \beta\mathbf{y}_i = 0$ for any $i \geq k$ and any $\alpha, \beta \in \mathbb{R}$.

We can define a basis on any linear subspace. For instance, think about the one-dimensional subspace $\mathbb{L} = \{(x, y) \in \mathbb{R}^2 : x + y = 0\}$, represented diagrammatically below.



Recall that a basis consists of a set of linearly independent vectors for which every vector in this space can be represented as a linear combination of the basis vectors. Now, this subspace is one-dimensional so we only need one vector to form a basis.

Satisfying linear independence is trivial as we are dealing with a single vector. Thus, any vector $\alpha(1, -1)^T$ for $\alpha \in \mathbb{R}\backslash 0$ is a basis for this subspace.

Proposition 28. *If $\mathbb{L}_1$ and $\mathbb{L}_2$ are subspaces of $\mathbb{R}^n$, then $\mathbb{L}_1 \cap \mathbb{L}_2$ is also a subspace.*

Proof. First, we check that $\mathbf{0} \in \mathbb{L}_1 \cap \mathbb{L}_2$ which is true if $\mathbf{0} \in \mathbb{L}_1$ and $\mathbf{0} \in \mathbb{L}_2$.

Next, we can similarly check that if $\mathbf{x}, \mathbf{y} \in \mathbb{L}_1$ and $\mathbf{x}, \mathbf{y} \in \mathbb{L}_2$, then $\mathbf{x}, \mathbf{y} \in \mathbb{L}_1 \cap \mathbb{L}_2$.

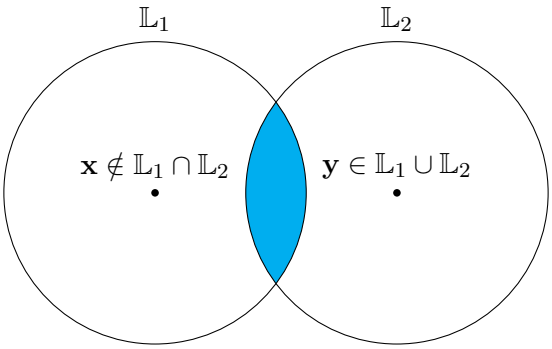
If it is true that $\alpha\mathbf{x}+\beta\mathbf{y} \in \mathbb{L}_1$ and $\alpha\mathbf{x}+\beta\mathbf{y} \in \mathbb{L}_2$, then it must be true that any linear combination $\alpha\mathbf{x} + \beta\mathbf{y} \in \mathbb{L}_1 \cap \mathbb{L}_2$. □

Note that we can easily extend this logic by defining $\mathbb{L}' = \mathbb{L}_1 \cap \mathbb{L}_2$ and observing that if $\mathbb{L}_3$ is a subspace, then $\mathbb{L}' \cap \mathbb{L}_3$ is a subspace. Hence, in general, the intersection of subspaces is a subspace.

However, this logic above does not apply to unions of subspaces. For instance, to see whether $\mathbb{L}_1 \cup \mathbb{L}_2$ is a subspace, first consider space $\mathbb{R}^2$. Let $\mathbb{L}_1 = \{(x, y) \in \mathbb{R}^2 : y = 0\}$ (i.e. the x-axis) and $\mathbb{L}_2 = \{(x, y) \in \mathbb{R}^2 : x = 0\}$ (i.e. the y-axis).

Clearly while $\mathbf{0} \in \mathbb{L}_1 \cup \mathbb{L}_2$, taking any two arbitrary points (e.g. (0,3) and (5,0)), we notice that while both these points are elements of $\mathbb{L}_1 \cup \mathbb{L}_2$, a linear combination of these points (e.g. add them together to obtain (5,3)) obviously does not lie in the union of the two axes.

For clarity, we may interpret the intersection or union of subspaces in terms of venn diagrams, as shown below.



Definition 21. *Let A be a real $n \times n$ matrix with eigenvalue λ. The union of eigenvectors associated with λ and $\boldsymbol{0}$, denoted $\mathbb{L}_\lambda = \{\boldsymbol{0} \in \mathbb{R}^n\} \cup \{\boldsymbol{x} \in \mathbb{R}^n : A\boldsymbol{x} = \lambda\boldsymbol{x}\}$, is a proper subspace of $\mathbb{R}^n$ called the **eigenspace** associated with λ. Note that if $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{L}_\lambda$, then $A(\alpha\boldsymbol{x} + \beta\boldsymbol{y}) = \alpha A\boldsymbol{x} + \beta A\boldsymbol{y} = \lambda(\alpha\boldsymbol{x} + \beta\boldsymbol{y})$ so $\alpha\boldsymbol{x} + \beta\boldsymbol{y} \in \mathbb{L}_\lambda$ for $\alpha, \beta \in \mathbb{R}$.*

Definition 22. *A set of vectors $\boldsymbol{x}_1, ..., \boldsymbol{x}_n \in \mathbb{R}^n$ generate a linear subspace called the **span***

$$\mathcal{C}(\boldsymbol{x}_1, ..., \boldsymbol{x}_n) = \{\boldsymbol{u} \in \mathbb{R}^n : \boldsymbol{u} = \gamma_1\boldsymbol{x}_1 + ... + \gamma_n\boldsymbol{x}_n\}, \quad \gamma_1, ..., \gamma_n \in \mathbb{R}.$$

The span is therefore the set of all linear combinations of a set of vectors. That is, the span is the smallest linear subspace that contains the set of vectors.

Or alternatively, the intersection of all subspaces that contain the set of vectors. In matrix form:

$$\mathcal{C}(\boldsymbol{x}_1, ..., \boldsymbol{x}_n) = \{X\boldsymbol{\gamma}\}, \quad X = (\boldsymbol{x}_1 ... \boldsymbol{x}_n), \quad \boldsymbol{\gamma} \in \mathbb{R}^n$$

Note that if $\mathbf{z}_1, \mathbf{z}_2 \in \mathcal{C}(X)$, then it means that $\mathbf{z}_1 = X\boldsymbol{\gamma}_1$ and $\mathbf{z}_2 = X\boldsymbol{\gamma}_2$ for some $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2 \in \mathbb{R}^n$. Then, it is the case that $\alpha\mathbf{z}_1 + \beta\mathbf{z}_2 = X(\alpha\boldsymbol{\gamma}_1 + \beta\boldsymbol{\gamma}_2) \in \mathcal{C}(X)$ for $\alpha, \beta \in \mathbb{R}$. Hence, if two vectors are in the span, then a linear combination of those two vectors is also in the span.

Using the definition of the span as the set of all linear combinations of a set of vectors, we can define the notion of a **column space** as the span of a matrix's column vectors. In other words, the set of every possible linear combination of its column vectors.

The **dimension** (i.e. in what space? For instance, $\mathbb{R}^n$ has n dimensions) of the linear combinations of the column vectors will be equal to the number of linearly independent column vectors, which is the rank.

Hence, we say that the dimension of the column space is equal to the rank, in turn equal to the number of linearly independent columns.

Analogously, the **row space** is the span of a matrix's row vectors, and the dimension of the row space is the rank.

Therefore, consider a set of vectors that form a basis. Their span must be a subspace $\mathbb{L}$.

Think of an easy example, say $(1 \;\; 0)^T$ and $(0 \;\; 1)^T$. Note that the span of these two vectors (the set of all linear combinations) is the space $\mathbb{R}^2$. Therefore, a basis in $\mathbb{R}^n$ is the smallest set which spans $\mathbb{R}^n$ i.e. any set of n linearly independent vectors that spans $\mathbb{R}^n$.

As an explicit example of determining if a set of vectors spans a set, consider the following vectors in $\mathbb{R}^3$: $(1 \;\; 3 \;\; 3)^T$, $(0 \;\; 0 \;\; 1)^T$ and $(1 \;\; 3 \;\; 1)^T$. First, we check for linear independence. That is, we can proceed through elementary operations and we find that $(1 \;\; 3 \;\; 3)^T$ and $(0 \;\; 0 \;\; 1)^T$ are linearly independent, and so the matrix $\begin{pmatrix} 1 & 0 \\ 3 & 0 \\ 3 & 1 \end{pmatrix}$ has rank 2.

Thus, $\mathcal{C} = \{\mathbf{u} \in \mathbb{R}^3 : \mathbf{u} = \delta_1(1 \;\; 3 \;\; 3)^T + \delta_2(0 \;\; 0 \;\; 1)^T\}$ is sufficient to span $\mathbb{R}^2$ but not $\mathbb{R}^3$.

Proposition 29. *Any basis in $\mathbb{R}^n$ has **cardinality** (i.e. number of vectors) n.*

Proof. The basis has k vectors. Suppose $k < n$. Then, the basis does not span the set $\mathbb{R}^n$ as there exists a $\mathbf{y} \in \mathbb{R}^n$ such that it cannot be represented as a linear combination of k vectors. Now suppose $k > n$. Then, the basis is not linearly independent as there exists at least one vector that is a linear combination of the others as n vectors is sufficient to span $\mathbb{R}^n$. By exhaustion, $k = n$. $\square$

Definition 23. *A basis for a subspace $\mathbb{L} \subseteq \mathbb{R}^n$ is a set of linearly independent vectors $\boldsymbol{x}_1, ..., \boldsymbol{x}_k \in \mathbb{R}^n$ such that $\forall \boldsymbol{y} \in \mathbb{L}$, $\exists \alpha_1, ..., \alpha_k \in \mathbb{R}$ such that $\boldsymbol{y} = \alpha_1 \boldsymbol{x}_1 + ... + \alpha_k \boldsymbol{x}_k = X\boldsymbol{\alpha}$.*

If and only if matrix X is full rank, the dimension of the space $\mathbb{L}$ is k (i.e. the dimension of a subspace is the number of basis vectors).

The subspace $\mathbb{L}$ is the span of the basis vectors so that $\mathbb{L} = \mathcal{C}(X)$. That is, if and only if matrix X is full rank, the span of the basis vectors is a subspace with dimension k and hence the basis vectors span $\mathbb{R}^n$.

Remember that a basis is not unique as any non-singular linear transformation of a basis is also a basis i.e. if we premultiply a basis by a non-singular matrix, the resulting matrix is also a basis.

Moreover, if matrix X has rank one less than the number of columns, then it means that the resulting linearly independent vectors span a subspace $\mathbb{L}$ with dimension $k - 1$. If there are two linearly dependent columns, the subspace has dimension $k - 2$ and so on.

Consider a real symmetric $n \times n$ matrix A with eigenvalues $\lambda_1, ..., \lambda_k$ where the eigenspace associated with λ_k is $\mathbb{L}_{\lambda_k} = \{\mathbf{x} \in \mathbb{R}^n : A\mathbf{x} = \lambda_k \mathbf{x}\}$.

For every eigenvalue λ_k, there are a number of eigenvectors captured by the subspace $\mathbb{L}_{\lambda_k}$. The dimension of $\mathbb{L}_{\lambda_k}$ is the number of repetitions of eigenvalue λ_k.

To see this, take $A = I_n$ which has 1 eigenvalue repeated n times. The eigenspace corresponding to its sole eigenvalue 1 is $\mathbb{L}_1 = \{\mathbf{x} \in \mathbb{R}^n\}$ i.e. every vector (which is by definition an eigenvector of the identity matrix). This means n basis vectors are required to span the space $\mathbb{L}_1$.

Now consider the other extreme. If all eigenvalues are distinct, then the eigenspace corresponding to a single eigenvalue λ_k is one-dimensional (just a line in $\mathbb{R}^n$). That means that the eigenspace corresponding to λ_k is spanned by just one eigenvector, which itself forms a basis.

Definition 24. *The **kernel (null space)** of a subspace $\mathbb{L} \subseteq \mathbb{R}^n$ is*

$$\mathcal{N}(\mathbb{L}) = \{\boldsymbol{u} \in \mathbb{R}^n : \langle \boldsymbol{v}, \boldsymbol{u} \rangle = 0, \ \forall \boldsymbol{v} \in \mathbb{L}\}$$

*This is also denoted $\mathbb{L}^\perp$ and called the **orthocomplement** of $\mathbb{L}$. Note that $(\mathbb{L}^\perp)^\perp = \mathbb{L}$.*

As an example, consider vectors (x, y, z). Assume $\mathbb{L}$ is formed where $x + y = 0$. For instance, $(1, -1, 0)$ and $(0, 0, 3)$. That is, we can write any vector in $\mathbb{L}$ as $(\alpha, -\alpha, \beta)$ with $\alpha, \beta \in \mathbb{R}$.

The orthocomplement $\mathbb{L}^\perp$ is formed by all vectors $(x, y, z)^T$ which are orthogonal to every vector that can be expressed as $(\alpha, -\alpha, \beta)$. We thus try to find $(x, y, z)(\alpha, -\alpha, \beta)^T = 0$ which implies that $x = y$ and $z = 0$, meaning that we can write $\mathbb{L}^\perp = \{(x, y, z) \in \mathbb{R}^3 : x = y, z = 0\}$, or vectors in general are $(\gamma, \gamma, 0)$ with $\gamma \in \mathbb{R}$.

Clearly, we can see that if we were to try to find $(\mathbb{L}^\perp)^\perp$, we would simply look for a set of vectors (x, y, z) orthogonal to $(\gamma, \gamma, 0)$, we would simply find the set of vectors constituting $\mathbb{L}$, thereby verifying $(\mathbb{L}^\perp)^\perp = \mathbb{L}$.

As another example, recall an eigenspace associated with λ_k (an eigenvalue for a real symmetric matrix A with eigenvalues $\lambda_1, ..., \lambda_k$) is $\mathbb{L}_{\lambda_k} = \{\mathbf{x} \in \mathbb{R}^n : A\mathbf{x} = \lambda_k \mathbf{x}\}$ which is a proper subspace of $\mathbb{R}^n$.

In Proposition 18, we proved that for distinct eigenvalues, their corresponding eigenvectors are mutually orthogonal. For $\mathbb{L}_{\lambda_j} = \{\mathbf{v} \in \mathbb{R}^n : A\mathbf{v} = \mu_k \mathbf{v}\}$, we know $\mathbb{L}_{\lambda_j}^\perp = \{\mathbf{y} \in \mathbb{R}^n : \langle \mathbf{y}, \mathbf{v} \rangle = 0, \quad \forall \mathbf{v} \in \mathbb{L}_{\lambda_k}\}$. Since $\mathbb{L}_{\lambda_k} = \{\mathbf{x} \in \mathbb{R}^n : A\mathbf{x} = \lambda_k \mathbf{x}\}$ is nothing but a set of all eigenvectors associated with $\lambda_k \neq \lambda_j$, it is a subset of all vectors orthogonal to $\mathbf{v}$ and hence $\mathbb{L}_{\lambda_k} \subseteq \mathbb{L}_{\lambda_j}$ for $k \neq j$. Note that $\mathbb{L}_{\lambda_k} = \mathbb{L}_{\lambda_j}$ for matrix A having only two distinct eigenvalues λ_k and λ_j.

Proposition 30. *Suppose that X is a $n \times k$ full rank matrix (rank k), then there is an orthonormal basis for $\mathcal{C}(X)$.*

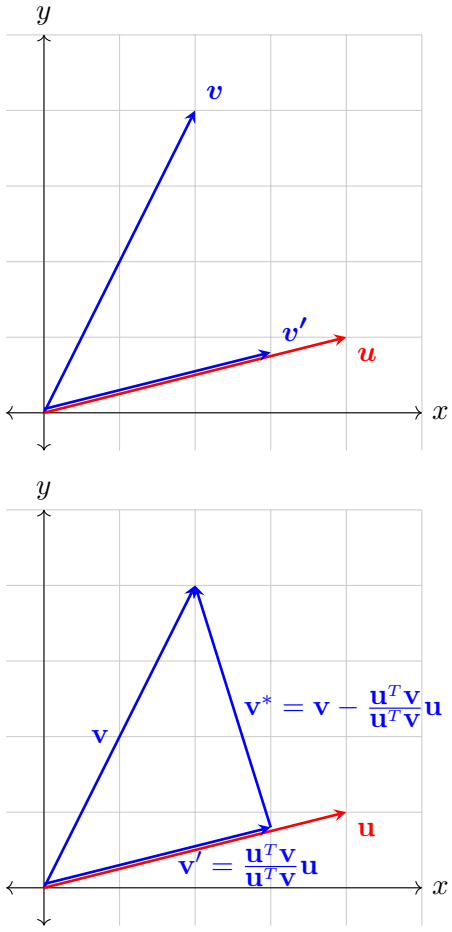
Proof. $X^T X$ is real symmetric and full rank by Proposition 26. Hence, there exists an eigendecomposition $X^T X = U \Lambda U^T$ so $U^T X^T X U = \Lambda$. Then U is a non-singular $k \times k$ matrix and we can define XU as an orthonormal basis for $\mathcal{C}(X)$. □

Notice that if we have a basis vector $\mathbf{y} \in \mathcal{C}(X)$, then $\mathbf{y} = X\mathbf{b}$ for some $\mathbf{b} \in \mathbb{R}^k$. Also, notice that $\mathbf{y} = XU(U^T \mathbf{b})$, meaning we can normalise by the length of the columns of XU to give us an orthonormal basis.

Observe that $\mathbb{L} \equiv \mathcal{C}(X)$ is a linear subspace of $\mathbb{R}^n$. The span of the linear subspace $\mathbb{L}$ and its orthocomplement $\mathbb{L}^\perp$ is the space $\mathbb{R}^n$, a result which will become clearer when we analyse projections. Therefore, we have the result which suggests that the dimension of $\mathbb{L}$ plus the dimension of its orthocomplement $\mathbb{L}^\perp$ must be the dimension of $\mathbb{R}^n$ as we must have n dimensions (i.e. n basis vectors) to span $\mathbb{R}^n$. As such, we know the dimension of $\mathbb{L}$ to be k as X is full rank, and the dimension of $\mathbb{R}^n$ is n. Hence, the dimension of the orthocomplement $\mathbb{L}^\perp$ must be $n - k$.

1.6 Projections onto a Linear Subspace

We have already talked about projections from one vector onto another. As a recap, consider two non-parallel vectors $\mathbf{u}$ and $\mathbf{v}$. A projection of $\mathbf{v}$ onto $\mathbf{u}$ (which is nothing but a projection onto the one-dimensional subspace generated by $\mathbf{u}$), denoted $\mathbf{v}' = \frac{\mathbf{u}^T\mathbf{v}}{\mathbf{u}^T\mathbf{u}}\mathbf{u}$, is parallel to $\mathbf{u}$. Once projected, this leaves the remainder of $\mathbf{v}$ as the residual $\mathbf{v}^* = \mathbf{v} - \frac{\mathbf{u}^T\mathbf{v}}{\mathbf{u}^T\mathbf{u}}\mathbf{u}$, which is orthogonal to $\mathbf{u}$ as $(\mathbf{v}^*)^T\mathbf{u} = \mathbf{v}^T\mathbf{u} - \frac{\mathbf{u}^T\mathbf{v}}{\mathbf{u}^T\mathbf{u}}\mathbf{u}^T\mathbf{u} = 0$.



However, things become different when we seek to project onto a subspace rather than a vector. Or rather, when we seek to project onto an n dimensional subspace rather than a one-dimensional subspace. In general, consider an $n \times k$ full-rank matrix $X = (\mathbf{x}_1, ..., \mathbf{x}_k)$ with $k \leq n$. The projection of a vector $\mathbf{y} \in \mathbb{R}^n$ onto $\mathcal{C}(X)$ is $\hat{\mathbf{y}} = X(X^TX)^{-1}X^T\mathbf{y}$. The projection $\hat{\mathbf{y}}$ lies in $\mathcal{C}(X)$ because $\hat{\mathbf{y}} = X\mathbf{b}$ for some $\mathbf{b} \in \mathbb{R}^k$, as we saw earlier.

Observe that any vector $\mathbf{y} - X(X^TX)^{-1}X^T\mathbf{y}$ is orthogonal to any vector $\mathbf{x}_1, ..., \mathbf{x}_k$ since we can see that $X^T(\mathbf{y} - X(X^TX)^{-1}X^T\mathbf{y}) = X^T\mathbf{y} - X^TX(X^TX)^{-1}X^T\mathbf{y} = X^T\mathbf{y} - X^T\mathbf{y} = 0$.

Definition 25. *When X is full rank, define $n \times n$ **projection matrices** $P_X = X(X^TX)^{-1}X^T$ and $M_X = I - X(X^TX)^{-1}X^T$ which project onto $\mathcal{C}(X)$ and onto the orthocomplement of $\mathcal{C}(X)$, denoted $\mathcal{C}^\perp(X)$. For any $\boldsymbol{y}$, we can uniquely write $\boldsymbol{y} = P_X\boldsymbol{y} + M_X\boldsymbol{y}$.*

We can compute that $P_XX = X(X^TX)^{-1}X^TX = X$ and $M_XX = X - XX^{-1}(X^T)^{-1}X^TX = 0$ from Definition 25.

We can also deduce that $M_XP_X\mathbf{y} = X(X^TX)^{-1}X^T\mathbf{y} - X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T\mathbf{y} = 0$.

A key characteristic of matrices P_X and M_X is that they are both symmetric and **idempotent** i.e. such that $P_X = P_X^2$ and $M_X = M_X^2$. That is, once applying the projection P_X once, we have a vector in $\mathcal{C}(X)$.

Applying the same projection twice yields the same outcome as the first time i.e. projecting a vector that is already in $\mathcal{C}(X)$ again will yield the same vector.

It follows from this that since P_X is symmetric it has an eigendecomposition $P_X = U\Lambda U^T$ and also $(P_X)^2 = (U\Lambda U^T)(U\Lambda U^T) = U\Lambda^2U^T$. The idempotent property gives us $\Lambda^2 = \Lambda$, implying $\lambda = 0$ or $\lambda = 1$. Exactly the same logic works with M_X which also has eigenvalues either 0 or 1.

Since we know that $det(A) = \prod_i \lambda_i$, we know that the determinant of all projection matrices is either 0 or 1.

Proposition 31. *If the determinant of a projection is 1, then it must be the identity matrix.*

Proof. Let a projection matrix be A and since it is idempotent, $A^2 = A$. Suppose $det(A) = 1$, meaning A is full rank and hence non-singular so A^{-1} exists. Then, $A = AAA^{-1} = A^2A^{-1} = AA^{-1} = I$. □

Hence, observe that M_X is not the identity matrix unless $X(X^TX)^{-1}X^T = 0$ which contradicts it being a projection, and hence $det(M_X) = 0$.

For P_X, if and only if X is non-singular and hence square, $X(X^TX)^{-1}X^T = XX^{-1}(X^T)^{-1}X^T = I$, in which case the determinant is 1. Otherwise (which is common since we do not know whether X is square, let alone invertible), its determinant is 0.

Recall that matrix X has columns consisting of k vectors each of which lives in $\mathbb{R}^n$. We can deduce that $tr(P_X) = tr(X(X^TX)^{-1}X^T) = tr(X^TX(X^TX)^{-1}) = tr(I_k) = \sum_i^k \lambda_i = k$, which means that k eigenvalues are equal to 1 (equally P_X has rank k) and the remaining $n-k$ eigenvalues are 0.

For M_X, we can derive the trace to be 0 in much the same way, implying that k eigenvalues are equal to 0 (i.e. M_K has rank zero) and $n-k$ are equal to 1, the opposite of P_X which is expected since M_X projects vectors onto the orthocomplement $\mathcal{C}^\perp(X)$. Our discussions above can be summarised formally as follows.

Proposition 32. ***Projection Theorem.*** *For every $\boldsymbol{y} \in \mathbb{R}^n$ and every subspace $\mathbb{L} \subseteq \mathbb{R}^n$, there exists a $\hat{\boldsymbol{y}} \in \mathbb{L}$ for which $\hat{\boldsymbol{y}} = \text{argmin}_{\boldsymbol{x}\in\mathbb{L}}||\boldsymbol{x}-\boldsymbol{y}||^2$.*

As described before, the defining property of $\hat{\mathbf{y}}$ is that the vector $\mathbf{y} - \hat{\mathbf{y}}$ be orthogonal to $\mathbb{L}$, meaning that for any $\mathbf{x} \in \mathbb{L}$, $\langle \mathbf{x}, \mathbf{y} - \hat{\mathbf{y}} \rangle = \mathbf{x}^T(\mathbf{y} - \hat{\mathbf{y}}) = \sum_{i=1}^n x_i(y_i - \hat{y}_i) = 0$.

We can rephrase the projection theorem as saying that every vector has an orthogonal decomposition with respect to some subspace and its orthocomplement.

That is, every vector $\mathbf{y}$ can be decomposed uniquely into a $\hat{\mathbf{y}} \in \mathbb{L}$ plus a $\hat{\boldsymbol{\epsilon}} \in \mathbb{L}^\perp$ such that $\mathbf{y} = \hat{\mathbf{y}} + \hat{\boldsymbol{\epsilon}} = P_X\mathbf{y} + M_X\mathbf{y}$.

This implies that every space can equally be decomposed into a subspace and the orthocomplement of the subspace, more formally described as $\mathbb{R}^n = \mathbb{L} \oplus \mathbb{L}^\perp$.

That is, $\mathbb{R}^n$ is spanned by all combinations of a subspace and the orthocomplement of that subspace. This requires that the sum of dimensions of the subspace and the orthocomplement of the subspace be equal to n, which implies $\mathbb{L}^\perp$ has dimension $n-k$ if $\mathbb{L}$ has dimension k.

Aside from using the projection formulae where we use $P_X\mathbf{y}$ to project onto $\mathcal{C}(X)$ or $M_X\mathbf{y}$ to project onto $\mathcal{C}^\perp(X)$, there is an alternative way to find the projection described in the steps below.

The method described below is specifically for projection of vector $\mathbf{y}$ onto a subspace, so we would ordinarily use P_X. This approach works whenever all vectors $\mathbf{x}_1, ..., \mathbf{x}_k$ are orthogonal to one another and we seek to find the orthogonal residual $\mathbf{y} - \hat{\mathbf{y}}$.

1. Project $\mathbf{y}$ onto $\mathbf{x}_1$ and take the residual $\mathbf{y}_1^*$.
2. Project $\mathbf{y}_1^*$ onto $\mathbf{x}_2$ and take residual $\mathbf{y}_2^*$. Note that $\mathbf{y}_1^* \perp \mathbf{x}_1$ and $\mathbf{x}_2 \perp \mathbf{x}_1$, we have $\mathbf{y}_2^* \perp \mathbf{x}_1$.

3. Repeat by projecting sequentially onto all vectors $\mathbf{x}_3, ..., \mathbf{x}_k$. The resulting residual $\mathbf{y}_k^*$ is orthogonal to all $\mathbf{x}_1, ..., \mathbf{x}_k$ and equals $\mathbf{y} - \hat{\mathbf{y}}$.

The procedure above relies on $\mathbf{x}_1, ..., \mathbf{x}_k$ being orthogonal. However, if they are not orthogonal, we can make a new orthonormal basis by taking orthogonal projections. The span of the new basis will still be $\mathcal{C}(X)$.

As an example, assume we have 2 vectors that live in $\mathbb{R}^3$ and so we have $X = (\mathbf{x}_1, \mathbf{x}_2) = \begin{pmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix}$ and $\mathbf{y} = (0 \;\; 4 \;\; 2)^T$ and we seek to find the projection of $\mathbf{y}$ onto $\mathcal{C}(X)$. First, note that matrix X has rank 2 and so is full rank. Hence, X spans $\mathbb{R}^2$.

As vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ are not orthogonal, we need to project before we can apply the method above. We consider the projection of $\mathbf{x}_2$ onto $\mathbf{x}_1$ giving us residual $\mathbf{x}_2^* = (-1 \;\; 0 \;\; 1)^T$ which is orthogonal to $\mathbf{x}_1$. Hence, we redefine matrix $X' = (\mathbf{x}_1, \mathbf{x}_2^*) = \begin{pmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}$.

Once again, the columns are linearly independent so the matrix is full rank and still spans $\mathbb{R}^2$.

Now, we can apply the method above. Start by projecting $\mathbf{y}$ onto $\mathbf{x}_1$, giving residual $\mathbf{y}_1^* = (-2 \; 2 \; 0)^T$, and then project $\mathbf{y}_1^*$ onto $\mathbf{x}_2^*$ with residual $\mathbf{y}_2^* = (-1 \; 2 \; -1)^T$. Note that $\mathbf{y}_2^*$ is orthogonal to both $\mathbf{x}_1$ and $\mathbf{x}_2^*$ as well as to $\mathbf{x}_2$.

Thus we have $\mathbf{y}_2^* = \mathbf{y} - \hat{\mathbf{y}}$ and so the projection of $\mathbf{y}$ onto $\mathcal{C}(X)$ is given by $\hat{\mathbf{y}} = \mathbf{y} - \mathbf{y}_2^* = (1 \; 2 \; 3)^T$. We would have obtained exactly the same answer if we had projected using the formula $\hat{\mathbf{y}} = P_X \mathbf{y}$.

Definition 26. ***Iterated Projection****. Suppose $\mathbb{L}_1 \subsetneq \mathbb{L} \subsetneq \mathbb{R}^n$. Then the projection of $\boldsymbol{y} \in \mathbb{R}^n$ onto $\mathbb{L}_1$ is equal to the projection of $\boldsymbol{y}$ onto $\mathbb{L}$ followed by the projection of that $\hat{\boldsymbol{y}}$ onto $\boldsymbol{L}_1$.*

As an explicit example, consider all vectors living in space $\mathbb{R}^3$. Let $\mathbb{L}_1 = \mathcal{C}(X_1) \subsetneq \mathbb{L} = \mathcal{C}(X)$ and $X_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$, $X = \begin{pmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix}$ and $\mathbf{y} = \begin{pmatrix} 0 \\ 4 \\ 2 \end{pmatrix}$. Note that X_1 has rank 1 (is full rank) so spans $\mathbb{R}$.

Iterated projection would suggest that the projection of $\mathbf{y}$ onto $\mathcal{C}(X_1)$ is equal to the projection of $\mathbf{y}$ onto $\mathcal{C}(X)$, denoted $\hat{\mathbf{y}}$, and then the projection of $\hat{\mathbf{y}}$ onto $\mathcal{C}(X_1)$, which is $(2 \;\; 2 \;\; 2)^T$.

1.7 Systems of Linear Equations

Let A be an $n \times k$ matrix and $\mathbf{y}$ be an $n \times 1$ vector. Consider the system of equations $A\mathbf{x} = \mathbf{y}$. The solution vector $\mathbf{x}$ will be $k \times 1$. However, we may have no solutions, many solutions or perhaps one unique solution. If and only if A is square and full rank, the unique solution is given by $\mathbf{x} = A^{-1}\mathbf{y}$.

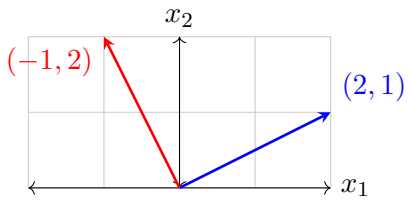
We say a system of linear equations $A\mathbf{x} = \mathbf{y}$ is **homogeneous** if and only if $\mathbf{y} = \mathbf{0}$. Suppose a vector $\mathbf{v}$ satisfies the homogeneous equation $A\mathbf{v} = \mathbf{0}$. Notice that this means $\mathbf{v} \in \mathcal{N}(A)$. That is, $\mathbf{v}$ is any vector in the kernel or null space of matrix A. To understand this further, recall that a kernel is simply the set of vectors in $\mathbb{R}^n$ which are orthogonal to every vector in the subspace. Here, we have essentially a subspace spanned by matrix A and so every vector in matrix A being orthogonal to $\mathbf{v}$ is represented by the homogeneous system $A\mathbf{v} = \mathbf{0}$.

Now suppose we seek to find a solution $\mathbf{x}$ to the system $A\mathbf{x} = \mathbf{y}$. It turns out that this solution can be written as $\mathbf{x} = \mathbf{x}' + \mathbf{x}_0$ where $\mathbf{x}'$ is any solution to the system $A\mathbf{x}' = \mathbf{y}$ and $\mathbf{x}_0 \in \mathcal{N}(A)$ is the **general solution** of the homogeneous system $A\mathbf{x}_0 = \mathbf{0}$. To verify this, consider $A(\mathbf{x}' + \mathbf{x}_0) = A\mathbf{x}' + A\mathbf{x}_0 = \mathbf{y} + \mathbf{0} = \mathbf{y}$. In reverse, we can show that any $\mathbf{x} - \mathbf{x}'$ is an element of the kernel by computing that $A(\mathbf{x} - \mathbf{x}') = A\mathbf{x} - A\mathbf{x}' = \mathbf{y} - \mathbf{y} = \mathbf{0}$. Consider a simple example where we try to find the solution to the following system of linear equations dependent on parameter $\gamma \in \mathbb{R}$:

$$2x_1 + x_2 = 3$$

$$4x_1 + 2x_2 = \gamma$$

Subtracting the first equation from the second gives $\gamma - 6 = 0$. Hence when $\gamma \neq 6$, no solution exists. With $\gamma = 6$ the first and second equations are equivalent so let's focus on $2x_1 + x_2 = 3$. One solution is $\mathbf{x}' = (0, 3)^T$. The general solution to the homogeneous system $2x_1 + x_2 = 0$ is $\mathbf{x}_0 = \alpha(-1, 2)^T$ with $\alpha \in \mathbb{R}$. Thus the overall general solution is $\mathbf{x} = (0, 3)^T + \alpha(-1, 2)^T$. We can write the equation as $\begin{pmatrix} 2 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0$, which implies that we are looking for a vector $(x_1, x_2)^T$ that is orthogonal to $(2, 1)^T$. The general solution to the homogeneous system is $\alpha(-1, 2)^T$ which is shown to be orthogonal below. Note that for any $\alpha > 1$ or $\alpha < 1$, we would have the orthogonal vector being of different magnitude but crucially pointing in the same direction, meaning it would still be orthogonal.



However, we often have more complex examples:

$$x_1 + 2x_2 + 3x_3 + 4x_4 = 4$$

$$2x_1 + 4x_2 + 4x_3 + 4x_4 = 4$$

$$3x_1 + 6x_2 + 5x_3 + 4x_4 = \gamma$$

We can write it in matrix form:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 4 & 4 \\ 3 & 6 & 5 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 4 \\ 4 \\ \gamma \end{pmatrix}$$

Performing row reduction reduces the system to:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \\ \gamma - 4 \end{pmatrix}$$

Clearly the first and second rows are linearly independent. The third has all zeros, so the matrix is rank 2. If $\gamma \neq 4$, no solution exists. If $\gamma = 4$, we must have at least one solution. Now let's focus on the case $\gamma = 4$, reducing the system to:

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \end{pmatrix}$$

Observe that during row reduction, we made elements below entry 1 in the first column and entry 1 in the third column all zero. Hence, we say that we used the first and third columns for row reduction. The largest upper triangular submatrix involving the first and third columns is $\begin{pmatrix} 1 & 3 \\ 0 & 1 \end{pmatrix}$. Note that x_1 and x_3 correspond to these columns. Hence, we find the first particular solution by simply setting $x_2, x_4 = 0$, meaning that $x_3 = 2$ and $x_1 = -2$, giving a particular solution to be $(-2, 0, 2, 0)^T$.

Next, we find the general solution to the following homogeneous system.

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Notice that the two rows of the matrix live in $\mathbb{R}^4$. The span of the two rows is a linear subspace of two dimensions since the matrix has rank 2. Therefore, the dimension of the orthocomplement of the matrix is 2. Hence, we need our general solution of the homogeneous system (which, remember, is just a subset of the null space of the matrix) to consist of the span of two linearly independent vectors orthogonal to the two rows.

To find the two vectors that constitute the general solution, first we look at setting $x_2 = 1$ and $x_4 = 0$ (recall that we did not use the corresponding columns of x_2 and x_4 in the row reduction). Hence, we obtain $(-2, 1, 0, 0)^T$ as one solution to the homogeneous system. For the other solution, we simply reverse the setup to $x_4 = 1$ and $x_2 = 0$ giving $(2, 0, -2, 1)^T$. Note that it is a general feature that the vectors constituting the general solution are orthogonal to one another.

Combining allows us to say that if $\gamma = 4$, the solution is $(-2, 0, 2, 0)^T + \alpha(-2, 1, 0, 0)^T + \beta(2, 0, -2, 1)^T$ for any $\alpha, \beta \in \mathbb{R}$. If $\gamma \neq 4$, no solutions exist.

More generally, we can summarise our procedure by saying:

1. Find out which rows of matrix A are linearly independent by performing row reduction.
2. If in the process of row reduction, a row of A is zero, the corresponding right hand side vector entry must be zero for there to be a solution. If not, no solution exists.

3. Find one solution to the system $A\mathbf{x} = \mathbf{y}$ by using the upper triangular submatrix formed from the columns used in the row reduction (i.e. columns for which all entries below a particular entry were made zero). Set all values corresponding to columns outside the submatrix to zero to obtain a particular solution $\mathbf{x}'$.

4. Next find the general solution to $A\mathbf{x}_0 = \mathbf{0}$. If matrix A has rank k and n columns, then there exist $n - k$ linearly independent vectors each of which is a solution to $A\mathbf{x} = \mathbf{0}$. The span of these $n - k$ vectors is the general solution to the homogeneous system. To find the general solution, set x_i for one column outside the submatrix to 1 and the rest to 0. Repeat for all $n - k$ columns.

5. Add $\mathbf{x}' + \mathbf{x}_0$ to obtain the solution to the system $A\mathbf{x} = \mathbf{y}$.

2 Static Optimisation

2.1 Topics in Real Analysis

Definition 27. *An ϵ-**neighbourhood** of vector $\boldsymbol{x} \in \mathbb{R}^n$ is given by $\mathcal{B}_\epsilon(\boldsymbol{x}) = \{\boldsymbol{x}' \in \mathbb{R}^n : ||\boldsymbol{x} - \boldsymbol{x}'|| < \epsilon\}$.*

Intuitively, the ϵ-neighbourhood is the set of all vectors whose distance from $\mathbf{x}$ is less than ϵ. Using the definition of norms, we can write that $\mathcal{B}_\epsilon(\mathbf{x}) = \{\mathbf{x}' \in \mathbb{R}^n : (\sum_{i=1}^n (x_i - x_i')^2)^{1/2} < \epsilon\}$.

Definition 28. *A **sequence** $\{x_k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n$ assigns to every positive integer an element $x_k \in \mathbb{R}^n$.*

We often simplify notation to $\{x_k\}$. For instance, we might write $\{x_k\}$ where $x_k = 1/k$. Since a sequence essentially maps every positive integer to an element x_k, we can see a sequence as nothing but a function whose domain is the set of natural numbers.

Definition 29. *A sequence is **bounded** if and only if $\exists r > 0$ such that $||x_k|| < r$ for all k.*

Intuitively, we can think of a bounded sequence as one in which it is possible to draw a "ball" of finite size around every element of the sequence.

Definition 30. *A **subsequence** $\{x_{k_m}\}_{m\in\mathbb{N}}$ is a sequence obtained by removing some elements of $\{x_k\}$ without changing the order of remaining elements, where m denotes the ranking in the subsequence.*

For instance, if a sequence $\{x_k\}$ is given by $x_k = 1/k$, then the sequence proceeds as $1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}...$,

while a subsequence may only apply for even k, meaning that x_{k_m} proceeds as $\frac{1}{2}, \frac{1}{4}, \frac{1}{6}...$ whilst not changing the order of $\{x_k\}$.

Definition 31. *A sequence* $\{x_k\}$ ***converges*** *to* $\boldsymbol{x} \in \mathbb{R}^n$ *if and only if* $\forall \epsilon > 0$, $\exists K \in \mathbb{N}$ *such that* $\forall k > K$, $x_k \in \mathcal{B}_\epsilon(\boldsymbol{x})$.

In other words, a sequence converges to $\mathbf{x}$ if all elements of the sequence after the Kth element lie within ϵ-distance of $\mathbf{x}$ for an arbitrarily sized ϵ. That is, if all elements beyond the Kth are such that $||x_k - \mathbf{x}|| < \epsilon$.

Often, we are interested in proving that a specific sequence converges. As an example, suppose that we are trying to prove that $\{x_k\}$, $x_k = 1/k$ converges to 0 i.e. $x_k \to 0$.

First, consider any $\epsilon > 0$. We are trying to prove that for all $k > K$, $x_k = 1/k < \epsilon$. Hence, define a K that would give precisely $1/K = \epsilon$, meaning that $K = 1/\epsilon$. Note for $k > K$, $1/k < 1/K = \epsilon$ and hence $|x_k - 0| < \epsilon$. Such a K exists for any $\epsilon > 0$ so $\{x_k\} \to 0$.

Proposition 33. *A sequence* $x_k = (x_{k1}, x_{k2}) \subset \mathbb{R}^2$ *converges to* $\boldsymbol{x} = (x_1, x_2)$ *if and only if* $x_{k1} \to x_1$ *and* $x_{k2} \to x_2$.

Proof. First, we prove the sufficiency of the statement. Assume $x_k \to \mathbf{x}$. This means $\forall \epsilon > 0$, $\exists K$ such that $\forall k > K$, $||x_k - \mathbf{x}|| < \epsilon$.

Observe that $||x_k - \mathbf{x}|| = [(x_{k1} - x_1)^2 + (x_{k2} - x_2)^2]^{1/2} \geq |x_{k1} - x_1|$ so if $||x_k - \mathbf{x}|| < \epsilon$, then it must be the case that $|x_{k1} - x_1| < \epsilon$. The same argument holds for $|x_{k2} - x_2| < \epsilon$.

Next, we prove the necessity of the statement. Now assume $x_{k1} \to x_1$ and $x_{k2} \to x_2$. This means that $\forall \epsilon' > 0$, $\exists K_1$ such that $\forall k > K_1$, $|x_{k1} - x_1| < \epsilon'$. Similarly, $\forall \epsilon' > 0$, $\exists K_2$ such that $\forall k > K_2$, $|x_{k2} - x_2| < \epsilon'$.

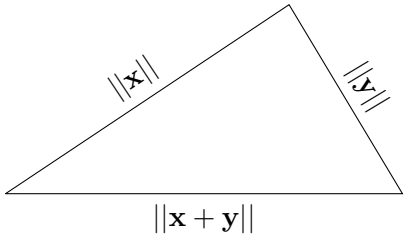
As before, $||x_k - \mathbf{x}|| = [(x_{k1} - x_1)^2 + (x_{k2} - x_2)^2]^{1/2}$ and since $|x_{k1} - x_1|, |x_{k2} - x_2| < \epsilon'$, then $||x_k - \mathbf{x}|| < \sqrt{2}\epsilon'$.

We can choose $\epsilon' = \dfrac{\epsilon}{\sqrt{2}}$ and $K_3 = \max\{K_1, K_2\}$. Hence, $\forall k > K_3$, $||x_k - \mathbf{x}|| < \epsilon$. Recall we defined ϵ' arbitrarily so without loss of generality.

We have proven both necessity and sufficiency. □

Note that the proof above can easily be extended to the case of $x_k = (x_{k1}, ..., x_{kn}) \subset \mathbb{R}^n$ converging to vector $\mathbf{x} = (x_1, ..., x_n)$ if and only if each of $x_{ki} \to x_i$ for all $i = 1, ..., n$ by recognising that $||x_k - \mathbf{x}|| = [\sum_{i=1}^{n}(x_{ki} - x_i)^2]^{1/2}$ and using the same approach as above.

Lemma 1. *The **triangle inequality** states that* $||\boldsymbol{x} + \boldsymbol{y}|| \leq ||\boldsymbol{x}|| + ||\boldsymbol{y}||$.



Lemma 1 is essentially a generalisation of the fact that $z \leq x + y$ for $z \geq x$ and $z \geq y$ where x, y and z are sides of a triangle. To verify it for the simple case of $\mathbb{R}$, note that $-|x| \leq x \leq |x|$ and $-|y| \leq y \leq |y|$. Adding gives $-(|x|+|y|) \leq x+y \leq |x|+|y|$. Since $|x+y| \leq |x|+|y| \iff -(|x|+|y|) \leq x + y \leq |x|+|y|$, we obtain Lemma 1.

Proposition 34. *Every converging sequence is bounded.*

Proof. Assume $\{x_k\} \subset \mathbb{R}^n$ is such that $x_k \to \mathbf{x}$. That means $\forall \epsilon > 0$, $\exists K$ such that $\forall k > K$, $||x_k - \mathbf{x}|| < \epsilon$.

Pick $\epsilon = 1$. Then, $\exists K$ such that $\forall k > K$, $||x_k - \mathbf{x}|| < 1$.

By Lemma 1, $||x_k|| = ||x_k - \mathbf{x} + \mathbf{x}|| \leq ||x_k - \mathbf{x}|| + ||\mathbf{x}||$. Hence, $||x_k|| < 1 + ||\mathbf{x}||$ for all $k > K$.

We have shown that $1 + ||\mathbf{x}||$ bounds $\{x_k\}$ for every $k > K$. However, for the remaining $k \leq K$, $||x_k||$ will always be no larger than the largest $||x_i||$ among all $i = 1, ..., k$.

Hence, let $B = max\{||\mathbf{x}_0||, ..., ||\mathbf{x}_k||, 1 + ||\mathbf{x}||\}$, and hence $||x_k|| \leq B$ for all $k \in \mathbb{N}$. □

The reverse statement (that a bounded sequence necessary converges) is not generally true. Consider $x_k = (-1)^k$ for which $|(-1)^k| \leq 1$ for any k. Hence, $x_k = (-1)^k$ is bounded. However, it does not converge to anything. Take $\epsilon = 1$. For any $K > 0$, we have either $x_{K+1} = -1$ or $x_{K+2} = -1$. That means either $|x_{K+1} - 1| > 1$ or $|x_{K+1} - 1| > 1$, which contradicts the definition of convergence.

Proposition 35. *If $x_k \to \boldsymbol{x}$ and $y_k \to \boldsymbol{y}$, then $x_k y_k \to \boldsymbol{xy}$.*

Proof. By Proposition 34, let $||x_k|| < r_x$ and $||y_k|| < r_y$ for all k.

By the triangle inequality, $||x_k y_k - \mathbf{xy}|| = ||x_k y_k - x_k\mathbf{y} + x_k\mathbf{y} - \mathbf{xy}|| = ||x_k(y_k - \mathbf{y}) + \mathbf{y}(x_k - \mathbf{x})|| \leq ||x_k||||y_k - \mathbf{y}|| + ||y_k||||x_k - \mathbf{x}||$.

There exists a K_X such that $\forall k > K_X$, $||x_k - \mathbf{x}|| < \dfrac{\epsilon}{2r_x}$. Similarly, $\forall k > K_y$, $||y_k - \mathbf{y}|| < \dfrac{\epsilon}{2r_y}$.

Hence, $\forall k > \max\{K_x, K_y\}$, $||x_k y_k - \mathbf{xy}|| < ||x_k|| \dfrac{\epsilon}{2r_y} + ||y|| \dfrac{\epsilon}{2r_x} < \dfrac{\epsilon}{2} + \dfrac{\epsilon}{2} = \epsilon$. □

Proposition 36. ***Bolzano-Weierstrass Theorem.*** *Every bounded sequence $\{x_k\} \subset \mathbb{R}^n$ has a converging subsequence $\{x_{k_m}\}$.*

We prove the theorem for $n = 1$. The logic is the same for $n \geq 2$ but harder to visualise.

Proof. If $\{x_k\}$ is bounded on $\mathbb{R}$, it means $\exists -r, r$ where $-r$ is a lower bound and r an upper bound.

Let an interval I_0 be a number line from $-r$ to r and $x_k \in I_0$ for all k. Then, split I_0 into 2 equally sized intervals. Without loss of generality, consider one of them denoted I_1.

I_1 contains infinitely many members of $\{x_k\}$. Hence, keep splitting intervals in half. Think about the sequences of lower bounds $\{a_m\}$ and upper bounds $\{b_m\}$ of sub-intervals. For $m = 0$, we have I_0 and so $a_m = -r$ and $b_m = r$.

Notice that $\{a_m\}$ weakly increases as m increases and $\{b_m\}$ weakly decreases as m increases. The length $|b_m - a_m| = \dfrac{2r}{2^m} \to 0$ as $m \to 0$. Hence, $\{a_m\}$ and $\{b_m\}$ converge to the same point $x \in \mathbb{R}$.

Now, pick any element x_{k_0} in interval I_0. More generally, in any interval I_m, pick an element x_{k_m} ranked after $x_{k_{m-1}}$ in the original sequence $\{x_k\}$.

That means $|x_{k_m} - x| \leq \dfrac{2r}{2^m}$ as any element of a subsequence will have distance from x no larger than the distance between the upper and lower bound of the interval, meaning $x_{k_m} \to x$. □

Definition 32. *A sequence $\{x_k\} \subset \mathbb{R}$ is a* ***Cauchy sequence*** *if and only if $\forall \epsilon > 0$, $\exists K$ such that $\forall n, m > K$, $||x_n - x_m|| < \epsilon$.*

That is, elements of a Cauchy sequence x_k become arbitrarily close to each other for large k.

Proposition 37. *A sequence $\{x_k\} \subset \mathbb{R}^n$ is converging if and only if it is Cauchy.*

Proof. We first prove the sufficiency of the statement. Assume $x_k \to \mathbf{x}$, meaning $\forall \epsilon > 0$, $\exists K$ such that $\forall k > K$, $||x_k - \mathbf{x}|| < \frac{\epsilon}{2}$. Equivalently, $\forall \epsilon > 0$, $\exists K$ such that $\forall m > K$, $||x_m - \mathbf{x}|| < \frac{\epsilon}{2}$.

Then $\forall n, m > K$, $||x_n - x_m|| = ||x_n - \mathbf{x} + \mathbf{x} - x_m|| \leq ||x_n - \mathbf{x}|| + ||\mathbf{x} - x_m|| < \epsilon$ by the triangle inequality, and so $\{x_k\}$ is Cauchy.

Next, we prove the necessity of the statement. This means we assume $\{x_k\}$ is Cauchy and prove that it converges. We do this in three stages, as shown below.

First, we prove that every Cauchy sequence is bounded i.e. that $\forall n$, $||x_n|| < C$.

Assume $\{x_k\}$ is Cauchy so $\forall \epsilon > 0$, $\exists K$ such that $\forall n, m > K$, $||x_n - x_m|| < \epsilon$. Observe that $||x_n|| = ||x_n - x_m + x_m|| \leq ||x_n - x_m|| + ||x_m||$ by the triangle inequality.

Set $\epsilon = 1$. Hence, $\exists K$ such that $\forall n, m > K$, $||x_n - x_m|| < 1$.

Set $m = K + 1$ giving $||x_n|| < 1 + ||x_{K+1}||$. Hence, $1 + ||x_{K+1}||$ is a bound for all $n > K$.

For all $n \in \mathbb{N}$, $\{x_k\}$ is bounded by $C = \max\{||x_1||, ..., ||x_K||, ||x_{K+1}|| + 1\}$, and so every Cauchy sequence is bounded.

Next, we argue that by the Bolzano-Weierstrass Theorem, every bounded sequence has a converging subsequence. Hence, every Cauchy sequence must have a converging subsequence.

We can now proceed onto the final part of the proof. Here, we argue that if a Cauchy sequence has a subsequence converging to $\mathbf{x}$, then the sequence must itself converge to $\mathbf{x}$.

Start by arguing that the existence of a converging subsequence implies $\forall \epsilon > 0$, $\exists N$ such that $\forall n > N$, $||x_{k_n} - \mathbf{x}|| < \frac{\epsilon}{2}$.

Next, the fact that $\{x_k\}$ is Cauchy implies $\forall \epsilon > 0$, $\exists K$ such that $\forall p, m > K$, $||x_p - x_m|| < \frac{\epsilon}{2}$.

Now, pick a sufficiently large $l > N$ such that $k_l > K$. That is, find a large enough element l of subsequence $\{x_{k_l}\}$ such that it converges to $\mathbf{x}$, and also make sure this choice of l is large enough so x_{k_l} is arbitrarily close to some x_p in the Cauchy sequence.

By the triangle inequality, $||x_p - \mathbf{x}|| = ||x_p - x_{k_l} + x_{k_l} - \mathbf{x}|| \leq ||x_p - x_{k_l}|| + ||x_{k_l} - \mathbf{x}|| < \epsilon$.

This implies that $\forall p > K$, $||x_p - \mathbf{x}|| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$, concluding the proof that a Cauchy sequence must converge.

We have proven both necessity and sufficiency of the statement, thus completing the proof. □

Definition 33. *A set $A \subset \mathbb{R}^n$ is **open** if $\forall \boldsymbol{x} \in A$, $\exists \epsilon > 0$ such that $\forall \boldsymbol{x}' \in \mathbb{R}^n$ such that $||\boldsymbol{x}' - \boldsymbol{x}|| < \epsilon$, one has $\boldsymbol{x}' \in A$.*

That is, for any vector in an open set, we can find an arbitrarily small neighbourhood around the vector that is also contained in the open set.

Consider the set $\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1\}$. This represents the set of all points in two-dimensional space that are within a circle of radius 1 centred around the origin, but not including the boundary.

For any vector $\mathbf{x}$ in this set, we have $||\mathbf{x}|| < 1$. Pick an $\epsilon = \frac{1 - ||\mathbf{x}||}{2}$. This means that any $\mathbf{x}'$ for which $||\mathbf{x}' - \mathbf{x}|| < \epsilon$ satisfies $||\mathbf{x}'|| \leq ||\mathbf{x}' - \mathbf{x}|| + ||\mathbf{x}||$ by the triangle inequality. Hence, we have $||\mathbf{x}'|| < \frac{1 - ||\mathbf{x}||}{2} + ||\mathbf{x}|| < 1$. This means $\mathbf{x}'$ is in the set, so the set is open.

A somewhat simpler example is the open unit interval $(0, 1)$ where we use () to indicate openness. Letting $\epsilon = \frac{\min\{x, 1 - x\}}{2}$ shows us that an open ball of such ϵ-radius lies within $(0, 1)$.

Definition 34. *A set $S \subset \mathbb{R}^n$ is **closed** if for any converging sequence $\{x_k\} \subset S$, its limit $\boldsymbol{x} \in S$.*

Intuitively, we can think of closed sets as containing their boundary. We are able to use such a definition to verify simple cases such as proving the set $\{(x_1, x_2) \in \mathbb{R}^2 : x_1 - x_2 = 1\}$ to be closed.

Assume this set to be open. That is, assume there to be a sequence $x_k = (x_{k1}, x_{k2})$ such that $x_{k1} - x_{k2} = 1$ for all k that converges to (x_1, x_2) such that $x_1 - x_2 \neq 1$. In other words, assume there to be a sequence whose limit point is not in the set.

Consider $\epsilon = |(x_1 - x_2) - 1|$. The sequence converges to (x_1, x_2) in each coordinate, meaning $\exists K$ such that $\forall k > K$, $|x_1 - x_{k1}| < \epsilon/2$ and $|x_2 - x_{k2}| < \epsilon/2$. By the triangle inequality, $|(x_1 - x_2) - 1| = |(x_1 - x_2) - (x_{k1} - x_{k2})| = |(x_1 - x_{k1}) - (x_2 - x_{k2})| \leq |x_1 - x_{k1}| + |x_2 - x_{k2}| < \epsilon$. This is a contradiction, meaning that the set is closed.

A simpler one-dimensional example occurs where we seek to prove that the unit interval $[0, 1]$

is closed, where [] is used to imply inclusion of boundary points. Again, we prove by contradiction by assuming $x_k \in [0,1]$ for all k but limit point $x \notin [0,1]$. First, consider $x > 1$. Pick $\epsilon = \dfrac{x-1}{2}$, meaning that $-\dfrac{x-1}{2} < x_k - x$ which implies $x_k > 1$, a contradiction. Then, consider limit points $x < 1$ by setting $\epsilon = \dfrac{|x|}{2}$. This means $x_k - x < \dfrac{x}{2}$ implying $x_k < 0$, also a contradiction. That means $x \in [0,1]$ and so $[0,1]$ is closed.

Proposition 38. *If $A_1, ..., A_n \subset \mathbb{R}^n$ are open, the unions of any number of A_is and the intersections of a finite number of A_is are open sets. If $B_1, ..., B_n \subset \mathbb{R}^n$ are closed, the unions of a finite number of B_is and the intersections of any number of B_is are closed sets.*

Definition 35. *For a set $A \subset \mathbb{R}^n$, if and only if its **complement** $\{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{x} \notin A\}$ is closed, then A is open. Similarly, if and only if the complement of a set is open, then the set is closed.*

Definition 35 implies that we could alternatively prove that $[0,1]$ is closed by proving that its complement $(-\infty, 0) \cup (1, \infty)$ is open. Consider the same definitions as before with $\epsilon = \dfrac{|x|}{2}$ for $x < 0$ and $\epsilon = \dfrac{x-1}{2}$ for $x > 1$ and we observe that an open ball of radius ϵ can be contained within the set.

From our definitions, note that it is possible for a set to be neither open nor closed. As an example, consider $[0,1)$ which is not open because 0 is contained in the set but setting $x = 0$ means that not all of an ϵ-neighbourhood can be contained within the set. It is not closed because we can consider a sequence with every element $x_k \in [0,1)$ with limit point 1 but 1 is not an element of the set. Alternatively, we could argue that the complement $(-\infty, 0) \cup [1, \infty)$ is neither open nor closed.

It is also possible for a set to be both open and closed. Consider $\mathbb{R}$. It is open because trivially we can pick any ϵ and deduce that an open ball around a point x is entirely contained in $\mathbb{R}$. With proving $\mathbb{R}$ is closed, we can proceed either by arguing that every limit point must by definition of a sequence in $\mathbb{R}$, or we could argue that the complement of $\mathbb{R}$ which is the empty set $\varnothing$ is trivially open as it has no elements (so no elements need to have an open ball around them for the set to be open, which is trivially true).

It is also the case that if we have a sequence $\{x_k\}$ such that $||x_{k+1} - x_k|| \to 0$, it is not true that $\{x_k\}$ is a converging sequence. That is, if the difference between neighbouring elements becomes

smaller and smaller towards zero, the sequence itself need not converge. As a simple disproof by counterexample, consider the sequence $x_k = sin\sqrt{k}$. The sequence oscillates to infinity but less often as $k \to \infty$. Therefore, $sin(\sqrt{k+1}) - sin(\sqrt{k})$ converges to zero because oscillations look almost flat for large enough k but the sequence itself continues to oscillate without convergence.

Definition 36. *A closed and bounded set is called* ***compact****.*

Proposition 39. *Any sequence $\{x_k\}$ in a compact set $S \subset \mathbb{R}^n$ has at least one converging subsequence with limit $\boldsymbol{x} \in S$.*

Proof. By the Bolzano-Weierstrass Theorem, every bounded sequence has a converging subsequence. As a compact set is bounded, such a converging subsequence must exist. Since the set is closed, the limit of the subsequence must lie in the set by definition. □

Definition 37. *For $A \subset \mathbb{R}^n$, a function $f : A \to \mathbb{R}$ is continuous if $\forall \boldsymbol{x} \in A$ and every sequence $\{x_k\} \subset A$ with $x_k \to \boldsymbol{x}$, we have $f(x_k) \to f(\boldsymbol{x})$.*

Alternative*: A function f is continuous at $\boldsymbol{x}_0 \in A$ if $\forall \epsilon > 0$, $\exists \delta > 0$ such that $\forall \boldsymbol{x}$ such that $||\boldsymbol{x} - \boldsymbol{x}_0|| < \delta$, one has $|f(\boldsymbol{x}) - f(\boldsymbol{x}_0)| < \epsilon$. A function is continuous in A if it is continuous $\forall \boldsymbol{x}_0 \in A$.*

Often, it is easier to pick one definition of continuity when undertaking proofs. For instance, consider the proof that $f(x) = x^2$ is continuous in $\mathbb{R}$. We can prove this as follows by using the alternative definition. The logic for proving a function is continuous is that we are free to select any $\epsilon > 0$ such that $|f(x') - f(x)| < \epsilon$ but for that specific ϵ, we must have a $\delta > 0$ ensuring $|x - x_0| < \delta$. Since we are free to choose ϵ, we can try one which works. Start by arguing $|f(x) - f(x_0)| = |x^2 - x_0^2| = |x - x_0||x_0 + x|$. This form is useful since we want to find a δ larger than $|x - x_0|$.

By the triangle inequality, $|x + x_0| = |x - x_0 + 2x_0| \leq |x - x_0| + 2|x_0|$. Now, we need a δ larger than $|x - x_0|$ so we can try $|x + x_0| < 1 + 2|x_0|$ as long as $\delta \leq 1$. This means that $|x - x_0||x_0 + x| \leq \delta(1 + 2|x_0|)$ which implies $\delta(1 + 2|x_0|) \leq \epsilon$ if $\delta \leq 1$. Since we now need δ to satisfy $\delta \leq \dfrac{\epsilon}{1 + 2|x_0|}$ but also be no larger than 1. Hence, the $\delta = \min\{1, \dfrac{\epsilon}{1 + 2|x_0|}\}$, proving our result as such a δ exists.

For an example of a discontinuity, consider $f(x) = 1/x$ at $x = 0$. It is discontinuous since the function is not defined. However, it is continuous on all $(0, \infty)$ and all $(-\infty, 0)$.

Note that if we have continuous functions $f, g : \mathbb{R} \to \mathbb{R}$, then $f \pm g$, fg, f/g and $f \circ g$ are also continuous where they are well-defined.

Proposition 40. ***Weierstrass Theorem.*** *Consider a continuous function $f : \mathbb{R}^n \to \mathbb{R}$ on compact set $A \subseteq \mathbb{R}^n$. Then, $\exists x^*, x^{**}$ such that $\forall \boldsymbol{x} \in A$, $f(x^*) \leq f(x) \leq f(x^{**})$.*

Proof. Let x^{**} denote the supremum (least upper bound) of f. That is, there exists a sequence $\{x_k\}$ such that $f(x_k) \to x^{**}$.

We do not know whether $\{x_k\}$ is convergent but since the set A is compact, by the Bolzano-Weierstrass Theorem, there must exist a convergent subsequence $\{x_{k_m}\}$.

Suppose that this subsequence is such that $x_{k_m} \to c \in A$. Then, it must be that $f(x_{k_m}) \to f(c)$. However, $f(x_{k_m})$ is a subsequence of $f(x_k)$, and so must converge to x^{**} (which we established as the last part of the proof of Proposition 37).

Hence, $f(c) = x^{**}$, meaning that f attains its supremum at x^{**}. The same argument can be made for f attaining its infimum at x^*, thus completing the proof. □

It is important to note that the Weierstrass Theorem is a sufficiency result. Just because a set is unbounded, not closed or the function is discontinuous doesn't mean that a maximum and minimum do not exist. Rather, the theorem just states that if we want to guarantee a maximum and minimum existing in the set, it must be a compact set and a continuous function.

2.2 The Hessian Matrix

Definition 38. *Assume $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable. The **gradient** of f at $\boldsymbol{x} = (x_1, ..., x_n) \in \mathbb{R}^n$ is row vector $\nabla f(x) = \{\frac{\partial f(\boldsymbol{x})}{\partial x_1}, ..., \frac{\partial f(\boldsymbol{x})}{\partial x_n}\}$.*

For any vector $\mathbf{x} \in \mathbb{R}^n$, we can write that

$$\nabla f(\mathbf{x}) = \lim_{h \to 0} \frac{||f(\mathbf{x}+h) - f(\mathbf{x})||}{||h||} \iff \lim_{h \to 0} \frac{||f(\mathbf{x}+h) - f(\mathbf{x}) - \nabla f(\mathbf{x})h||}{||h||} = 0$$

The gradient intuitively tells us the direction in which f increases "fastest".

For instance, assume we have $f(x_1, x_2) = 4x_1 + 3x_2$. Then, the direction of fastest increase is the direction of vector $(4, 3)^T$. Consider unit vector $(1, 0)^T$. Now, suppose we increase f in this

direction, giving us $f(\mathbf{x}+(1,0)^T)-f(x)=4(x_1+1)+3x_2-4x_1-3x_2=4$. Hence, the change in the value of the function from increasing in direction $(1,0)^T$ is 4. However, by increasing in the direction of ∇f, we need a vector proportional to $(4,3)^T$. Since $(1,0)^T$ was of unit norm, we normalise vector $(4,3)^T$ to also have unit norm. Observe that $(4,3)^T$ has norm 5 so we take $\frac{1}{5}(4,3)^T = (\frac{4}{5},\frac{3}{5})^T$. Hence, $f(\mathbf{x}+(\frac{4}{5},\frac{3}{5})^T)-f(\mathbf{x}) = 4(x_1+\frac{4}{5})+3(x_2+\frac{3}{5})-4x_1-3x_2 = 5$. We cannot move in a direction of any vector that gives a change in the function greater than 5.

Definition 39. *For $f:\mathbb{R}^n\to\mathbb{R}$, if all second-order partial derivatives exist and are continuous, then we can write them as a symmetric $n\times n$ matrix called a **Hessian Matrix** at any $\boldsymbol{x}\in\mathbb{R}^n$:*

$$\boldsymbol{H}=\begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1\partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1\partial x_n} \\ \frac{\partial^2 f}{\partial x_2\partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n\partial x_1} & \frac{\partial^2}{\partial x_n\partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

Proposition 41. ***Young's Theorem.*** *Assuming all $\dfrac{\partial^2 f}{\partial x_i\partial x_j}$ are continuous, we have $\dfrac{\partial^2 f}{\partial x_i\partial x_j} = \dfrac{\partial^2 f}{\partial x_j\partial x_i}$. This guarantees the Hessian to be symmetric.*

Young's Theorem may not hold if the second order partial derivatives are not continuous. For instance, consider $f(x,y)=\dfrac{xy(x+y)(x-y)}{x^2+y^2}$ for $(x,y)\neq(0,0)$ and $f(0,0)=0$. Then, $\dfrac{\partial f(x,y)}{\partial x} = \dfrac{x^4y+4x^2y^3-y^5}{(x^2+y^2)^2}$ for $(x,y)\neq 0$ and $\dfrac{\partial f(0,0)}{\partial x}=0$. Set $x=0$ gives $\dfrac{\partial f(0,y)}{\partial x}=-y$ and $\dfrac{\partial^2 f(0,0)}{\partial y\partial x} = -1$. However, reversing the order gives $\dfrac{\partial^2 f(0,0)}{\partial x\partial y}=1$.

Definition 40. *A set $A\subset\mathbb{R}^n$ is **convex** if and only if $\forall \boldsymbol{x},\boldsymbol{y}\in A$ and any $\alpha\in[0,1]$, $\alpha\boldsymbol{x}+(1-\alpha)\boldsymbol{y}\in A$. If the same conditions guarantee $\alpha\boldsymbol{x}+(1-\alpha)\boldsymbol{y}\in A_{int}$ where A_{int} is the interior (not the boundary) of the set A for $\alpha\in(0,1)$, then the set is **strictly convex**.*

Definition 41. *A function $A:\mathbb{R}$ is **concave** if and only if $\forall \boldsymbol{x},\boldsymbol{y}\in A$ and any $\alpha\in(0,1)$, $f(\alpha\boldsymbol{x}+(1-\alpha)\boldsymbol{y})\geq\alpha f(\boldsymbol{x})+(1-\alpha)f(\boldsymbol{y})$. For **convex** f, the inequality holds with $\leq$. Function f is **strictly concave** or **strictly convex** if the definitions hold with strict inequality.*

We can verify the convexity or concavity of functions from first principles using these definitions. Consider the set $A = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 + x_2 = 1\}$. Let (y_1, z_1) and (y_2, z_2) be in A. That means $y_1 + z_1 = 1$ and $y_2 + z_2 = 1$. Hence, $\alpha y_1 + \alpha z_1 = \alpha$ and $(1-\alpha)y_2 + (1-\alpha)z_2 = 1 - \alpha$. Adding these two gives $[\alpha y_1 + (1-\alpha)y_2] + [\alpha z_1 + (1-\alpha)z_2] = 1$, verifying that A is convex.

Also, it is important to distinguish between convex (concave) sets and convex (concave) functions. Note that while the function $f(x) = x^2$ is convex (rather, strictly convex), the set $B = \{(x, y) \in \mathbb{R}^2 : y = x^2\}$ is not convex. This can be seen by taking any two points and any α. Say for example, $(0, 0)$ and $(1, 1)$ with $\alpha = 0.5$. Clearly $(\frac{1}{2}, \frac{1}{2}) \notin B$.

Definition 42. *Function $f : A \to \mathbb{R}$ reaches its **global maximum** (respectively **global minimum**) at $\boldsymbol{x}^* \in A$ if and only if $f(\boldsymbol{x}^*) \geq f(\boldsymbol{x})$ (respectively $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x})$) for all $\boldsymbol{x} \in A$.*

Definition 43. *Function $f : A \to \mathbb{R}$ reaches its **local maximum** (respectively **local minimum**) at $\boldsymbol{x}^* \in A$ if and only if $\exists \epsilon > 0$ such that $f(\boldsymbol{x}^*) \geq f(\boldsymbol{x})$ (respectively $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x})$) for all $\boldsymbol{x} \in A$ such that $||\boldsymbol{x}^* - \boldsymbol{x}|| < \epsilon$.*

Definition 44. *A point $P \in \mathbb{R}^n$ is **stationary** for $f : \mathbb{R}^n \to \mathbb{R}$ if and only if $\nabla f(P) = \boldsymbol{0}$.*

Definition 45. *A point $P \in \mathbb{R}^n$ is **critical** for $f : \mathbb{R}^n \to \mathbb{R}$ if and only if $\nabla f(P) = \boldsymbol{0}$ or $\nabla f(P)$ does not exist. If f is differentiable, this eliminates the possibility that ∇f does not exist so a point is critical if and only if it is stationary.*

Proposition 42. *All local maxima and minima of f defined on an open set in $\mathbb{R}^n$ are critical.*

Proposition 42 critically relies on the local maxima or minima being defined on an open set. The openness of the set intuitively means it does not include its boundary, so we use this definition to eliminate the possibility of boundary solutions, which are critical because f is non-differentiable at those points.

Importantly, we must also consider the second-order conditions of an optimisation problem. For example, for a twice continuously differentiable function $f(x)$ of one variable, if $f'(x) = 0$ and $f''(x) > 0$, then x is a local minimum. If $f'(x) = 0$ and $f''(x) < 0$, then x is a local maximum. We can generalise this notion for a multivariable function $f : \mathbb{R}^n \to \mathbb{R}$.

For multivariable functions, instead of $f'(x) = 0$, we have $\nabla f(\mathbf{x}) = 0$. And instead of $f''(x) > 0$ (or < 0), we have $\mathbf{H}(\mathbf{x})$ to be positive definite (or negative definite).

From our discussion of eigenvalues and eigenvectors, we can summarise the following.

Proposition 43. *$\boldsymbol{H}(\boldsymbol{x})$ is positive definite $\iff$ All eigenvalues of $\boldsymbol{H}(\boldsymbol{x})$ are strictly positive $\implies$ $\boldsymbol{x}$ is a local minimum. $\boldsymbol{H}(\boldsymbol{x})$ is positive definite for all $\boldsymbol{x}$ $\implies$ f is strictly convex for all $\boldsymbol{x}$.*

Note that we cannot say that if f is strictly convex, $\mathbf{H}(\mathbf{x})$ must be positive definite. Consider the counterexample of $f(x) = x^4$ which is clearly strictly convex everywhere, however its Hessian at $x = 0$ is not positive definite.

Proposition 44. *$\boldsymbol{H}(\boldsymbol{x})$ is negative definite $\iff$ All eigenvalues of $\boldsymbol{H}(\boldsymbol{x})$ are strictly negative $\implies$ $\boldsymbol{x}$ is a local maximum. $\boldsymbol{H}(\boldsymbol{x})$ is negative definite for all $\boldsymbol{x}$ $\implies$ f is strictly concave for all $\boldsymbol{x}$.*

Proposition 45. *$\boldsymbol{H}(\boldsymbol{x})$ is positive semi-definite $\iff$ All eigenvalues of $\boldsymbol{H}(\boldsymbol{x})$ are non-negative $\iff$ f is convex.*

Proposition 46. *$\boldsymbol{H}(\boldsymbol{x})$ is negative semi-definite $\iff$ All eigenvalues of $\boldsymbol{H}(\boldsymbol{x})$ are non-positive $\iff$ f is concave.*

Proposition 47. *The eigenvalues of $\boldsymbol{H}(\boldsymbol{x})$ are both positive and negative $\implies$ $\boldsymbol{x}$ is a **saddle point**.*

If we encounter a positive semi-definite or negative semi-definite Hessian, we may suspect $\mathbf{x}$ to be a local minimum or maximum but we need more tests. We cannot conclude anything substantive.

Thus far, we have observed that we can look at the eigenvalues of the Hessian to categorise it as positive or negative definite or semi-definite. If we obtain definiteness, we can conclude something substantive about a point being a local maxima or minima. However, we will now show that there is (usually) a faster way to assess the definiteness of a Hessian matrix than looking at its eigenvalues. First, we define some core concepts.

Definition 46. *A **leading principal submatrix** of order k of an $n \times n$ matrix is obtained by deleting the last $n - k$ rows and columns. Note that the matrix itself is also a leading principal submatrix of order n.*

Definition 47. *A **leading principal minor** is a determinant of a leading principal submatrix.*

Proposition 48. ***Sylvester's Criterion**. A square matrix A is:*

Positive definite $\iff$ All leading principal minors are strictly positive.

Negative definite $\iff$ Odd leading principal minors are negative and even leading principal minors are positive.

Indefinite $\iff$ *One of its even leading principal minors is negative, or any two of its odd leading principal minors have different signs* $\implies$ $\boldsymbol{x}$ *is a saddle point.*

Sometimes Sylvester's Criterion tells us nothing. For instance, take any Hessian matrix with its first entry 0. Then, every leading principal minor will be zero, which is not useful. Then, looking at its eigenvalues looks like a good option.

However, if we have a function whose second partial derivatives are all zero, the Hessian will be a zero matrix. The eigenvalues in this case are all zero, giving us an inconclusive test. Similarly, Sylvester's Criterion doesn't work since all leading principal minors are zero.

2.3 Static Optimisation with Equality Constraints

Assume we have functions f, $\{g_i\}_{i=1,...,m} : \mathbb{R}^n \to \mathbb{R}$ that are twice continuously differentiable. Consider the following maximisation problem with m constraints:

$$\max_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \text{ subject to } g_i(\mathbf{x}) = b_i, \quad i = 1, ..., m$$

The **constraint set** is $C = \{\mathbf{x} \in \mathbb{R}^n : g_i(\mathbf{x}) = b_i, \quad i = 1, ..., m\}$. We say that the point $\mathbf{x}^* \in C$ is a **local constrained maximiser** if there exists an open neighbourhood $\mathcal{B}_\epsilon(\mathbf{x})$ of $\mathbf{x}^*$ such that $f(\mathbf{x}^*) \geq f(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{B}_\epsilon(\mathbf{x}) \cap C$. It is a **global constrained maximiser** if this is true for all $\mathbf{x} \in C$. Analogous definitions apply for minimisers.

First, we need to check a technical condition called the **constraint qualification (CQ)**. We say that the CQ is satisfied at $\mathbf{x} \in C$ if the matrix of gradients $D\mathbf{g}(\mathbf{x}) = \begin{pmatrix} \nabla g_1(\mathbf{x}) \\ \vdots \\ \nabla g_m(\mathbf{x}) \end{pmatrix}$ is of rank m. That is, the gradients $\nabla g_i(\mathbf{x})$ are all linearly independent, meaning that any $\nabla g_i(\mathbf{x}) \neq \mathbf{0}$.

If the CQ fails, then we may fail to find some extreme points using the **Lagrange Method**. Typically, we deal with this problem by testing the points directly.

Proposition 49. *Suppose* $\boldsymbol{x}^*$ *is a local constrained extremum and the CQ is satisfied at* $\boldsymbol{x}^*$*. Then, there exist* ***Lagrange multipliers*** $\lambda_i \in \mathbb{R}$*,* $i = 1, ..., m$ *such that*

$$\nabla f(\boldsymbol{x}^*) = \sum_{i=1}^{m} \lambda_i \nabla g_i(\boldsymbol{x}^*)$$

The theorem says that at a local constrained extremum (maximiser or minimiser), the gradient of the objective function must be a linear combination of the gradients of constraint functions.

For $\mathbf{x} \in \mathbb{R}^n$, $\boldsymbol{\lambda} \in \mathbb{R}^m$, the **Lagrangian** is given by

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) - \sum_{i=1}^{m} \lambda_i (g_i(x) - b_i)$$

Then, we say that $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ is a critical point of the Lagrangian if and only if the CQ is satisfied at $\mathbf{x}^*$ and

$$\frac{\partial \mathcal{L}}{\partial x_j} = 0, \quad j = 1, ..., n$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_i} = 0 \quad i = 1, ..., m$$

If $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ is a critical point of the Lagrangian, then it is a possible candidate for the global constrained maximiser or minimiser. Note that at points $\mathbf{x}$ where the CQ failed, we should evaluate $f(\mathbf{x})$ as the Lagrangian will not pick up these points. We can then evaluate f at all of the critical points of the Lagrangian and the greatest value(s) of f correspond to the global constrained maximiser(s). The least value(s) correspond to the global constrained minimiser(s).

Recall that when we have a continuous function on a compact set, the Weierstrass Theorem guarantees the existence of a global constrained maximiser and minimiser. Hence, it is useful to check whether the set is compact and the function is continuous prior to solving the problem.

2.4 Bordered Hessians

The Lagrangian was able to identify global constrained maximisers and minimisers providing the CQ holds. However, if we are trying to find a sufficient condition for a local constrained maximiser or minimiser, we must take a different approach.

Definition 48. *Consider the Lagrange function $\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) - \sum_{i=1}^{m} \lambda_i (g_i(\boldsymbol{x}) - b_i)$. We may test $\boldsymbol{x}^*$ for local constrained maxima or minima with the following $(m+n) \times (m+n)$ matrix called a* ***Bordered Hessian****:*

$$\boldsymbol{H\mathcal{L}}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*) = \begin{pmatrix} 0 & \dots & 0 & \frac{\partial g_1}{\partial x_1} & \dots & \frac{\partial g_1}{\partial x_n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \frac{\partial g_m}{\partial x_1} & \dots & \frac{\partial g_m}{\partial x_n} \\ \frac{\partial g_1}{\partial x_1} & \dots & \frac{\partial g_m}{\partial x_1} & \frac{\partial^2 \mathcal{L}}{\partial x_1^2} & \dots & \frac{\partial^2 \mathcal{L}}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_1}{\partial x_n} & \dots & \frac{\partial g_m}{\partial x_n} & \frac{\partial^2 \mathcal{L}}{\partial x_n \partial x_1} & \dots & \frac{\partial^2 \mathcal{L}}{\partial x_n^2} \end{pmatrix} = \begin{pmatrix} 0_m & D\boldsymbol{g} \\ (D\boldsymbol{g})^T & D^2_{\boldsymbol{x}}\mathcal{L} \end{pmatrix}$$

Denote the kth leading principal minor of $\mathbf{H}\mathcal{L}$ as $|\mathbf{H}_k|$. If we are to consider the effect on f of any infinitesimally small change $d\mathbf{x} = (dx_1, ..., dx_n)$, observe that we are bound by constraint $g_i(\mathbf{x}) = b_i$. Hence, we are only able to consider variations $d\mathbf{x}$ orthogonal to $D\mathbf{g}(\mathbf{x})$ i.e. on the space $Z = \{\mathbf{z} \in \mathbb{R}^n : (D\mathbf{g})\mathbf{z} = 0\}$.

Since definiteness of a bordered Hessian reveals some information about the convexity or concavity of a function, the notion of definiteness is intimately related to the shape of the function. Hence, when we consider the shape of f, we consider variations $d\mathbf{x}$ and so it only makes sense if notions of definiteness are defined on space Z.

Proposition 50. *Assume $D\boldsymbol{g}$ has full rank m. That is, the CQ holds. Then:*

$D^2_{\boldsymbol{x}}\mathcal{L}$ is negative definite on Z if and only if $(-1)^{k+m}|\boldsymbol{H}_k| > 0$ for every $k = 2m+1, ..., n+m$.

$D^2_{\boldsymbol{x}}\mathcal{L}$ is positive definite on Z if and only if $(-1)^m|\boldsymbol{H}_k| > 0$ for every $k = 2m+1, ..., n+m$.

If and only if there are minors $|\boldsymbol{H}_k| \neq 0$ for $k = 2m+1, ..., n+m$ whose pattern of signs contradict both positive or negative definiteness, then $D^2_{\boldsymbol{x}}\mathcal{L}$ is indefinite on Z.

Note that if we have $m = 0$, Proposition 50 is simply Sylvester's Criterion. With $m > 0$ constraints, we check only the $n - m$ largest minors.

Consider the example of finding the extrema of $f(x, y, z) = x^2 + y^2 + z^2$ subject to $z - xy = 2$.

First, the function is defined everywhere and is continuous on all (x, y, z). Next, we check if the constraint set is compact. The constraint set is closed as there are no non-boundary points on the function $z - xy = 2$. To consider boundedness, take a subspace where $z = 0$. Then, $y = -2/x$ which is unbounded, and so $z - xy = 2$ is not compact. Hence, we cannot say that a global constrained maximum or minimum must exist.

The CQ is that $[-1,-1,1] \neq [0,0,0]$ which is everywhere satisfied. Therefore, the Lagrangian will identify all the critical points of the function.

The Lagrangian is $\mathcal{L} = x^2 + y^2 + z^2 - \lambda(z - xy - 2)$. The first-order conditions are:

$$\frac{\partial \mathcal{L}}{\partial x} = 2x + \lambda y = 0$$

$$\frac{\partial \mathcal{L}}{\partial y} = 2y + \lambda x = 0$$

$$\frac{\partial \mathcal{L}}{\partial z} = 2z - \lambda = 0$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = -(z - xy - 2) = 0$$

Substitute the third condition to obtain $x + zy = 0$, $y + zx = 0$ and $z = 2 + xy$. First, if $x = 0$, we easily have critical point $(x, y, z, \lambda) = (0, 0, 2, 4)$.

If $x \neq 0$, we can divide the first equation by x and substitute the second equation to obtain $z = \pm 1$. If $z = 1$, we have $y = \pm 1$ giving $(1, -1, 1, 2)$ and $(-1, 1, 1, 2)$. If $z = -1$, we have $x = y$ which contradicts the constraint as $xy = -3$ meaning one of x or y must be positive. Hence, our critical points are $(0, 0, 2, 4)$, $(1, -1, 1, 2)$ and $(-1, 1, 1, 2)$. Observe that $f(0, 0, 2) = 4$, $f(1, -1, 1) = f(-1, 1, 1) = 3$. Since the Weierstrass Theorem did not hold, we couldn't conclude whether we have a global constrained maximum or minimum.

However, intuitively think about $z - xy = 2$. Since the objective function is increasing in all arguments, the global constrained maximum asks whether it possible to pick an (x, y, z) sufficiently large such that no larger values can be picked which satisfy $z - xy = 2$.

We are asking whether $(0, 0, 2)$ is a global constrained maximum. However, we can pick arbitrarily large numbers e.g. $(x, y, z) = (2, 499, 1000)$ which satisfy the constraint. Hence, we have no global constrained maximum.

Next, consider a ball $\mathcal{B}_r(0)$ where r is the radius such that $r^2 > 3$. Every point outside this ball would have f larger than 3. Hence, the global constrained minimum lies in this ball.

Since we know that the set $\{(x, y, z) \in \mathbb{R}^3 : \{z - xy = 2\} \cap \{x^2 + y^2 + z^2 \leq 3\}\}$ is compact and the function is continuous, the Weierstrass Theorem holds, guaranteeing us that a global constrained minimum exists. Hence, critical points $(1, -1, 1)$ and $(-1, 1, 1)$ are global constrained minima.

To assess for local maxima or minima, we use the bordered Hessian. Observe that $D^2\mathcal{L} = D^2 f - \lambda D^2 g$. Hence, we have:

$$\begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix} - \lambda \begin{pmatrix} 0 & -1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 2 & \lambda & 0 \\ \lambda & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

The matrix of constraints is given by $[-y, -x, 1]$ and so the bordered Hessian is:

$$\begin{pmatrix} 0 & -y & -x & 1 \\ -y & 2 & \lambda & 0 \\ -x & \lambda & 2 & 0 \\ 1 & 0 & 0 & 2 \end{pmatrix}$$

Now, recall the conditions noted in Proposition 50 but for $m = 1$. For a local constrained minimum, we need $|\mathbf{H}_3|, |\mathbf{H}_4| < 0$. For a local constrained maximum, we need $|\mathbf{H}_3| > 0$ and $|\mathbf{H}_4| < 0$. Our bordered Hessian for $(0, 0, 2, 4)$ is:

$$\begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 2 & 4 & 0 \\ 0 & 4 & 2 & 0 \\ 1 & 0 & 0 & 2 \end{pmatrix}$$

We don't need to compute $|\mathbf{H}_3|$ as $|\mathbf{H}_4| = 12 > 0$ which contradicts both definitions of positive and negative definiteness. Hence, we have an indefinite matrix and $(0, 0, 2, 4)$ is a saddle point.

We can repeat for the other two critical points $(1, -1, 1)$ and $(-1, 1, 1)$, yielding the result that both are local constrained minima.

2.5 The Separating Hyperplane Theorem

Definition 49. *$\boldsymbol{x} \in \mathbb{R}^n$ is a **limit point** of $A \subset \mathbb{R}^N$ if there exists a sequence $\{x_k\} \subset A$ that converges to $\boldsymbol{x}$.*

Hence, a set that is closed must contain all its limit points.

Definition 50. *A **closure** of a set A is a set $\bar{A}$ containing all limit points of A.*

If A is an open set, then $\bar{A} = A \cup A_{bdry}$ where A_{bdry} is the set of boundary points of A.

Consider a $f : \mathbb{R} \to \mathbb{R}$ defined on bounded set $X \subset \mathbb{R}$ and its closure $\bar{X}$. Now, let $Y = \{y \in \mathbb{R} : \exists x \in X : f(x) = y\}$ and $Y' = \{y \in \mathbb{R} : \exists x \in \bar{X} : f(x) = y\}$.

That is, Y represents the set of all values that f takes on X while Y' represents the set of values that f takes on $\bar{X}$.

If we have f to be continuous, then $Y = Y'$. This is because the graph of a continuous function in $\mathbb{R}$ must be closed, and so must contain all its limit points.

However, if f is discontinuous, this need not hold. To see this, consider an f that is undefined at a point $\hat{x}$ where f reaches a discontinuity. Thus, $f(\hat{x})$ does not exist so is not an element of Y. However, if $\hat{x}$ is a limit point of f, then $f(\hat{x})$ must be in Y'.

Proposition 51. ***Separating Hyperplane Theorem**. Let $A, B \subseteq \mathbb{R}^n$ be disjoint, non-empty and convex. Then, $\exists \boldsymbol{v} \in \mathbb{R}^n \backslash \boldsymbol{0}$ and $c \in \mathbb{R}$ such that $\boldsymbol{x}^T \boldsymbol{v} \geq c$ and $\boldsymbol{y}^T \boldsymbol{v} \leq c$ for all $\boldsymbol{x} \in A, \boldsymbol{y} \in B$. That is, hyperplane $\{\boldsymbol{z} \in \mathbb{R}^n : \boldsymbol{z}^T \boldsymbol{v} = c\}$ separates A and B.*

The Separating Hyperplane Theorem says that if we take any two non-empty, disjoint and convex sets, we can find a hyperplane that separates them. Think of $\mathbb{R}^2$ with two such sets. Both sets will live in $\mathbb{R}^2$ but the hyperplane will be one-dimensional i.e. just a line. In $\mathbb{R}^3$, the sets will live in three-dimensions but the hyperplane separating them will be a two-dimensional plane.

We can generalise this by saying that a hyperplane is an $n-1$ dimensional subspace of $\mathbb{R}^n$.

The main focus of this section is to establish two lemmas which are used in the proof of the Separating Hyperplane Theorem but are of interest by themselves.

Lemma 2. *Let $A \subseteq \mathbb{R}^n$ be non-empty and closed. Let $\boldsymbol{x} \in \mathbb{R}^n \backslash A$. Then, $\exists \boldsymbol{a} \in A$ such that $\forall \boldsymbol{b} \in A$, $||\boldsymbol{x} - \boldsymbol{b}|| \geq ||\boldsymbol{x} - \boldsymbol{a}||$.*

Proof. Define closed ball $B_\epsilon(\mathbf{x})$ with $B_\epsilon(\mathbf{x}) \cap A \neq \varnothing$.

Then, consider the problem $\min_{\mathbf{z}} ||\mathbf{x} - \mathbf{z}||$ on $z \in B_\epsilon(\mathbf{x}) \cap A$.

The intersection of two closed sets is closed and so $B_\epsilon(\mathbf{x}) \cap A$ is compact.

By the Weierstrass Theorem, there exists a global constrained minimum $\mathbf{a} \in B_\epsilon(\mathbf{x}) \cap A$. $\square$

Lemma 3. *The point $\boldsymbol{a} \in A$ established in Lemma 2 is the unique closest point in A to $\boldsymbol{x}$.*

Proof. Proceed by contradiction.

Assume $\exists \mathbf{b} \in A$ such that $||\mathbf{x}-\mathbf{a}||=||\mathbf{x}-\mathbf{b}|| \leq ||\mathbf{x}-\mathbf{c}||$ for all $\mathbf{c} \in A$.

Since A is convex, $\dfrac{\mathbf{a}+\mathbf{b}}{2} \in A$.

By the triangle inequality, $||\mathbf{x} - \dfrac{\mathbf{a}+\mathbf{b}}{2}|| = \dfrac{1}{2}||\mathbf{x}-\mathbf{a}+\mathbf{x}-\mathbf{b}|| \leq \dfrac{1}{2}||\mathbf{x}-\mathbf{a}|| + \dfrac{1}{2}||\mathbf{x}-\mathbf{b}||$.

Since $\mathbf{a}$ and $\mathbf{b}$ are the closest points to $\mathbf{x}$, the inequality above cannot be strict otherwise we have a contradiction.

Hence, $||\mathbf{x}-\mathbf{a}|| = \mu||\mathbf{x}-\mathbf{b}||$ for $|\mu| = 1$.

If $\mu = -1$, then $\mathbf{x}-\mathbf{a} = -\mathbf{x}+\mathbf{b} \implies \mathbf{x} = \dfrac{\mathbf{a}+\mathbf{b}}{2} \in A$ which is a contradiction.

Hence, $\mu = 1$ and so $\mathbf{a} = \mathbf{b}$ so $\mathbf{a}$ is the unique closest point. □

There is an additional lemma (quite involved) which proves that a separating hyperplane exists to separate a vector and a subset of $\mathbb{R}^n$. The statement and proof are in the footnotes.[8]

---

[8]

Lemma 4. *Let $A \subseteq \mathbb{R}^n$ be non-empty and closed. Let $\boldsymbol{x} \in \mathbb{R}^n \backslash A$. Then, there exists a separating hyperplane (that is, a vector $\boldsymbol{v} \in \mathbb{R}^n \backslash \boldsymbol{0}$ and scalar $c \in \mathbb{R}$ such that $\boldsymbol{x}^T\boldsymbol{v} < c$ and $\forall \boldsymbol{z} \in A$, $\boldsymbol{z}^T\boldsymbol{v} \geq c$).*

Proof. Let $\mathbf{a}$ minimise $||\mathbf{x}-\mathbf{z}||$ on $\mathbf{z} \in B_\epsilon(\mathbf{x}) \cap A$. Now let $\mathbf{v} = \mathbf{a}-\mathbf{x}$ and $c = \mathbf{a}^T\mathbf{v}$. Observe that $\mathbf{x}^T\mathbf{v} = (\mathbf{a}-\mathbf{v})^T\mathbf{v} = \mathbf{a}^T\mathbf{v} - \mathbf{v}^T\mathbf{v} = c - ||\mathbf{v}||^2 < c$, completing the first part.

Consider any $\mathbf{y} \in A$. As A is convex, the set $\{\mathbf{z} : \mathbf{z} = t\mathbf{y} + (1-t)\mathbf{a}\}$ for $t \in (0,1)$ lives in A. Let $\mathbf{a}$ is the closest element of A to $\mathbf{x} \notin A$. So for all $t > 0$, $||\mathbf{a}-\mathbf{x}||^2 \leq ||t\mathbf{y} + (1-t)\mathbf{a} - \mathbf{x}||^2 = ||t(\mathbf{y}-\mathbf{x}) + (1-t)(\mathbf{a}-\mathbf{x})||^2$.

We have $||t(\mathbf{y}-\mathbf{x}) + (1-t)(\mathbf{a}-\mathbf{x})||^2 = \langle t(\mathbf{y}-\mathbf{x}) + (1-t)(\mathbf{a}-\mathbf{x}), t(\mathbf{y}-\mathbf{x}) + (1-t)(\mathbf{a}-\mathbf{x})\rangle = \langle t(\mathbf{y}-\mathbf{x}), t(\mathbf{y}-\mathbf{x})\rangle + 2\langle t(\mathbf{y}-\mathbf{x}), (1-t)(\mathbf{a}-\mathbf{x})\rangle + \langle (1-t)(\mathbf{a}-\mathbf{x}), (1-t)(\mathbf{a}-\mathbf{x})\rangle$.

Thus, $||t(\mathbf{y}-\mathbf{x}) + (1-t)(\mathbf{a}-\mathbf{x})||^2 = t^2||\mathbf{y}-\mathbf{x}||^2 + 2t(1-t)(\mathbf{y}-\mathbf{x})^T(\mathbf{a}-\mathbf{x}) + (1-t)^2||\mathbf{a}-\mathbf{x}||^2$.

Hence, $||\mathbf{a}-\mathbf{x}||^2 \leq t^2||\mathbf{y}-\mathbf{x}||^2 + 2t(1-t)(\mathbf{y}-\mathbf{x})^T(\mathbf{a}-\mathbf{x}) + (1-t)^2||\mathbf{a}-\mathbf{x}||^2$.

Subtract $||\mathbf{a}-\mathbf{x}||^2$ and divide by t to obtain $0 \leq t||\mathbf{y}-\mathbf{x}||^2 + 2(1-t)(\mathbf{a}-\mathbf{x})^T(\mathbf{y}-\mathbf{x}) - (2-t)||\mathbf{a}-\mathbf{x}||^2$.

As $t \to 0$, we get $0 \leq 2(\mathbf{a}-\mathbf{x})^T(\mathbf{y}-\mathbf{x}) - 2||\mathbf{a}-\mathbf{x}||^2 = 2[(\mathbf{a}-\mathbf{x})^T\mathbf{y} - (\mathbf{a}-\mathbf{x})^T\mathbf{a}] = 2[\mathbf{v}^T\mathbf{y} - c]$.

Thus for all $\mathbf{y} \in A$, $\mathbf{y}^T\mathbf{v} \geq c$, completing the second part.

We have proven both parts of the statement so the proof is complete. □

Using the lemmas we have established, the final proof of the Separating Hyperplane Theorem (again, quite involved) is in the footnotes.[9]

2.6 Static Optimisation with Inequality Constraints

Assume functions $f,\ \{g_i\}_{i=1,...,m} : \mathbb{R}^n \to \mathbb{R}$ are twice continuously differentiable. Consider the following maximisation problem with m constraints:

$$\max_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \ \text{ subject to } \ g_i(x) \leq b_i, \ \ i = 1, ..., m$$

The constraint set is given by $C = \{\mathbf{x} \in \mathbb{R}^n : g_i(\mathbf{x}) \leq b_i, i = 1, ..., m\}$. The point $\mathbf{x}^*$ is a local constrained maximiser if there exists an open neighbourhood $\mathcal{B}_\epsilon(\mathbf{x})$ of $\mathbf{x}^*$ such that $f(\mathbf{x}^*) \geq f(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{B}_\epsilon(\mathbf{x}) \cap C$. It is a global constrained maximiser if this is true for all $\mathbf{x} \in C$. Again, we have analogous definitions for minimisers.

We call the constraint $g_i(\mathbf{x})$ **binding** if and only if it holds at equality and **non-binding** if it holds as strict inequality. The CQ that we discussed from before is now modified such that it holds whenever the binding constraints at $\mathbf{x}^*$ are linearly independent.

Once again, if the CQ fails at specific points, we must check those points directly as the Lagrangian will not identify them.

---

[9]

Proof of Separating Hyperplane Theorem. Given A and B being disjoint, nonempty and convex, let $K = A + (-B) = \{\mathbf{x} - \mathbf{y} : \mathbf{x} \in A, \mathbf{y} \in B\}$ be the **Minkowski sum** of the two sets.

For $\mathbf{l}_1, \mathbf{l}_2 \in K$, let $\mathbf{x}, \mathbf{x}' \in A$ and $\mathbf{y}, \mathbf{y}' \in B$ such that $\mathbf{l}_1 = \mathbf{x} + \mathbf{x}'$ and $\mathbf{l}_2 = \mathbf{y} + \mathbf{y}'$. We know $t\mathbf{l}_1 + (1-t)\mathbf{l}_2 = t(\mathbf{x} + \mathbf{x}') + (1-t)(\mathbf{y} + \mathbf{y}') = (t\mathbf{x} + (1-t)\mathbf{y}) + (t\mathbf{x}' + (1-t)\mathbf{y}') \in K$. Hence, K is convex so $\bar{K}$ (the closure) is closed and convex.

By Lemmas 2 and 3, there exists a unique point in a convex set that is the closest point to any point outside the set. Define $\mathbf{a} \in A$ as that closest point for set A and $\mathbf{x} \in B$ to be that point for set B. Let $\mathbf{v} = \mathbf{a} - \mathbf{x} \in K$.

By Lemma 4, there is a separating hyperplane between any point outside a convex set and the set itself. Thus by extension a separating hyperplane separates A and B. So $\mathbf{v}$ has unique minimum norm of any point in K.

Since $\bar{K}$ is convex, for any $\mathbf{n} \in K$, line segment $\mathbf{v} + t(\mathbf{n} - \mathbf{v})$ lies in $\bar{K}$ for all $t \in [0, 1]$. Hence, $||\mathbf{v}||^2 \leq ||\mathbf{v} + t(\mathbf{n} - \mathbf{v})||^2 = ||\mathbf{v}||^2 + 2t\langle \mathbf{v}, \mathbf{n} - \mathbf{v}\rangle + t^2||\mathbf{n} - \mathbf{v}||^2$.

For $t \in (0, 1]$, we may manipulate the above expression similarly to the proof of Lemma 4 to obtain $0 \leq 2\langle \mathbf{v}, \mathbf{n} - \mathbf{v}\rangle + t||\mathbf{n} - \mathbf{v}||^2$. Taking $t \to 0$ gives $\langle \mathbf{n}, \mathbf{v}\rangle \geq ||\mathbf{v}||^2$. Since $\mathbf{n} \in K$, we have $\langle \mathbf{x} - \mathbf{y}, \mathbf{v}\rangle \geq ||\mathbf{v}||^2$.

Since we require $\mathbf{v} \neq \mathbf{0}$, the proof is complete because $\inf_{\mathbf{x}\in A} \langle \mathbf{x}, \mathbf{v}\rangle \geq ||\mathbf{v}||^2 + \sup_{\mathbf{y}\in B} \langle \mathbf{y}, \mathbf{v}\rangle$. □

Proposition 52. *Suppose $\boldsymbol{x}^*$ is a local constrained maximiser and the CQ holds at $\boldsymbol{x}^*$. Then, there exist Lagrange multipliers $\lambda_i \in \mathbb{R}$ for $i = 1, ..., m$ such that*

$$\nabla f(\boldsymbol{x}^*) = \sum_{i=1}^{m} \lambda_i \nabla g_i(\boldsymbol{x}^*)$$

where $\lambda_i \geq 0$ and $\lambda_i(b_i - g_i(\boldsymbol{x}^*)) = 0$ for $i = 1, ..., m$.

The expression $\lambda_i(b_i - g_i(\mathbf{x}^*)) = 0$ is called a **complementary slackness** condition and implies that $\lambda_i = 0$ if the constraint is non-binding.

It is important to note that a binding constraint could also have $\lambda_i = 0$. Consider the intuition as follows. Suppose $\mathbf{x}^*$ is an unconstrained maximiser. Then, impose $g(\mathbf{x}) \leq b_i$ such that $g(\mathbf{x}^*) = b_i$ i.e. the constraint binds at the unconstrained maximiser. The associated multiplier $\lambda^* = 0$ since the unconstrained objective function is maximised by $\mathbf{x}^*$ and the constraint makes no difference.

If we seek to solve a minimisation problem, this is the same as maximising $-f$ and so the conditions are the same except $\lambda_i \leq 0$.

We can write the Lagrangian as before:

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) - \sum_{i=1}^{m} \lambda_i (g_i(\mathbf{x}) - b_i)$$

We can solve for the following first-order conditions, which are called **Kuhn-Tucker Conditions**:

$$\frac{\partial \mathcal{L}}{\partial x_j} = 0, \quad j = 1, ..., n$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_i} \geq 0, \quad \lambda_i \geq 0, \quad \lambda_i(b_i - g_i(\mathbf{x})) = 0 \quad \text{for all} \quad i = 1, ..., m$$

When we solve constrained optimisation problems with inequality constraints, it is again useful to see if the Weierstrass Theorem applies. That is, we should check whether we have a continuous function on a compact domain. We know how to check for continuity of functions. Typically, we can do this by observation if the function is defined across the domain, rather than by rigorous proof. In the case of compactness, we need to check two criteria: closedness and boundedness.

When we worked with equality constraints, closedness was easy to check as any graph of a continuous function in $\mathbb{R}^n$ must be closed as it solely consists of its set of boundary points.

However, now, we are dealing with inequalities and often spaces rather than graphs of continuous functions. Hence, we need to check for closedness more carefully.

Proposition 53. *Assume $g(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$ is continuous and exists for all $\boldsymbol{y} \in D \subset \mathbb{R}^n$ for closed set D. Then, the set $\{\boldsymbol{x} : g(\boldsymbol{x}) \leq b\} \cap D$ is closed for any scalar $b \in \mathbb{R}$.*

Proof. Consider a sequence $\{x_k\} \in \{\mathbf{x} : g(\mathbf{x}) \leq b\} \cap D$ converging to $\mathbf{x}^* \in \mathbb{R}^n$. Since D is closed, its limit $\mathbf{x}^* \in D$. Since g is continuous and the constraint holds for any x_k, $g(\mathbf{x}^*) \leq b$. Hence, $\mathbf{x}^* \in \{\mathbf{x} : g(\mathbf{x}) \leq b\} \cap D$. This holds for any limit point $\mathbf{x}^*$ and so the proof is complete. □

Proposition 53 argues that we can restrict our attention to a subset. For instance, we may define D to be a closed ball in $\mathbb{R}^n$ if we know that the larger space is not compact. This helps us establish boundedness and closedness, meaning we can artificially construct compact subspaces by ignoring the areas where g_i is not defined or f clearly doesn't reach the desired extremum.

Also note that if constraint set C is determined by several constraints, each of which determines a closed set, then the intersection of these constraints is also closed. However, we cannot perform illegal operations on g_1 in space D. This means we cannot divide by zero or take logarithms of negative numbers, for example.

In many optimisation problems, we have **non-negativity constraints** which do not allow for negative values to be taken. If we were to use the standard Kuhn-Tucker conditions, we would have multiple constraints and multiple Lagrange multipliers, complicating the analysis. However, we can modify the Lagrangian with a simple trick as follows.

Suppose we have $x_j \geq 0$, which is equivalent to $-x_j \leq 0$.

Proposition 52 gives us $\dfrac{\partial f}{\partial x_j} = \sum_{i \neq j} \lambda_i \dfrac{\partial g_i}{\partial x_j} - \lambda_j$ for $\lambda_j \geq 0$ and $\lambda_j \dfrac{\partial \mathcal{L}}{\partial \lambda_j} = 0$.

Rewriting gives $\dfrac{\partial f}{\partial x_j} \leq \sum_{i \neq j} \lambda_i \dfrac{\partial g_i}{\partial x_j}$ with equality if $x_j > 0$ i.e. $\lambda_j = 0$ if the non-negativity constraint is non-binding.

Hence, we write the **modified Lagrangian** as

$$\mathcal{L} = f(\mathbf{x}) - \sum_{i=1}^{m} \lambda_i (g_i(\mathbf{x}) - b_i)$$

$$\frac{\partial \mathcal{L}}{\partial x_j} \leq 0 \quad (= 0 \quad \text{if} \quad x_j > 0)$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_i} \geq 0, \quad \lambda_i \geq 0, \quad \lambda_i(b_i - g_i(\mathbf{x})) = 0 \quad \text{for all} \quad i = 1, ..., m$$

Note that in the event that we have a minimisation problem as opposed to a maximisation problem, we reverse the sign of the multiplier to $\lambda_i \leq 0$ and the signs of the $\dfrac{\partial \mathcal{L}}{\partial x_j} \geq 0$ with equality if $x_j > 0$. This follows from the fact that minimisation is simply maximising $-f$.

Consider an example of finding the extrema of $f(x, y) = xy$ subject to $2x^2 + y^2 \leq b$ for $b \geq 0$ and $x, y \geq 0$. First, xy is continuous and $2x^2 + y^2 \leq b$ is compact. Restricting to $x, y \geq 0$ also gives a compact set. Hence, the Weierstrass Theorem holds and we are guaranteed the existence of a global constrained maximum and minimum.

Next, we check the CQ. We require $\begin{pmatrix} 4x & 2y \\ -1 & 0 \\ 0 & -1 \end{pmatrix}$ to be such that the gradients of binding constraints are linearly independent. We will come back to this once we consider the Lagrangian.

Writing the modified Lagrangian gives us

$$\mathcal{L} = xy - \lambda(2x^2 + y^2 - b)$$

Our Kuhn-Tucker conditions for maximisation are

$$\mathcal{L}_x = y - 4\lambda x \leq 0 \ (= 0 \ \text{if} \ x > 0)$$

$$\mathcal{L}_y = x - 2\lambda y \leq 0 \ (= 0 \ \text{if} \ y > 0)$$

$$\mathcal{L}_\lambda = b - 2x^2 - y^2 \geq 0, \quad \lambda \geq 0, \quad \lambda(b - 2x^2 - y^2) = 0$$

First, consider $\lambda = 0$. That means $y = 0$ and $x = 0$ giving $(0, 0, 0)$, feasible for $b \geq 0$.

Next, consider $\lambda > 0$, which means $2x^2 + y^2 = b$. If $x = 0$, then $y = 0$, only possible with $b = 0$ (same as if $y = 0$ implying $x = 0$). That is, if $b = 0$, then we may have $(0, 0, \lambda)$ where $\lambda > 0$.

If $x, y > 0$, that means $y^2 = 2x^2$ and so $x = \pm\dfrac{\sqrt{b}}{2}$ and $y = \pm\dfrac{\sqrt{2b}}{2}$ giving four possible solutions

$(\frac{\sqrt{b}}{2}, \frac{\sqrt{2b}}{2}, \frac{\sqrt{2}}{4})$, $(\frac{\sqrt{b}}{2}, -\frac{\sqrt{2b}}{2}, -\frac{\sqrt{2}}{4})$, $(-\frac{\sqrt{b}}{2}, \frac{\sqrt{2b}}{2}, -\frac{\sqrt{2}}{4})$ and $(-\frac{\sqrt{b}}{2}, -\frac{\sqrt{2b}}{2}, \frac{\sqrt{2}}{4})$. When $b = 0$, all of them are possible giving $(0, 0, \frac{\sqrt{2}}{4})$ and $(0, 0, -\frac{\sqrt{2}}{4})$. When $b > 0$, only $(\frac{\sqrt{b}}{2}, \frac{\sqrt{2b}}{2}, \frac{\sqrt{2}}{4})$ is possible.

We now check whether we are missing any solutions because the CQ did not hold. Recall the gradient matrix $\begin{pmatrix} 4x & 2y \\ -1 & 0 \\ 0 & -1 \end{pmatrix}$ needed to be such that the gradients of binding constraints were linearly independent. This is only violated when $2x^2 + y^2 = b$ but $x = y = 0$, which requires $b = 0$. However, we've already dealt with this case in the critical points we found.

Testing the value of f at each critical point clearly gives us that $(\frac{\sqrt{b}}{2}, \frac{\sqrt{2b}}{2})$ is the global constrained maximum.

For the minimisation problem, we solve

$$y - 4\lambda x \geq 0 \ (= 0 \text{ if } x > 0)$$

$$x - 2\lambda y \geq 0 \ (= 0 \text{ if } y > 0)$$

$$b - 2x^2 - y^2 \geq 0, \quad \lambda \leq 0, \quad \lambda(b - 2x^2 - y^2) = 0$$

First, consider $\lambda < 0$ which requires $2x^2 + y^2 = b$. If $x = 0$, then $y = 0$ and if $y = 0$, then $x = 0$. Both such cases are contradictions unless $b = 0$ and so $(0, 0, \lambda)$ for $\lambda < 0$ holds only when $b = 0$.

If $y = 0$, then $x = \pm\frac{\sqrt{b}}{2}$. If $b = 0$, this reduces to $(0, 0, \lambda)$ for $\lambda < 0$. If $b > 0$, $x = \frac{\sqrt{b}}{2}$ implying $y = 4\lambda x = 0$ suggesting $\lambda = 0$ which is a contradiction. The same applies if $x = 0$.

If $x, y > 0$, then $y^2 = 2x^2$ and so $x = \pm\frac{\sqrt{b}}{2}$. If $b = 0$, we are reduced to $x = 0$ which implies $y = 0$, a contradiction. If $b > 0$, $x = \frac{\sqrt{b}}{2}$ giving $y = 2\lambda\sqrt{b}$. Since $\lambda < 0$, $y < 0$ which is a contradiction.

Hence, we have exhausted all possibilities with $\lambda < 0$ and found one critical point at $(0, 0, \lambda)$ which works when $b = 0$ only.

Then, consider $\lambda = 0$. Then, if $x = y = 0$, then we have a critical point $(0, 0, 0)$ for all $b \geq 0$.

If $x = 0, y > 0$, $2\lambda y = 0$ meaning $\lambda = 0$ and hence $(0, y, 0)$ for $y > 0$ is possible for all $b \geq 0$. Also, $(x, 0, 0)$ for $x > 0$ is possible for $b \geq 0$.

Finally, if $x, y > 0$, then $y = 4\lambda x$ and $x = 2\lambda y$. As $\lambda = 0$, $x = y = 0$ which is a contradiction.

As such, we have found that if $b = 0$, we have the critical points $(0, 0, \lambda)$ for any $\lambda < 0$, $(0, 0, 0)$, $(x, 0, 0)$ for $x > 0$ and $(0, y, 0)$ for $y > 0$.

If however $b > 0$, then we have the latter three: $(0, 0, 0)$, $(x, 0, 0)$ for $x > 0$ and $(0, y, 0)$ for $y > 0$.

Note that since $x, y > 0$ is always a contradiction for the minimisation problem, the function takes value f equal to zero at all of these points, giving us infinite number of global constrained minima.

In any case, if we seek to check for local maxima or minima, we must use the bordered Hessian with the binding constraints only.

Proposition 54. ***Theorem of Multipliers***. *Assume $\boldsymbol{b} = (b_1, ..., b_m) \in \mathbb{R}^m$ is such that the problem $\max_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x})$ subject to $g_i(\boldsymbol{x}) \leq b_i$ for all $i = 1, ..., m$ has a solution $\boldsymbol{x}(\boldsymbol{b})$. The value function of this solution is $F(\boldsymbol{b}) \equiv f(\boldsymbol{x}(\boldsymbol{b}))$ where F is continuous.*

Suppose in an open neighbourhood of $b \in \mathbb{R}^m$, the set of binding constraints is unaltered and $F(b)$ is differentiable. Then, for every $i = 1, ..., m$, $\dfrac{\partial F(\boldsymbol{b})}{\partial b_i} = \lambda_i$.

Think about Proposition 54 with the example of the maximisation problem of $\max_{x,y} f(x, y) = xy$ subject to $2x^2 + y^2 \leq b$ for $b \geq 0$ and $x, y \geq 0$. We obtained from the Kuhn-Tucker conditions that $F(b) = \dfrac{b\sqrt{2}}{4}$, giving us that $\lambda = \dfrac{\sqrt{2}}{4}$. Using the theorem of multipliers, we could have just directly looked at $\lambda = \dfrac{\sqrt{2}}{4}$ to get the same answer.

We may generalise the theorem of multipliers to cases where we optimise f which depends on parameters $\mathbf{q} = (q_1, ..., q_s) \in \mathbb{R}^s$.

That is, we have the problem of optimising $f(\mathbf{x}, \mathbf{q})$ subject to $g_i(\mathbf{x}, \mathbf{q}) \leq b_i$ for $i = 1, ..., m$. Denote the value function $F(\mathbf{q}) \equiv f(\mathbf{x}(\mathbf{q}), \mathbf{q})$. That is, $\mathbf{x}(\mathbf{q})$ is the constraint maximiser of the problem when the set of parameters is $\mathbf{q}$.

Assume F is well-defined around some $\bar{\mathbf{q}}$ and we are interested in how F changes as $\mathbf{q}$ varies in the neighbourhood of $\bar{\mathbf{q}}$.

Proposition 55. ***Envelope Theorem.*** *Assume F is differentiable at $\bar{\boldsymbol{q}}$ and $\lambda_1, ..., \lambda_m \in \mathbb{R}$ are Lagrange multipliers associated with $\boldsymbol{x}(\bar{\boldsymbol{q}})$. Then*

$$\frac{\partial F(\bar{\boldsymbol{q}})}{\partial q_j} = \frac{\partial f(\boldsymbol{x}(\bar{\boldsymbol{q}}), \bar{\boldsymbol{q}})}{\partial q_j} - \sum_{i=1}^{m} \lambda_i \frac{\partial g_i(\boldsymbol{x}(\bar{\boldsymbol{q}}), \bar{\boldsymbol{q}})}{\partial q_j}, \quad \textit{for} \;\; j = 1, ..., s$$

Proof. We prove the case of $n = m = s = 1$. The logic is the same for $n, m, s > 1$. The chain rule gives $\dfrac{\partial F(\bar{q})}{\partial q} = \dfrac{\partial f(x(\bar{q}), \bar{q})}{\partial x}\dfrac{\partial x(\bar{q})}{\partial q} + \dfrac{\partial f(x(\bar{q}), \bar{q})}{\partial q}$. The Lagrange method is such that $\dfrac{\partial f(x(\bar{q}), \bar{q})}{\partial x} = \lambda \dfrac{\partial g(x(\bar{q}), \bar{q})}{\partial x}$.

If g is non-binding, $\lambda = 0$ giving the result $\dfrac{\partial F(\bar{q})}{\partial q} = \dfrac{\partial f(x(\bar{q}), \bar{q})}{\partial q}$.

If g binds at $\mathbf{x}$, then $\dfrac{dg(x(\bar{q}), \bar{q})}{dx} = \dfrac{\partial g(x(\bar{q}), \bar{q})}{\partial x}\dfrac{dx(\bar{q})}{dq} + \dfrac{\partial g(x(\bar{q}), \bar{q})}{\partial q} = 0$ meaning $\dfrac{\partial f(x(\bar{q}), \bar{q})}{\partial x} = \lambda \dfrac{\partial g(x(\bar{q}), \bar{q})}{\partial x} = 0$ and so $\dfrac{\partial F(\bar{q})}{\partial q} = \dfrac{\partial f(x(\bar{q}), \bar{q})}{\partial q}$. □

In the context of constrained optimisation problems, the value function F is nothing but the Lagrangian. Hence, by applying the Envelope Theorem to our example of $\max_{x,y} f(x, y) = xy$ subject to $2x^2 + y^2 \leq b$ for $b \geq 0$ and $x, y \geq 0$, observe that $\dfrac{\partial \mathcal{L}}{\partial b} = \lambda$. Computing the value function directly gave us $F(b) = \dfrac{b\sqrt{2}}{4}$, whose derivative is λ. In fact, our example is such that f and g are independent of any parameters, and as such, we are left with the case of the theorem of multipliers.

2.7 Application: Regulation

Suppose a regulator R delegates the production of q units of a good to firm f. The value to R is $S(q)$ with $S' > 0$ but $S'' < 0$ and $S(0) = 0$.

The production cost of f is unobservable by R but it is common knowledge that the marginal cost $\theta \in \{\theta_1, \theta_2\}$ with $\theta_2 - \theta_1 > 0$.

Total cost for firm i is $C_i(q) = \theta_i q$ for $i = \{1, 2\}$ where 1 denotes the efficient firm and 2 the inefficient firm. The marginal cost is knowable to f.

First, we assume that R knows θ. This is the **first-best** case.

f obtains compensation t_1 or t_2 for producing quantity q_1 or q_2 conditional on marginal cost parameter θ_1 or θ_2.

R solves for each i:

$$\max_{q_i,t_i} S(q_i) - t_i \text{ subject to } t_i - C_i(q_i) \geq 0$$

The constraint in this problem is called a **participation constraint**. The firm can get zero utility by not participating so the utility the firm receives must be weakly greater than zero.

Note that we do not require the Lagrangian to solve this problem. The constraints must bind. Observe that if the constraint was non-binding, we could reduce t_i feasibly while raising $S(q_i) - t_i$.

Substituting for the constraint gives $\max_{q_i} S(q_i) - C_i(q_i)$ with first-order condition $S'(q_i^*) = \theta_i$. We may solve for q_i^* and check that the utility of firm i is exactly zero.

If we denote the social surplus $V_i = S(q_i) - t_i$, we find that since S is concave, the more efficient firm produces more i.e. $q_1^* > q_2^* > 0$ (assuming both types produce strictly positive quantities). The optimal transfers are determined by $t_i^* = \theta_i q_i^*$. The social surplus is such that $V_1^* > V_2^*$.

Next, consider a different setup. R does not know θ but f needs to break even. Hence, R offers a menu of contracts $\{q_i, t_i\}_{i=1,2}$ with the goal of inducing f to self-select the contract that matches the type of firm.

We must therefore introduce a new set of constraints. The efficient firm must prefer (q_1, t_1) and the inefficient firm must prefer (q_2, t_2). We call these **incentive compatibility constraints**. We still have participation constraints as before.

Now, R believes he faces an efficient firm with probability p and an inefficient firm with probability $1 - p$ (we disregard the possibility of more than 2 types of firms or firms which cannot be easily categorised).

R has a new objective function which is to maximise expected social surplus. We call this setup the **second-best**.

We can write R's problem as:

$$\max_{\{q_i,t_i\}_{i=1,2}} p(S(q_1) - t_1) + (1-p)(S(q_2) - t_2)$$

subject to

$$t_1 - \theta_1 q_1 \geq 0, \quad t_2 - \theta_2 q_2 \geq 0$$

$$t_1 - \theta_1 q_1 \geq t_2 - \theta_1 q_2, \quad t_2 - \theta_2 q_2 \geq t_1 - \theta_2 q_2$$

The problem above seems rather complicated. However, before we think about using a Lagrangian, we can simplify the constraints.

First, observe that $\theta_1 < \theta_2$ and so $t_1 - \theta_1 q_1 \geq t_2 - \theta_1 q_2 \geq t_2 - \theta_2 q_2 \geq 0$. That is, we can ignore $t_1 - \theta_1 q_1 \geq 0$ as this is implied.

Next, observe that if we decrease t_1, the regular is better off. We can decrease t_1 feasibly until $t_1 - \theta_1 q_1 = t_2 - \theta_1 q_2$.

Then, think about what happens if we decrease t_1 and t_2 at the same rate. The incentive compatibility constraints do not change, so we are only bound by $t_2 - \theta_2 q_2 \geq 0$ which must be binding (recall we ignore $t_1 - \theta_1 q_1 \geq 0$).

Hence, our constraints are now as follows

$$t_2 = \theta_2 q_2, \quad t_1 - \theta_1 q_1 = t_2 - \theta_1 q_2, \quad t_2 - \theta_2 q_2 \geq t_1 - \theta_2 q_2$$

However, we can simplify even further!

Take our two incentive compatibility constraints (allow for $t_1 - \theta_1 q_1 \geq t_2 - \theta_1 q_2$ to be non-binding) and add them to give $(\theta_2 - \theta_1) q_1 \geq (\theta_2 - \theta_1) q_2$.

Since $\theta_2 > \theta_1$, we have $q_1 \geq q_2$. We have reduced our two incentive compatibility constraints into one **monotonicity constraint**.

Note that if a monotonicity constraint holds and one of the incentive compatibility constraints binds, then the other incentive compatibility constraint holds.

Thus, we can check the monotonicity constraint and ignore $t_2 - \theta_2 q_2 \geq t_1 - \theta_2 q_2$.

Our constraints are now

$$t_2 = \theta_2 q_2, \quad t_1 - \theta_1 q_1 = t_2 - \theta_1 q_2, \quad q_1 \geq q_2$$

Two of our constraints are binding. Hence, we substitute them into the objective function, giving a reduced problem

$$\max_{q_1,q_2} p(S(q_1) - \theta_1 q_1 - (\theta_2 - \theta_1)q_2) + (1-p)(S(q_2) - \theta_2 q_2) \text{ subject to } q_1 \geq q_2$$

If we solve it ignoring the monotonicity constraint and the constraint happens to hold, then we will have an optimum. This way, we ignore the Lagrangian and simplify the problem using the first-order conditions.

First-order conditions over q_1 and q_2 give

$$S'(q_1) = \theta_1$$

$$S'(q_2) = \theta_2 + \frac{p}{1-p}(\theta_2 - \theta_1)$$

Since $S'' < 0$ and $\theta_2 > \theta_1$, observe that $S'(q_2) > S'(q_1)$ implying that $q_1 \geq q_2$ holds. Hence, our first-order conditions characterise the constrained solution to the regulator's problem.

We can describe a few characteristics of the second-best solution and contrast them with the first-best case.

First, the efficient firm produces the efficient quantity in the second-best case. In both the first-best and second-best cases, q_1^* is the same. This is called the **no distortion at the top** property. Intuitively, there is no incentive for the inefficient firm to pretend to be efficient as it would need to produce more and be paid less than in the first-best. Hence, there is no need to distort the efficient firm's quantity.

Second, $t_1 = \theta_1 q_1 + (\theta_2 - \theta_1)q_2 > \theta_1 q_1$. That means the efficient firm receives an **information rent** of $(\theta_2 - \theta_1)q_2$. This is because the efficient firm has an incentive to pretend to be inefficient in order to reimbursed at rate θ_2 instead of θ_1. Hence, the efficient firm must be paid not to lie.

The information rent $(\theta_2 - \theta_1)q_2$ depends on q_2 and $\theta_2 - \theta_1$. The greater θ_2 is relative to θ_1, the greater the incentive for the efficient firm to lie, so the more information rent it must receive to stop this. If $q_2 = 0$, pretending to be the inefficient firm produces nothing and is not reimbursed. Hence, the information rent is zero. However, if q_2 is very large, pretending to be the inefficient firm is very attractive to the efficient firm and so the information rent paid must be higher.

In the first-best, the inefficient firm produced according to condition $S'(q_2) = \theta_2$. Now, in the second-best, it produces according to $S'(q_2) = \theta_2 + \dfrac{p}{1-p}(\theta_2 - \theta_1)$. Since $S'' < 0$, q_2^* in the first-best must have been larger than q_2^* in the second-best. In other words, with asymmetric information, the distortion for the inefficient firm is larger.

The size of the distortion is $\dfrac{p}{1-p}(\theta_2 - \theta_1)$. If θ_2 is relatively much higher, there is an incentive to for R to try and reduce the information rent to the efficient firm by asking the inefficient firm to produce less. If p is higher, it increases the chances that R faces an efficient firm (so raises the probability of paying an information rent). Hence, there is an incentive for R tro try and ask the inefficient firm to produce less (since distortion is not costly).

Note that the inefficient firm receives no rent. It simply recoups its cost as given by $t_2 = \theta_2 q_2$.

Notice that in our optimisation problem, we neglected the non-negativity constraints $q_1, q_2 \geq 0$. There is a unique interpretation of the cases $q_1 = 0$ and $q_2 = 0$ here; we call them **shut-down** cases. That is, both the efficient firm and the inefficient firm can always choose to leave the contract and produce nothing, earning zero utility.

For the efficient firm, we can rewrite the first-order condition as $S'(q_1) \leq \theta_1$ with equality if $q_1 > 0$. Hence, if $S'(0) \leq \theta_1$ then $q_1 = 0$.

For the inefficient firm, we have $S'(q_2) \leq \theta_2 + \dfrac{p}{1-p}(\theta_2 - \theta_1)$ with equality if $q_2 > 0$. If we have $q_2 = 0$, then only the efficient firm will asked to produce but will receive no information rent.

3 Dynamic Optimisation

3.1 Topics in Differential Equations

Definition 51. *An nth-order* ***ordinary differential equation (ODE)*** *is a relation between a function $y(x)$ of some variable x and its nth order derivatives:*

$$F(x, y, \frac{dy}{dx}, \frac{d^2y}{dx^2}, ..., \frac{d^ny}{dx^n}) = 0$$

Talking of a solution to an nth-order ordinary differential equation means finding a n-times differentiable function y that satisfies the ODE.

In general, most ODEs require numerical methods to solve. However, we will look at a few types of ODEs with explicit formulae for solutions.

Definition 52. *An nth-order **linear ODE** is an ODE such that it is defined as a linear polynomial in the function $y(x)$ and its derivatives:*

$$a_0(x)y + a_1(x)\frac{dy}{dx} + a_2(x)\frac{d^2y}{dx^2}+, ..., +a_n(x)\frac{d^ny}{dx^n} = b(x)$$

*where $a_i(x)$, $\forall i = \{0, ..., n\}$ and $b(x)$ are differentiable functions. If $b(x) = 0$, then the linear ODE is **homogeneous**.*

The solution to a linear ODE, denoted $y(x)$, is the sum of the **general solution** $y_0(x)$ to the related homogeneous equation and a **partial solution** $y_p(x)$ such that $y(x) = y_0(x)+y_p(x)$. Hence, we can break down the procedure of finding a solution into the two parts: first, find the general solution; and second, find a partial solution. For now, we'll focus on the general solution for linear homogeneous ODEs with constant coefficients such that $a_i(x)$, $\forall i = \{0, ..., n\}$ are constants.

To find the general solution of the associated homogeneous linear ODE with constant coefficients, set $b(x) = 0$ such that:

$$a_0y + a_1\frac{dy}{dx} + a_2\frac{d^2y}{dx^2}+, ..., +a_n\frac{d^ny}{dx^n} = 0$$

Note that we write the coefficients a_i to indicate that they are no longer functions of x. We then look at solutions of the form $e^{\lambda x}$. Substituting for y and dividing throughout by $e^{\lambda x}$ yields:

$$a_0 + a_1\lambda + a_2\lambda^2+, ..., +a_n\lambda^n = 0$$

This is a characteristic polynomial. Using the factor theorem recursively[10] gives:

$$a_n\prod_{j=1}^{n}(\lambda - \lambda_j) = a_n(\lambda - \lambda_1)(\lambda - \lambda_2)...(\lambda - \lambda_{n-1})(\lambda - \lambda_n) = 0$$

When n is large, numerical methods are typically used to find the roots λ_i, $\forall i = \{1, ..., n\}$. However, for some problems, we can make guesses of the values of these roots. We write the general

[10]Let $P(\lambda) = a_0 + a_1\lambda + a_2\lambda^2+, ..., +a_n\lambda^n$. The factor theorem states that if λ_1 is a root of $P(\lambda)$, then we can write $P(\lambda) = (\lambda - \lambda_1)Q(\lambda)$. Furthermore, as the coefficients a_i, $\forall i = \{0, ..., n\}$ are either real or complex numbers, by the **Fundamental Theorem of Algebra**, $P(\lambda)$ has a real or complex root. Hence, applying the factor theorem recursively yields the expression.

solution of a linear, homogeneous ODE with constant coefficients as follows:

$$y(x) = C_1 e^{\lambda_1 x} + C_2 e^{\lambda_2 x} +, ..., + C_n e^{\lambda_n x} = \sum_{j=1}^{n} C_j e^{\lambda_j x}$$

where $C_1, ..., C_n$ are constants and all roots are unique and real. Thus, a general solution to an nth order differential equation has n constants.

In the case that we have one root λ_i repeated m times, then we can write the repeated term of the general solution as $e^{\lambda_i x}[C_i + C_{i+1}x + C_{i+2}x^2 +, ..., + C_{i+(m-1)}x^{m-1}]$.

Now, what if the roots λ_i are not real? For example, if we have $\lambda^2 + 1 = 0$? Then, we have roots $\pm i$ where $i = \sqrt{-1}$. In general, if our polynomial has no real solutions, then the roots will be complex numbers $a + bi$ where $a, b \in \mathbb{R}$. If so, we may apply **Euler's formula** [11] $e^{ix} = \cos(x) + i\sin(x)$ to give the complex term $e^{(a+bi)x} = e^{ax}[\cos(bx) + i\sin(bx)]$.

When we deal with polynomials with real coefficients and obtain complex solutions, we get them in pairs $\lambda = a \pm bi$ and so $y(x) = D_1 e^{(a+bi)x} + D_2 e^{(a-bi)x} = e^{ax}[D_1(\cos(bx) + i\sin(bx)) - D_2(\cos(bx) + i\sin(bx))] = e^{ax}[C_1\cos(bx) + C_2\sin(bx)]$ is our general solution.

For repetitions of pairs of complex roots m times, we may write $e^{ax}[\cos(bx)(C_1 + C_2x + ... + C_m x^{m-1}) + \sin(bx)(D_1 + D_2x + ... + D_m x^{m-1})]$.

The next step is to find a partial solution of the ODE

$$a_0(x)y + a_1(x)\frac{dy}{dx} + a_2(x)\frac{d^2y}{dx^2} +, ..., + a_n(x)\frac{d^n y}{dx^n} = b(x)$$

Assume $b(x)$ has the form $b(x) = e^{\gamma x}Q_r(x)$ where Q_r is a polynomial of order r. Then, if γ is not a root of the polynomial, the partial solution takes a form $e^{\gamma x}P_r(x)$. If γ is a root of the polynomial repeated m times, then the partial solution takes a form $x^m e^{\gamma x}P_r(x)$.

Similarly, if $b(x)$ has the particular form $e^{\gamma x}Q_r(x)\cos(\phi x)$ or $e^{\gamma x}Q_r(x)\sin(\phi x)$ and $\gamma \pm i\phi$ is not a root, then the partial solution takes a form $e^{\gamma x}[P_{1r}(x)\cos(\phi x) + P_{2r}(x)\sin(\phi x)]$ with P_{1r}, P_{2r} being polynomials of order at most r. If $\gamma \pm i\phi$ is a root repeated m times, then the solution is $e^{\gamma x}x^m[P_{1r}(x)\cos(\phi x) + P_{2r}(x)\sin(\phi x)]$.

[11] To see intuitively that Euler's formula works, observe the Taylor polynomial of $e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3} + ...$ and those of $\sin(x) = x - \frac{x^3}{3} + \frac{x^5}{5} + ...$ and $\cos(x) = 1 - \frac{x^2}{2} + \frac{x^4}{4}...$, and so we can write $e^{ix} = 1 + (ix) + \frac{(ix)^2}{2} + \frac{(ix)^3}{3} + \frac{(ix)^4}{4}... = 1 + (ix) - \frac{x^2}{2} - i\frac{x^3}{3} + \frac{x^4}{4}... = \cos(x) + i\sin(x)$, giving Euler's formula.

Consider an explicit example of the differential equation $\frac{d^3y}{dx^3}+4\frac{d^2y}{dx^2}+\frac{dy}{dx}+4y = 34\sin(x)$. The resulting characteristic polynomial is $\lambda^3 + 4\lambda^2 + \lambda + 4 = 0$. Since the equation is cubic, we have to guess one solution. Then, we can factorise the remaining quadratic. Our roots are -4 and $\pm i$. Hence, we have general solution $y_0(x) = C_1e^{-4x} + C_2\cos(x) + C_3\sin(x)$.

Next, we look for a partial solution. Note that $b(x) = sin(x)$, which corresponds to root $\lambda = \pm i$ repeated once. Hence, we may use a partial solution $x(a\cos(x)+b\sin(x))$. This partial solution has first derivative $a\sin(x) + ax\cos(x) + b\cos(x) - bx\sin(x)$, second derivative $2a\cos(x) - ax\sin(x) - 2b\sin(x) - bx\cos(x)$ and third derivative $-3a\sin(x) - ax\cos(x) - 3b\cos(x) + bx\sin(x)$. Substituting into the ODE and making coefficients correspond to the right hand side gives us $a = -1$ and $b = -4$.

Thus the full solution to the differential equation is $y(x) = C_1e^{-4x} + C_2\cos(x) + C_3\sin(x) - x(\sin(x) + 4\cos(x))$.

For many first-order differential equations, we can solve using the method of **integrating factors**. Consider the following ODE

$$\frac{dy}{dx} + u(x)y = v(x)$$

where u, v are differentiable functions. Multiplying both sides by an integrating factor $e^{\int u(x).dx}$ and then integrating gives us

$$e^{\int u(x).dx}y = \int e^{\int u(x).dx}v(x).dx$$

This gives

$$y(x) = [e^{\int u(x).dx}]^{-1}\int e^{\int u(x).dx}v(x).dx$$

As an explicit example, consider the equation $\frac{dy}{dx} - \frac{1}{x}y = -xe^{-x}$. Our integrating factor is $e^{\int -1/x.dx} = e^{-ln(x)} = \frac{1}{x}$. Hence, multiply every term by $\frac{1}{x}$ to give $\frac{1}{x}\frac{dy}{dx} - \frac{1}{x^2}y = -e^{-x}$.

The left-hand side simplifies down to $\frac{y}{x} = -\int e^{-x}.dx = e^{-x} + c$ and hence we write the solution as $y = xe^{-x} + cx$.

We occasionally have differential equations which are **separable**, so they do not require any of

the techniques we have discussed so far. For instance, if we have a differential equation of the form $u(x)v(t)\frac{dx}{dt} = r(x)w(t)$, we can simply rearrange to $\frac{u(x)}{r(x)}dx = \frac{w(t)}{v(t)}dt$ and integrate to obtain our solution of x in terms of t.

Definition 53. *An **initial value problem** is a differential equation $\frac{dy}{dx} = f(x, y(x))$ with $f : \Omega \subset \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ with Ω being an open set of $\mathbb{R} \times \mathbb{R}^n$, together with a point in the domain of f denoted the **initial condition** (x_0, y_0).*

A solution to an initial value problem is a function y that is a solution to the differential equation that satisfies $y(x_0) = y_0$. We may have initial conditions for both y and its derivatives. That is, y is a vector of function f and its first n derivatives.

Sometimes, we have differential equations of the form $y' + P(x)y = Q(x)y^n$ where P, Q are polynomials. This is called a **Bernoulli differential equation** and has a few special properties. First, when $n = 0$, we are reduced to a first-order linear differential equation. When $n = 1$, the equation becomes separable. However, when $n \geq 2$, by substituting $u = y^{1-n}$, we can transform the differential equation into one that is linear. Typically, we require this step to much simplify the analysis.

As an example of a Bernoulli differential equation, consider $\frac{dx}{dt} = \alpha x(1 - \frac{x}{\beta})$ with initial condition $x(0) = x_0$. By writing $\frac{dx}{dt} - \alpha x = -\frac{\alpha}{\beta}x^2$, observe we substitute $u = \frac{1}{x}$.

The substitution and simplification noting that $\frac{du}{dt} = \frac{du}{dx}\frac{dx}{dt}$ gives $\frac{du}{dt} + \alpha u = \frac{\alpha}{\beta}$, which is linear and separable by writing $\frac{du}{dt} = \alpha(\frac{1}{\beta} - u)$.

Rearranging gives us $\int \frac{\beta}{1 - \beta u}du = \int \alpha t.dt$ and integrating on both sides gives $-ln(1 - \beta u) = \alpha t + c$. Some manipulation gives $u = \frac{1}{\beta}(1 - e^{-\alpha t - c})$. Substituting back for x and manipulating gives $x = \frac{\beta e^{\alpha t + c}}{e^{\alpha t + c} - 1}$.

Then finally, substituting for the initial condition gives us an explicit value for c that results in

the solution $x(t) = \dfrac{\beta e^{\alpha t}(\dfrac{x_0}{x_0 - \beta})}{e^{\alpha t}(\dfrac{x_0}{x_0 - \beta}) - 1}$.

3.2 Systems of First-Order Differential Equations

Consider two variables $x_1(t)$ and $x_2(t)$ that change over time and are described by the following system

$$\dot{x}_1 = ax_1 + bx_2 + r_1$$

$$\dot{x}_2 = cx_1 + dx_2 + r_2$$

with constants $a, b, c, d, r_1, r_2 \in \mathbb{R}$. We may write the system in matrix form

$$\dot{\mathbf{x}} = A\mathbf{x} + \mathbf{r}$$

where $\mathbf{x} = (x_1, x_2)^T$, $\mathbf{r} = (r_1, r_2)^T$ and $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$.

Note that the ODEs do not explicit depend on t and hence they are called **autonomous**. For each $x(t)$, we can identify all future values $x(s > t)$.

Definition 54. *An **equilibrium vector** $\bar{\boldsymbol{x}}$ yields $\dot{\boldsymbol{x}} = \boldsymbol{0}$. That is, $\boldsymbol{x}(t)$ does not change with t in equilibrium.*

We are interested in the behaviour of $\mathbf{x}(t)$ around equilibrium. If $det(A) \neq 0$ (i.e. A is full rank), there exists a unique equilibrium $\bar{\mathbf{x}} = -A^{-1}\mathbf{r}$. By introducing $z_1(t) = x_1(t) - \bar{x}_1$ and $z_2(t) = x_2(t) - \bar{x}_2$, the system can be rewritten as $\dot{\mathbf{z}} = A\mathbf{z}$. Substituting variables allows us to study $\dot{\mathbf{x}} = A\mathbf{x}$ around equilibrium $(0, 0)$.

Let's look for the following type of solution $\mathbf{x}(t) = e^{\lambda t}\mathbf{v}$. Substituting into $\dot{\mathbf{x}} = A\mathbf{x}$ gives us $\lambda e^{\lambda t}\mathbf{v} = Ae^{\lambda t}\mathbf{v}$ which reduces to $\lambda\mathbf{v} = A\mathbf{v}$, or the eigenvectors and eigenvalues of A.

Thus we say that assuming there exist two linearly independent eigenvectors $\mathbf{u}_1$ and $\mathbf{u}_2$ corresponding to eigenvalues λ_1 and λ_2, the general solution is $\mathbf{x}(t) = C_1 e^{\lambda_1 t}\mathbf{u}_1 + C_2 e^{\lambda_2 t}\mathbf{u}_2$.

Recall that since non-zero orthogonal vectors are linearly independent, we can establish that there will always exist two linearly independent eigenvectors if we have a real symmetric A.

We are interested in the **phase diagram** which shows how, given some initial value $\mathbf{x}(0) = \mathbf{x}_0$, $\mathbf{x}(t)$ evolves over time and sketches the curve along which $\mathbf{x}(t) = (x_1(t), x_2(t))^T$ moves. The shape of the phase diagram depends on eigenvalues λ.

If the eigenvalues are of different signs, we know that we have a saddle point at $\bar{\mathbf{x}}$.

If $\lambda_1 = \lambda_2$ and are real, then we have $\mathbf{x}(t) = \mathbf{x}(0)e^{\lambda t}$. The equilibrium is called a **proper node**. With $\lambda > 0$, observe that $\mathbf{x}(t)$ increases indefinitely with t and so the equilibrium is unstable. With $\lambda < 0$, it is **globally asymptotically stable**.

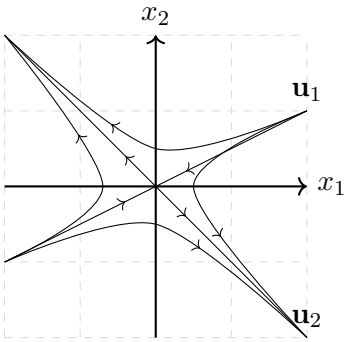
If we have $\lambda_1 \neq \lambda_2$, the equilibrium is called an **improper node**. If $\lambda_1, \lambda_2 > 0$, the equilibrium is unstable. If $\lambda_1, \lambda_2 < 0$, the equilibrium is globally asymptotically stable.

However, if we have complex roots i.e. $\lambda_{1,2} = a \pm bi$, then the solution looks like $\mathbf{x}(t) = e^{at}[C_1 \cos(bt)\mathbf{v}_1 + C_2 \sin(bt)\mathbf{v}_2]$ with $\mathbf{v}_1, \mathbf{v}_2$ being some constant vectors. That is, as t increases, $\mathbf{x}(t)$ converges to $(0, 0)$ if $a < 0$ (called a **stable focus**) and diverges if $a > 0$ (an **unstable focus**). There is rotation around the equilibrium so the diagram looks like a spiral for $a \neq 0$ and an ellipsoid for $a = 0$. We can determine the direction of rotation by picking a point $\mathbf{x} \neq \mathbf{0}$ and looking at the direction of $\dot{\mathbf{x}}$. The spiral can be narrow along some direction and wide along another. It is quite difficult to determine the exact shape.

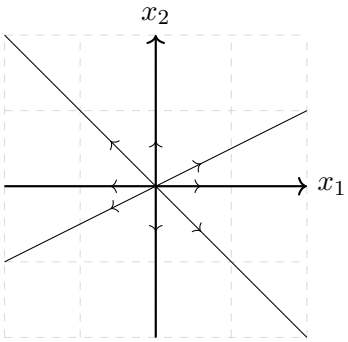
Let's look at a few examples of the different scenarios above, and sketch the phase diagrams.

First, consider $\dot{x}_1 = -2x_2$ and $\dot{x}_2 = -x_1 + x_2$. The equilibrium point is $(x_1, x_2) = (0, 0)$. The matrix $A = \begin{pmatrix} 0 & -2 \\ -1 & 1 \end{pmatrix}$ and its eigenvalues are $\lambda_1 = -1$ and $\lambda_2 = 2$. Corresponding eigenvectors are $\mathbf{u}_1 = (2, 1)^T$ and $\mathbf{u} = (-1, 1)^T$ respectively. Hence, we have general solution $\mathbf{x}(t) = C_1 e^{-t}(2, 1)^T + C_2 e^{2t}(-1, 1)^T$. Clearly, the eigenvalues have different signs so we have $(0, 0)$ being a saddle point.

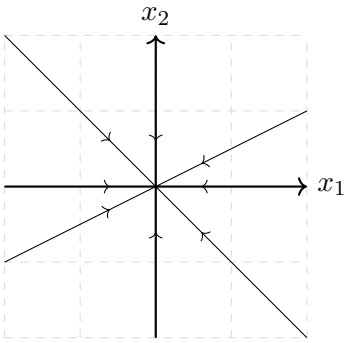
From $(0, 0)$, we have two straight lines representing the eigenvector directions $\mathbf{u}_1$ and $\mathbf{u}_2$. Since $\lambda_1 < 0$, we will have convergence along $\mathbf{u}_1$ and since $\lambda_2 > 0$, divergence along $\mathbf{u}_2$. For all other points, the limit will be close to the line containing $\mathbf{u}_2$. The diagram shows the saddle at $(0, 0)$.



In the event that we have $\lambda_1 = \lambda_2$ and $\lambda > 0$ i.e. an unstable proper node, then our diagram looks as follows.

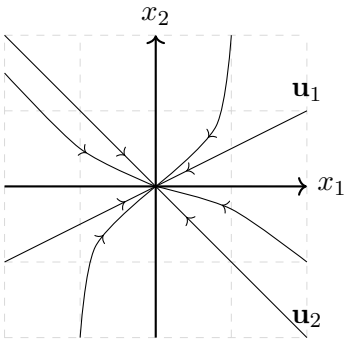


If $\lambda_1 = \lambda_2$ but $\lambda < 0$ i.e. we have a stable proper node, then we have the opposite case.



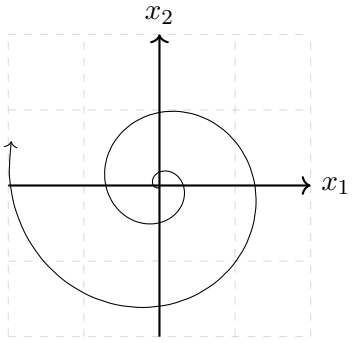
Next, consider the system $\dot{x}_1 = -4x_1 + 2x_2$ and $\dot{x}_2 = x_1 - 5x_2$. The relevant matrix has eigenvalues -3 and -6 and respective eigenvectors $\mathbf{u}_1 = (2,1)^T$ and $\mathbf{u}_2 = (-1,1)^T$. The general solution is thus $\mathbf{x}(t) = C_1 e^{-3t}(2,1)^T + C_2 e^{-6t}(-1,1)^T$.

Note there is convergence along both eigenvectors. However, the rate of convergence along $\mathbf{u}_2$ is faster. Hence, the arrows all converge to $(0,0)$ and touch the line containing $\mathbf{u}_1$, making $(0,0)$ a globally asymptotically stable improper node.



Consider the following system $\dot{x}_1 = x_1 + x_2$ and $\dot{x}_2 = -2x_1 + x_2$ which has relevant matrix with eigenvalues $\lambda = 1 \pm i\sqrt{2}$, meaning we have general solution $\mathbf{x}(t) = e^{at}[C_1 \cos(bt)\mathbf{v}_1 + C_2 \sin(bt)\mathbf{v}_2]$ for some constant $\mathbf{v}_1, \mathbf{v}_2$. We don't typically find the constant vectors as this can be extremely tricky.

Notice that the real parts of the eigenvalues are positive and so we have an unstable focus represented as a spiral going away from $(0, 0)$. To assess the direction of this spiral, pick any point e.g. $(x_1, x_2) = (1, 0)$. Observe that at this point, $\dot{x}_1 > 0$ but $\dot{x}_2 < 0$ and so we are moving in a clockwise direction.



3.3 Hamiltonians and Continuous-Time Optimisation

In this subsection, we work exclusively in continuous time $t \in [0, T]$. Often, we may approximate discrete time periods as continuous, allowing us to have nicer mathematical properties and predictions. Here, we consider the control problem of optimal consumption in a dynamical system.

The **state variable** $s(t)$ describes the state of the system and changes continuously over time. The **control variable** $c(t)$ is chosen by agents in the model and may experience jumps over time. We use differential equations to describe the change of state variables over time and this rate depends on control variables which are easily adjustable. We are interested both in the optimal choices of

control variables and how state variables evolve over time.

The goal is to maximise some objective function $\max_{c(t)} J = \int_0^T v(s(t), c(t), t).dt$ subject to a constraint that $\dot{s}(t) = f(s(t), c(t), t)$. That is, we have a time interval $[0, T]$ in which a flow payoff v exists in each period. We maximise the integral, which is the aggregate payoff. We also have boundary conditions $s(0) = s_0$ and $s(T) = s_T$. Often, we may have that s_T is not fixed or $T \to \infty$. In general, we often need to ensure that we restrict the control variable to make sure, for instance, that it cannot be negative.

We define the **Hamiltonian** as

$$H(s(t), c(t), \pi(t), t) \equiv v(s(t), c(t), t) + \pi(t) f(s(t), c(t), t)$$

where the first term is some "myopic" objective function and the second term is the constraint weighted by the **costate variable** $\pi(t)$ that is analogous to a Lagrange multiplier in static problems.

Proposition 56. ***Pontryagin's Maximum Principle.*** *Assume $v(s(t), c(t), t)$ and $f(s(t), c(t), t)$ are continuously differentiable with no constraints on $c(t)$. Then, necessary conditions for optimisation are such that $c(t)$ maximises $H(s(t), c(t), \pi(t), t)$*

$$\frac{\partial H(s(t), c(t), \pi(t), t)}{\partial c(t)} = 0$$

and the state and costate variables change over time as follows

$$\dot{s}(t) = \frac{\partial H(s(t), c(t), \pi(t), t)}{\partial \pi(t)}$$

$$\dot{\pi}(t) = -\frac{\partial H(s(t), c(t), \pi(t), t)}{\partial s(t)}$$

After obtaining solutions, substitute the boundary conditions.

If functions $v(s(t), c(t), t)$ and $f(s(t), c(t), t)$ are concave in $s(t)$ and $c(t)$, then these conditions are sufficient i.e. the solution to the Hamiltonian provides the optimum. Equivalently, if the Hamiltonian is concave in $s(t)$ and $c(t)$ for all t, the solution provides the optimum.

Consider the example of $\max_{c(t)} \int_0^1 ln(c(t)4s(t))dt$ subject to $\dot{s}(t) = 4(1-c(t))s(t)$ with $s(0) = 1$

and $s(1) = e^2$. We start by setting up the Hamiltonian

$$H = ln(4) + ln(c(t)) + ln(s(t)) + \pi(t)(4(1 - c(t))s(t))$$

Then, applying the maximum principle gives

$$\frac{1}{c(t)} = 4\pi(t)s(t)$$

$$\dot{s}(t) = 4(1 - c(t))s(t)$$

$$\dot{\pi}(t) = -\frac{1}{s(t)} - \pi(t)4(1 - c(t))$$

Substituting the first equation into the remaining two gives $\dot{s}(t) = 4s(t) - \frac{1}{\pi(t)}$ and $\dot{\pi}(t) = -4\pi(t)$. Solving the second differential equation gives us $\pi(t) = \pi(0)e^{-4t}$ and substituting into the first differential equation gives $\dot{s}(t) = 4s(t) - \frac{1}{\pi(0)e^{-4t}}$.

Solving by the method of integrating factors and using the boundary constraints gives us $s(t) = (1 - \frac{t}{1 - e^{-2}})e^{4t}$. Then, we can find that $c(t) = \frac{1}{\frac{4}{1-e^{-2}} - 4t}$, giving our optimal level of consumption over time.

Often, we have no constraint on $s(T)$. If this is the case, then we impose $\pi(T) = 0$ which is caled the **transversality condition**. This gives us the constraint required to find an explicit solution.

Let's think about the optimal consumption problem again. Imagine that we have an objective $\int_0^T U(c(t))dt$ subject to $\dot{s}(t) = -c(t)$. The Hamiltonian is

$$H = U(c(t)) - \pi(t)c(t)$$

which by the maximum principle gives conditions

$$U'(c(t)) = \pi(t)$$

$$\dot{s}(t) = -c(t)$$

$$\dot{\pi}(t) = 0$$

We can solve the third differential equation trivially to yield $\pi(t) = c_1$ where $c_1 \in \mathbb{R}$. Hence, we have $U'(c(t)) = c_1$ or that the marginal utility of consumption is constant. If U is strictly increasing and strictly concave, then U' is a one-to-one correspondence and so a unique inverse U^{-1} exists such that we have $U'(c(t)) = c_1 \implies c(t) = (U')^{-1}c_1$.

However, we are yet to consider optimality. The above statement would hold if U was strictly increasing but strictly convex. However, this would not correspond to a maximum (recalling that we require a concave Hamiltonian for the first-order conditions to correspond to a maximum).

References

1. Andreu Mas-Colell, Michael D. Whinston & Jerry R. Green, Microeconomic Theory (Oxford University Press, 1995)
2. Knut Sydsæter, Topics in Mathematical Analysis for Economists (Academic Press, 1981)
3. Michael Carter, Foundations of Mathematical Economics (MIT Press, 2001)
4. Carl P. Simon and Lawrence Blume, Mathematics for Economists (Norton, 1994)
5. Knut Sydsæter and Peter J. Hammond, Further Mathematics for Economic Analysis (Prentice-Hall, 1995)