for $x = ax_1 + bx_2$. But since T^{-1} exists, $T^{-1}(y_1) = x_1$ and $T^{-1}(y_2) = x_2$ and $T^{-1}(ay_1 + by_2) = x = aT^{-1}(y_1) + bT^{-1}(y_2)$. $\square$

Theorem 43. *Let $T \in L(\mathbb{X}, \mathbb{Y})$. T is one-to-one if, and only if, $T(x) = \theta$ implies $x = \theta$.*

Proof. ($\Longrightarrow$) Recall from Remark 8 that because T is linear we have that $T(\theta) = \theta$. Also, if T is one-to-one, for any other $x \in \mathbb{X}$, $T(x) \neq \theta$. Hence, $T(x) = \theta$ only when $x = \theta$.
($\Longleftarrow$) If $T(x) = \theta$ implies $x = \theta$, then the $null(T) = \{\theta\}$. For any two different $x, x' \in \mathbb{X}$, that is, $x - x' \neq \theta$, we have by linearity that $T(x) \neq T(x')$. Consequently, T is one-to-one. $\square$

Definition 45. *Vector spaces $\mathbb{X}$ and $\mathbb{Y}$ are said to be isomorphic if there exists a linear function $T \in L(\mathbb{X}, \mathbb{Y})$ that has an inverse T^{-1}. In this case, any such T is called an isomorphism.*

Theorem 44. *Let $\mathbb{X}$ and $\mathbb{Y}$ be real vector spaces and $\{x_1, \cdots, x_n\}$ be a basis for $\mathbb{X}$.*

1. *If $\{y_1, \cdots, y_n\} \in \mathbb{Y}$, then there exists $\mathcal{T} \in L(\mathbb{X}, \mathbb{Y})$ such that $y_i = \mathcal{T}(x_i)$ for $i = 1, \cdots, n$.*

2. *If $T, f \in L(\mathbb{X}, \mathbb{Y})$ and $T(x_i) = f(x_i) = y_i$ for $i = 1, \cdots, n$, then $T(x) = f(x)$ for all $x \in \mathbb{X}$.*

Proof. 1. Consider an arbitrary (not necessarily linear) function $T : \{x_1, \cdots, x_n\} \to \mathbb{Y}$ with $T(x_i) = y_i$ for $i = 1, \cdots, n$. Since, $\{x_1, \cdots, x_n\}$ is a basis for $\mathbb{X}$, any $x \in \mathbb{X}$ can be written as $x = \sum_{i=1}^n a_i x_i$, where $a_i \neq 0$ for some i. Now, define an extension of T to all of $\mathbb{X}$ as $\mathcal{T}(x) = \sum_{i=1}^n a_i T(x_i) = \sum_{i=1}^n a_i y_i$. Note that for $w, v \in \mathbb{X}$

$$x = aw + bv = a\sum_{i=1}^n c_i x_i + b\sum_{i=1}^n d_i x_i = \sum_{i=1}^n (ac_i + bd_i) x_i.$$

Hence, $\mathcal{T}(x) = \sum_{i=1}^n (ac_i + bd_i) T(x_i) = a\sum_{i=1}^n c_i y_i + b\sum_{i=1}^n d_i y_i = a\mathcal{T}(w) + b\mathcal{T}(v)$. Hence, $\mathcal{T}$ is linear and entirely determined by $T : \{x_1, \cdots, x_n\} \to \{y_1, \cdots, y_n\}$.

2. Let $x = \sum_{i=1}^{n} a_i x_i$, and since f, T are linear functions, we have

$$\begin{aligned} T(x) &= \sum_{i=1}^{n} a_i T(x_i), \\ f(x) &= \sum_{i=1}^{n} a_i f(x_i), \end{aligned}$$

and by assumption $T(x_i) = f(x_i)$. Consequently, $T(x) = f(x)$ for all $x \in \mathbb{X}$. □

Theorem 45. *Let $\mathbb{X}$ and $\mathbb{Y}$ be real vector spaces. Then, $\mathbb{X}$ and $\mathbb{Y}$ are isomorphic if, and only if, $dim(\mathbb{X}) = dim(\mathbb{Y})$.*

Proof. ($\Longrightarrow$) Let $\{x_1, \cdots, x_n\}$ be a basis for $\mathbb{X}$, then $dim(\mathbb{X}) = n$. $\mathbb{X}$ and $\mathbb{Y}$ are isomorphic if there exist $T \in L(\mathbb{X}, \mathbb{Y})$ with an inverse T^{-1}. First, we show that $\{T(x_1), \cdots, T(x_n)\}$ is a basis for $\mathbb{Y}$. That is, we show that a) $\{T(x_1), \cdots, T(x_n)\}$ is a linearly independent collection of vectors and b) any $y \in \mathbb{Y}$ can be written as $\sum_{i=1}^{n} T(x_i) a_i$. If this is the case, then $dim(\mathbb{Y}) = n$. For a), note first that $\sum_{i=1}^{n} a_i T(x_i) = \theta$ implies $T(\sum_{i=1}^{n} a_i x_i) = \theta$. Since T^{-1} exists, we have $\sum_{i=1}^{n} a_i x_i = T^{-1}(\theta) = \theta$, but given that $\{x_1, \cdots, x_n\}$ is a basis for $\mathbb{X}$ it must be that $a_i = 0$ for all i if the equality is to hold. Hence, $\sum_{i=1}^{n} a_i T(x_i) = \theta$ implies $a_i = 0$ for all i, establishing the linear independence of $\{T(x_1), \cdots, T(x_n)\}$. For b), note that for all $y \in \mathbb{Y}$ there exists one, and only one, $x \in \mathbb{X}$ such that $y = T(x)$. But $x = \sum_{i=1}^{n} a_i x_i$, hence by linearity $y = T(\sum_{i=1}^{n} a_i x_i) = \sum_{i=1}^{n} a_i T(x_i)$.

($\Longleftarrow$) Now, suppose $\{x_1, \cdots, x_n\}$ and $\{y_1, \cdots, y_n\}$ are bases for $\mathbb{X}$ and $\mathbb{Y}$. By Theorem 44, there exist $\mathcal{T} \in L(\mathbb{X}, \mathbb{Y})$ such that $\mathcal{T}(x_i) = y_i$. Hence, all there is to show is that $\mathcal{T}$ has an inverse $\mathcal{T}^{-1}$. Suppose $x' = \sum_{i=1}^{n} a_i x_i$ and $x'' = \sum_{i=1}^{n} b_i x_i$ are such that $\mathcal{T}(x') = \mathcal{T}(x'')$, then $\sum_{i=1}^{n} a_i \mathcal{T}(x_i) = \sum_{i=1}^{n} b_i \mathcal{T}(x_i)$ which implies that $\sum_{i=1}^{n} (a_i - b_i)\mathcal{T}(x_i) = \sum_{i=1}^{n} (a_i - b_i) y_i = \theta$. But since $\{y_1, \cdots, y_n\}$ is a basis for $\mathbb{Y}$, it must be that $a_i = b_i$ for all i, and therefore $x' = x''$. So, $\mathcal{T}$ is one-to-one, and we need only show that it is onto to conclude it has an inverse. From Theorem 39 (part 2), $\{\mathcal{T}(x_1), \cdots, \mathcal{T}(x_n)\}$ generates $\mathcal{T}(\mathbb{X})$ which is a subspace of $\mathbb{Y}$.

But $\{\mathcal{T}(x_1), \cdots, \mathcal{T}(x_n)\} = \{y_1, \cdots, y_n\}$ which is a basis for $\mathbb{Y}$. Hence, $\mathcal{T}(\mathbb{X}) = \mathbb{Y}$ and $\mathcal{T}$ is onto. $\square$

10.1 Vector coordinates

Let $\mathbb{X}$ be a vector space of finite dimension n and $\{v_1, \cdots, v_n\}$ be a basis for $\mathbb{X}$. Then, for every $x \in \mathbb{X}$ there exists a unique collection of scalars $\{a_1, \cdots, a_n\}$ such that $x = \sum_{i=1}^n a_i v_i$. We call a_i the i^{th} coordinate of x given $\{v_1, \cdots, v_n\}$. We define the coordinate vector function as $c_v : \mathbb{X} \to \mathbb{R}^n$, such that

$$c_v(x) = c_v\left(\sum_{i=1}^n a_i v_i\right) = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} = a.$$

Remark 10. *1. Since for any two vectors $x \neq x'$ and for a fixed basis $\{v_1, \cdots, v_n\}$, $c_v(x) \neq c_v(x')$, c_v is one-to-one.*

2. For every $a \in \mathbb{R}^n$, we have that $\sum_{i=1}^n a_i v_i$ is an element of $\mathbb{X}$ (by the definition of a vector space and the fact that $\{v_1, \cdots, v_n\}$ is a basis), thus c_v is an onto function.

3. If $x, y \in \mathbb{X}$ then $x = \sum_{i=1}^n a_i v_i$ and $y = \sum_{i=1}^n b_i v_i$. Hence, if c, d are scalars and $z = cx + dy$ we have that $z = \sum_{i=1}^n c a_i v_i + \sum_{i=1}^n d b_i v_i = \sum_{i=1}^n (c a_i + d b_i) v_i$. Thus, $c_v(z) = ca + db = c c_v(x) + d c_v(y)$. Hence, c_v is a linear function.

Let $f \in L(\mathbb{X}, \mathbb{Y})$, and $\{v_1, \cdots, v_m\}$ and $\{w_1, \cdots, w_n\}$ are basis for $\mathbb{X}$ and $\mathbb{Y}$. Then, for any $x \in \mathbb{X}$ we can write $x = \sum_{i=1}^m a_i v_i$ and because of linearity we can write $f(x) = \sum_{i=1}^m a_i f(v_i)$. Since $f(v_i) \in \mathbb{Y}$ we can write $f(v_i) = \sum_{j=1}^n b_{ji} w_j$ and $c_w(f(v_i)) = \begin{pmatrix} b_{1i} \\ \vdots \\ b_{ni} \end{pmatrix} = b_{.i}$. Since c_w is linear,

$$c_w(f(x)) = c_w\left(\sum_{i=1}^m a_i f(v_i)\right) = \sum_{i=1}^m a_i c_w(f(v_i)) = \sum_{i=1}^m a_i b_{.i} = \sum_{i=1}^m \begin{pmatrix} a_i b_{1i} \\ \vdots \\ a_i b_{ni} \end{pmatrix}.$$

Note that for fixed x (a is fixed), $f(x)$ is entirely determined by its action on v_i. This, in turn, is determined by $b_{.i}$. Hence, we can say that $f(x)$ is determined by the collection of coordinates $\{b_{.1}, \cdots, b_{.m}\}$. Thus, we can associate with $f(x)$ the following object

$$B := \begin{pmatrix} b_{.1} & \cdots & b_{.m} \end{pmatrix} = \begin{pmatrix} b_{11} & \cdots & b_{1m} \\ \vdots & \vdots & \vdots \\ b_{n1} & \cdots & b_{nm} \end{pmatrix} = \begin{pmatrix} c_w(f(v_1)) & \cdots & c_w(f(v_m)) \end{pmatrix}.$$

We call this the matrix B associated with f and we write the function

$$M(f) : L(\mathbb{X}_m, \mathbb{Y}_n) \to \mathcal{M}^{n\times m},$$

where $\mathcal{M}^{n\times m}$ is the set containing matrices with n rows and m columns. Hence, we think of $M(\cdot)$ as a function that maps linear functions to a space of $n \times m$ matrices. The subscripts on $\mathbb{X}$ and $\mathbb{Y}$ in $L(\mathbb{X}_m, \mathbb{Y}_n)$ denote the dimension of the spaces.

We note that the set of matrices $\mathcal{M}^{n\times m}$ can itself be viewed as a vector space by defining addition and scalar multiplication for $Q, N \in \mathcal{M}^{n\times m}$ and $a \in \mathbb{R}$ as

1. $Q + N$ is the matrix S with element $S_{ij} = Q_{ij} + N_{ij}$
2. aQ is the matrix P with element $P_{ij} = aQ_{ij}$
3. θ, the null vector is a matrix with element $\theta_{ij} = 0$

The next theorem establishes that M is linear, one-to-one and onto. One-to-one and onto means that for any two $f, g \in L(\mathbb{X}_m, \mathbb{Y}_n)$ and $f \neq g$ we have $M(f) \neq M(g)$ and for every $M \in \mathcal{M}^{n\times m}$ there exists one, and only one, $f \in L(\mathbb{X}_m, \mathbb{Y}_n)$. Consequently, $L(\mathbb{X}_m, \mathbb{Y}_n)$ and $\mathcal{M}^{n\times m}$ are isomorphic.

Theorem 46. *Let $f, g \in L(\mathbb{X}_m, \mathbb{Y}_n)$ and $h = \alpha f + \beta g$ for scalars α and β. Then, 1. $M(h) = \alpha M(f) + \beta M(g)$; 2. M is one-to-one and onto.*

Proof. 1. Since $L(\mathbb{X}_m, \mathbb{Y}_n)$ is a vector space, $h \in L(\mathbb{X}_m, \mathbb{Y}_n)$. Let $x = \sum_{i=1}^m a_i v_i$ and observe that by Remark 10-3 (linearity of the coordinate vector function) and linearity of f and

g we have $c_w(h(x)) = \alpha c_w(f(x)) + \beta c_w(g(x)) = \alpha c_w\left(\sum_{i=1}^m a_i f(v_i)\right) + \beta c_w\left(\sum_{i=1}^m a_i g(v_i)\right) = \alpha \sum_{i=1}^m a_i c_w(f(v_i)) + \beta \sum_{i=1}^m a_i c_w(g(v_i))$. Recall that by definition

$$M(h) = \begin{pmatrix} c_w(h(v_1)) & \cdots & c_w(h(v_m)) \end{pmatrix},$$

hence

$$\begin{aligned} M(h) &= \begin{pmatrix} \alpha c_w(f(v_1)) & \cdots & \alpha c_w(f(v_m)) \end{pmatrix} + \begin{pmatrix} \beta c_w(g(v_1)) & \cdots & \beta c_w(g(v_m)) \end{pmatrix} \\ &= \alpha M(f) + \beta M(g). \end{aligned}$$

2. First, note that $M(f) = M(g)$ implies that $f(v_i) = g(v_i)$ for all $i = 1, \cdots, m$. By Theorem 44-2 this implies $f = g$. Second, $f = g$ implies $f(v_i) = g(v_i)$ for all $i = 1, \cdots, m$ since the basis are in $\mathbb{X}$. Consequently, $M(f) = M(g)$. Thus, M is one-to-one.

Now, let $Q \in \mathcal{M}^{n\times m}$ and write $Q = \begin{pmatrix} b_{.1} & \cdots & b_{.m} \end{pmatrix}$. Note that $b_{.j} = \begin{pmatrix} b_{1j} \\ \vdots \\ b_{nj} \end{pmatrix}$ is an array of n scalars, and if $\{w_i\}_{i=1}^n$ is a basis for $\mathbb{Y}_n$, for some $y_j \in \mathbb{Y}_n$ we have $y_j = \sum_{i=1}^n w_i b_{ij}$ since $dim(\mathbb{Y}_n) = n$. Then, if $\{v_1, \cdots, v_m\}$ is a basis for $\mathbb{X}$, by Theorem 44 there exists $\mathcal{T} \in L(\mathbb{X}_m, \mathbb{Y}_n)$ such that $\mathcal{T}(v_j) = y_j = \sum_{i=1}^n w_i b_{ij}$. Also, by Theorem 39-2, $\{\mathcal{T}(v_1), \cdots, \mathcal{T}(v_m)\}$ is a basis for the image of $\mathcal{T}$. Thus, any $y \in \mathbb{Y}_n$ can be written as $y = \sum_{j=1}^m a_j \mathcal{T}(v_j)$, and by linearity of $\mathcal{T}$, $y = \mathcal{T}\left(\sum_{j=1}^m a_j v_j\right) = \mathcal{T}(x)$, which is entirely characterized by Q. $\square$

In the following example we will take $\mathbb{X}_n = \mathbb{R}^n$ and choose the unit-coordinate vectors as the components of the basis.

Example 19. *Let $f \in L(\mathbb{R}^m, \mathbb{R}^n)$ and choose the unit-coordinate vectors as the components of the basis for $\mathbb{R}^m$ and $\mathbb{R}^n$. That is, the basis for $\mathbb{R}^m$ is the collection $\{e_1, \ldots, e_m\}$ with*

$$e_i^T = \begin{pmatrix} 0 & \cdots & 1 & \cdots & 0 \end{pmatrix}$$

where 1 appears at the i^{th} position of the m-tuple. Similarly, the basis for $\mathbb{R}^n$ is the collection $\{u_1, \ldots, u_n\}$ with $u_i^T = \begin{pmatrix} 0 & \cdots & 1 & \cdots & 0 \end{pmatrix}$ where 1 appears at the i^{th} position of the n-

tuple. For $x \in \mathbb{R}^m$ we have $x = \sum_{i=1}^m x_i e_i$

$$y = f(x) = f\left(\sum_{i=1}^m x_i e_i\right) = \sum_{i=1}^m x_i f(e_i).$$

Now, $f(e_i) \in \mathbb{R}^n$, therefore $f(e_i) = \sum_{j=1}^n f_j(e_i) u_j = \begin{pmatrix} f_1(e_i) & f_2(e_i) & \cdots & f_n(e_i) \end{pmatrix}^T$. Hence,

$$\begin{aligned}
f(x) &= \sum_{i=1}^m x_i \begin{pmatrix} f_1(e_i) & f_2(e_i) & \cdots & f_n(e_i) \end{pmatrix}^T \\
&= \begin{pmatrix} \begin{pmatrix} f_1(e_1) & f_1(e_2) & \cdots & f_1(e_m) \end{pmatrix} x \\ \begin{pmatrix} f_2(e_1) & f_2(e_2) & \cdots & f_2(e_m) \end{pmatrix} x \\ \vdots \\ \begin{pmatrix} f_n(e_1) & f_n(e_2) & \cdots & f_n(e_m) \end{pmatrix} x \end{pmatrix} = \begin{pmatrix} f_1(e_1) & f_1(e_2) & \cdots & f_1(e_m) \\ f_2(e_1) & f_2(e_2) & \cdots & f_2(e_m) \\ & & \vdots & \\ f_n(e_1) & f_n(e_2) & \cdots & f_n(e_m) \end{pmatrix} x \\
&:= Fx,
\end{aligned}$$

where F is an $n \times m$ matric with typical element $F_{ji} := f_j(e_i)$

Example 20. *As in the previous example let $f \in L(\mathbb{R}^m, \mathbb{R}^n)$ and $g \in L(\mathbb{R}^n, \mathbb{R}^p)$ where the domain of g is the range of f. Let $(g \circ f)(x) = g(f(x))$ for $x \in \mathbb{R}^m$ and note that $(g \circ f)(x) \in L(\mathbb{R}^m, \mathbb{R}^p)$ (why?). Let $\{e_1, \cdots, e_m\}$, $\{u_1, \cdots, u_n\}$ and $\{w_1, \cdots, w_p\}$ be the unit coordinate basis for $\mathbb{R}^m$, $\mathbb{R}^n$ and $\mathbb{R}^p$. Then, $f(x) = f\left(\sum_{i=1}^m x_i e_i\right) = \sum_{i=1}^m x_i f(e_i)$, $f(e_i) = \sum_{j=1}^n f_j(e_i) u_j$ and $F = [F_{ji}]_{j=1,i=1}^{n,m}$ is the matrix of f. Also, $g(y) = g\left(\sum_{j=1}^n y_j u_j\right) = \sum_{j=1}^n y_j g(u_j)$, $g(u_j) = \sum_{k=1}^p g_k(u_j) w_k$ and $G = [G_{ki}]_{k=1,i=1}^{p,n}$ is the matrix of g. Now, $(g \circ f)(x) = (g \circ f)(\sum_{i=1}^m x_i e_i) = \sum_{i=1}^m (g \circ f)(e_i) x_i$ and*

$$\begin{aligned}
(g \circ f)(e_i) = g(f(e_i)) = g\left(\sum_{j=1}^n f_j(e_i) u_j\right) &= \sum_{j=1}^n f_j(e_i) g(u_j) = \sum_{j=1}^n f_j(e_i) \sum_{k=1}^p g_k(u_j) w_k \\
&= \sum_{k=1}^p \left(\sum_{j=1}^n f_j(e_i) g_k(u_j)\right) w_k.
\end{aligned}$$

Hence, the matrix of $(g \circ f)$ is given by $\Pi = \left[\sum_{j=1}^n f_j(e_i) g_k(u_j)\right]_{k=1,i=1}^{p,m}$. We define this matrix to be the (Cayley) product of the matrices G and F and we write $\Pi = GF$. Thus, the well known formula for multiplication of matrices derives from the composition of two linear functions.

Example 21. *1. Let $B \subset \mathbb{R}^K$ and $T \in L(B, \mathbb{R}^n)$. Let $X \in \mathcal{M}^{n\times K}$ be the matrix associated with T, such that, as in Example 19 we write $T(b) = Xb$ where $b \in B$. Let $\{e_k\}_{k=1}^K$ be a basis for $\mathbb{R}^K$, where e_k is the unit coordinate vector of dimension $K \times 1$. $X = (X_{.1} \ \cdots \ X_{.K})$ and $T(e_k) = Xe_k = X_{.k}$. The rank of T is the cardinality of $\{X_{.1}, \cdots, X_{.K}\} = K$. Recall that the null$(T) = \{a : Xa = \theta\}$ and by Theorem 41 we have that dim(null$(T)) = 0$.*
2. Note that if $y = Xb$ and $y = X\beta$, we can conclude that $X(b - \beta) = \theta$ and this implies $b = \beta$ if the columns of X are linearly independent or, equivalently, rank$(T) = K$.
3. If $Y = T(\beta) + \epsilon$, Y is the element of a linear variety, not element of the image of T. A very interesting question is whether or not there exists an element in the image of T that is "closest" to Y. For example, is there a solution for the minimization problem

$$\min_{\beta} \|Y - T(\beta)\|_E?$$

10.2 Bounded linear functions

We start by establishing that linear functions defined on finite dimensional spaces are always continuous.

Theorem 47. *Let $(\mathbb{X}, \|\cdot\|_{\mathbb{X}})$ and $(\mathbb{Y}, \|\cdot\|_{\mathbb{Y}})$ be normed vector spaces with dim$(\mathbb{X}) = n \in \mathbb{N}$. $f \in L(\mathbb{X}, \mathbb{Y})$ implies that f is continuous on $\mathbb{X}$.*

Proof. Since $dim(\mathbb{X}) = n \in \mathbb{N}$, there exists a basis $\{b_i\}_{i=1}^n$ and a collection of scalars $\{s_i\}_{i=1}^n$ such that any $x \in \mathbb{X}$ can be written as $x = \sum_{i=1}^n s_i b_i$. Linearity of f implies $f(x) = \sum_{i=1}^n s_i f(b_i)$ and by the properties of norms,

$$\|f(x)\|_{\mathbb{Y}} \leq \sum_{i=1}^n |s_i| \|f(b_i)\|_{\mathbb{Y}}.$$

The set of real numbers $\{\|f(b_i)\|_{\mathbb{Y}}\}_{i=1}^n$ is finite and we define $M := \max_{1\leq i\leq n} \|f(b_i)\|_{\mathbb{Y}}$. Hence,

$$\|f(x)\|_{\mathbb{Y}} \leq M \sum_{i=1}^n |s_i|.$$

Now, recall that if f is linear $f(\theta) = \theta$ and continuity at any point $x_0 \in \mathbb{X}$ establishes continuity on $\mathbb{X}$. As in Theorem 35 define the norm $\|x\|_1 = \sum_{i=1}^n |s_i|$ on $\mathbb{X}$ and take $x_0 = \theta$. Then, we have

$$\|f(x) - f(x_0)\|_{\mathbb{Y}} = \|f(x) - f(\theta)\|_{\mathbb{Y}} = \|f(x)\|_{\mathbb{Y}} \leq M \sum_{i=1}^n |s_i| = M\|x\|_1.$$

But in finite dimensional spaces, from Theorem 35, any two norms are equivalent. Hence, there exists $C > 0$ such that $\|x\|_1 \leq C\|x\|_{\mathbb{X}}$, which gives

$$\|f(x)\|_{\mathbb{Y}} \leq M\|x\|_1 \leq MC\|x\|_{\mathbb{X}}.$$

Hence, for any $\epsilon > 0$ whenever $\|x\|_{\mathbb{X}} < \epsilon/MC$, $\|f(x)\|_{\mathbb{Y}} \leq \epsilon$. $\square$

We now provide a definition of boundedness of linear functions.

Definition 46. *Let $f : (\mathbb{X}, \|\cdot\|_{\mathbb{X}}) \to (\mathbb{Y}, \|\cdot\|_{\mathbb{Y}})$ be linear. f is bounded if there exists a real scalar $C > 0$ such that for all $x \in \mathbb{X}$ we have*

$$\|f(x)\|_{\mathbb{Y}} \leq C\|x\|_{\mathbb{X}}.$$

The norm of f, denoted by $\|f\|$ is defined as $\|f\| := \inf\{C : \|f(x)\|_{\mathbb{Y}} \leq C\|x\|_{\mathbb{X}}$ for all $x \in \mathbb{X}\}$.

Remark 11. *If $(\mathbb{X}, \|\cdot\|_{\mathbb{X}})$ is not finite dimensional in Theorem 47, it is still true that $\|f(x) - f(\theta)\|_{\mathbb{Y}} = \|f(x)\|_{\mathbb{Y}}$, and if $\|f(x)\|_{\mathbb{Y}} \leq C\|x\|_{\mathbb{Y}}$, that is f is bounded, f will be continuous at θ and consequently everywhere.*

Remark 12. *By definition $\|f(x)\|_{\mathbb{Y}} \leq \|f\|\|x\|_{\mathbb{X}}$ for all $x \in \mathbb{X}$. Consequently, for all $x \neq \theta$ we have $\frac{\|f(x)\|_{\mathbb{Y}}}{\|x\|_{\mathbb{X}}} \leq \|f\|$ and by definition of supremum (the least upper bound on a set of numbers in $\mathbb{R}$) we have*

$$\|f\| = \sup_{\|x\|_{\mathbb{X}} \neq 0} \frac{\|f(x)\|_{\mathbb{Y}}}{\|x\|_{\mathbb{X}}}.$$

Also, letting $x' = x/\|x\|_{\mathbb{X}}$ we have that $\|x'\|_{\mathbb{X}} = 1$ and substituting x with $x'\|x\|_{\mathbb{X}}$ we have

$$\frac{\|f(x)\|_{\mathbb{Y}}}{\|x\|_{\mathbb{X}}} = \frac{\|f(x'\|x\|_{\mathbb{X}})\|_{\mathbb{Y}}}{\|x'\|x\|_{\mathbb{X}}\|_{\mathbb{X}}} = \frac{\|x\|_{\mathbb{X}}\|f(x')\|_{\mathbb{Y}}}{\|x'\|_{\mathbb{X}}\,\|x\|_{\mathbb{X}}} = \|f(x')\|_{\mathbb{Y}},$$

where the second equality follows from the properties of norms and linearity of f. Thus,

$$\|f\| = \underset{\|x\|_{\mathbb{X}}\neq 0}{sup} \frac{\|f(x)\|_{\mathbb{Y}}}{\|x\|_{\mathbb{X}}} = \sup_{\|x\|_{\mathbb{X}}=1} \|f(x)\|_{\mathbb{Y}}.$$

Theorem 48. *A linear function $f : (\mathbb{X}, \|\cdot\|_{\mathbb{X}}) \to (\mathbb{Y}, \|\cdot\|_{\mathbb{Y}})$ is bounded if, and only if, f is continuous on $\mathbb{X}$.*

Proof. Let f be bounded. Then, there exists a real scalar $C > 0$ such that for all $x \in \mathbb{X}$ we have $\|f(x)\|_{\mathbb{Y}} \leq C\|x\|_{\mathbb{X}}$. Consider a sequence $\{x_n\}_{n\in\mathbb{N}} \in \mathbb{X}$ such that $x_n \to \theta$, i.e., $\|x_n - \theta\|_{\mathbb{X}} \to 0$. Then, $\|f(x_n)\|_{\mathbb{Y}} \leq C\|x_n\|_{\mathbb{X}} \to 0$. Since $f(\theta) = \theta$ and $\|f(x_n) - f(\theta)\|_{\mathbb{Y}} \to 0$, f is continuous at θ. Then, by Theorem 38, f is continuous on $\mathbb{X}$.

Let f be continuous on $\mathbb{X}$. In particular, there exists $\delta > 0$ such that $\|f(x)\|_{\mathbb{Y}} \leq 1$ whenever $\|x\|_{\mathbb{X}} \leq \delta$. Now, for all $x \in \mathbb{X}$ such that $x \neq \theta$ we have that $\|x\|_{\mathbb{X}} \neq 0$ and we define $x' = \frac{\delta x}{\|x\|_{\mathbb{X}}}$. Then, by the properties of norms we have $\|x'\|_{\mathbb{X}} = \delta$. By linearity of f we have that $f(x) = f\left(\frac{\delta x}{\|x\|_{\mathbb{X}}}\right)\frac{\|x\|_{\mathbb{X}}}{\delta}$ and, consequently, $\|f(x)\|_{\mathbb{Y}} = \frac{\|x\|_{\mathbb{X}}}{\delta}\|f\left(\frac{\delta x}{\|x\|_{\mathbb{X}}}\right)\|_{\mathbb{Y}} \leq \frac{\|x\|_{\mathbb{X}}}{\delta}$, where the last inequality follows from $\|f(x)\|_{\mathbb{Y}} \leq 1$. This establishes that f is bounded with $C = 1/\delta$. □

We denote by $B(\mathbb{X}, \mathbb{Y})$ be the collection of all bounded linear functions from the normed space $(\mathbb{X}, \|\cdot\|_{\mathbb{X}})$ to the normed space $(\mathbb{Y}, \|\cdot\|_{\mathbb{Y}})$.

Theorem 49. *$B(\mathbb{X}, \mathbb{Y})$ is a normed vector space with norm $\|f\|$ given in Definition 46.*

Proof. Let $f, g \in B(\mathbb{X}, \mathbb{Y})$ and define $s(x) = f(x) + g(x)$, $p(x) = af(x)$ and the null vector to be the function $n(x) = 0$ for all $x \in \mathbb{X}$. We must show that $s(x)$ and $p(x)$ are bounded (we already know they are linear). By the properties of norms and the fact that f and g are bounded, we have

$$\|s(x)\|_{\mathbb{Y}} \leq \|f(x)\|_{\mathbb{Y}} + \|g(x)\|_{\mathbb{Y}} \leq (\|f\| + \|g\|)\|x\|_{\mathbb{X}}.$$

Thus, s is bounded and $\|s\| = \|f\| + \|g\|$. Also,

$$\|p(x)\|_{\mathbb{Y}} = |a|\|f(x)\|_{\mathbb{Y}} \leq |a|\|f\|\|x\|_{\mathbb{X}}.$$

Thus, p is bounded and $\|p\| = |a|\|f\|$. That $n(x) = 0$ is linear and bounded is obvious. $\square$

Recall that if $\mathbb{X}$ is a finite dimensional space, every linear function defined on $\mathbb{X}$ is continuous, and by Theorem 48 it is bounded. The next theorem says that the norm of a (bounded) linear function $f : (\mathbb{R}^m, \|\cdot\|_E) \to (\mathbb{R}^n, \|\cdot\|_E)$ is related to the maximum element of its matrix and the dimension of the spaces.

Theorem 50. *Let f be a linear function from the normed space $(\mathbb{R}^m, \|\cdot\|_E)$ to the normed space $(\mathbb{R}^n, \|\cdot\|_E)$ where $\|x\|_E = (\sum_{i=1}^m x_i^2)^{1/2}$ and $\|y\|_E = (\sum_{i=1}^n y_i^2)^{1/2}$. Let A be the matrix of f with typical element a_{ij} and $\mu := \max_{i=1,\cdots,n;\, j=1,\cdots,m} |a_{ij}|$. Then,*

$$\mu \leq \|f\| \leq \mu\sqrt{mn}.$$

Proof. For all $x \in \mathbb{R}^m$ we can write $y = Ax$, where $y_i = \sum_{j=1}^m a_{ij}x_j$. By the Cauchy-Schwarz inequality, we have

$$|y_i| = \left|\sum_{j=1}^m a_{ij}x_j\right| \leq \left(\sum_{j=1}^m a_{ij}^2\right)^{1/2}\left(\sum_{j=1}^m x_j^2\right)^{1/2} = \left(\sum_{j=1}^m a_{ij}^2\right)^{1/2}\|x\|_E \leq (m\mu^2)^{1/2}\|x\|_E$$

Then,

$$\begin{aligned}\|f(x)\|_E = \|y\|_E = \left(\sum_{i=1}^n y_i^2\right)^{1/2} &= \left(\sum_{i=1}^n |y_i|^2\right)^{1/2} \leq \left(\sum_{i=1}^n (m\mu^2)\|x\|_E^2\right)^{1/2} = (n(m\mu^2)\|x\|_E^2)^{1/2} \\ &= \sqrt{nm}\mu\|x\|_E.\end{aligned}$$

Thus, by definition of $\|f\|$ as the infimum of the sets of constants that satisfies $\|f(x)\|_E \leq C\|x\|_E$ for all x, we have $\|f\| \leq \sqrt{nm}\mu$.

Now, let $\{e_j\}_{j=1}^m$ be the basis composed of unit coordinate vectors associated with the matrix A and $n_j = \|f(e_j)\|_E$. Note that $n_j = \|Ae_j\|_E = \|A_{\cdot j}\|_E$ where $A_{\cdot j}$ is the j^{th} column of A. Hence, $n_j = (\sum_{i=1}^n a_{ij}^2)^{1/2}$. Let $\bar{n} = \max_{j=1,\cdots,m} n_j$. By definition of $\|f\|$ it must be that $\|f\| \geq \bar{n} = \max_{j=1,\cdots,m} (\sum_{i=1}^n a_{ij}^2)^{1/2}$. But $(\sum_{i=1}^n a_{ij}^2)^{1/2} \geq \max_{j=1,\cdots,m} |a_{ij}|$ since $\max_{j=1,\cdots,m} |a_{ij}| = \max_{j=1,\cdots,m} (a_{ij}^2)^{1/2} \leq \sqrt{\sum_{i=1}^n a_{ij}^2} = n_j$. Hence, $\|f\| \geq \bar{n} = \max_{j=1,\cdots,m} n_j \geq \max_{j=1,\cdots,m} \max_{i=1,\cdots,n} |a_{ij}| = \mu$. $\square$

Remark 13. *1. The proof of the previous theorem relies on the Cauchy-Schwarz Inequality (in $\mathbb{R}^n$). To see its validity let $a_1, \cdots, a_n, b_1, \cdots, b_n$ be a real numbers. Then, for all $\lambda \in \mathbb{R}$ consider*

$$\sum_{i=1}^n (a_i - \lambda b_i)^2 = \sum_{i=1}^n a_i^2 + \lambda^2 \sum_{i=1}^n b_i^2 - 2\lambda \sum_{i=1}^n b_i a_i.$$

In particular, if $\sum_{i=1}^n b_i^2 \neq 0$, put $\lambda = \frac{\sum_{i=1}^n a_i b_i}{\sum_{i=1}^n b_i^2}$. Then,

$$\sum_{i=1}^n a_i^2 - \frac{(\sum_{i=1}^n a_i b_i)^2}{\sum_{i=1}^n b_i^2} \geq 0 \text{ and consequently } \left(\sum_{i=1}^n a_i b_i\right)^2 \leq \left(\sum_{i=1}^n b_i^2\right)\left(\sum_{i=1}^n a_i^2\right).$$

Hence, $|\sum_{i=1}^n a_i b_i| \leq \left(\sum_{i=1}^n b_i^2\right)^{1/2} \left(\sum_{i=1}^n a_i^2\right)^{1/2}$.

2. If $(f \circ g)(x) = f(g(x))$ and f and g are bounded linear functions, then $\|(f \circ g)(x)\| = \|f(g(x))\| \leq \|f\|\|g(x)\| \leq \|f\|\|g\|\|x\|$ for all x. Therefore, $\|f \circ g\| \leq \|f\|\|g\|$.

When f is a linear function from the normed space $(\mathbb{X}, \|\cdot\|_{\mathbb{X}})$ to itself, we call f a linear operator. In many instances, we are interested in linear operators on $(\mathbb{R}^n, \|\cdot\|_E)$. If we take as a basis for $\mathbb{R}^n$ the collection of unit coordinate vectors, then $f(x) = y = Ax$ where A is the $n \times n$ matrix associated with f (why?). If f has an inverse, then by definition

$$f^{-1}(y) = f^{-1}(f(x)) = (f^{-1} \circ f)(x) = x.$$

If the matrix associated with f^{-1} is denoted by A^{-1}, then $f^{-1}(y) = A^{-1}y = A^{-1}(Ax) = A^{-1}Ax = I_n x = x$. The matrix I_n is associated with the identity operator, i.e., $\mathcal{I}(x) = x$. The identity operator is a composition of two linear functions, and since $\mathbb{R}^n$ is a finite dimensional space, f is continuous, and consequently bounded. So, by the last remark $\mathcal{I}$ is bounded. Thus, noting that $\|\mathcal{I}(x)\| = \|x\|$, there exists $C > 0$ such that for all $x \in \mathbb{X}$, $\|\mathcal{I}(x)\| = \|x\| \leq C\|x\|$, giving $\|\mathcal{I}\| = 1$. Then,

$$1 = \|f^{-1} \circ f\| \leq \|f^{-1}\|\|f\| \text{ and consequently } \|f^{-1}\| \geq \frac{1}{\|f\|}.$$

Hence, invertible linear operators are those associated with a matrix A^{-1} such that $AA^{-1} = A^{-1}A = I_n$ and we call A^{-1} the inverse of A. Since invertibility means that f is one-to-one and onto, and given that by Theorem 41

$$dim(\mathbb{R}^n) = dim(im(f)) + dim(null(f))$$

we have that $dim(null(f)) = 0$ and $rank(f) = n$.

Theorem 51. *Let f, g be invertible linear operators on the normed space $(\mathbb{R}^n, \|\cdot\|_E)$. Then, $f \circ g$ is invertible and $(f \circ g)^{-1} = g^{-1} \circ f^{-1}$.*

Proof. Since, f and g are linear, so is $(f \circ g)$ from Example 20 . Given that g is linear $g(\theta) = \theta$, and since it is invertible (one-to-one and onto) $g^{-1}(\theta) = \theta$ and for any $x \neq \theta$, $g(x) \neq \theta$. But since f is invertible $f(g(x)) \neq \theta$ for $x \neq \theta$. Consequently, $(f \circ g)$ is invertible. Now, let $y = (f \circ g)(x) = f(g(x))$ and since f is invertible, $f^{-1}(y) = g(x)$. Furthermore, since g is invertible, $g^{-1}(f^{-1}(y)) = (g^{-1} \circ f^{-1})(y) = x$. Consequently, $(f \circ g)^{-1} = g^{-1} \circ f^{-1}$. □

Remark 14. 1. *If $M(f)$ is the matrix of f and $M(g)$ is the matrix of g, then we know that $(f \circ g)$ has matrix $M(f)M(g)$ and $(f \circ g)^{-1}$ has matrix $M(g)^{-1}M(f)^{-1}$. Hence, $(M(f)M(g))^{-1} = M(g)^{-1}M(f)^{-1}$.*

2. *The collections of all invertible linear operators on $(\mathbb{R}^n, \|\cdot\|_E)$ is, of course, a subset of the linear operators on $(\mathbb{R}^n, \|\cdot\|_E)$. Hence, a norm on $L(\mathbb{R}^n, \mathbb{R}^n)$ applies to the collection of invertible linear operators.*

Theorem 52. *Let f be a linear (and bounded) operator on the normed space $(\mathbb{R}^n, \|\cdot\|_E)$ and $\mathcal{I}$ be the identity operator.*

1. *If $\|f\| < 1$, then $\mathcal{I} - f$ is invertible and $\|(\mathcal{I}-f)^{-1}\| \leq \frac{1}{1-\|f\|}$.*

2. *If $\|\mathcal{I} - f\| < 1$ then f is invertible.*

Proof. 1. Let $x \neq \theta$ be an element of $\mathbb{R}^n$.

$$\begin{aligned} \|(\mathcal{I}-f)(x)\|_E = \|\mathcal{I}(x) - f(x)\|_E &\geq \|x\|_E - \|f(x)\|_E \text{ by Lemma 1} \\ &\geq \|x\|_E - \|f\|\|x\|_E \text{ by boundedness of } f \\ &= \|x\|_E(1-\|f\|) > 0 \text{ since } \|f\| < 1. \end{aligned}$$

$\|(\mathcal{I}-f)(x)\|_E > 0$ implies $x \neq f(x)$ for all $x \neq \theta$, which implies that $\mathcal{I} - f$ is invertible. Now, let $(\mathcal{I}-f)(x) = y$, $x = (\mathcal{I}-f)^{-1}(y)$ and note that

$$\|(\mathcal{I}-f)\left((\mathcal{I}-f)^{-1}(y)\right)\|_E = \|(\mathcal{I}-f)\circ(\mathcal{I}-f)^{-1}(y)\|_E = \|y\|_E.$$

Hence, $\|y\|_E = \|(\mathcal{I}-f)(x)\|_E \geq (1-\|f\|)\|(\mathcal{I}-f)^{-1}(y)\|$ which implies that $\|(\mathcal{I}-f)^{-1}(y)\| \leq \frac{\|y\|_E}{(1-\|f\|)}$, and consequently $\frac{\|(\mathcal{I}-f)^{-1}(y)\|}{\|y\|_E} \leq \frac{1}{1-\|f\|}$. Since $\frac{1}{1-\|f\|}$ is an upper bound for $\frac{\|(\mathcal{I}-f)^{-1}(y)\|}{\|y\|_E}$, $\|(\mathcal{I}-f)^{-1}\| \leq \frac{1}{1-\|f\|}$.

2. Let $g(x) = \mathcal{I}(x) - f(x)$. Then, $f(x) = \mathcal{I}(x) - g(x)$. From part 1, $\|g\| < 1$ implies $\mathcal{I} - g$ is invertible, which is equivalent to $\|\mathcal{I} - f\| < 1$ and this implies that f is invertible. □

Theorem 53. *Let f, g be (bounded) linear operators on $(\mathbb{R}^n, \|\cdot\|_E)$. If f is invertible and $\|f - g\| < \frac{1}{\|f^{-1}\|}$, then g is invertible. Furthermore, $\|g^{-1}\| \leq \frac{\|f^{-1}\|}{1-\|f^{-1}\circ(f-g)\|}$.*

Proof. Note that

$$\begin{aligned} g(x) &= f(x) - (f(x) - g(x)) \\ &= f(x) - f(f^{-1}(f(x) - g(x))) \text{ since } f \text{ is invertible} \\ &= f\left(x - f^{-1}(f(x) - g(x))\right) \text{ since } f \text{ is linear} \\ &= \left(f \circ (\mathcal{I} - f^{-1} \circ (f - g))\right)(x). \end{aligned}$$

Now, $\|f^{-1} \circ (f - g)\| \leq \|f^{-1}\|\|f - g\|$ from Remark 13.2 and $\|f^{-1}\|\|f - g\| < 1$ since $\|f - g\| < 1/\|f^{-1}\|$ by assumption. Consequently, $\mathcal{I} - f^{-1} \circ (f - g)$ is invertible and $g = (f \circ (\mathcal{I} - f^{-1} \circ (f - g)))$ is the composition of two invertible functions. Hence, by Theorem 51, g is invertible and $g^{-1} = (\mathcal{I} - f^{-1} \circ (f - g))^{-1} \circ f^{-1}$. Then, $\|g^{-1}\| \leq \|(\mathcal{I} - f^{-1} \circ (f - g))^{-1}\|\|f^{-1}\| \leq \frac{1}{1 - \|f^{-1} \circ (f-g)\|}\|f^{-1}\|$ from Theorem 52. $\square$

Definition 47. *Let $I(\mathbb{R}^n, \|\cdot\|_E)$ be a collection of linear bounded invertible operators on $(\mathbb{R}^n, \|\cdot\|_E)$ and $i(f)$ be a function that assigns to any linear bounded invertible operator f its inverse. That is, $i : I(\mathbb{R}^n, \|\cdot\|_E) \to I(\mathbb{R}^n, \|\cdot\|_E)$ with $i(f) = f^{-1}$.*

Theorem 54. *Let i be as in the previous definition. i is continuous.*

Proof. Let $f, g \in I(\mathbb{R}^n, \|\cdot\|_E)$. We must show that for all $\epsilon > 0$, there exits $\delta > 0$ such that whenever $\|f - g\| < \delta$ we have $\|f^{-1} - g^{-1}\| < \epsilon$. If $\|f - g\| < 1/(2\|f^{-1}\|)$, then by Theorem 53, g is invertible. Then, $\|g^{-1}\| \leq \frac{\|f^{-1}\|}{1 - \|f^{-1} \circ (f-g)\|}$ and $\|f^{-1} \circ (f - g)\| < 1/2$. Hence, $\|g^{-1}\|| \leq 2\|f^{-1}\|$. Now, $g^{-1} - f^{-1} = (\mathcal{I} - f^{-1} \circ g) \circ g^{-1} = f^{-1} \circ (f - g) \circ g^{-1}$ and

$$\|g^{-1} - f^{-1}\| \leq \|f^{-1}\|\|(f - g) \circ g^{-1}\| \leq \|f^{-1}\|\|(f - g)\|\|g^{-1}\| \leq 2\|f^{-1}\|^2\|(f - g)\|.$$

For any $\epsilon > 0$, put $\delta = (\epsilon/2)\|f^{-1}\|^{-2}$, then $\|f - g\| < \delta$ which implies that $\|g^{-1} - f^{-1}\| < 2\|f^{-1}\|^2(\epsilon/2)\|f^{-1}\|^{-2} = \epsilon$. $\square$

11 Matrices

The matrices associated with linear functions can be manipulated in interesting and very revealing ways.

11.1 Transposition, trace, vectorization and rank

It is useful to think of the rows and columns of a matrix associated with a linear function $f : (\mathbb{R}^m, \|\cdot\|_m) \to (\mathbb{R}^n, \|\cdot\|_n)$ as elements of a vector space. Hence, if $A \in \mathcal{M}^{n \times m}$ we can write

$$A = \begin{pmatrix} A_{1\cdot} \\ \vdots \\ A_{n\cdot} \end{pmatrix} \text{ or } A = \begin{pmatrix} A_{\cdot 1} & \cdots & A_{\cdot m} \end{pmatrix}$$

where $A_{i\cdot} = (A_{i1} \cdots A_{im}) \in \mathbb{R}^m$ and $A_{\cdot j} = (A_{1j} \cdots A_{nj})^T \in \mathbb{R}^n$.

The column space of the matrix A is the set of all $y \in \mathbb{R}^n$ such that $y = \sum_{j=1}^m s_j A_{\cdot j}$ where $\{s_j\}_{j=1}^m$ is a collection of scalars. Clearly, the dimension of the column space of A is $\leq m$. This space is sometimes referred to as the span of A and denoted by span(A).

When $n = m$ we say that A is a square matrix and when, in this case, $A_{ij} = A_{ji}$ we say that the matrix A is a symmetric matrix. If the square matrix A is such that $A_{ij} = 0$ for all $i \neq j$ we say that A is a diagonal matrix. The (main) diagonal of A is the $n \times 1$ vector with elements A_{ii}. If the square matrix A is such that $A_{ij} = 0$ for all $i > j(i < j)$ we say that the matrix A is an upper (lower) triangular matrix.

The transposition of A is denoted by A^T and is the matrix with typical element A_{ji}. Clearly, for A symmetric we have $A = A^T$. If $A \in \mathcal{M}^{n \times m}$ we define $vec : \mathcal{M}^{n \times m} \to \mathbb{R}^{nm}$ as

$$vec(A) = \begin{pmatrix} A_{\cdot 1} \\ \vdots \\ A_{\cdot m} \end{pmatrix}.$$

The trace of a square matrix is the sum of the elements in its main diagonal, i.e. $tr(A) = \sum_{i=1}^n A_{ii}$. Often, it is convenient to "partition" a matrix into sub-matrices. For example, we may write

$$A = \begin{pmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{pmatrix}$$

where $A_{1,1}$ is $n_1 \times m_1$, $A_{1,2}$ is $n_1 \times m_2$, $A_{2,1}$ is $n_2 \times m_1$, $A_{2,2}$ is $n_2 \times m_2$ with $n_1 + n_2 = n$ and $m_1 + m_2 = m$.

If $s \in \mathbb{R}$ we say that sA is the matrix with typical element sA_{ij}. If two matrices have the same row and column dimensions, $A + B$ is the matrix with typical element $A_{ij} + B_{ij}$. $C = AB$, the product of A and B was introduced in Example 20. It is easy to verify that

1. $(A+B)^T = A^T + B^T$, $(AB)^T = B^T A^T$,

2. $tr(A) = tr(A^T)$, $tr(sA) = s\,tr(A)$ for $s \in \mathbb{R}$, $tr(A+B) = tr(A) + tr(B)$,

3. $A(B+C) = AB + AC$, $(B+C)A = BA + CA$, $(AB)C = A(BC)$, where in general $AB \neq BA$,

4. $AI_m = I_n A = A$, $tr(AB) = tr(BA)$.

As a matter of terminology, the rank of a matrix A, denoted by $r(A)$, is the cardinality of the largest number of independent rows or columns of A. Hence, we immediately have that

$$r(A) \leq \min\{n, m\}.$$

We now prove that $r(A)$=dim(span(A)), i.e., the number of linearly independent columns of A.

Theorem 55. *Let $A \in \mathcal{M}^{n\times m}$. Then, $r(A)$=dim(span(A))=dim(span(A^T)).*

Proof. Write $A = \begin{pmatrix} A_{1\cdot} \\ \vdots \\ A_{n\cdot} \end{pmatrix}$ so that $A^T = \begin{pmatrix} A_{1\cdot}^T & \cdots & A_{n\cdot}^T \end{pmatrix}$ and suppose dim(span(A^T))= $k \leq n$. Then,

$$A_{j\cdot}^T = \sum_{l=1}^{k} s_{lj} e_l \text{ for } j = 1, \cdots, n \text{ where } s_{lj} \in \mathbb{R}$$

and $\{e_l\}_{l=1}^k$ is a basis for span(A^T). Select the i-th element of each $A_{j\cdot}^T$ and form the vector

$$\begin{pmatrix} A_{1i} \\ \vdots \\ A_{ni} \end{pmatrix} = \begin{pmatrix} \sum_{l=1}^k e_{li} s_{l1} \\ \vdots \\ \sum_{l=1}^k e_{li} s_{ln} \end{pmatrix} = e_{1i} \begin{pmatrix} s_{11} \\ \vdots \\ s_{1n} \end{pmatrix} + \cdots + e_{ki} \begin{pmatrix} s_{k1} \\ \vdots \\ s_{kn} \end{pmatrix}$$
$$= \sum_{l=1}^{k} e_{li} s_l \text{ where } s_l = \begin{pmatrix} s_{l1} \\ \vdots \\ s_{ln} \end{pmatrix}.$$

Since $\begin{pmatrix} A_{1i} \\ \vdots \\ A_{ni} \end{pmatrix}$ is the i-th column of A, $\text{span}(A) \subseteq \text{span}\ \{s_1, \cdots, s_k\}$ and

$$\dim(\text{span}(A)) \leq k = \dim(\text{span}(A^T)).$$

Repeating the same argument with A^T gives

$$\dim(\text{span}(A^T)) \leq k = \dim(\text{span}((A^T)^T)) = \dim(\text{span}(A)).$$

Hence, $r(A) = \dim(\text{span}(A)) = \dim(\text{span}(A^T))$. □

Remark 15. *1. We say that a matrix is of full rank if its rank is equal to the smallest of its dimension. If that is not the case we say that a matrix is rank deficient. A square full rank matrix is called nonsingular. If not, it is called singular. It is easy to show that for $c \in \mathbb{R} - \{0\}$, $r(cA) = r(A)$*

2. *Let $A \in \mathcal{M}^{n\times m}$ with $r(A) = r \leq \min\{n, m\}$. From Theorem 46, there is one, and only one, linear function $f : \mathbb{R}^m \to \mathbb{R}^n$ such that $y = f(x) := Ax$. Since, the collection of unit coordinate vectors $\{e_j\}_{j=1}^m$ is a basis for $\mathbb{R}^m$, from Theorem 39, we have that $\{f(e_j) = A_{\cdot j}\}_{j=1}^m$ spans the $im(f)$. The rank of f is the number of linearly independent vectors in $\{A_{\cdot j}\}_{j=1}^m$, which by assumption is r. By Theorem 41*

$$dim(\mathbb{R}^m) = dim(im(f)) + dim(null(f)) \iff m = r + (m - r).$$

3. *It follows immediately from the previous theorem that the rows and columns of a matrix $A \in \mathcal{M}^{n\times m}$ with $r(A) = r \leq \min\{n, m\}$ can be re-arranged (by repeated permutation of its rows and columns) to produce the matrix*

$$A_{FP} = \begin{pmatrix} B_{1,1} & B_{1,2} \\ B_{2,1} & B_{2,2} \end{pmatrix}$$

where the sub-matrix $B_{1,1} \in \mathcal{M}^{r\times r}$ is of full rank and the other sub-matrices are of conformable dimensions. A_{FP} is called the full rank partitioning of A.

If $A \in \mathcal{M}^{n\times n}$ is nonsingular, for the unit coordinate vector $e_i \in \mathbb{R}^n$, we have that there exists x_i such that $Ax_i = e_i$ for $i = 1, \cdots, n$. Hence, we may write $AX = I_n$ where the i-th column of X is x_i. X is called the right-inverse of A. But X is itself a nonsingular element of $\mathcal{M}^{n\times n}$ (why?), so it has a right-inverse we call Y, such that $XY = I_n$. Thus, $AXY = Y$ which gives $A = Y$ and $XA = I_n$. Hence, pre- or post-multiplying A by X gives I_n. In this case, X is called the inverse of A and denoted A^{-1} as earlier in these notes.

Theorem 56. *1. Let $A \in \mathcal{M}^{n\times m}$ with $r(A) = r \leq \min\{n, m\}$ and $X \in \mathcal{M}^{n\times n}$ and $Y \in \mathcal{M}^{m\times m}$ be nonsingular. Then, $r(XA) = r(AY) = r$.*

2. Let $A \in \mathcal{M}^{n\times n}$ be nonsingular, then $(A^{-1})^T = (A^T)^{-1}$.

3. Let $A \in \mathcal{M}^{m\times n}$ and $B \in \mathcal{M}^{n\times p}$. Then, $r(AB) \leq \min\{r(A), r(B)\}$.

4. Let $A, B \in \mathcal{M}^{m\times n}$. Then, $r(A + B) \leq r(A) + r(B)$.

Proof. 1. Let $x \in null(A)$. Then, $Ax = \theta$ which gives $XAx = \theta$. Thus, $x \in null(XA)$. Conversely, if $x \in null(XA)$, then $XAx = \theta$. Since, X is nonsingular, it has an inverse and $X^{-1}XAx = \theta \iff Ax = \theta$. Thus, $x \in null(A)$ and we conclude that $null(A) = null(XA)$. But since A and XA have the same number of columns $r(A) = r(XA)$. Now, $(AY)^T = Y^T A^T$. Since $r(A^T) = r$ and Y^T is nonsingular, we use the first part of the result to conclude that $r(AY) = r$.

2. By definition $AA^{-1} = I_n$. Then, $(AA^{-1})^T = (A^{-1})^T A^T = I_n$. Alternatively, by definition $(A^T)^{-1}A^T = I_n$. Consequently,

$$(A^{-1})^T A^T = (A^T)^{-1}A^T \iff \left((A^{-1})^T - (A^T)^{-1}\right) A^T = \theta,$$

and since A^T is nonsingular, it must be that $(A^{-1})^T = (A^T)^{-1}$.

3. If $x \in null(B)$ then $Bx = \theta$ and $ABx = \theta$. Thus, $null(B) \subseteq null(AB)$. Then,

$r(null(B)) \leq r(null(AB))$. Since, B and AB have the same number of columns,

$$r(B) + r(null(B)) = r(AB) + r(null(AB))$$
$$\iff r(AB) - r(B) = r(null(B)) - r(null(AB)) \leq 0$$
$$\implies r(AB) \leq r(B).$$

Now, $r(AB) = r((AB)^T) = r(B^T A^T) \leq r(A^T) = r(A)$, where the inequality follows from the first step of the proof of 3. Thus, $r(AB) \leq \min r(A), r(B)$.

4. $A + B = \begin{pmatrix} A & B \end{pmatrix} \begin{pmatrix} I_n \\ I_n \end{pmatrix}$. From item 3., $r(A + B) \leq r(\begin{pmatrix} A & B \end{pmatrix})$, which is the the smallest number of linearly independent columns in $\begin{pmatrix} A & B \end{pmatrix}$). But, this is smaller than the number of linearly independent columns in A plus the number of linearly independent columns in B. Thus, $r(A + B) \leq r(A) + r(B)$. □

Remark 16. *If $A \in \mathcal{M}^{m \times n}$, the matrix $A^T A$ is often encountered. It is called a Gramian matrix. Note that if $x \in null(A)$ then $Ax = \theta$ and $A^T Ax = \theta$. Also, if $x \in null(A^T A)$ then $A^T Ax = \theta$ and $x^T A^T Ax = (Ax)^T Ax = \theta$ if, and only if, $Ax = \theta$. But in this case, $x \in null(A)$. Hence, $null(A) = null(A^T A)$ and since A and $A^T A$ have the same number of columns, we have $r(A) = r(A^T A)$ (also $r(AA^T) = r(A)$).*

11.2 The determinant

Consider the set S_n of all $n! = 1 \cdot 2 \cdots n$ permutations of the n-tuple $(1, 2, \ldots, n)$. For example, if $n = 2$ we have $S_2 = \{(1, 2), (2, 1)\}$ where $\#S_2 = 2! = 2$ and if $n = 3$ we have $S_3 = \{(1, 2, 3), (1, 3, 2), (3, 1, 2), (3, 2, 1), (2, 1, 3), (2, 3, 1)\}$ where $\#S_3 = 3! = 6$. Each permutation is obtained by exchanging the adjacent elements of the original n-tuple a certain number of times (or not at all, as is the case of the original n-tuple). For example, if $n = 2$ the permutation $(2, 1)$ was obtained from one exchange. In the case of $n = 3$ the permutation $(1, 3, 2)$ was obtained from one exchange, but the permutation $(3, 2, 1)$ was obtained after 3 exchanges of adjacent elements. We represent this as

$$(1,2,3) \to (1,3,2) \to (3,1,2) \to (3,2,1)$$

where each arrow represents and exchange. Let sgn : $S_n \to \{-1,1\}$ be a function defined as $\text{sgn}(p) = (-1)^x$ where x is the number of exchanges needed to obtain the permutation p. For example,

$$\text{sgn} : S_2 \to \{-1,1\} \text{ is given by } \begin{cases} (-1)^0 = 1 \text{ if } p = (1,2) \\ (-1)^1 = -1 \text{ if } p = (2,1) \end{cases}$$

and

$$\text{sgn} : S_3 \to \{-1,1\} \text{ is given by } \begin{cases} (-1)^0 = 1 \text{ if } p = (1,2,3) \\ (-1)^1 = -1 \text{ if } p = (1,3,2),(2,1,3) \\ (-1)^2 = 1 \text{ if } p = (3,1,2),(2,3,1) \\ (-1)^3 = -1 \text{ if } p = (3,2,1) \end{cases}$$

Definition 48. *The determinant of $A \in \mathcal{M}^{n\times n}$ is a function $det : \mathcal{M}^{n\times n} \to \mathbb{R}$ defined as*

$$det(A) = \sum_{p\in S_n} sgn(p) A_{1j_1} A_{2j_2} \cdots A_{nj_n}$$

where $p = (j_1, j_2, \cdots, j_n)$ and A_{ij} is the (i,j) element of A.

Example 22. *If $n = 2$, $det(A) = A_{11}A_{22} - A_{12}A_{21}$. If $n = 3$, $det(A) = A_{11}A_{22}A_{33} + A_{13}A_{21}A_{32} + A_{12}A_{23}A_{31} - A_{11}A_{23}A_{32} - A_{12}A_{21}A_{33} - A_{13}A_{22}A_{31}$.*

Theorem 57. *Let $A \in \mathcal{M}^{n\times n}$ be nonsingular. Then, 1. $det(cA) = c^n det(A)$, 2. $det(A^T) = det(A)$, 3. if A is diagonal, lower or upper triangular $det(A) = \prod_{i=1}^n A_{ii}$*

Proof. 1. Since cA is the matrix obtained by multiplying every element of A by c,

$$\det(cA) = \sum_{p\in S_n} \text{sgn}(p) cA_{1j_1} cA_{2j_2} \cdots cA_{nj_n} = c^n \sum_{p\in S_n} \text{sgn}(p) A_{1j_1} A_{2j_2} \cdots A_{nj_n} = c^n \det(A).$$

2. Now, $\det(A^T) = \sum_{p\in S_n} \text{sgn}(p) A_{j_1 1} A_{j_2 2} \cdots A_{j_n n}$. For any permutation $p \in S_n$ the order of the factors in the product $A_{j_1 1} A_{j_2 2} \cdots A_{j_n n}$ can be re-arranged and ordered by the first index

without altering the product. Doing this for all permutations and observing that sgn(p) is not altered by any of these rearrangements gives

$$\det(A^T) = \sum_{p \in S_n} \text{sgn}(p) A_{1j_1} A_{2j_2} \cdots A_{nj_n} = \det(A).$$

3. This follows directly from the observation that for diagonal, lower or upper triangle matrices $A_{1j_1} A_{2j_2} \cdots A_{nj_n} = 0$ for all permutations except the permutation $\{j_1, \cdots, j_n\} = \{1, \cdots, n\}$. $\square$

It is convenient to express the determinant of A as linear combinations of the determinants of matrices of smaller dimensions. To this end, the square matrix $(n-1 \times n-1)$ obtained from deleting the i-th row and j-th column of A will be denoted by A^{ij}. The co-factor associated with the element A_{ij}, denoted $A_{(ij)}$, is

$$A_{(ij)} = (-1)^{i+j} \det(A^{ij})$$

and its minor is $\det(A^{ij})$. It can be shown that for $A \in \mathcal{M}^{n \times n}$ and any $i = 1, \cdots, n$

$$\det(A) = \sum_{j=1}^{n} A_{ij} A_{(ij)}. \tag{7}$$

Similarly, for any $j = 1, \cdots, n$, $\det(A) = \sum_{i=1}^{n} A_{ij} A_{(ij)}$. These are called Laplace expansions of the determinant of A.

The adjoint of A, denoted by Adj(A) is defined as $\text{Adj(A)} = ([A_{(ij)}])^T$. This matrix has the interesting property that $A\text{Adj(A)} = \text{Adj(A)}A = \det(A) I_n$.

We now give some properties of the determinant for $A \in \mathcal{M}^{n \times n}$. They are easy (but tedious) to verify.

1. If $A = \begin{pmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{pmatrix}$ where $A_{1,1}$ is $n_1 \times n_1$, $A_{1,2}$ is $n_1 \times n_2$, $A_{2,1}$ is $n_2 \times n_1$, $A_{2,2}$ is $n_2 \times n_2$ and either $A_{1,2}$ or $A_{2,1}$, or both, are matrices containing only zeros, then $\det(A) = \det(A_{1,1}) \det(A_{2,2})$.

2. If $A_{1,1}$ and $A_{2,2}$ are nonsingular

$$A^{-1} = \begin{pmatrix} A_{1,1}^{-1} + A_{1,1}^{-1}A_{1,2}Z^{-1}A_{2,1}A_{1,1}^{-1} & -A_{1,1}^{-1}A_{1,2}Z^{-1} \\ -Z^{-1}A_{2,1}A_{1,1}^{-1} & Z^{-1} \end{pmatrix}$$

where $Z = A_{2,2} - A_{2,1}A_{1,1}^{-1}A_{1,2}$. Z is called the Schur complement of $A_{1,1}$.

3. $\det(AB) = \det(A)\det(B)$ but in general $\det(A+B) \neq \det(A) + \det(B)$

4. If A is nonsingular $\det(A) = 1/\det(A^{-1})$. Thus, $\det(A) = 0$ if, and only if, A is singular

5. For a nonsingular matrix A, $A^{-1} = (\det(A))^{-1}\mathrm{Adj}(A)$.

We close this section with a useful type of product of matrices.

Definition 49. *Let $A \in \mathcal{M}^{m\times n}$ and $B \in \mathcal{M}^{p\times q}$. The Kronecker product of A and B is denoted by $A \otimes B \in \mathcal{M}^{mp\times nq}$ with*

$$A \otimes B := \begin{pmatrix} A_{11}B & A_{12}B & \cdots & A_{1n}B \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1}B & A_{m2}B & \cdots & A_{mn}B \end{pmatrix}.$$

It can be easily checked that:

1. $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$

2. $(A \otimes B)(C \otimes D) = AC \otimes BD$

3. $(A \otimes B)^T = A^T \otimes B^T$

4. $\mathrm{tr}(A \otimes B) = \mathrm{tr}(A)\mathrm{tr}(B)$

5. $\mathrm{vec}(ABC) = (C^T \otimes A)\mathrm{vec}(B)$

6. If $A \in \mathcal{M}^{n\times n}$ and $B \in \mathcal{M}^{m\times m}$, $\det(A \otimes B) = \det(A)^n \det(B)^m$.

12 Eigenvalues and eigenvectors

Definition 50. *Let $A \in \mathcal{M}^{n\times n}$, $x \in \mathbb{R}^n - \{\theta\}$ and $\lambda \in \mathbb{R}$ or $\mathbb{C}$, such that $Ax = \lambda x$. We say that λ is an eigenvalue of A and x is an eigenvector of A corresponding to the eigenvalue λ. (λ, x) is called an eigenpair.*

Remark 17.

1. *Given one eigenpair (λ, x) we can immediately conclude that many others exist. Indeed, for $a \in \mathbb{R} - \{0\}$*

$$aAx = A(ax) = a\lambda x = \lambda(ax) \text{ and if } y := ax,\ Ay = \lambda y.$$

Hence, (λ, y) is an eigenpair. It is often the case that eigenvectors are normalized such that $\|x\| = 1$.

2. *$Ax = \lambda x$ implies $(A - \lambda I_n)x = \theta$. An obvious choice of x that satisfies this equation is $x = \theta$ (trivial solution), but since in the previous definition we have $x \neq \theta$, x will satisfy this equation (non-trivial solution) only if $A - \lambda I_n$ is not invertible, i.e., det$(A - \lambda I_n) = 0$. If $A = [A_{ij}]_{i=1,j=1}^{n,n}$ we have*

$$det(A - \lambda I_n) = det\begin{pmatrix} A_{11}-\lambda & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22}-\lambda & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn}-\lambda \end{pmatrix} = 0. \tag{8}$$

3. *If A is a diagonal, upper or lower triangular matrix det$(A - \lambda I_n) = \prod_{i=1}^{n}(A_{ii} - \lambda)$. This equals zero for $\lambda = A_{ii}$. The associated eigenvector is the unit coordinate vector.*

Let us take a closer look at equation (8), known in this context as the characteristic equation. First consider the case where $n = 2$. Then,

$$\begin{aligned}\det(A - \lambda I) &= \lambda^2 - (A_{11} + A_{22})\lambda + (A_{11}A_{22} - A_{12}A_{21}) \\ &= (-1)^2\lambda^2 + (-1)^1\lambda \sum_{\{i_1\}\in C_1^2} \det(A_{\{i_1\}}) + \det(A).\end{aligned}$$

Here, $A_{\{i_1\}}$ denotes the matrix obtained by keeping the row and column i_1 in the matrix A and C_1^2 is the set of combinations (2 choose 1) from the set $\{1, 2\}$. That is, C_1^2 has two elements: 1 and 2. For the case $n = 3$, we obtain

$$\det(A - \lambda I) = (-1)^3\lambda^3 + (-1)^2\lambda^2 \sum_{\{(i_1)\}\in C_1^3} \det(A_{\{i_1\}})$$
$$+ (-1)^1\lambda^1 \sum_{\{(i_1,i_2)\}\in C_2^3} \det(A_{\{i_1,i_2\}}) + \det(A).$$

where $A_{\{i_1,i_2\}}$ denotes the matrix obtained by keeping the rows and columns i_1 and i_2 in the matrix A. After similar algebraic manipulation and induction, we obtain for $n \in \mathbb{N}$

$$\det(A - \lambda I) = (-\lambda)^n + (-\lambda)^{n-1} \sum_{\{(i_1)\}\in C_1^n} \det(A_{\{i_1\}}) + (-\lambda)^{n-2} \sum_{\{(i_1,i_2)\}\in C_2^n} \det(A_{\{i_1,i_2\}})$$
$$\cdots + (-\lambda) \sum_{\{(i_1,i_2,\ldots,i_{n-1})\}\in C_{n-1}^n} \det(A_{\{i_1,i_2,\ldots,i_{n-1}\}}) + \det(A). \tag{9}$$

Equation (9) shows that $\det(A - \lambda I)$ is an n-degree polynomial in λ. As such, the characteristic equation has n solutions not necessarily real or distinct. For example, if $n = 2$

$$\lambda_1 = \frac{1}{2}\left(\mathrm{tr}(A) + \sqrt{\mathrm{tr}(A)^2 - 4\det(A)}\right) \text{ and } \lambda_2 = \frac{1}{2}\left(\mathrm{tr}(A) - \sqrt{\mathrm{tr}(A)^2 - 4\det(A)}\right).$$

The set containing the distinct eigenvalues of A is called the spectrum of A and is denoted by $\sigma(A)$. Given an eigenvalue λ_i, an eigenvector x_i can be obtained by solving the equation

$$Ax_i = \lambda_i x_i \Leftrightarrow (A - \lambda_i I_n)x_i = \theta. \tag{10}$$

That is, the eigenvectors associated with λ_i are the elements in the null space of $(A - \lambda_i I_n)$. The space of solutions for $(A - \lambda_i I_n)x_i = \theta$ for $i = 1, \cdots, n$ is called the eigenspace of A belonging to λ_i.

Since $\det(A - \lambda_i I_n) = 0$, $r(A - \lambda_i I_n) < n$. The dimension of the eigenspace associated with the eigenvalue λ_i, which we denote by g_i, is called the "geometric multiplicity" of λ_i.

Recall from Theorem 41 that if we think of $A - \lambda_i I_n$ as the matrix associated with a linear function f on $\mathbb{R}^n$, we have $dim(\mathbb{R}^n) = dim(im(f)) + dim(null(f))$ or

$$n = r(A - \lambda_i I_n) + g_i \iff r(A - \lambda_i I_n) = n - g_i.$$

In general, the geometric multiplicity of λ_i is different from the algebraic multiplicity λ_i. The latter refers to how many times the λ_i is repeated as a root for equation (9).

Theorem 58. *The eingenspace of A belonging to the eigenvalue λ is a vector space.*

Proof. Let x, y be eigenvectors associated with λ. Then, $Ax = \lambda x$ and $Ay = \lambda y$. Now, let $z = ax + by$ for $a, b \in \mathbb{R}$. Then,

$$Az = A(ax + by) = a\lambda x + b\lambda y = \lambda(ax + by) = \lambda z.$$

Hence, z is an eigenvector associated with λ. □

Theorem 59. *If λ is an eigenvalue of A, then: 1. λ^n is an eigenvalue for A^n, for $n \in \mathbb{N}$; 2. λ^{-1} is an eigenvalue of A^{-1}.*

Proof. 1. We prove this by using the principle of induction. Take $n = 2$: if $Ae = \lambda e$, then $A^2 e = \lambda A e = \lambda^2 e$. Now, assume that $A^n e = \lambda^n e$. Then, $A^{n+1} e = AA^n e = A\lambda^n e = \lambda^n A e = \lambda^{n+1} e$. 2. $Ae = \lambda e$ implies $A^{-1}Ae = \lambda A^{-1} e = e$ which implies $\frac{1}{\lambda} e = A^{-1} e$. □

Theorem 60. *Let $A \in \mathcal{M}^{n \times n}$ with typical element A_{ij} and $\lambda_1, \dots, \lambda_n$ be eingenvalues. Then, 1. $det(A) = \prod_{i=1}^n \lambda_i$; 2. $tr(A) = \sum_{i=1}^n A_{ii} = \sum_{i=1}^n \lambda_i$.*

Proof. 1. From equation (9) $\det(A - \lambda I)$ is,

$$p(\lambda) = (-1)^n \lambda^n + (-1)^{n-1} \lambda^{n-1} c_{n-1} + \cdots + (-1)\lambda c_1 + \det(A). \tag{11}$$

where $c_1, \dots, c_{n-1}$ are the coefficients in (9). If $\lambda_1, \cdots, \lambda_n$ are eigenvalues of A, i.e., solutions for $p(\lambda) = 0$, we can write

$$\begin{aligned} p(\lambda) &= (\lambda_1 - \lambda) \cdots (\lambda_n - \lambda) \\ &= (-1)^n \lambda^n + (-1)^{n-1} \lambda^{n-1} \sum_{i_1 \in C_1^n} \lambda_{i_1} + (-1)^{n-2} \lambda^{n-2} \sum_{(i_1, i_2) \in C_2^n} \lambda_{i_1} \lambda_{i_2} \\ &+ (-1)^1 \lambda^1 \sum_{(i_1, i_2, \dots, i_{n-1}) \in C_{n-1}^n} \lambda_{i_1} \lambda_{i_2} \cdots \lambda_{i_{n-1}} + \lambda_1 \cdots \lambda_n. \end{aligned} \tag{12}$$

Comparing the coefficients on the polynomials in equations (9), (11) and (12) completes the proof. $\square$

Theorem 61. *Let $A \in \mathcal{M}^{n \times n}$ and $\lambda_1, \dots, \lambda_k$ where $k \leq n$ be a set of distinct eigenvalues. Let (λ_j, x_j) for $j = 1, 2, \dots k$ be eigenpairs. Then, the set $\{x_j\}_{j=1}^k$ is linearly independent.*

Proof. The proof is by contradiction. Suppose $\{x_j\}_{j=1}^k$ is linearly dependent. Then, there exist $S \subset \{1, 2, \dots, k\}$ with $\#S := c < k$ such that $\{x_j\}_{j \in S}$ is linearly independent. Let the corresponding distinct eigenvalues be $\{\lambda_j\}_{j \in S}$. Then, there exist scalars t_j such that $x_k = \sum_{j \in S} t_j x_j$ where $k \notin S$. Now, multiply the last equation by $A - \lambda_k I$, so that

$$\theta = (A - \lambda_k I) x_k = \sum_{j \in S} t_j (A - \lambda_k I) x_j = \sum_{j \in S} t_j (A x_j - \lambda_k x_j) = \sum_{j \in S} t_j (\lambda_j - \lambda_k) x_j.$$

Since the eigenvalues are distinct $\lambda_j - \lambda_k \neq 0$. Consequently, the collection $\{x_j\}_{j \in S}$ is linearly dependent, a contradiction. $\square$

If all eigenvalues λ_j of A are distinct, i.e. $n = k$, then the set of eigenvectors $\{x_j\}_{j=1}^n$ is linearly independent. Let $X \in \mathcal{M}^{n \times n}$ be the matrix $X = \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix}$, which in this case is of full rank, $r(X) = n$. Letting

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix},$$

by definition we have

$$AX = X\Lambda \iff A = X\Lambda X^{-1}.$$

The equation $A = X\Lambda X^{-1}$ is called the diagonal factorization of A. Clearly, if all eigenvalues of A are distinct, A can be diagonally factorized. But as the next theorem shows, having all eigenvalues of A being distinct is not necessary for the diagonal factorization of A.

Theorem 62. *Let $A \in \mathcal{M}^{n\times n}$ with distinct eigenvalues $\lambda_1, \ldots, \lambda_k$ where $k < n$ and each having algebraic multiplicity $m_1, \ldots, m_k$. If X is the matrix of eigenvectors of A, then*

$$r(A - \lambda_i x) = n - m_i \iff X \text{ is invertible.}$$

Proof. ($\Leftarrow$) $\Lambda - \lambda_i I_n$ is a diagonal matrix with exactly m_i zeros on its diagonal. Hence, $r(\Lambda - \lambda_i I_n) = n - m_i$. If X is invertible,

$$X(\Lambda - \lambda_i I_n)X^{-1} = (X\Lambda - \lambda_i X)X^{-1} = (AX - \lambda_i X)X^{-1} = A - \lambda_i I_n.$$

Since multiplication of any matrix by another matrix of full rank does not alter its rank, $r(A - \lambda_i I_n) = n - m_i$.

($\Rightarrow$) If $r(A - \lambda_i I_n) = n - m_i$ the rank of the null space of $A - \lambda_i I$ is m_i. That is, there exists m_i linearly independent eigenvectors of A associated with λ_i. Denote these eigenvectors by $\{w_l\}_{l=1}^{m_i}$. These eigenvectors are columns of X. Now, let v_j be an eigenvector associated with $\lambda_j \neq \lambda_i$ and assume v_j is linearly dependent on $\{w_l\}_{l=1}^{m_i}$. Then, there exists a sequence of non-zero scalars $\{t_l\}_{l=1}^{m_i}$ such that $v_j = \sum_{l=1}^{m_i} t_l w_l$ and

$$Av_j = A\sum_{l=1}^{m_i} t_l w_l = \sum_{l=1}^{m_i} t_l A w_l = \lambda_i \sum_{l=1}^{m_i} t_l w_l = \lambda_i v_j$$

contradicting the assumption that v_j is not associated with λ_i. Thus, it must be that v_j is independent of $\{w_l\}_{l=1}^{m_i}$. □

Remark 18. *1. If λ_i and λ_j are distinct eigenvalues of a symmetric matrix A, and if x_i and x_j are eigenvectors associated with λ_i and λ_j, then $x_i^T x_j = 0$. This is promptly* *verified by noting that*

$$x_j^T A x_i = \lambda_i x_j^T x_i \text{ and } x_i^T A x_j = \lambda_j x_i^T x_j$$

and by symmetry $(x_j^T A x_i)^T = x_i^T A x_j$. *Thus,* $\lambda_i x_j^T x_i = \lambda_j x_i^T x_j$. *But since* $\lambda_i \neq \lambda_j$, *it must be that* $x_j^T x_i = 0$. *That is,* x_i *and* x_j *are "orthogonal" to each other.*

2. *It can be shown, see Hadley (1977) pp. 243-245, that if* A *is a symmetric matrix and* λ_i *is an eigenvalue with multiplicity* m_i, *there will be* m_i *linearly independent orthogonal eigenvectors associated with* λ_i.

3. *The previous two remarks show that if* A *is a symmetric matrix with eigenvalues* $\lambda_1, \ldots, \lambda_n$, *possibly not all distinct, there exists a sequence of associated linearly independent orthogonal eigenvectors* $\{x_1, \ldots, x_n\}$. *Thus, every symmetric matrix can be diagonally factorized.*

Theorem 63. *Let* $A \in \mathcal{M}^{n \times n}$ *be diagonizable with eigenpairs* $\{(\lambda_i, x_i)\}_{i=1}^n$. *1. The* $r(A)$ *is equal to the number of nonzero eigenvalues, 2. The eigenvalues of an idempotent matrix are either* 0 *or* 1. *In this case,* $r(A) = tr(\Lambda) = tr(A)$.

Proof. 1. Since $X^{-1}AX = \Lambda$ and X is nonsingular, $r(X^{-1}AX) = r(A)$. Hence, $r(\Lambda) = r(A)$. But since Λ is a diagonal matrix containing the eigenvalues of A, $r(A)$ is the number of nonzero elements in the diagonal of Λ.

2. $Ax = \lambda x$ and since A is idempotent $Ax = \lambda A x = \lambda^2 x$. Thus,

$$\lambda x = \lambda^2 x \implies \lambda(\lambda - 1)x = 0.$$

Since $x \neq \theta$, it must be that $\lambda(\lambda - 1) = 0$ which implies $\lambda = 0$ or 1. $\square$

13 Quadratic forms

In Example 13, for $x, y \in \mathbb{R}^n$ we defined the inner product $\langle x, y\rangle = \sum_{i=1}^n x_i y_i$ or

$$\langle x, y\rangle = x^T y.$$

Of course, $x^T y = x^T I_n y$ which can be viewed as a special case of

$$x^T Ay \text{ for } A \in \mathcal{M}^{n\times n} \text{ when } A = I_n. \tag{13}$$

The expression $x^T Ay$ is called a bilinear form and when $x = y$, i.e., when we have

$$x'Ax = \sum_{i=1}^n \sum_{j=1}^n A_{ij} x_i x_j,$$

we refer to it as a quadratic form. It is common to define quadratic forms for symmetric matrices A. In fact, this is with loss of generality, as made evident in the next theorem.

Theorem 64. *For every $A \in \mathcal{M}^{n\times n}$ there exists a matrix $A_s = \frac{1}{2}(A + A^T) \in \mathcal{M}^{n\times n}$ such that*

$$x^T A_s x = x^T Ax \text{ where } A_s \text{ is a symmetric matrix.}$$

Proof. For any $A \in \mathcal{M}^{n\times n}$ let $M = A + A^T$. Clearly M is symmetric ($M^T = (A + A^T)^T = A^T + A = M$). Now,

$$x^T Mx = x^T Ax + x^T A^T x = x^T Ax + (x^T A^T x)^T = x^T Ax + x^T Ax = 2\, x^T Ax.$$

Hence, $\frac{1}{2} x^T Mx = x^T Ax$ and we define $A_s = \frac{1}{2}(A + A^T)$. □

The following definition is useful in ordering matrices.

Definition 51. *Let $A \in \mathcal{M}^{n\times n}$ be symmetric.*

1. *A is said to be positive semi-definite if $x^T Ax \geq 0$ for all $x \in \mathbb{R}^n$,*

2. A is said to be positive definite if $x^T Ax > 0$ *for all* $x \in \mathbb{R}^n - \{\theta\}$*,*

3. A is said to be negative semi-definite if $x^T Ax \leq 0$ *for all* $x \in \mathbb{R}^n$*,*

4. A is said to be negative definite if $x^T Ax < 0$ *for all* $x \in \mathbb{R}^n - \{\theta\}$*.*

Theorem 65. *Let* $A \in \mathcal{M}^{n\times n}$ *be symmetric. A is positive definite if, and only if, all eigenvalues are positive.*

Proof. ($\Leftarrow$) Since A is symmetric, the matrix formed by eigenvectors X is invertible. Thus, $\{X_{.j}\}_{j=1}^n$ is a linearly independent collection of vectors. Hence, any vector $y \in \mathbb{R}^n - \{\theta\}$ can be written as $y = \sum_{j=1}^n X_{.j}s_j$ where $s_j \in \mathbb{R}$. Then,

$$Ay = A\sum_{j=1}^n X_{.j}s_j = \sum_{j=1}^n AX_{.j}s_j = \sum_{j=1}^n \lambda_j s_j X_{.j} \text{ and}$$

$$y^T Ay = \sum_{i=1}^n X_{.i}^T s_i \sum_{j=1}^n \lambda_j s_j X_{.j} = \sum_{i=1}^n \sum_{j=1}^n \lambda_j s_j s_i X_{.i}^T X_{.j}.$$

Since A is symmetric $X_{.i}^T X_{.j} = 0$ for $i \neq j$, hence

$$y^T Ay = \sum_{j=1}^n \lambda_j s_j^2 X_{.j}^T X_{.j} = \sum_{j=1}^n \lambda_j (s_j X_{.j})^T (s_j X_{.j}).$$

Since $(s_j X_{.j})^T (s_j X_{.j}) > 0$, if $\lambda_j > 0$ we have $y^T Ay > 0$.

($\Rightarrow$) $AX_{.i} = \lambda_i X_{.i} \implies X_{.i}^T AX_{.i} = \lambda_i X_{.i}^T X_{.i}$. Since $X_{.i}^T X_{.i} > 0$, we have that for $X_{.i}^T AX_{.i} > 0$ it must be that $\lambda_i > 0$. □

For any any $A \in \mathcal{M}^{n\times n}$, the upper sub-matrices of A are given by

$$A_1 := A_{11}, A_2 := \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, A_3 := \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix}, \cdots, A_n := A.$$

Theorem 66. *Let* $A \in \mathcal{M}^{n\times n}$ *be symmetric. A is positive definite if, and only if,* $det(A_j) > 0$ *for* $j = 1, \ldots, n$*, where* A_j *are upper sub-matrices of A.*

Proof. Exercise. □

14 Banach spaces

We start by defining Cauchy sequences.

Definition 52. *A sequence $x_1, x_2, \cdots$ in a normed vector space $(\mathbb{X}, \|\cdot\|_{\mathbb{X}})$ is said to be a Cauchy sequence if $\|x_n - x_m\|_{\mathbb{X}} \to 0$ as $n, m \to \infty$.*

Recall that this means that for all $\epsilon > 0$, there exists N_ϵ such that for all $n, m \geq N_\epsilon$ we have $\|x_n - x_m\|_{\mathbb{X}} < \epsilon$. Note that for any $x \in \mathbb{X}$ we have $\|x_n - x_m\|_{\mathbb{X}} = \|x_n - x + x - x_m\|_{\mathbb{X}} \leq \|x_n - x\|_{\mathbb{X}} + \|x - x_m\|_{\mathbb{X}}$. Hence, if $x_1, x_2, \cdots$ converges to x, then $\|x_n - x_m\|_{\mathbb{X}} \to 0$. Thus, every convergent sequence is a Cauchy sequence. However, Cauchy sequences may not converge.

Definition 53. *A normed vector space $(\mathbb{X}, \|\cdot\|_{\mathbb{X}})$ is complete if every Cauchy sequence in $\mathbb{X}$ has a limit in $\mathbb{X}$. That is, every Cauchy sequence in the space converges to an element in the space. A complete, normed vector space is called a Banach space.*

Lemma 2. *A Cauchy sequence is bounded.*

Proof. Let N be such that for all $n \geq N$ we have $\|x_n - x_N\| < 1$. Hence, for $n \geq N$ we have $\|x_n\|_{\mathbb{X}} = \|x_n - x_N + x_N\|_{\mathbb{X}} \leq \|x_n - x_N\|_{\mathbb{X}} + \|x_N\|_{\mathbb{X}} < 1 + \|x_N\|_{\mathbb{X}}$. □

Theorem 67. *Let $\{x_1, x_2, \cdots\}$ be a sequence in $(\mathbb{X}, \|\cdot\|_{\mathbb{X}})$. If x_L is a limit point of $\{x_1, x_2, \cdots\}$, then there exists a subsequence $\{x_{n(1)}, x_{n(2)}, \cdots\}$ of $\{x_1, x_2, \cdots\}$ such that $x_{n(k)} \to x_L$ as $n(k) \to \infty$.*

Proof. Since x_L is a limit point, for all $\epsilon > 0$ and for all N, there exists $n > N$ such that $\|x_n - x_L\| < \epsilon$. Consider $B(x_L, 1/k) = \{x : \|x - x_L\| < 1/k\}$, where $k \in \mathbb{N}$. Since x_L is a limit point, for all k there exists $n(k)$ such that $x_{n(k)} \in B(x_L, 1/k)$. From Remark 6 there are infinitely many points in every ball centered at x_L, hence it is always possible to choose $n(k) < n(k+1)$ for all $k = 1, 2, \ldots$, so that $x_{n(k)}$ is indeed a subsequence. As $k \to \infty$, $1/k \to 0$, which implies that $x_{n(k)} \to x_L$. □

Example 23. *Let* $x_n = \begin{cases} -1 & \text{if } n = 1,3,5,\dots \\ n^{-1} & \text{if } n = 2,4,6,\dots \end{cases}$. *This sequence does not converge in* $\mathbb{R}$ *as there is no* $x \in \mathbb{R}$ *such that all (except finitely many)* $x_n \in B(x,\epsilon)$ *for any* $\epsilon > 0$. *Note that* $x_L = 0$ *is a limit point of the sequence since every* $B(0,\epsilon)$ *contains an element of the sequence different from* 0. *Furthermore, the subsequence* $x_{2n} = n^{-1} \to 0$.

The following remark is for a sequence taking values in $\mathbb{R}$.

Remark 19. *We say that* $\{x_n\}_{n\in\mathbb{N}}$ *in* $\mathbb{R}$ *is bounded if there exists* $a, r \in \mathbb{R}$ *with* $r > 0$ *such that for all* n, $x_n \in B(a,r) = \{y : |y-a| \leq r\}$. *That is* $a - r \leq x_n \leq a + r$. *Note that if* $a \geq 0$ *then* $|a| = a$ *and* $a - r \geq -r \geq -a - r$. *Thus,* $|a| - r \geq -r \geq -|a| - r$ *and*

$$-|a| - r \leq |a| - r \leq x_n \leq |a| + r.$$

Hence, $x_n \in B(0, |a|+r)$. *Alternatively, if* $a \leq 0$ *then* $|a| = -a$ *we have*

$$-|a| - r \leq x_n \leq -|a| + r \leq |a| + r$$

and consequently $x_n \in B(0, |a| + r)$.

Theorem 68. *Let* $(\mathbb{X}, \|\cdot\|)$ *be a normed vector space with metric* $d_{\mathbb{X}}(x,y) = \|x-y\|$ *and* $S \subset \mathbb{X}$ *have infinite many elements.* x_L *is a limit point of* S *if, and only if, there exists a sequence of distinct points* $\{x_n\}_{n\in\mathbb{N}} \subset S$ *such that* $\lim_{n\to\infty} x_n = x_L$.

Proof. ($\Longrightarrow$) x_L a limit point of S means that for all $\epsilon > 0$ there exists $x \in S$, distinct from x_L such that $d_{\mathbb{X}}(x, x_L) < \epsilon$. Then, choosing $\epsilon = 1 = \frac{1}{2^0}$, there exists $x_1 \in S$ distinct from x_L such that $d_{\mathbb{X}}(x_1, x_L) < \frac{1}{2^0}$. Next, choose $\epsilon = d_{\mathbb{X}}(x_1, x_L)/2$, then there exists $x_2 \in S$, distinct from x_L, such that $d_{\mathbb{X}}(x_2, x_L) < d_{\mathbb{X}}(x_1, x_L)/2 < 1/2^{2-1}$. Next, choose $\epsilon = d_{\mathbb{X}}(x_2, x_L)/2$, then there exists $x_3 \in S$, distinct from x_L, such that $d_{\mathbb{X}}(x_3, x_L) < d_{\mathbb{X}}(x_2, x_L)/2 < 1/2^{3-1}$. Repeating the argument $n \in \mathbb{N}$ times, gives $d_{\mathbb{X}}(x_n, x_L) < 1/2^{n-1}$. Hence, there exists a sequence of distinct points in S that converges to x_L.

($\Longleftarrow$) $x_n \to x$ as $n \to \infty$ means that for all $\epsilon > 0$ there exists $N_\epsilon \in \mathbb{N}$ such that for all $n \geq N_\epsilon$ we have $d_{\mathbb{X}}(x_n, x_L) < \epsilon$ or $B(x_L, \epsilon)$ contains infinitely many elements of $\{x_1, x_2, \cdots\} \subset S$. Hence, $B(x_L, \epsilon) \bigcap S \neq \emptyset$ and x_L is a limit point. $\square$

The next theorem is called the Bolzano-Weierstrass Theorem.

Theorem 69. *Every bounded infinite set $S \subset \mathbb{R}$ has a limit point.*

Proof. Since S is bounded, there exists $a, b \in \mathbb{R}$ such that $S \subset [a, b]$. Set $c = (a + b)/2$ and consider $[a, c]$ and $[c, b]$. It must be that either $[a, c]$ or $[c, b]$ (or both) contain infinitely many elements of S. Without loss of generality suppose $[a, c]$ contains infinitely many elements of S and label $[a, c] = [a_1, b_1]$. Let $c_1 = (a_1 + b_1)/2$ and define $[a_1, c_1]$ or $[c_1, b_1]$. By the same reasoning, either $[a_1, c_1]$ or $[c_1, b_1]$ (or both) contain infinitely many elements of S. Say it is $[c_1, b_1]$ and label $[c_1, b_1] = [a_2, b_2]$. Repeating this reasoning $n \in \mathbb{N}$ time gives

1. $[a, b] \supset [a_1, b_1] \supset [a_2, b_2] \supset \cdots \supset [a_n, b_n] \cdots$
2. $a \leq a_1 \leq a_2 \leq \cdots$ and $b \geq b_1 \geq b_2 \geq \cdots$
3. $b_1 - a_1 = (b - a)/2$, $b_2 - a_2 = (b - a)/4$, $\cdots$ $b_n - a_n = (b - a)/2^n$.

Since a_n is monotonically increasing and bounded, $a_n \to \bar{a} = \sup\limits_{n \in \mathbb{N}} a_n$. Also, since b_n is monotonically decreasing and bounded, $b_n \to \underline{b} = \inf\limits_{n \in \mathbb{N}} b_n$ and $a_n - b_n \to 0$ as $n \to \infty$. Then,

$$|b_n - \bar{a}| = |b_n - a_n + a_n - \bar{a}| \leq |b_n - a_n| + |a_n - \bar{a}|.$$

By item 3 and the fact that $a_n \to \bar{a}$ we have that $b_n \to \bar{a}$. Since limits are unique $\bar{a} = \underline{b}$.

Now, let $\bar{a} = \underline{b} = x_L$, we will show that x_L is a limit point of S. For any $\epsilon > 0$ consider $B(x_L, \epsilon) = (x_L - \epsilon, x_L + \epsilon)$. For n sufficiently large $[a_n, b_n] \subset (x_L - \epsilon, x_L + \epsilon)$ and consequently any ball centered at x_L contains infinitely many points of S. Hence, $B(x_L, \epsilon) \bigcap S \neq \emptyset$, establishing that x_L is a limit point of S. $\square$

Remark 20. *The reasoning used in the proof of the Bolzano-Weierstrass Theorem extends without difficulty to the case where $S \subset \mathbb{R}^n$. Hence, we have the following theorem, which is stated without proof (see Apostol (1974), page 54).*

Theorem 70. *Every bounded infinite set $S \subset \mathbb{R}^n$ has a limit point.*

Hence, if x_L is such a limit point of S, using Theorem 68 we can conclude that there exists a sequence $\{x_1, x_2, \cdots\} \in S$ such that $\lim_{n\to\infty} x_n = x_L$. Now, let S be itself a sequence and write $S = \{x_1, x_2, \cdots\}$

Theorems 18 and 19 imply that every bounded sequence $\{x_n\}_{n\in\mathbb{N}} \in \mathbb{R}$ has a converging subsequence. Of course, the point to which this subsequence converges is a limit point of $\{x_n\}_{n\in\mathbb{N}}$. Put differently, every bounded sequence $\{x_n\}_{n\in\mathbb{N}} \in \mathbb{R}$ has a limit point in $\mathbb{R}$. This is a version the classic Bolzano-Weierstrass Theorem. It can be easily extended to $\mathbb{R}^m$ with the Euclidean norm. We state it next.

Below are examples of Banach spaces.

1. $(\mathbb{R}, |\cdot|)$ is a Banach space. To verify we need to show that every Cauchy sequence $\{x_n\}_{n\in\mathbb{N}} \in \mathbb{R}$ converges to some $x \in \mathbb{R}$. Since $\{x_n\}_{n\in\mathbb{N}}$ is Cauchy, it is bounded. Since it is bounded it has a limit point $x \in \mathbb{R}$. By the triangle inequality $|x_n - x| \leq |x_n - x_m| + |x_m - x|$. Note that by the Cauchy property, there exists N_ϵ such that for $n, m > N_\epsilon$, $|x_n - x_m| < \epsilon$ and by definition of a limit point $|x_m - x| < \epsilon$. Hence, $|x_n - x| < 2\epsilon$.

2. It follows from 1. and the Bolzano-Weierstrass Theorem that $\mathbb{R}^m$ for $m \in \mathbb{N}$ with the Euclidean norm is a Banach space.

3. $C[0,1]$, the space of continuous functions on $[0,1]$ with $\|x(t)\|_{\mathbb{X}} = \max_{0\leq t\leq 1}|x(t)|$.

4. l_p for $1 \leq p < \infty$, the space of scalar sequences $\{x_1, x_2, \cdots\}$ for which $\sum_{i=1}^{\infty} |x_i|^p < \infty$ whenever $1 \leq p < \infty$ with norm $\|x\|_p = \left(\sum_{i=1}^{\infty} |x_i|^p\right)^{1/p}$. When $p = \infty$, $\|x\|_\infty = \sup_i |x_i|$.

5. If $\mathbb{X}$ and $\mathbb{Y}$ are Banach spaces and $\mathbb{W} = \mathbb{X} \times \mathbb{Y}$ be the cartesian product of the spaces, with norm $\|w\|_{\mathbb{W}} = \|x\|_{\mathbb{X}} + \|y\|_{\mathbb{Y}}$. $\mathbb{W}$ is a Banach space.

The Bolzano-Weierstrass Theorem can then be used to prove Cantor's Intersection Theorem, an important result that is helpful characterizing compact sets. These sets play an important role in optimization.

Theorem 71. *(Cantor's Intersection Theorem) Let $\{Q_1, Q_2, \cdots\}$ be a countable collection of nonempty sets in $\mathbb{R}^n$ such that: a) $Q_{k+1} \subseteq Q_k$ and b) each set Q_k is closed and Q_1 is bounded. Then, $\bigcap_{k=1}^{\infty} Q_k$ is closed and nonempty.*

Proof. Let $I = \bigcap_{k=1}^{\infty} Q_k$. Since each Q_k is closed, so is I. We now need to show that there exists a point $x \in I$. If one of the sets Q_k is finite, then the proof is trivial given that the sequence $\{Q_1, Q_2, \cdots\}$ is composed of nonempty sets. Thus, let's assume each Q_k contains infinitely many points. Let $A = \{x_1, x_2, \cdots\}$ be a collection of distinct points such that $x_k \in Q_k$. Since $A \subset Q_1$ and Q_1 is bounded we have that A is bounded. But every bounded sequence has a convergent subsequence that is bounded, which has a limit. Thus, A has a limit point, call it x. We will show that $x \in I$ by showing that it is in every Q_k. Since the Q_k are closed they contain their limit points, so it suffices to show that x is a limit point of every Q_k. But any ball centered at x contains infinitely many points of A since all except (possibly) a finite number of elements of A belong to Q_k and this ball also contains infinitely many points of Q_k. So, x is a limit point of every Q_k. $\square$

Theorem 72. *Let $(\mathbb{X}, d)$ be a metric space and $A \subset \mathbb{X}$. x_L is a limit point of A if, and only if, there exists a sequence in $A - \{x_L\}$ that converges to x_L.*

Proof. ($\Leftarrow$) Let $\{a_n\}_{n=1,2,\ldots} \in A - \{x_L\}$ such that $a_n \to x_L$, then for all $\epsilon > 0$ there exists N_ϵ such that $d(a_n, x_L) < \epsilon$ for all $n > N_\epsilon$. Since $B(x_L, \epsilon) = \{x : d(x, x_L) < \epsilon\}$, $B(x_L, \epsilon) \bigcap (A - \{x_L\}) \neq \emptyset$ for the given ϵ, since some a_n are arbitrarily close to x_L. Therefore, by definition x_L is a limit point of A.

($\Rightarrow$) x_L is a limit point of A implies that for all $r > 0$, $B(x_L, r) \bigcap (A - \{x_L\}) \neq \emptyset$. We will show that we can construct a sequence in $A - \{x_L\}$ that converges to x_L. Put $r_1 = 1$, then there exists $a_1 \in B(x_L, r_1 = 1) \bigcap (A - \{x_L\})$. Next, $r_2 = \frac{1}{2}d(a_1, x_L) \leq \frac{1}{2} = \frac{1}{2^{2-1}}$ there exists $a_2 \in S(x_L, r_2) \bigcap A - \{x_L\}$. Similarly, $r_3 = \frac{1}{2}d(a_2, x_L) \leq \frac{1}{4} = \frac{1}{2^{3-1}}$ there exists $a_3 \in B(x_L, r_3) \bigcap A - \{x_L\}$, and so on. Hence, there exists $\{a_1, a_2, \cdots\}$ such that $d(a_n, x_L) \leq \frac{1}{2^{n-1}} \to 0$ as $n \to \infty$. Therefore, $a_n \to x_L$. $\square$

Theorem 73. *$A \subset (\mathbb{X}, d)$ is closed if, and only if, every convergent sequence in A has its limit in A. That is, if $\{x_1, x_2, \cdots\}$ is a sequence in a closed set A and $x_n \to x$, then $x \in A$.*

Proof. ($\Leftarrow$) If every convergent sequence in A has a limit in A, then this is true for all convergent sequences $\{x_1, x_2, \cdots\}$ for which $x_n \neq x$ for all n. Therefore, A contains all of its limit points, but since a limit point is a closure point, A contains all of its closure points, and by definition it is a closed set.

($\Rightarrow$) Assume that A is closed and let $\{x_1, x_2, \cdots\}$ be a sequence in A such that $x_n \to x$. If $x_n \neq x$ for all n, then x is a limit point of A, and since A is closed, $x \in A$. If $x_n = x$ for some n, since $x_n \in A$, then $x \in A$. $\square$

Theorem 74. *Let $(\mathbb{X}, \|\cdot\|_{\mathbb{X}})$ be a Banach space. $A \subset \mathbb{X}$ is complete if, and only if, A is closed.*

Proof. ($\Rightarrow$) A complete implies that every Cauchy sequence converges to an element in A. Then this is true for all sequences $\{x_1, x_2, \cdots\}$ for which $x_n \neq x$ for all n. Therefore, A contains all of its limit points, but since a limit point is a closure point, A contains all of its closure points, and by definition it is a closed set.

($\Leftarrow$) Let $\{x_1, x_2, \cdots\}$ be a sequence in A such that $x_n \to x$. If $x_n \neq x$ for all n, then x is a limit point of A, and since A is closed, $x \in A$. If $x_n = x$ for some n, since $x_n \in A$, then $x \in A$. $\square$

Theorem 75. *Let $f : (\mathbb{X}, d_{\mathbb{X}}) \to (\mathbb{Y}, d_{\mathbb{Y}})$ be a function. f is continuous if, and only if, for all subsets C closed in $\mathbb{Y}$, $f^{-1}(C)$ is closed in $\mathbb{X}$.*

Proof. ($\Rightarrow$) We will verify that $f^{-1}(C)$ contains all of its limit points when f is continuous and C closed. Let x_L be a limit point of $f^{-1}(C)$. Then, by Theorem 72 there exists a sequence $\{x_1, x_2, \cdots\}$ in $f^{-1}(C) - \{x_L\}$ that converges to x_L. Note that, since f is continuous $f(x_n) \to f(x_L)$. By construction, $f(x_n) \in C$ for all n, and by assumption C is closed. Therefore, by Theorem 73, $f(x_n)$ has its limit $f(x_L) \in C$. But this means that $x_L \in f^{-1}(C)$, therefore $f^{-1}(C)$ is closed.

($\Leftarrow$) We will prove that f discontinuous at some x_0 implies that for some C closed in $\mathbb{Y}$, $f^{-1}(C)$ is not closed in $\mathbb{X}$. f discontinuous implies that there exists x_0 such that $x_n \to x_0$ with $f(x_n) \nrightarrow f(x_0)$. That is, there exists $r > 0$ such that $d_{\mathbb{Y}}(f(x_n), f(x_0)) \geq r$ for all n. Now, let C be the closure of $\{f(x_n)\}$. C is, consequently, closed. Since, $f(x_n) \in C$ we have that $x_n \in f^{-1}(C)$ for all n, and since $x_n \to x_0$, x_0 is a limit point of $f^{-1}(C)$. However, $f(x_0) \notin C$ ($f(x_0)$ is not a closure point of $\{f(x_n)\}$), for $B(f(x_0), \epsilon) \bigcap \{f(x_n)\} = \emptyset$ for all n and $\epsilon < r$. Hence, $x_0 \notin f^{-1}(C)$ and $f^{-1}(C)$ not closed. $\square$

Corollary 4. *Let $f : (\mathbb{X}, d_{\mathbb{X}}) \to (\mathbb{Y}, d_{\mathbb{Y}})$ be a function. f is continuous if, and only if, for all subsets C open in $\mathbb{Y}$, $f^{-1}(C)$ is open in $\mathbb{X}$.*

The last theorem and corollary characterize continuous functions without explicitly describing the involved norms and metrics. In this sense, we can provided an axiomatic approach to continuity. Let $(\mathbb{X}, \tau_{\mathbb{X}})$ and $(\mathbb{Y}, \tau_{\mathbb{Y}})$ be two topological spaces. A function $f : (\mathbb{X}, \tau_{\mathbb{X}}) \to (\mathbb{Y}, \tau_{\mathbb{Y}})$ is continuous if for every $Y \in \tau_{\mathbb{Y}}$, $f^{-1}(Y) \in \tau_{\mathbb{X}}$.

15 Compactness and extreme values

Definition 54. *Let I be an index set and $U = \{U_i : i \in I\}$ be in $(\mathbb{X}, d_{\mathbb{X}})$. U is a cover for $A \subset \mathbb{X}$ if $A \subseteq \bigcup_{i \in I} U_i$. If U_i's are open for all i, U is called an open cover.*

Definition 55. *$A \subset \mathbb{X}$ is compact if for all open covers U of A, we can find a finite subcollection of U, say $\{U_i\}_{i=1}^n$ such that $A \subseteq \bigcup_{i=1}^n U_i$.*

Example 24. *$(0,1)$ is not compact. $U = \{(1/n, 1) : n \geq 2\}$ is an open cover for $(0,1)$, but no finite subcollection of U will cover $(0,1)$.*

It is important to note that compactness does not result from being able to cover a set with a finite open collection of sets. Clearly, $(0,1)$ can be covered by $(0, 0.51) \bigcup (0.5, 1)$. What is critical in the definition is that associated with every open cover there is a finite subcollection that is *also* a cover for the set under study.

15.1 Characterizing compactness in $\mathbb{R}^n$

We will give useful characterizations for compact sets in $\mathbb{R}^n$. We will give first some necessary and important supporting results.

Theorem 76. *Let $G = \{A_1, A_2, \cdots\}$ denote the countable collection of all balls in $\mathbb{R}^n$ with rational radius and centers with rational coordinates. Let $x \in \mathbb{R}^n$ and S be open in $\mathbb{R}^n$ with $x \in S$. Then, at least one of the balls in G contains x and is contained in S. That is, $x \in A_k \subseteq S$ for some A_k in G.*

Proof. Recall that the set of rational numbers is countable, hence G is countable and there exists $B(x,r) \subseteq S$. We will find $y \in S$ with rational coordinates and find $B(y,q) \in G$ such that $B(y,q) \subseteq B(x,r)$ and $x \in B(y,q)$. Let $\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ and let $y_k \in \mathbb{Q}$ such that $|y_k - x_k| < r/4n$ for all $k = 1, \cdots, n$. Then, $\|y - x\| \leq |y_1 - x_1| + \cdots + |y_n - x_n| < \frac{r}{4n} n = \frac{r}{4}$ (recall that $\|x\|^2 = \sum_{i=1}^n |x_i|^2 = (\sum_{i=1}^n |x_i|)^2 - 2\sum_{i<j} |x_i||x_j| \leq (\sum_{i=1}^n |x_i|)^2$, and $\|x\| \leq \sum_{i=1}^n |x_i|$). Let $q \in \mathbb{Q}$ such that $\frac{r}{4} < q < \frac{r}{2}$. Then, $x \in B(y,q)$ and $B(y,q) \subseteq B(x,r) \subset S$. But $B(y,q) \in G$ which proves the theorem. □

Theorem 77. *(Lindelöf Covering Theorem) Assume $A \subseteq \mathbb{R}^n$ and let $\mathcal{O}$ be an open cover of A. Then, there exists a countable sub-collection of $\mathcal{O}$ which also covers A.*

Proof. Let $x \in A$, then there exists an open set S in $\mathcal{O}$ such that $x \in S$. By Theorem 76 there exists a ball A_k such that $x \in A_k \subseteq S$. Of course, there may be infinitely many such A_k, but we choose only one of these, say the one with smallest index $m(x)$. Then, the set of balls $A_{m(x)}$ obtained as x varies over all elements of A is a countable collection of open sets which covers A. To get a countable sub-collection of $\mathcal{O}$, correlate with each $A_{m(x)}$ one of the sets $\mathcal{O}$ in U that contains $A_{m(x)}$. □

The next theorem, known as the Heine-Borel Theorem, establishes that closed and bounded sets in $\mathbb{R}^n$ are compact. The proof is an example of the usefulness of Cantor's Intersection Theorem.

Theorem 78. *Let $\mathcal{O}$ be an open cover of a closed an bounded set $A \subset \mathbb{R}^n$. Then, a finite sub-collection of $\mathcal{O}$ covers A.*

Proof. By Theorem 77 there exists a countable sub-collection of $\mathcal{O}$ which covers A. Describe this collection as $\mathcal{O}_s = \{I_1, I_2, \cdots\}$ and consider $S_m = \bigcup_{k=1}^{m} I_k$ for m finite. S_m is open, since the finite union of open sets is open. We will show that for some m, $A \subseteq S_m$. Let $Q_1 = A$, $Q_2 = A \bigcap (\mathbb{R}^n - S_1)$, $Q_3 = A \bigcap (\mathbb{R}^n - S_2), \cdots$. If we can show that for some m, $A \bigcap (\mathbb{R}^n - S_m) = \emptyset$, then for that m, $A \subseteq S_m$. By construction, $S_1 \subseteq S_2 \subseteq \cdots$ and $Q_1 \supseteq Q_2 \supseteq \cdots$. By assumption, A is closed, and since $\mathbb{R}^n - S_m$ is closed (recall that the complement of an open set is closed), $Q_m = A \bigcap (\mathbb{R}^n - S_m)$ is closed for all m, since the finite intersection of closed sets is closed. Since by construction $Q_m \subset A$, and since A is bounded by assumption, Q_m is bounded for all m. Suppose no set Q_m is empty, then by Cantor's Intersection Theorem, $\bigcap_{k=1}^{\infty} Q_k \neq \emptyset$. This means that there exists $x \in A$ that is an all Q_m, and therefore, outside of all S_m, viz., $x \notin \bigcup_{k=1}^{\infty} S_k$. But this is impossible, since $A \subseteq \bigcup_{k=1}^{\infty} S_k$. Therefore, some Q_m must be empty, which completes the proof. □

The following theorem provides a full characterization of compact sets in Euclidean spaces. In particular, in these vector spaces compactness is equivalent to closeness and boundedness of sets.

Theorem 79. *Let $A \subseteq (\mathbb{R}^n, d_E)$ where d_E is the metric associated with the Euclidean norm. The following three statements are equivalent: 1. A is compact; 2. A is closed and bounded; 3. Every infinite subset of A has a limit point in A.*

Proof. In Theorem 78 we showed that 2 implies 1. We now show that 1 implies 2. If $a \in A$, then the collection of open balls $\mathcal{B} = \{B(a,k) : k = 1, 2, \cdots\}$ is an open covering for A. Since A is compact, there exists a finite sub-collection that covers A. Therefore, A is bounded.

Suppose A is not closed. Then, by Theorem 73, there exists a convergent sequence in A whose limit is not in A, i.e., there exists a limit point x_L of A that is not in A. For all $x \in A$, let $r(x) = \frac{1}{2}\|x - x_L\|$. Note that $r(x) > 0$ since $x \neq x_L$ ($x_L \notin A$) and consider the collection $\mathcal{B} = \{B(x, r(x)) : x \in A\}$. $\mathcal{B}$ is an open cover for A, and since A is compact, there exists a finite subcollection of $\mathcal{B}$ that covers A. Thus, $A \subseteq \bigcup_{k=1}^{m} B(x_k; r_k)$. Let $r = min\{r_1, r_2, \cdots, r_m\}$. Note that if $x \in B(x_L; r)$ then $\|x - x_L\| < r \leq r_k$ and consequently by the triangle inequality $\|x_L - x_k\| \leq \|x_L - x\| + \|x - x_k\|$, or $\|x - x_k\| \geq \|x_L - x_k\| - \|x_L - x\| = 2r_k - \|x_L - x\| > r_k$ since $\|x - x_L\| < r_k$. Hence, $x \notin B(x_k, r_k)$ for all k and $B(x_L, r) \bigcap \bigcup_{k=1}^{m} B(x_k, r_k) = \emptyset$, which contradicts x_L being a limit point. Hence, A must be closed.

Since 1 and 2 are equivalent, the proof will be complete establishing that 2 and 3 are equivalent. This is left as an exercise. □

Example 25. *Intervals such as $[a, b]$, with $a < b$ and $a, b \in \mathbb{R}$, are closed and bounded and consequently compact when we endow $\mathbb{R}$ with the Euclidean norm; 2. It is not hard to show that $\times_{i=1}^{n}[a_i, b_i] \subset \mathbb{R}^n$ is also closed and bounded and compact when $\mathbb{R}^n$ is endowed with the norm $\|x\| = \left(\sum_{i=1}^{n} x_i^2\right)^{1/2}$.*

Theorem 80. *Let $f : (\mathbb{X}, d_{\mathbb{X}}) \to (\mathbb{Y}, d_{\mathbb{Y}})$ be a continuous function on $A \subset \mathbb{X}$, where A is compact. Then, $f(A)$ is compact in $\mathbb{Y}$. In particular, $f(A)$ is closed and bounded.*

Proof. Let $\mathcal{O}$ be an open covering for $f(A)$. We will show that a finite sub-collection of $\mathcal{O}$ covers $f(A)$. Let $f(A) \subseteq \bigcup_{C \in \mathcal{O}} C$ and note that $f^{-1}(C)$ is open on $(\mathbb{X}, d_{\mathbb{X}})$ by continuity of f and Corollary 4. Now, $\{f^{-1}(C) : C \in \mathcal{O}\}$ is an open cover for A, and since A is compact, there exists a a finite sub-collection $\{C_i\}_{i=1}^{p}$, $p \in \mathbb{N}$ that covers A. Therefore,

$$A \subseteq f^{-1}(C_1) \bigcup f^{-1}(C_2) \bigcup \cdots \bigcup f^{-1}(C_p).$$

Hence,

$$\begin{aligned} f(A) &\subseteq f\left(f^{-1}(C_1) \bigcup f^{-1}(C_2) \bigcup \cdots \bigcup f^{-1}(C_p)\right) \\ &= f(f^{-1}(C_1)) \bigcup f(f^{-1}(C_2)) \bigcup \cdots \bigcup f(f^{-1}(C_p)) \\ &\subseteq C_1 \bigcup C_2 \bigcup \cdots \bigcup C_p. \text{ by Homework 1} \end{aligned}$$

Hence, $f(A)$ is compact. $f(A)$ compact implies $f(A)$ is closed and bounded (from Theorem 79). We note here that the structure of $\mathbb{R}^n$ was not used in Theorem 79 to establish that 1 implies 2. $\square$

On Definition 46 we defined what is meant by boundedness of a **linear** function f. We now define what is meant by boundedness of a function f.

Definition 56. *The function $f : (\mathbb{X}, \|\cdot\|_{\mathbb{X}}) \to (\mathbb{Y}, \|\cdot\|_{\mathbb{Y}})$ is bounded on $A \subseteq \mathbb{X}$ if there exists $C > 0$ such that $\|f(x)\|_{\mathbb{Y}} < C$ for all $x \in A$.*

Theorem 81. *Let $f : (\mathbb{X}, \|\cdot\|_{\mathbb{X}}) \to (\mathbb{R}^n, \|\cdot\|_E)$. If f is continuous on $A \subseteq \mathbb{X}$ and A is compact, then f is bounded on A.*

Proof. Since A is compact, $f(A)$ is compact by the previous theorem. Therefore, $f(A)$ is closed and bounded, and consequently f is bounded on A. $\square$

Remark 21. *In the case where* $f : (\mathbb{X}, \|\cdot\|_{\mathbb{X}}) \to (\mathbb{R}, |\cdot|)$*,* $f(A)$ *is a closed and bounded subset of* $\mathbb{R}$*, therefore* $f(A)$ *has an infimum and a supremum. Hence,*

$$\inf f(A) := f(x_i) \leq f(x) \leq \sup f(A) := f(x_s) \text{ for all } x \text{ and some } x_i, x_s \in A.$$

But $f(x_i)$ *is a closure point for* $f(A)$*, hence* $f(x_i) \in f(A)$*. Similarly,* $f(x_s) \in f(A)$*. Hence, there exist minimum and maximum for* f *over* A*.*

This remark leads to the following corollary to the previous theorem. It is known as Weierstrass Theorem and provides for the existence of maxima or minima.

Corollary 5. *(Weierstrass Theorem) Let* A *be a compact set in a metric space* $(\mathbb{X}, d_{\mathbb{X}})$ *and* $f : A \to \mathbb{R}$ *be continuous. Then,* f *is bounded in* A *and attains both its minimum and maximum in the set. That is, there exists* $x_{min}, x_{max} \in A$ *such that* $f(x_{max}) = \sup\ f(A)$ *and* $f(x_{min}) = \inf\ f(A)$*.*

Theorem 82. *Let* $f : (\mathbb{X}, \|\cdot\|_{\mathbb{X}}) \to (\mathbb{Y}, \|\cdot\|_{\mathbb{Y}})$*. Suppose* f *is one-to-one and onto, so that the inverse function* f^{-1} *exists. If* $\mathbb{X}$ *is compact and* f *is continuous, then* f^{-1} *is continous.*

Proof. f^{-1} is continuous if, and only if, for every closed set $C \subset \mathbb{X}$ has image $f(C)$ closed in $\mathbb{Y}$. Since C is closed and $\mathbb{X}$ is compact we have that C is compact.

To see this, let $\mathcal{O}$ be any open cover of C. Since C is closed, its complement $C^c = \mathbb{X} - C$ is open. Thus, $\mathcal{O} \bigcup (\mathbb{X} - C)$ is an open covering for $\mathbb{X}$. But since $\mathbb{X}$ is compact, it contains a finite subcover which we can assume contains $\mathbb{X} - C$. That is, there exist $A_1, A_2, \cdots A_p$ and

$$\mathbb{X} \subseteq A_1 \bigcup \cdots A_p \bigcup \mathbb{X} - C.$$

This collection also covers C, and since $\mathbb{X} - C$ contains no element of C, $U_{i=1}^{p} A_i$ covers C.

Now, since C is compact $f(C)$ is compact and, consequently, closed. $\square$

The next theorem shows that continuity of a function on a compact set implies uniform continuity of the function on the set. That is, for any $\epsilon > 0$ there exists a $\delta_\epsilon > 0$ depending only on ϵ - and not on x that assures ϵ-proximity in the range of the function whenever there is δ-proximity in the domain.

Example 26. *1. Let $f(x) = x^{-1}$ for $x \in (0,1]$ and let $x_0 \in (0,1]$. For $x \neq x_0$ note that*

$$|f(x) - f(x_0)| = \left|x^{-1} - x_0^{-1}\right| = \frac{1}{xx_0}\,|x - x_0|\,.$$

Hence, for any $\epsilon > 0$, $|f(x) - f(x_0)| < \epsilon$ if $\frac{1}{xx_0}\,|x - x_0| < \epsilon$ or $|x - x_0| < xx_0\epsilon := \delta(x_0, \epsilon)$, where δ depends on x_0 and ϵ.

2. Let $f(x) = x^2$ for $x \in (0,1]$ and let $x_0 \in (0,1]$. For $x \neq x_0$ note that

$$|f(x) - f(x_0)| = \left|x^2 - x_0^2\right| = |x - x_0|\,(x + x_0) \leq 2\,|x - x_0| \;\; \textit{since } x, x_0 \leq 1.$$

Hence, for any $\epsilon > 0$, $|f(x) - f(x_0)| < \epsilon$ if $2\,|x - x_0| < \epsilon$ or $|x - x_0| < \epsilon/2 := \delta(\epsilon)$

Theorem 83. *(Heine's Theorem) Let $f : (\mathbb{X}, d_{\mathbb{X}}) \to (\mathbb{Y}, d_{\mathbb{Y}})$ and A be a compact subset of $\mathbb{X}$. Suppose f is continuous on A, then f is uniformly continuous on A.*

Proof. Let $\mathcal{B} = \{B(a, r_a/2) : a \in A\}$. Then, $\mathcal{B}$ is an open cover for A. Since A is compact there exists a finite sub-collection of $\mathcal{B}$ which covers A. Let's write $A \subseteq \bigcup_{k=1}^{m} B(a_k, r_k/2)$ where m is finite. Let $\delta := \min\limits_{1\leq k\leq m}\{r_k/2\}$. Since f is continuous on A, for all $\epsilon > 0$ there exists r_k (depending on a_k) such that $d(f(x), f(a_k)) = \|f(x) - f(a_k)\| < \epsilon/2$ whenever $x \in B(a_k, r_k)$. Let $x, x_0 \in A$ such that $d(x, x_0) = \|x - x_0\| < \delta := \min\limits_{1\leq k\leq m}\{r_k/2\}$. Then, there exists $B(a_k, r_k/2)$ such that $x \in B(a_k, r_k/2)$ implies $d(f(x), f(a_k)) < \epsilon/2$. Hence,

$$\begin{aligned} d(x_0, a_k) &\leq d(x_0, x) + d(x, a_k) \\ &< \delta + d(x, a_k) \\ &< \delta + r_k/2 \leq r_k/2 + r_k/2 = r_k. \end{aligned}$$

Hence, $x_0 \in B(a_k, r_k)$ and therefore $d(f(x_0), f(a_k)) < \epsilon/2$. But since

$$\begin{aligned} d(f(x), f(x_0)) &\leq d(f(x), f(a_k)) + d(f(a_k), f(x_0)) \\ &< \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \end{aligned}$$

This completes the proof, since for arbitrary $x, x_0 \in A$ such that $d(x, x_0) < \delta$ implies $d(f(x), f(x_0)) < \epsilon$. $\square$

16 Differential Calculus

We will provide a short review of differentiability as introduced in elementary Calculus and then give some more advanced notions of differentiability. We start with functions defined on $\mathbb{R}$.

16.1 f defined on $\mathbb{R}$

We start with functions defined on a subset of $\mathbb{R}$.

Definition 57. *Let (a, b) be an open interval in $\mathbb{R}$ and $f : (a, b) \to \mathbb{R}$. f is differentiable at $c \in (a, b)$ if*

$$\lim_{x \to c} \frac{f(x) - f(c)}{x - c}$$

exists ($x \neq c$). We denote such limit, which clearly depends on c, by $f^{(1)}(c)$. It is called the derivative of f at c.

It is worth remembering what is meant by the type of limit that appears in the definition. For fixed c with $c \neq x$ we think of the ratio $r_c(x) := \frac{f(x)-f(c)}{x-c}$ as a function on $(a, b) - \{c\}$. If c is a limit point of (a, b), then the limit statement in the definition means that for all $\epsilon > 0$, there exists a $\delta(\epsilon) > 0$ such that $|r_c(x) - f^{(1)}(c)| < \epsilon$ whenever $|x - c| < \delta(\epsilon)$.

Remark 22. 1. *Since limits are unique, we are defining a new function $f^{(1)}$ in (a, b).*

2. *As a matter of notation, the limit in the definition is alternatively denoted by $\frac{d}{dx} f(c)$ or $f'(c)$.*

3. *It is clear that by letting* $h = x - c \neq 0$ *we can equivalently define* $f^{(1)}(c) = \lim_{h\to 0} \frac{f(c+h)-f(c)}{h}$. *As mentioned above, this means that for all* $\epsilon > 0$ *there exist* $\delta(\epsilon) > 0$ *such that whenever* $|h| < \delta(\epsilon)$

$$|f(c+h) - f(c) - hf^{(1)}(c)| < |h|\epsilon$$

4. *When* f *is differentiable at all* $c \in (a,b)$*, we say that* f *is differentiable on* (a,b)*.*

Theorem 84. *Let* (a,b) *be an interval in* $\mathbb{R}$ *and* $f : (a,b) \to \mathbb{R}$ *be differentiable at* $x_0 \in (a,b)$*. Then, there exists* $f^* : (a,b) \to \mathbb{R}$ *continuous at* x_0 *which satisfies*

$$f(x) - f(x_0) = (x - x_0)f^*(x) \text{ for all } x \in (a,b) \tag{14}$$

with $f^*(x_0) = f^{(1)}(x_0)$*. Conversely, if there exists* f^* *which is continuous at* x_0 *and satisfies (14), then* f *is differentiable at* x_0 *and* $f^{(1)}(x_0) = f^*(x_0)$*.*

Proof. ($\Rightarrow$) If $f^{(1)}(x_0)$ exists, define for all $x \in (a,b)$ the following function

$$f^*(x) = \begin{cases} \frac{f(x)-f(x_0)}{x-x_0}, & \text{if } x \neq x_0 \\ \\ f^{(1)}(x_0), & \text{if } x = x_0. \end{cases}$$

Note that $f^*(x) - f^*(x_0) = \frac{f(x)-f(x_0)}{x-x_0} - f^{(1)}(x_0)$ for $x \neq x_0$. Therefore, $\lim_{x\to x_0}(f^*(x) - f^*(x_0)) = 0$, which shows that f^* is continuous at x_0 and clearly satisfies (14).

($\Leftarrow$) If $f(x) - f(x_0) = (x - x_0)f^*(x)$ for all $x \in (a,b)$ then we can write $\frac{f(x)-f(x_0)}{x-x_0} = f^*(x)$ for $x \neq x_0$ and

$$\lim_{x\to x_0} \frac{f(x) - f(x_0)}{x - x_0} = \lim_{x\to x_0} f^*(x).$$

By continuity of f^* at x_0, $\lim_{x\to x_0} f^*(x) = f^*(x_0) = \lim_{x\to x_0} \frac{f(x)-f(x_0)}{x-x_0} = f^{(1)}(x_0)$. □

Remark 23. 1. *Note that* $f(x_0 + h) - f(x_0) = \frac{f(x_0+h)-f(x_0)}{h}h$ *and if* f *is differentiable at* x_0*, then*

$$\lim_{h\to 0} f(x_0 + h) - f(x_0) = f^{(1)}(x_0) \times 0 = 0$$

Therefore, $\lim_{h\to 0} f(x_0+h) = f(x_0)$*, which implies that* f *is continuous at* x_0*, by Theorem 37. Hence, differentiability at* x_0 *implies continuity at* x_0*.*

2. *If $p(x) = f(x)g(x)$ where f, g are differentiable on (a,b). Then,*

$$\begin{aligned}\frac{p(x+h)-p(x)}{h} &= \frac{1}{h}\left(f(x+h)g(x+h) - f(x)g(x) - f(x)g(x+h)\right.\\ &\left.+f(x)g(x+h)\right)\\ &= g(x+h)\frac{f(x+h)-f(x)}{h} + f(x)\frac{g(x+h)-g(x)}{h}\\ p^{(1)}(x) = \lim_{h\to 0}\frac{p(x+h)-p(x)}{h} &= g(x)f^{(1)}(x) + f(x)g^{(1)}(x)\end{aligned}$$

where $\lim_{h\to 0} g(x+h) = g(x)$ by continuity of g.

3. *Using the same arguments as in item 2, if f and g are differentiable, it can be easily verified that*

   if $s(x) = f(x) + g(x)$, then $s^{(1)}(x) = f^{(1)}(x) + g^{(1)}(x)$,

   if $p(x) = x^n$ with $n \in \mathbb{N}$, then $p^{(1)}(x) = nx^{n-1}$,

   if $q(x) = \frac{f(x)}{g(x)}$ for $g(x) \neq 0$, then $q^{(1)}(x) = \frac{1}{g(x)^2}\left(f^{(1)}(x)g(x) - g^{(1)}(x)f(x)\right)$.

The next theorem, which gives the derivative of a composite function, is known as the chain rule for differentiation.

Theorem 85. *Let (a,b) be an interval in $\mathbb{R}$, $f : (a,b) \to \mathbb{R}$ and $g : I := f((a,b)) \to \mathbb{R}$ and consider $h = (g \circ f) : (a,b) \to \mathbb{R}$. Assume that there exists $x_0 \in (a,b)$ such that $f(x_0) \in \overset{\circ}{I}$. If f is differentiable at x_0 and g is differentiable at $f(x_0)$, then $(g \circ f)$ is differentiable at x_0 and $(g \circ f)^{(1)}(x_0) = g^{(1)}(f(x_0))f^{(1)}(x_0)$.*

Proof. By Theorem 84, there exist f^* such that

$$f(x) - f(x_0) = (x - x_0)f^*(x) \text{ for all } x \in (a,b) \text{ with } f^*(x_0) = f^{(1)}(x_0).$$

Similarly, $g(y) - g(f(x_0)) = (y - f(x_0))g^*(f(x))$ for all $y \in T$, where T is some open sub-interval of $f((a,b))$ containing $f(x_0)$. Here, g^* is continuous at $f(x_0)$ and $g^*(f(x_0)) =$

$g^{(1)}(f(x_0))$. Choosing $x \in (a,b)$ such that $y = f(x) \in T$ we have

$$\begin{aligned}(g \circ f)(x) - (g \circ f)(x_0) &= g(f(x)) - g(f(x_0)) \\ &= (f(x) - f(x_0))g^*(f(x)) = (x - x_0)f^*(x)g^*(f(x)).\end{aligned}$$

Therefore, $\frac{(g\circ f)(x)-(g\circ f)(x_0)}{x-x_0} = f^*(x)g^*(f(x))$ and

$$\lim_{x\to x_0} \frac{(g \circ f)(x) - (g \circ f)(x_0)}{x - x_0} = f^{(1)}(x_0)g^{(1)}(f(x_0)).$$

□

The definition of a derivative at a point x_0 required that the point be interior to an interval and that $f^{(1)}(x_0) \in \mathbb{R}$. It is desirable to extend the definition to closed sets and to allow for the possibility of $f^{(1)}(x_0) \in \mathbb{R} \bigcup \{-\infty\} \bigcup \{\infty\}$.

Definition 58. *Let f be defined on $[a,b]$ and assume that f is continuous at x_0. f is said to have a right-derivative at x_0, denoted by $f_+^{(1)}(x_0)$, if*

$$f_+^{(1)}(x_0) := \lim_{x\downarrow x_0} \frac{f(x) - f(x_0)}{x - x_0} \in \mathbb{R} \bigcup \{-\infty\} \bigcup \{\infty\}.$$

Similarly, f is said to have a left-derivative at x_0, denoted by $f_-^{(1)}(x_0)$, if

$$f_-^{(1)}(x_0) := \lim_{x\uparrow x_0} \frac{f(x) - f(x_0)}{x - x_0} \in \mathbb{R} \bigcup \{-\infty\} \bigcup \{\infty\}.$$

If $x_0 \in (a,b)$ (the interior of $[a,b]$), we say that $f^{(1)}(x_0) = \pm\infty$ if $f_+^{(1)}(x_0) = f_-^{(1)}(x_0) = \pm\infty$.

The classical geometric interpretation of $f^{(1)}(x_0)$ as the slope of the line that is tangent to $f(x_0)$ is preserved by the last definition in the case where the tangent line is vertical. Also, note that continuity of f at x_0 is now *assumed*.

16.2 Derivatives and extrema

Definition 59. *Let $(a,b) \subset \mathbb{R}$ and $f : (a,b) \to \mathbb{R}$. We say that x_0 is a local maximum for f if, there exists $r > 0$ such that for all $x \in B(x_0, r)$, $f(x_0) \geq f(x)$. Similarly, we say that x_0 is a local minimum for f if, there exists $r > 0$ such that for all $x \in B(x_0, r)$, $f(x_0) \leq f(x)$.*

Theorem 86. *Let (a, b) be an open interval in $\mathbb{R}$ and $f : (a, b) \to \mathbb{R}$. Assume that for some $x_0 \in (a, b)$, $f^{(1)}(x_0) > 0$ or $f^{(1)}(x_0) = \infty$. Then, there exists $\epsilon > 0$ such that $B(x_0, \epsilon) \subseteq (a, b)$ such that $f(x) > f(x_0)$ if $x > x_0$ and $f(x) < f(x_0)$ if $x < x_0$.*

Proof. If $0 < f^{(1)}(x_0) < \infty$ then by Theorem 84, there exist f^* such that

$$f(x) - f(x_0) = (x - x_0)f^*(x) \text{ for all } x \in (a, b) \text{ with } f^*(x_0) = f^{(1)}(x_0),$$

for all $x \in (a, b)$ with f^* continuous at x_0 and $f^*(x_0) = f^{(1)}(x_0) > 0$. Since f^* is continuous at x_0, there exists $B(x_0, \epsilon) \subset (a, b)$ such that for $x \in B(x_0, \epsilon)$, $f^*(x)$ has the same sign of $f^*(x_0)$, i.e., $f^*(x) > 0$. This means that $f(x) - f(x_0)$ has the same sign of $x - x_0$.

If $f^{(1)}(x_0) = \infty$, there exists $B(x_0, \epsilon) \subset (a, b)$ such that for all $x \in B(x_0, \epsilon)$

$$0 < M < \frac{f(x) - f(x_0)}{x - x_0}, \text{ for } x \neq x_0.$$

Thus, in the ball, $\frac{f(x)-f(x_0)}{x-x_0} > 0$. Therefore, the conclusion follows similarly to the case where $f^{(1)}(x_0) < \infty$. ∎

Theorem 87. *Let $f : (a, b) \to \mathbb{R}$ be differentiable on (a, b). If x_0 is a local maximum for f, then $f^{(1)}(x_0) = 0$.*

Proof. Since x_0 is a local maximum $f(x_0 + h) - f(x_0) \leq 0$ for all $|h| < \delta$ for some $\delta > 0$. Therefore,

$$\frac{f(x_0 + h) - f(x_0)}{h} \leq 0 \text{ for } h \in (0, \delta)$$

and

$$\frac{f(x_0 + h) - f(x_0)}{h} \geq 0 \text{ for } h \in (-\delta, 0).$$

Then,

$$\lim_{h \to 0^+} \frac{f(x_0 + h) - f(x_0)}{h} \leq 0 \tag{15}$$

and

$$\lim_{h \to 0^-} \frac{f(x_0 + h) - f(x_0)}{h} \geq 0. \tag{16}$$

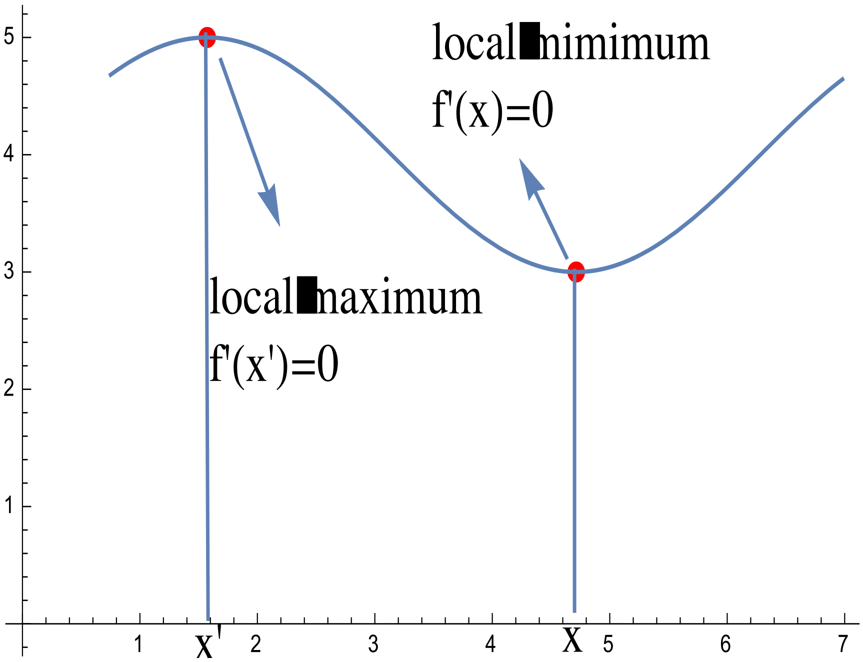

Figure 3: Graphical representation of a function that has derivatives at a local minimum at x and a local maximum at x'

But since f is differentiable at x_0, viz., $\lim\limits_{h\to 0}\frac{f(x_0+h)-f(x_0)}{h}$ exists as $f^{(1)}(x_0)$, the limits in (15) and (16) must coincide. This can happen only at 0, therefore $f^{(1)}(x_0)=0$. $\square$

Remark 24. *The same is true for a local minimum. That is, if x_0 is a local minimum for f on (a,b), then $f^{(1)}(x_0)=0$.*

Theorem 88. *(Rolle's Theorem) Let $f:[a,b]\to\mathbb{R}$ be differentiable on (a,b). Assume that f is continuous at a and b. If $f(a)=f(b)$ there exists at least one $x\in(a,b)$ such that $f^{(1)}(x)=0$.*

Proof. Since f is continuous on a compact set, there exist x_i and x_s on $[a,b]$ such that $f(x_i)=\min\limits_{x\in[a,b]} f(x)$, $f(x_s)=\max\limits_{x\in[a,b]} f(x)$. Now, assume that $f^{(1)}(x)\neq 0$ on (a,b). By Theorem 87 x_i and x_s are not in (a,b) because, if they were, $f^{(1)}(x_i)=f^{(1)}(x_s)=0$. Hence, both are at the endpoints. Since $f(a)=f(b)$, then $f(x_i)=f(x_s)$ and f is a constant on $[a,b]$.

However, in this case $f^{(1)}(x) = 0$, a contradiction. $\square$

Rolle's Theorem can be used to prove the very important Mean Value Theorem.

Theorem 89. *(Mean Value Theorem) Let $f : [a,b] \to \mathbb{R}$ be differentiable on (a,b) and continuous at both end points. Then, there exists $x \in (a,b)$ such that $f(b)-f(a) = f^{(1)}(x)(b-a)$.*

Proof. Let

$$\phi(y) = f(y) - f(a) - \frac{f(b) - f(a)}{b - a}(y - a). \tag{17}$$

Note that $\phi(b) = \phi(a) = 0$ and for all $x \in (a,b)$

$$\frac{\phi(y) - \phi(x)}{y - x} = \frac{f(y) - f(x)}{y - x} - \frac{f(b) - f(a)}{b - a} \tag{18}$$

and

$$\lim_{y \to x} \frac{\phi(y) - \phi(x)}{y - x} = f^{(1)}(x) - \frac{f(b) - f(a)}{b - a} \tag{19}$$

since f is differentiable on (a,b). Therefore, ϕ satisfies all assumptions on Theorem 88 and, consequently, there exists $x \in (a,b)$ such that $\phi^{(1)}(x) = 0$, and from (19)

$$\phi^{(1)}(x) = f^{(1)}(x) - \frac{f(b) - f(a)}{b - a} = 0,$$

which implies $f(b) - f(a) = f^{(1)}(x)(b - a)$. $\square$

Since limits are unique, $f^{(1)}$ is a function with domain $S \subset (a,b)$ where f is differentiable. If $f^{(1)}$ is differentiable at $c \in S$, then we write

$$f^{(2)}(c) := \lim_{h \to 0} \frac{f^{(1)}(c + h) - f^{(1)}(c)}{h}$$

and call $f^{(2)}(c)$ the second derivative of f at c. When it exists, the n^{th} derivative of f at c is denoted by $f^{(n)}(c)$.