Background Material

These notes contain a summary of background material from linear algebra and calculus. Much of the content should be familiar to some degree, and the purpose is to bring it back to attention. Important concepts are **highlighted** in the notes.

1 Asymptotic notation

An **algorithm** is a sequence of instructions carried out by a computer. Important features of an algorithm are computation time (measured as the number of operations or the number of iterations needed to reach a solution) and accuracy. One is mainly interested in the **orders of magnitude** of these quantities, and not so much in their exact values. A convenient notation for this purpose is the asymptotic O-notation.

algorithm

orders of magnitude

Let $f, g\colon \mathbb{R} \to \mathbb{R}$ be two functions. Then:

- $f(n) \in O(g(n))$ as $n \to \infty$ if there exists a constant $C > 0$ and an integer n_0 such that $|f(n)| \leq C|g(n)|$ for $n > n_0$;

$O(g(n)), O(g(x))$

- $f(x) \in O(g(x))$ as $x \to 0$ if there exists a constant $C > 0$ and a real number $\varepsilon > 0$ such that $|f(x)| \leq C|g(x)|$ for $|x| < \varepsilon$.

We omit "$n \to \infty$" or "$x \to 0$" when it is clear from the context. One often finds statements such as $f(n) = O(g(n))$ or $f(x) = 1 + x + O(x^2)$; the first is equivalent to $f(n) \in O(g(n))$, while the second should be read as $f(x) = 1 + x + g(x)$ for a function $g(x) \in O(x^2)$. The following examples illustrate the O-notation.

- $\sqrt{n} + n^2 \in O(n^2)$ as $n \to \infty$
- $n^5 \in O(e^n)$ as $n \to \infty$
- $10^{100} \in O(1)$ as $n \to \infty$
- $x^3 \in O(x^2)$ as $x \to 0$
- $\sin(x) \in O(x)$ as $x \to 0$
- $e^x = 1 + x + O(x^2)$ as $x \to 0$

The notation $f(n) \in \Omega(g(n))$ as $n \to \infty$ means that $g(n) \in O(f(n))$ as $n \to \infty$, and $f(n) \in o(g(n))$ as $n \to \infty$ means that for *all* $C > 0$ there exists n_0 such that $|f(n)| < C|g(n)|$ for $n > n_0$. If $g(n) \neq 0$ for sufficiently large n, this is equivalent to $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$. One defines $\Omega(g(x))$ and $o(g(x))$ as $x \to 0$ analogously.

2 Linear Algebra

vector
transpose

We restrict to linear algebra over the field of real numbers $\mathbb{R}$, as this is the setting that is of most interest in optimization. A **vector** in $\mathbb{R}^n$ and its **transpose** are written as

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = (x_1, \dots, x_n)^\top, \quad \boldsymbol{x}^\top = (x_1, \dots, x_n),$$

coordinates

with **coordinates** $x_i \in \mathbb{R}$ for $1 \leq i \leq n$. The zero vector is denoted by $\mathbf{0}$, while $\boldsymbol{e}$ is the vector with every coordinate equal to 1. If $\lambda_1, \lambda_2 \in \mathbb{R}$ and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, then $\lambda_1 \boldsymbol{x} + \lambda_2 \boldsymbol{y}$ is the vector with coordinates $\lambda_1 x_i + \lambda_2 y_i$ for $1 \leq i \leq n$.

inner product

In $\mathbb{R}^n$ we have the Euclidean (or standard) **inner product** (or scalar product)

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

bilinear

The Euclidean inner product is **bilinear**: for $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathbb{R}^n$ and $\alpha, \beta \in \mathbb{R}$,

$$\langle \alpha \boldsymbol{x}_1 + \beta \boldsymbol{x}_2, \boldsymbol{y} \rangle = \alpha \langle \boldsymbol{x}_1, \boldsymbol{y} \rangle + \beta \langle \boldsymbol{x}_2, \boldsymbol{y} \rangle, \quad \langle \boldsymbol{x}, \alpha \boldsymbol{y}_1 + \beta \boldsymbol{y}_2 \rangle = \alpha \langle \boldsymbol{x}, \boldsymbol{y}_1 \rangle + \beta \langle \boldsymbol{x}, \boldsymbol{y}_2 \rangle,$$

orthogonal

symmetric ($\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{y}, \boldsymbol{x} \rangle$) and satisfies $\langle \boldsymbol{x}, \boldsymbol{x} \rangle \geq 0$, with equality if and only if $\boldsymbol{x} = \mathbf{0}$. Two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are called **orthogonal**, if $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$.

Example 2.1. The vectors $(1, 1)^\top$ and $(1, -1)^\top$ are orthogonal in $\mathbb{R}^2$, while $(1, 1)^\top$ and $(2, -1)^\top$ are not.

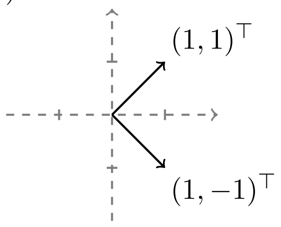


Figure 2.1: Orthogonal vectors

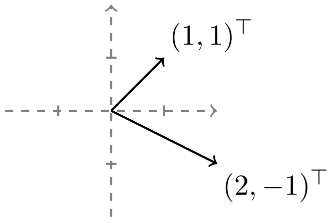


Figure 2.2: Non-orthogonal vectors

Linear subspaces

linear subspace

A **linear subspace** is a subset $V \subseteq \mathbb{R}^n$ such that for any $\boldsymbol{x}, \boldsymbol{y} \in V$ and for all $\alpha, \beta \in \mathbb{R}$, $\alpha \boldsymbol{x} + \beta \boldsymbol{y} \in V$. In particular, the sets $\{\mathbf{0}\}$ and $\mathbb{R}^n$ are linear subspaces.

Example 2.2. The linear subspaces of $\mathbb{R}^2$ are $\{\mathbf{0}\}$, lines through the origin, and $\mathbb{R}^2$. The linear subspaces of $\mathbb{R}^3$ are $\{\mathbf{0}\}$, lines and planes through the origin, and $\mathbb{R}^3$.

linear combination

A **linear combination** of vectors $\boldsymbol{x}_1, \dots, \boldsymbol{x}_k \in \mathbb{R}^n$ is an expression of the form $\boldsymbol{x} = \sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i$, where $\lambda_i \in \mathbb{R}$ for $1 \leq i \leq k$. The set of linear combinations

$$V = \operatorname{span}\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_k\} := \left\{ \sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i : \lambda_i \in \mathbb{R} \right\}$$

forms a linear subspace of $\mathbb{R}^n$. It is the intersection of all linear subspaces that contain $\boldsymbol{x}_1, \dots, \boldsymbol{x}_k$. The vectors $\boldsymbol{x}_1, \dots, \boldsymbol{x}_k$ are **linearly independent** if $\sum_{i=1}^k \lambda_i \boldsymbol{x}_i = 0$ implies $\lambda_1 = \cdots = \lambda_k = 0$. A minimal set of vectors that span a linear subspace V is called a **basis** of this subspace, and the number of elements in a basis is the **dimension** of the linear subspace. The elements of a basis are always linearly independent, and a maximal linearly independent set in a vector subspace V is a basis. If $\{\boldsymbol{b}_1, \dots, \boldsymbol{b}_k\}$ is a basis of a subspace V, then every $\boldsymbol{x} \in V$ has a *unique* representation $\boldsymbol{x} = \sum_{i=1}^k \lambda_i \boldsymbol{b}_i$. A basis is **orthogonal** if $\langle \boldsymbol{b}_i, \boldsymbol{b}_j \rangle = 0$ for $i \neq j$, and **orthonormal** if in addition $\langle \boldsymbol{b}_i, \boldsymbol{b}_i \rangle = 1$ for $1 \leq i \leq k$. The unique expression of $\boldsymbol{x} \in V$ as linear combination of an orthonormal basis $\{\boldsymbol{b}_1, \dots, \boldsymbol{b}_k\}$ of V is given by

linearly independent

basis

dimension

orthonormal basis

$$\boldsymbol{x} = \sum_{i=1}^k \langle \boldsymbol{x}, \boldsymbol{b}_i \rangle \boldsymbol{b}_i.$$

The **standard basis** of $\mathbb{R}^n$ is the orthonormal basis $\{\boldsymbol{e}_1, \dots, \boldsymbol{e}_n\}$, where $\boldsymbol{e}_i$ has a 1 in the i-th coordinate and 0 elsewhere.

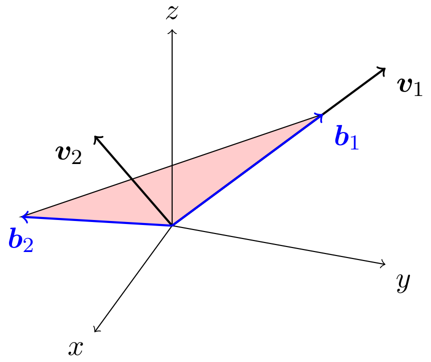
standard basis

Example 2.3. The vectors $\boldsymbol{v}_1 = (0,1,1)^\top$ and $\boldsymbol{v}_2 = (1,0,1)^\top$ span a linear subspace of $\mathbb{R}^3$ of dimension 2, but they are not orthogonal. The vectors

$$\boldsymbol{b}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \boldsymbol{b}_2 = \frac{1}{\sqrt{3}} \begin{pmatrix} \sqrt{2} \\ -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$$

form an orthonormal basis of V. The vector $\boldsymbol{x} = (1,1,2)^\top$ lives in V, and its representation in terms of $\{\boldsymbol{b}_1, \boldsymbol{b}_2\}$ is

$$\boldsymbol{x} = \frac{3}{\sqrt{2}} \boldsymbol{b}_1 + \sqrt{\frac{3}{2}} \boldsymbol{b}_2.$$



The **direct sum** of two vector subspaces $V, W \subset \mathbb{R}^n$ with $V \cap W = \{\boldsymbol{0}\}$ is

direct sum $\oplus$

$$V \oplus W = \{\boldsymbol{v} + \boldsymbol{w} : \boldsymbol{v} \in V, \boldsymbol{w} \in W\}.$$

orthogonal complement $\perp$

The **orthogonal complement** of a subspace $V \subseteq \mathbb{R}^n$ is the set

$$V^\perp = \{\boldsymbol{x} \in \mathbb{R}^n : \forall \boldsymbol{y} \in V \colon \langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0\}.$$

The vector space $\mathbb{R}^n$ is the direct sum of V and its orthogonal complement,

$$\mathbb{R}^n = V \oplus V^\perp. \tag{1}$$

If $V = \operatorname{span}\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_k\}$, then $V^\perp = \{\boldsymbol{y} : \langle \boldsymbol{x}_1, \boldsymbol{y} \rangle = \cdots \langle \boldsymbol{x}_k, \boldsymbol{y} \rangle = 0\}$; to check whether a vector $\boldsymbol{y}$ is in the orthogonal complement of V we therefore only need to check whether $\boldsymbol{y}$ is orthogonal to a spanning set (for example, a basis) of V.

Example 2.4. The vector $\boldsymbol{x} = (-1, -1, 1)^\top$ is orthogonal to the basis $\{\boldsymbol{b}_1, \boldsymbol{b}_2\}$ from Example 2.3. It is therefore orthogonal to the whole plane $V = \operatorname{span}\{\boldsymbol{b}_1, \boldsymbol{b}_2\}$ spanned by these vectors. The orthogonal complement of V is the line $\{\lambda \boldsymbol{x} : \lambda \in \mathbb{R}\}$.

direct product $\times$

The **direct product** of two vector subspaces $V \subseteq \mathbb{R}^n$, $W \subseteq \mathbb{R}^m$ is defined as

$$V \times W = \{(\boldsymbol{v}, \boldsymbol{w}) \in \mathbb{R}^{n+m} : \boldsymbol{v} \in V, \boldsymbol{w} \in W\}.$$

where $(\boldsymbol{v}, \boldsymbol{w})$ is the vector whose first n coordinates coincide with $\boldsymbol{v}$, and the last m coordinates coincide with $\boldsymbol{w}$. In particular, $\mathbb{R}^n \times \mathbb{R}^m = \mathbb{R}^{n+m}$.

Linear maps

matrix

An $m \times n$ **matrix**

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix},$$

linear map

represents a **linear map** from $\mathbb{R}^n$ to $\mathbb{R}^m$ by means of

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}, \quad y_i = \sum_{j=1}^n a_{ij} x_j.$$

For example,

$$\begin{pmatrix} 2 & 1 & 0 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 3 \\ 5 \end{pmatrix}.$$

The columns of a matrix are vectors, and we sometimes write

$$\boldsymbol{A} = (\boldsymbol{a}_1, \dots, \boldsymbol{a}_n)$$

block matrix

for the matrix whose columns are given by the vectors $\boldsymbol{a}_i$. If $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ and $n = m + k$, then $\boldsymbol{A}$ can be written as **block matrix**,

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix},$$

with $\boldsymbol{A}_{11} \in \mathbb{R}^{m\times m}$, $\boldsymbol{A}_{22} \in \mathbb{R}^{k\times k}$, $\boldsymbol{A}_{12} \in \mathbb{R}^{m\times k}$ and $\boldsymbol{A}_{21} \in \mathbb{R}^{k\times m}$. The sum and difference of matrices of the same size are defined component-wise.

The $n \times n$ matrix $\mathbf{1}$ is the matrix with 1 on the diagonal and 0 elsewhere, while $\mathbf{0}$ is the matrix consisting of only zeros. A matrix is **diagonal** if all the off-diagonal elements are 0, **lower-triangular** if all the elements above the diagonal are 0, and **upper-triangular** if all the elements below the diagonal are 0. A **block-diagonal** matrix is a block matrix, with all blocks outside the main diagonal consisting of zero-matrices $\mathbf{0}$.

diagonal, **triangular**, **block-diagonal**

The **transpose** $\boldsymbol{A}^\top$ is the matrix with entries $a'_{ij} := a_{ji}$. It is the matrix $\boldsymbol{A}$ mirrored on the diagonal from top left to bottom right. A matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is called **symmetric** if $\boldsymbol{A}^\top = \boldsymbol{A}$. The set of symmetric matrices in $\mathbb{R}^{n\times n}$ is denoted by $\mathcal{S}^n$.

transpose

symmetric

The **product** of an $m \times p$ matrix $\boldsymbol{A}$ with a $p \times n$ matrix $\boldsymbol{B}$,

product

$$\boldsymbol{C} = \boldsymbol{A}\boldsymbol{B},$$

is the $m \times n$ matrix $\boldsymbol{C}$ whose (i, j)-th entry is given by

$$c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj}.$$

It represents a composition of maps $\mathbb{R}^n \to \mathbb{R}^p \to \mathbb{R}^m$. The number of columns of $\boldsymbol{A}$ has to equal the number of rows of $\boldsymbol{B}$ for this definition to make sense. Products of block matrices or of block matrices with vectors can be carried out block-wise. If, for example, $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2)^\top$ with $\boldsymbol{x}_1 \in \mathbb{R}^{1\times m}$ and $\boldsymbol{x}_2 \in \mathbb{R}^{1\times k}$, then

$$\begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{A}_{11}\boldsymbol{x}_1 + \boldsymbol{A}_{12}\boldsymbol{x}_2 \\ \boldsymbol{A}_{21}\boldsymbol{x}_1 + \boldsymbol{A}_{22}\boldsymbol{x}_2 \end{pmatrix}.$$

The matrix $\mathbf{1}$ satisfies $\mathbf{1}\boldsymbol{A} = \boldsymbol{A}$ and $\boldsymbol{A}\mathbf{1} = \boldsymbol{A}$, whenever the dimensions are such that this is defined. In general, even if $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$, $\boldsymbol{A}\boldsymbol{B} \neq \boldsymbol{B}\boldsymbol{A}$.

Example 2.5. Let

$$\boldsymbol{A} = \begin{pmatrix} 1 & 2 \\ 1 & 4 \end{pmatrix}, \quad \boldsymbol{B} = \begin{pmatrix} 2 & 3 \\ 3 & 2 \end{pmatrix}.$$

Then

$$\boldsymbol{A}\boldsymbol{B} = \begin{pmatrix} 8 & 7 \\ 14 & 11 \end{pmatrix}, \quad \boldsymbol{B}\boldsymbol{A} = \begin{pmatrix} 5 & 16 \\ 5 & 14 \end{pmatrix}.$$

If we consider a vector $\boldsymbol{x} \in \mathbb{R}^n$ as an $n \times 1$ matrix and the transpose as an $1 \times n$ matrix, then for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n \cong \mathbb{R}^{n\times 1}$ we have

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^\top \boldsymbol{y}.$$

The transpose of a product satisfies $(\boldsymbol{A}\boldsymbol{B})^\top = \boldsymbol{B}^\top \boldsymbol{A}^\top$. From this it follows that for any matrix, $\boldsymbol{A}^\top \boldsymbol{A}$ is symmetric. For any matrix $\boldsymbol{A}$ we have

$$\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x} \rangle = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = (\boldsymbol{A}^\top \boldsymbol{x})^\top \boldsymbol{x} = \langle \boldsymbol{A}^\top \boldsymbol{x}, \boldsymbol{x} \rangle.$$

It follows from this that if a matrix is symmetric, then it is also self-adjoint, which means that $\langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{x}\rangle = \langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x}\rangle$.

rank
kernel
image

The **rank** of a matrix $\boldsymbol{A}$, $\mathrm{rk}(\boldsymbol{A})$, is the maximum number of linearly independent rows or columns of $\boldsymbol{A}$. The **kernel** and **image** of $\boldsymbol{A}$ are the linear subspaces

$$\ker \boldsymbol{A} := \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{x} = 0\}, \quad \mathrm{im}\, \boldsymbol{A} = \{\boldsymbol{A}\boldsymbol{x} : \boldsymbol{x} \in \mathbb{R}^n\}.$$

The dimensions are given by $\dim\ker \boldsymbol{A} = n - \mathrm{rk}(\boldsymbol{A})$ and $\dim\mathrm{im}\, \boldsymbol{A} = \mathrm{rk}(\boldsymbol{A})$. While $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ represents a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$, the transpose $\boldsymbol{A}^\top$ represents a map in the other direction, and the image of $\boldsymbol{A}^\top$ coincides with the orthogonal complement of the kernel of $\boldsymbol{A}$, $(\ker \boldsymbol{A})^\perp = \mathrm{im}\, \boldsymbol{A}^\top$. In particular, in view of (1) we have the direct sum decomposition

$$\mathbb{R}^n = \ker \boldsymbol{A} \oplus \mathrm{im}\, \boldsymbol{A}^\top.$$

linear equations

A **system of linear equations**

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1 \\ \vdots \qquad\quad \vdots \quad &\quad \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= b_m \end{aligned}$$

is written as a matrix vector product

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \tag{2}$$

where the $m \times n$ matrix $\boldsymbol{A}$ is defined as above, and $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{b} \in \mathbb{R}^m$. If the columns of $\boldsymbol{A}$ are linearly independent, then the system of equations can have at most one solution, and otherwise it has infinitely many solutions (this is the case if $n > m$). If $n = m$, then the system (2) has a unique solution if and only if the matrix $\boldsymbol{A}$ is **invertible** or **non-singular**. This is the case if the rows of $\boldsymbol{A}$ (or equivalently, the columns of $\boldsymbol{A}$) are linearly independent. If $\boldsymbol{A}$ is not invertible, it is called **singular**.

invertible, **singular**

inverse

If $\boldsymbol{A}$ is invertible, there exists a matrix $\boldsymbol{A}^{-1}$ (the **inverse**) such that

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \mathbf{1}.$$

The solution of (2) is then given by $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$. The following conditions on a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ are equivalent:

1. $\boldsymbol{A}$ is invertible,
2. $\mathrm{rk}(\boldsymbol{A}) = n$,
3. $\ker \boldsymbol{A} = \{\mathbf{0}\}$,
4. $\mathrm{im}\, \boldsymbol{A} = \mathbb{R}^n$,
5. the rows of $\boldsymbol{A}$ are linearly independent,

6. the columns of $\boldsymbol{A}$ are linearly independent,
7. $\det(\boldsymbol{A}) \neq 0$,

where the determinant is

$$\det(\boldsymbol{A}) = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) a_{1\sigma(1)} \cdots a_{n\sigma(n)},$$

and S_n is the group of permutations of $[n] = \{1, \dots, n\}$, with $\operatorname{sgn}(\sigma)$ the sign of the permutation (parity of the number of inversions).

Example 2.6. For two- and three-dimensional matrices

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \quad \boldsymbol{B} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix},$$

the determinants are

$$\begin{aligned} \det(\boldsymbol{A}) &= a_{11}a_{22} - a_{12}a_{21}, \\ \det(\boldsymbol{B}) &= a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31}). \end{aligned}$$

A matrix $\boldsymbol{Q}$ is **orthogonal** if $\boldsymbol{Q} = (\boldsymbol{q}_1, \dots, \boldsymbol{q}_n)$, with $\langle \boldsymbol{q}_i, \boldsymbol{q}_j \rangle = \delta_{ij}$, and **orthogonal**

$$\delta_{ij} = \begin{cases} 0 & \text{if } i \neq j, \\ 1 & \text{if } i = j. \end{cases}$$

As the (i,j)-th entry of $\boldsymbol{Q}^\top \boldsymbol{Q}$ are given by $\langle \boldsymbol{q}_i, \boldsymbol{q}_j \rangle$, the orthogonality of $\boldsymbol{Q}$ can succinctly be characterized by the requirement $\boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{1}$. In particular, $\boldsymbol{Q}^\top = \boldsymbol{Q}^{-1}$, and the columns (and rows) of an orthogonal matrix form an orthonormal basis of $\mathbb{R}^n$. Orthogonal matrices have the property that $\langle \boldsymbol{Q}\boldsymbol{x}, \boldsymbol{Q}\boldsymbol{y} \rangle = \langle \boldsymbol{x}, \boldsymbol{y} \rangle$. From this it follows that orthogonality of vectors is preserved under orthogonal transformations. The determinant of an orthogonal matrix is $\det(\boldsymbol{Q}) = 1$. As the product of orthogonal matrices is again orthogonal, the set of orthogonal $n \times n$ matrices forms a group, commonly denoted by $O(n)$.

Example 2.7. Consider the three matrices,

$$\boldsymbol{A} = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}, \quad \boldsymbol{B} = \begin{pmatrix} 2 & 3 \\ 3 & 2 \end{pmatrix}, \quad \boldsymbol{C} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}.$$

The matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric, while $\boldsymbol{C}$ is not. The matrices $\boldsymbol{B}$ and $\boldsymbol{C}$ are invertible, with inverse

$$\boldsymbol{B}^{-1} = \begin{pmatrix} -0.4 & 0.6 \\ 0.6 & -0.4 \end{pmatrix}, \quad \boldsymbol{C}^{-1} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}.$$

The matrix $\boldsymbol{A}$ is not invertible, since the second column is a multiple of the first. The kernel of $\boldsymbol{A}$ the linear span of $(-2, 1)^\top$. The matrix $\boldsymbol{C}$ is orthogonal (this can be seen by checking that the columns or rows are orthonormal, or looking at the expression of the inverse above).

Eigenvalues

eigenvector

A vector $\boldsymbol{u} \neq \mathbf{0}$ is an **eigenvector** of $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, if there exists a $\lambda \in \mathbb{C}$ such that

$$\boldsymbol{A}\boldsymbol{u} = \lambda \boldsymbol{u}.$$

eigenvalue

Such a number λ is called an **eigenvalue** of $\boldsymbol{A}$. Note that the eigenvectors are only defined up to scaling: if $\boldsymbol{u}$ is an eigenvector, then so is $\lambda\boldsymbol{u}$ for any non-zero $\lambda \in \mathbb{R}$.

characteristic polynomial

From the definition of the determinant, the function $\lambda \mapsto \det(\lambda\mathbf{1} - \boldsymbol{A})$ is a polynomial of degree at most n, called the **characteristic polynomial** of $\boldsymbol{A}$. The eigenvalues are the roots of this polynomial,

$$\det(\lambda\mathbf{1} - \boldsymbol{A}) = 0.$$

determinant
trace

The eigenvalues can be complex numbers, and appear in complex conjugate pairs. If the matrix $\boldsymbol{A}$ is symmetric, then the eigenvalues are all real numbers. Two important quantities, the **determinant** and the **trace** of a matrix (corresponding, up to sign, to the highest and lowest coefficient of the characteristic polynomial) can be expressed in terms of the eigenvalues:

$$\det(\boldsymbol{A}) = \lambda_1 \cdots \lambda_n, \quad \operatorname{trace}(\boldsymbol{A}) := a_{11} + \cdots + a_{nn} = \lambda_1 + \cdots + \lambda_n.$$

A matrix has a zero eigenvalue if and only if it is singular. Eigenvalues may occur with multiplicity.

Norms

norm

A **norm** in $\mathbb{R}^n$ is a function $\|\cdot\|$ that satisfies the following three properties

1. $\|\boldsymbol{x}\| \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$, and $\boldsymbol{x} = 0$ if and only if $\boldsymbol{x} = \mathbf{0}$;
2. $\|\lambda\boldsymbol{x}\| = |\lambda|\|\boldsymbol{x}\|$ for $\lambda \in \mathbb{R}$ and $\boldsymbol{x} \in \mathbb{R}^n$;
3. $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.

Three important examples of norms are the following:

1. The 1-norm: $\|\boldsymbol{x}\|_1 = \sum_{i=1}^n |x_i|$;
2. The 2-norm: $\|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$;
3. The ∞-norm: $\|\boldsymbol{x}\|_\infty = \max_{1\leq i\leq n} |x_i|$.

Example 2.8. Let $\boldsymbol{x} = (2, -3, 4)^\top$. The $\|\boldsymbol{x}\|_1 = 9$, $\|\boldsymbol{x}\|_2 = \sqrt{29}$, and $\|\boldsymbol{x}\|_\infty = 4$.

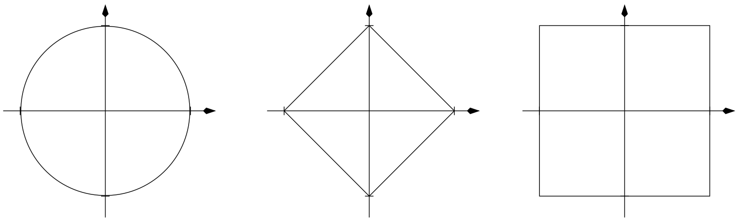
2-norm

Note that the 2-norm, also called **Euclidean norm**, can be defined as

$$\|\boldsymbol{x}\|_2^2 = \boldsymbol{x}^\top\boldsymbol{x} = \langle \boldsymbol{x}, \boldsymbol{x} \rangle,$$

that is, it is the norm induced by the Euclidean inner product. From this it follows that the 2-norm does not change under orthogonal transformations: if $\boldsymbol{Q} \in O(n)$ and $\boldsymbol{x} \in \mathbb{R}^n$, then $\|\boldsymbol{Qx}\|_2 = \|\boldsymbol{x}\|_2$. Orthogonal transformations in $\mathbb{R}^2$ and $\mathbb{R}^3$ correspond to rotations and reflections, so it is intuitively clear that these don't change distances.

The **unit sphere** with respect to a norm is the set $\{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| = 1\}$, and the (closed) **unit ball** is the set $\{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\| \leq 1\}$. The unit spheres with respect to the 2-norm, the 1-norm and the ∞-norm in $\mathbb{R}^2$ are shown in the following diagram. **unit sphere**, **ball**



The unit sphere with respect to the 2-norm in $\mathbb{R}^n$ is usually denoted by S^{n-1}.

The 1-, 2- and ∞-norms are equivalent, in the sense that they can be bounded in terms of each other. In particular,

$$\|\boldsymbol{x}\|_\infty \leq \|\boldsymbol{x}\|_2 \leq \sqrt{n}\|\boldsymbol{x}\|_\infty, \quad \|\boldsymbol{x}\|_\infty \leq \|\boldsymbol{x}\|_1 \leq n\|\boldsymbol{x}\|_\infty. \tag{3}$$

The inner product and the 2-norm are related the **Cauchy-Schwarz** inequality, **Cauchy-Schwarz**

$$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \leq \|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2,$$

with equality if and only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are linearly dependent. As a consequence of the Cauchy-Schwarz inequality we get

$$-1 \leq \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2} \leq 1.$$

The **angle** between vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is the number $\theta \in [0, 2\pi)$ such that **angle**

$$\cos(\theta) = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2}.$$

If $\boldsymbol{x}$ and $\boldsymbol{y}$ are orthogonal, then $\cos(\theta) = 0$ and $\theta \in \{\pi/2, 3\pi/2\}$.

Norms are an important device to measure the size of vectors. In order to measure the amount by which a linear transformation (matrix) distorts vectors, we need the concept of **matrix norms**. A matrix norm is a function on the set of matrices that is a norm when considering a matrix as a vector, and in addition satisfies the condition **matrix norm**

$$\|\boldsymbol{AB}\| \leq \|\boldsymbol{A}\|\|\boldsymbol{B}\|.$$

The most important examples are given by the **operator norms**. Given a vector norm **operator norm**

$\|\boldsymbol{x}\|$, the associated matrix norm is defined as

$$\|\boldsymbol{A}\| = \max_{\boldsymbol{x}\neq \boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \max_{\boldsymbol{x}\colon \|\boldsymbol{x}\|\leq 1} \|\boldsymbol{A}\boldsymbol{x}\|.$$

The matrix norms $\|\boldsymbol{A}\|_1, \|\boldsymbol{A}\|_2, \|\boldsymbol{A}\|_\infty$ are the operator norms that arise when using the 1-, 2- and ∞-norms. They can conveniently characterized as follows

- $\|\boldsymbol{A}\|_1 = \max_j \sum_{i=1}^n |a_{ij}|$;
- $\|\boldsymbol{A}\|_\infty = \max_i \sum_{j=1}^n |a_{ij}|$;
- $\|\boldsymbol{A}\|_2 = \sqrt{\lambda_{\max}(\boldsymbol{A}^\top \boldsymbol{A})}$.

Here, $\lambda_{\max}$ denotes the largest eigenvalue of the symmetric matrix $\boldsymbol{A}^\top\boldsymbol{A}$. If $\boldsymbol{A}$ is symmetric, then $\boldsymbol{A}^\top\boldsymbol{A} = \boldsymbol{A}^2$, and the eigenvalues of $\boldsymbol{A}^2$ are the squares of the eigenvalues of $\boldsymbol{A}$. It follows that for symmetric $\boldsymbol{A}$, $\|\boldsymbol{A}\|_2 = \lambda_{\max}(\boldsymbol{A})$.

Example 2.9. Let

$$\boldsymbol{A} = \begin{pmatrix} 1 & -2 \\ -2 & 1 \end{pmatrix}.$$

Then $\|\boldsymbol{A}\|_1 = \|\boldsymbol{A}\|_\infty = 3$ and $\|\boldsymbol{A}\|_2 = 3$.

dual norm

A special case is the **dual norm** of a vector norm: to a given norm is defined as

$$\|\boldsymbol{x}\|^* = \max_{\boldsymbol{y}\colon \|\boldsymbol{y}\|\leq 1} \langle \boldsymbol{x}, \boldsymbol{y}\rangle.$$

This is the operator norm of $\boldsymbol{x}^\top$, considered as a $1\times d$ matrix. The dual of the dual norm is the norm itself. The dual norm of the 2-norm is again the 2-norm, while the 1-norm and the ∞-norm are dual to each other.

Frobenius norm

In addition to the operator norms, an important matrix norm is the **Frobenius norm** of a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$,

$$\|\boldsymbol{A}\|_F = \sqrt{\sum_{i,j=1}^n a_{ij}^2}.$$

This is just the 2-norm of $\boldsymbol{A}$ interpreted as a vector in $\mathbb{R}^{n^2}$. The 2-norm and the Frobenius norm have the important property of being **orthogonal invariant**, which means that for any $\boldsymbol{Q} \in O(n)$,

orthogonal invariant

$$\|\boldsymbol{Q}\boldsymbol{A}\|_2 = \|\boldsymbol{A}\boldsymbol{Q}\|_2 = \|\boldsymbol{A}\|_2, \quad \|\boldsymbol{Q}\boldsymbol{A}\|_F = \|\boldsymbol{A}\boldsymbol{Q}\|_F = \|\boldsymbol{A}\|_F.$$

Orthogonal invariance allows to simplify a matrix without changing the norm.

Positive semidefinite matrices

If $\boldsymbol{A} \in \mathcal{S}^n$ is a symmetric matrix and $\boldsymbol{u} \in \mathbb{R}^n$ an eigenvector with $\|u\|_2 = 1$ and corresponding eigenvalue λ, then $\boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{u} = \lambda \boldsymbol{u}^\top \boldsymbol{u} = \lambda$. In particular, the largest and smallest values of an eigenvalue are given by

$$\lambda_1 = \max_{\boldsymbol{u}\colon \|\boldsymbol{u}\|_2=1} \boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{u}, \quad \lambda_n = \min_{\boldsymbol{u}\colon \|\boldsymbol{u}\|_2=1} \boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{u}.$$

A symmetric matrix $\boldsymbol{A}$ is called **positive semidefinite**, written $\boldsymbol{A} \succeq \boldsymbol{0}$, if for all non-zero $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \geq 0$, and **positive definite**, written $\boldsymbol{A} \succ \boldsymbol{0}$, if $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} > 0$ for all $\boldsymbol{x} \neq \boldsymbol{0}$. Equivalently, a symmetric matrix is positive semidefinite if all its eigenvalues are non-negative, and positive definite if they are all positive. The set of positive semidefinite symmetric matrices in $\mathbb{R}^{n\times n}$ is denoted by $\mathcal{S}^n_+$, while the set of positive definite matrices is $\mathcal{S}^n_{++}$.

positive (semi)definite

An **inner product** (or scalar product) on $\mathbb{R}^{n\times n}$ is a function

inner product

$$\langle \cdot, \cdot \rangle \colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}, \quad (\boldsymbol{x}, \boldsymbol{y}) \mapsto \langle \boldsymbol{x}, \boldsymbol{y} \rangle$$

that is bilinear (linear in each of the two arguments), symmetric ($\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{y}, \boldsymbol{x} \rangle$), and satisfies $\langle \boldsymbol{x}, \boldsymbol{x} \rangle \geq 0$, with $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = 0$ if and only if $\boldsymbol{x} = \boldsymbol{0}$. The standard inner product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{x}^\top \boldsymbol{y}$ is an example, and the notation $\langle \cdot, \cdot \rangle$ usually refers to this product. More generally, every matrix $\boldsymbol{A} \in \mathcal{S}^n_{++}$ defines an inner product by

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\boldsymbol{A}} := \langle \boldsymbol{x}, \boldsymbol{A}\boldsymbol{y} \rangle = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{y}.$$

The associated norm is $\|\boldsymbol{x}\|_{\boldsymbol{A}} = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\boldsymbol{A}}}$. The unit sphere with respect to this norm,

$$E = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = 1\},$$

is an **ellipsoid**, where the 2-norms of the largest and smallest axes are the largest and smallest eigenvalues of $\boldsymbol{A}^{-1}$.

ellipsoid

Matrix decompositions

Matrices can be represented as products of simpler matrices. Important examples are:

QR

1. QR **decomposition**. A matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ can be written as

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R},$$

where $\boldsymbol{Q} \in \mathbb{R}^{m\times m}$ is an orthogonal, and $\boldsymbol{R} \in \mathbb{R}^{m\times n}$ an upper triangular matrix. Gram-Schmidt orthogonalisation produces such a decomposition.

LU

2. LU **decomposition**. A square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ can be written as

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}$$

where $\boldsymbol{L} \in \mathbb{R}^{n\times n}$ is a lower triangular, and $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ an upper triangular matrix. Gaussian eliminations produces such a decomposition.

Symmetric eigenvalue decomposition

3. **Symmetric eigenvalue decomposition**. A symmetric matrix $\boldsymbol{A} \in \mathcal{S}^n$ can be written as
$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top,$$
where $\boldsymbol{Q}$ is an orthogonal matrix, with the eigenvectors as rows, and $\boldsymbol{\Lambda}$ a *diagonal* matrix with the eigenvalues on the diagonal.

Cholesky

4. **Cholesky decomposition**. A positive definite symmetric matrix $\boldsymbol{A} \in \mathcal{S}^n_{++}$ can be factored as
$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^\top,$$
with $\boldsymbol{L} \in \mathbb{R}^{n\times n}$ lower-triangular and with strictly positive diagonal entries.

SVD

One of the most powerful matrix decompositions is the **singular value decomposition** (SVD). It states that any $m \times n$ matrix $\boldsymbol{A}$ can be written as
$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top,$$

singular value

where $\boldsymbol{U}$ is a $m \times m$ orthogonal matrix, $\boldsymbol{V}$ an $n \times n$ orthogonal matrix, and $\boldsymbol{\Sigma}$ is a diagonal matrix with the **singular values** $\sigma_1 \geq \cdots \geq \sigma_{\min\{m,n\}}$ on the diagonal. The singular values are the square roots of the eigenvalues of $\boldsymbol{A}^\top\boldsymbol{A}$. The singular values are related to the matrix 2-norm and Frobenius norm of $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ as follows:
$$\|\boldsymbol{A}\|_2 = \sigma_1(\boldsymbol{A}), \quad \|\boldsymbol{A}\|_F = \sqrt{\sum_{i=1}^n \sigma_i^2(\boldsymbol{A})}.$$
If $\boldsymbol{A}$ is symmetric, the singular values are the absolute values of the eigenvalues of $\boldsymbol{A}$.

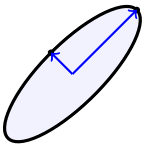
Matrix decompositions can help reduce a problem into one involving simpler (orthogonal, triangular) matrices. For example, to solve $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, one can first compute $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$, solve the simpler system of equations $\boldsymbol{Q}\boldsymbol{y} = \boldsymbol{b}$ by computing $\boldsymbol{y} = \boldsymbol{Q}^\top\boldsymbol{b}$, and then solve the triangular system $\boldsymbol{R}\boldsymbol{x} = \boldsymbol{y}$ by back-substitution.

Example 2.10. Consider the matrix $\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$. The QR decomposition is
$$\boldsymbol{Q} = \frac{1}{\sqrt{5}}\begin{pmatrix} -2 & 1 \\ 1 & 2 \end{pmatrix}, \quad \boldsymbol{R} = \frac{1}{\sqrt{5}}\begin{pmatrix} -5 & 4 \\ 0 & 3 \end{pmatrix},$$
the LU decomposition is
$$\boldsymbol{L} = \begin{pmatrix} 1 & 0 \\ -1/2 & 1 \end{pmatrix}, \quad \boldsymbol{U} = \begin{pmatrix} 2 & -1 \\ 0 & 3/2 \end{pmatrix},$$
the symmetric eigenvalue decomposition and the SVD are given by
$$\boldsymbol{\Lambda} = \boldsymbol{\Sigma} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}, \quad \boldsymbol{Q} = \boldsymbol{U} = \boldsymbol{V} = \begin{pmatrix} \cos(\pi/4) & \sin(\pi/4) \\ -\sin(\pi/4) & \cos(\pi/4) \end{pmatrix},$$

and the Cholesky decomposition is given by

$$\boldsymbol{L} = \frac{1}{\sqrt{2}} \begin{pmatrix} 2 & 0 \\ -1 & \sqrt{3} \end{pmatrix}.$$

The eigenvalue decomposition shows how to visualize the ellipse $\{\boldsymbol{Ax} : \|\boldsymbol{x}\|_2 = 1\}$:



Applying the transformation $\boldsymbol{Ax} = \boldsymbol{Q\Lambda Q}^\top \boldsymbol{x}$ corresponds to rotating the vector $\boldsymbol{x}$ clockwise by an angle of $\pi/4$, then stretching by a factor of 3 in the x-direction, and then rotating back.

3 Calculus

We write $C([a,b]) = C^0([a,b])$ for the set of continuous functions on an interval $[a,b]$, and for $k \geq 1$ we write $C^k([a,b])$ for the set of functions continuous on $[a,b]$, and whose first k derivatives $f', \dots, f^{(k)}$ exist and are continuous on (a,b). In the above definition we allow $a = -\infty$ or $b = \infty$. If $a, b \in \mathbb{R}$, then any function $f \in C([a,b])$ is bounded. The **infimum** (largest lower bound) and **supremum** (smallest upper bound) of a function f on an interval $[a,b]$ are defined as

infimum
supremum

$$\inf_{x\in[a,b]} f(x) = \max\{y \in \mathbb{R} : \forall x \in [a,b], f(x) \geq y\},$$
$$\sup_{x\in[a,b]} f(x) = \min\{y \in \mathbb{R} : \forall x \in [a,b], f(x) \leq y\}.$$

Again, we allow the "values" $-\infty$ and ∞. If the infimum is attained (i.e., there exists x^* such that $f(x^*) = \inf_{x\in[a,b]} f(x)$), then we write $\min_{x\in[a,b]} f(x)$, and similarly $\max$ if the supremum is attained. Any $f \in C([a,b])$ for $a, b \in \mathbb{R}$ attains its infimum and supremum on $[a,b]$.

Three important concepts are the **Intermediate Value Theorem**, the **Mean Value Theorem**, and the **Taylor expansion**.

Intermediate Value Theorem

Theorem (Intermediate Value Theorem). If $f \in C([a,b])$ and if y satisfies

$$\inf_{x\in[a,b]} f(x) \leq y \leq \sup_{x\in[a,b]} f(x),$$

then there exists $\xi \in [a,b]$ such that $f(\xi) = y$. In particular, the infimum and supremum are attained.

Mean Value Theorem

Theorem (Mean Value Theorem). Let $f \in C^1([a,b])$ and let $x, x_0 \in (a,b)$ with $x \neq x_0$. Then there exists a number $\xi \in (x_0, x)$ (or (x, x_0) if $x < x_0$) such that

$$f(x) = f(x_0) + f'(\xi)(x - x_0).$$

This can also be written as

$$f'(\xi) = \frac{f(x) - f(x_0)}{x - x_0}.$$

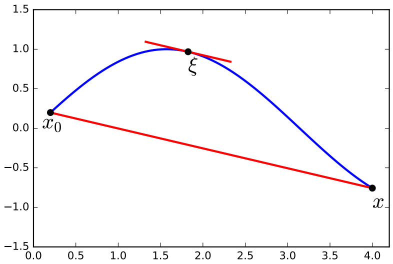


Figure 3.1: MVT: there exists a point at which the slope (derivative) is the same as that of the secant connecting $(x_0, f(x_0))$ and $(x, f(x))$.

The Mean Value Theorem is a special case of the Taylor expansion.

Taylor series

Theorem (Taylor expansion). Let $f \in C^{(n)}([a,b])$ and let $x, x_0 \in (a,b)$ with $x \neq x_0$. Then there exists $\xi \in (x, x_0)$ (or (x, x_0) if $x < x_0$) such that

$$\begin{aligned} f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2 + \cdots \\ + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)^{n+1} \end{aligned}$$

The first $(n+1)$ terms of the above sum can be seen as an approximation to the function f that becomes more accurate as n increases. The last term is known as the *truncation error* in numerical approximation.

As an example, consider the Taylor expansion of the sine function at $x_0 = 0$,

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots$$

The Taylor approximation to different orders is illustrated in the following figure.

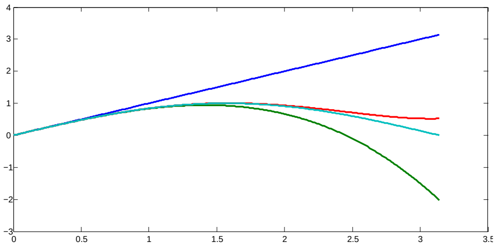


Figure 3.2: Taylor expansion of $\sin(x)$.

A special case is of the Mean Value Theorem is **Rolle's Theorem**.

Rolle's Theorem

Theorem (Rolle's Theorem)**.** Let $f \in C^1([a,b])$ with $f(a) = f(b)$. Then there exists a number $\xi \in (a,b)$ such that $f'(\xi) = 0$.

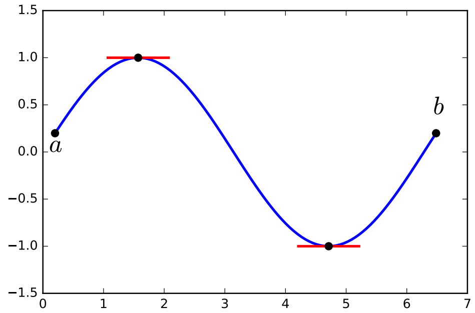


Figure 3.3: Rolle's Theorem: there exist points with “flat" slope (derivative zero).

The intuition is that if you walk across mountains and arrive at a point with the same elevation as where you started, then there must be places on the way where the slope is 0, that is, a local maximum or minimum. Of importance is also the following variant of the the Mean Value Theorem.

Integral MVT

Theorem (**Integral Mean Value Theorem**)**.** Let $f, g \in C([a,b])$ and assume that $f(x)$ does not change sign on $[a,b]$. Then there exists a $\xi \in (a,b)$ such that

$$\int_a^b f(x)g(x)\, dx = g(\xi) \int_a^b f(x)\, dx.$$

Topology

open ball

The **open ball** of radius ε around $\boldsymbol{p} \in \mathbb{R}^n$ is defined as

$$B^n(\boldsymbol{p}, \varepsilon) = \{\boldsymbol{x} : \|\boldsymbol{x} - \boldsymbol{p}\|_2 < \varepsilon\}.$$

open and closed set

closure, **boundary**, **interior**

We write $B^n := B^n(\mathbf{0}, 1)$ for the (open) **unit ball**. A subset $U \subseteq \mathbb{R}^n$ is called **open** if for every $p \in U$ there exists an $\varepsilon > 0$ such that $B(\boldsymbol{p}, \varepsilon) \subset U$. A set C is **closed** if $\overline{C} = \mathbb{R}^n \backslash C$ is open. The **closure** $\mathrm{cl}\, S$ of a set $S \subseteq \mathbb{R}^n$ is the intersection of all closed sets containing S, while the **interior** $\mathrm{int} S$ is the union of all open sets contained in S. The **boundary** of S is defined as $\mathrm{bd}\, S = \mathrm{cl}\, S \backslash \mathrm{int} S$. For example, the boundary of the open unit ball is the **unit sphere**

unit sphere, ball

$$\mathrm{bd}\, B^n = S^{n-1} = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_2 = 1\}.$$

neighbourhood

The superscript $n-1$ refers to the fact that this set is a manifold of dimension $n-1$. A **neighbourhood** N of a point $\boldsymbol{x} \in \mathbb{R}^n$ is a set such that there exists an open set U with $\boldsymbol{x} \in U \subseteq N$. An **open neighbourhood** is a neighbourhood that is open. Note that if $U_1 \subseteq \mathbb{R}^n$ and $U_2 \subseteq \mathbb{R}^m$ are open sets, then the product $U_1 \times U_2 \subseteq \mathbb{R}^{n+m}$ is also open.

Any subset $S \subseteq \mathbb{R}^n$ inherits a topological structure from $\mathbb{R}^n$, where the open sets in S are the sets of the form $U \cap S$, with $U \subseteq \mathbb{R}^n$ open. If a set S is contained in a lower-dimensional linear subspace $V \subset \mathbb{R}^n$, say, with $\dim V = k < n$, then S is always closed. However, it can be open *relative to its linear span*,

$$\operatorname{span}(S) = \left\{ \sum_{i=1}^{k} \lambda_i \boldsymbol{x}_i : \lambda_1, \dots, \lambda_k \in \mathbb{R}, \boldsymbol{x}_1, \dots, \boldsymbol{x}_k \in S \right\}.$$

relatively open, **relatively closed**

We call a set S **relatively open** or **relatively closed** if it is open or closed in the induced topology on $\operatorname{span}(S)$. Another way of defining this notion goes as follows. Let $\{\boldsymbol{b}_1, \dots, \boldsymbol{b}_k\}$ be an orthonormal basis of $\operatorname{span}(S)$, with $\boldsymbol{B} = (\boldsymbol{b}_1, \dots, \boldsymbol{b}_k)$, and consider the map

$$\varphi_{\boldsymbol{B}} \colon \mathbb{R}^k \to \operatorname{span}(S), \quad \varphi_{\boldsymbol{B}}(\boldsymbol{x}) = \sum_{i=1}^{k} x_i \boldsymbol{b}_i.$$

Then a set S is relatively open or relatively closed if the preimage

$$\varphi_{\boldsymbol{B}}^{-1}(S) = \{\boldsymbol{x} : \varphi_{\boldsymbol{B}}(\boldsymbol{x}) \in S\}$$

relative closure, **relative interior**

is open or closed in $\mathbb{R}^k$. Based on these notions, one defines the **relative closure** $\operatorname{relcl} S$ and **relative interior** $\operatorname{relint} S$ just as before.

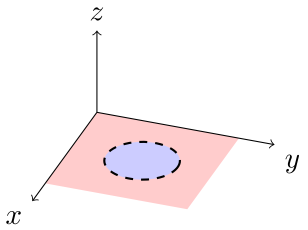


Figure 3.4: The disk without boundary on the xy-plane is relatively open, and is the relative interior of the disk with boundary.

bounded

compact

continuous

A subset $S \subseteq \mathbb{R}^n$ is **bounded** if there exists number $M > 0$ such that $\|\boldsymbol{x}\|_2 < M$ for all $\boldsymbol{x} \in S$. Invoking the equivalence of norms (3) one sees that this definition does not depend on the norm chosen. A set $K \subseteq \mathbb{R}^n$ is called **compact** if it is closed and bounded. Equivalently, every cover of K with open sets contains a finite subcover. A function $f \colon \mathbb{R}^n \to \mathbb{R}^m$ is **continuous** if for every open set $U \subset \mathbb{R}^m$,

$$f^{-1}(U) := \{\boldsymbol{x} \in \mathbb{R}^n : f(\boldsymbol{x}) \in U\}$$

is an open subset of $\mathbb{R}^n$. A function defined on a subset $S \subseteq \mathbb{R}^n$ is said to be continuous if it is continuous on the induced topology. The set of continuous functions

$f\colon S \to \mathbb{R}^m$ is denoted by $C(S,\mathbb{R}^m) = C^0(S,\mathbb{R}^m)$. If $f \in C(K,\mathbb{R})$, where K is compact, then f is bounded, and attains its infimum and supremum there: there exist $\boldsymbol{x}_*, \boldsymbol{x}^* \in K$ such that

$$\inf_{\boldsymbol{x}\in K} f(\boldsymbol{x}) = f(\boldsymbol{x}_*), \quad \sup_{\boldsymbol{x}\in K} f(\boldsymbol{x}) = f(\boldsymbol{x}^*).$$

Lipschitz

A weaker notion is that of a **Lipschitz continuous** function. A function $f\colon S \to \mathbb{R}^m$ is called Lipschitz continuous with Lipschitz constant $L > 0$, if for all $\boldsymbol{x}, \boldsymbol{y} \in S$,

$$\|f(\boldsymbol{x}) - f(\boldsymbol{y})\|_2 \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_2.$$

sequence
convergence

Notions about continuity of functions can be conveniently stated in terms of sequences. A **sequence** of points $\{\boldsymbol{x}_k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n$ (for short, $\{\boldsymbol{x}_k\}$) **converges** to $\boldsymbol{x} \in \mathbb{R}^n$ as $k \to \infty$ with respect to a norm $\|\cdot\|$, written $\boldsymbol{x}_k \to \boldsymbol{x}$, if the sequence of numbers $\|\boldsymbol{x}_k - \boldsymbol{x}\|$ converges to 0,

$$\lim_{k\to\infty} \|\boldsymbol{x}_k - \boldsymbol{x}\| = 0.$$

Formally, this means that for every $\varepsilon > 0$ there exists an index k_0, such that for all $k > k_0$, $\|\boldsymbol{x}_k - \boldsymbol{x}\| < \varepsilon$. From the equivalence of norms (3) it follows that if a sequence converges with respect to one norm, it also converges with respect to the other norms.

subsequence
limit point
Cauchy sequence
Banach space

A **subsequence** of a sequence $\{\boldsymbol{x}_k\}$ is an infinite subset of S. A **limit point** for a sequence $S = \{\boldsymbol{x}_k\}$ is a point $\boldsymbol{x}$ that is the limit of an subsequence of S. Formally, for every $\varepsilon > 0$ there exists a k_0 such that $\|\boldsymbol{x}_k - \boldsymbol{x}\| < \varepsilon$ *for some* $k > k_0$. A sequence $\{\boldsymbol{x}_k\}$ is called a **Cauchy sequence** if for every $\varepsilon > 0$ there exists an index $k_0 > 0$, such that for all $k, \ell > k_0$, $\|\boldsymbol{x}_k - \boldsymbol{x}_\ell\|_2 < \varepsilon$. The vector space $\mathbb{R}^n$ with the 2-norm (or any other norm) is a **Banach space**, which means that every Cauchy sequence contains a convergent subsequence.

All the topological notions discussed earlier have an interpretation in terms of sequences and limits:

1. A set C is closed if and only if for every sequence $\{x_k\} \subset C$, all limit points are in C;

2. The closure of a set S is the set of all limit points of sequences in S;

3. A set K is compact, if and only if every sequence of points $\{x_k\}$ in K has a limit point in K.

continuous at $\boldsymbol{x}^*$

Given a function $f\colon \Omega \to \mathbb{R}^m$, where $\Omega \subseteq \mathbb{R}^n$, and $\boldsymbol{x} \in \Omega$, then f is **continuous at $\boldsymbol{x}^*$** if

$$\lim_{\boldsymbol{x}\to\boldsymbol{x}^*} f(\boldsymbol{x}) = f(\boldsymbol{x}^*).$$

Formally, for every $\varepsilon > 0$ there exists a $\delta > 0$ such that whenever $\|\boldsymbol{x} - \boldsymbol{x}^*\| < \delta$, $\|f(\boldsymbol{x}) - f(\boldsymbol{x}^*)\| < \varepsilon$. This means that for every sequence of points $\{\boldsymbol{x}_k\}$ with $\lim_{k\to\infty} \boldsymbol{x}_k = \boldsymbol{x}^*$, the sequence $f(\boldsymbol{x}_k) \to f(\boldsymbol{x}^*)$ as $k \to \infty$ with respect to some norm on $\mathbb{R}^m$.

Differentiable functions

differentiable

A function $f\colon \mathbb{R}^n \to \mathbb{R}^m$ is called (Fréchet) **differentiable** at $\boldsymbol{x}_0 \in \mathbb{R}^n$ if there exists a linear map $\boldsymbol{J}f(\boldsymbol{x}_0)\colon \mathbb{R}^n \to \mathbb{R}^m$, such that

$$\lim_{\boldsymbol{h}\to\boldsymbol{0}} \frac{\|f(\boldsymbol{x}_0+\boldsymbol{h}) - f(\boldsymbol{x}_0) - \boldsymbol{J}f(\boldsymbol{x}_0)\boldsymbol{h}\|_2}{\|\boldsymbol{h}\|_2} = 0.$$

Jacobian

A function is differentiable on an open subset $U \subseteq \mathbb{R}^n$ if it is differentiable at every $\boldsymbol{x} \in U$. If $f(\boldsymbol{x}) = (f_1(\boldsymbol{x}), \dots, f_m(\boldsymbol{x}))^\top$ is differentiable, then all the partial derivatives exist, and $\boldsymbol{J}f(\boldsymbol{x}_0)$ is represented by the **Jacobian matrix**

$$\boldsymbol{J}f(\boldsymbol{x}_0) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix},$$

continuously differentiable

gradient

where the partial derivatives are evaluated at $\boldsymbol{x}_0$. If all the partial derivatives exist and are continuous in a neighbourhood of $\boldsymbol{x}_0$ (called **continuously differentiable**), then f is differentiable at $\boldsymbol{x}_0$.

If $m = 1$, then $J(\boldsymbol{x}_0)$ is the transpose of the **gradient** $\nabla f(\boldsymbol{x}_0)$ of f at $\boldsymbol{x}_0$,

$$\nabla f(\boldsymbol{x}_0) = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n}\right)^\top.$$

The gradient points in the direction in which f increases the most.

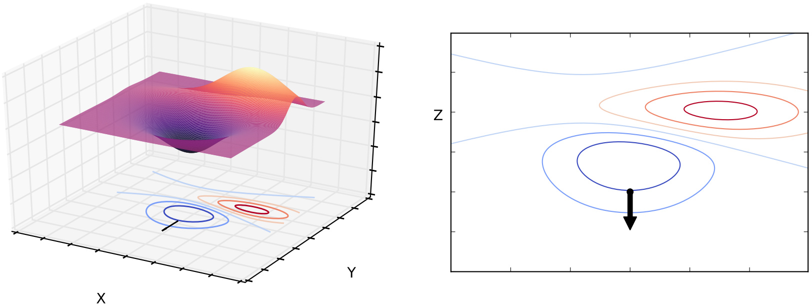


Figure 3.5: A surface, level sets, and the gradient

level set

A convenient way to visualise a function $f\colon \mathbb{R}^2 \to \mathbb{R}$ is through **level sets** $\{\boldsymbol{x} \in \mathbb{R}^2 : f(\boldsymbol{x}) = c\}$. For each $c \in \mathbb{R}$, such a level set defines a curve in $\mathbb{R}^2$, the curve on which the function value does not change. The gradient is always orthogonal to the level set, pointing in the direction in which f increases the most (see Figure 3.5).

Hessian

If the gradient, considered as a map $\mathbb{R}^n \to \mathbb{R}^n$, is itself differentiable at $\boldsymbol{x}_0$, then the Jacobian matrix of the gradient is called the **Hessian matrix**,

$$\nabla^2 f(\boldsymbol{x}_0) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{pmatrix}$$

Since

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i},$$

the Hessian is a symmetric matrix.

directional derivative

The **directional derivative** $D_{\boldsymbol{x}_0} f(\boldsymbol{x}_0)$ of a function $f\colon \mathbb{R}^n \to \mathbb{R}^m$ in direction $\boldsymbol{v} \in \mathbb{R}^n$ at $\boldsymbol{x}_0$ is defined as

$$D_{\boldsymbol{v}} f(\boldsymbol{x}_0) = \lim_{h \to 0} \frac{f(\boldsymbol{x}_0 + h\boldsymbol{v}) - f(\boldsymbol{x}_0)}{h}.$$

In the special case where $\boldsymbol{v} = \boldsymbol{e}_i$, we obtain the partial derivative with respect to x_i,

$$\frac{\partial f}{\partial x_i}(\boldsymbol{x}_0) = D_{\boldsymbol{e}_i} f(\boldsymbol{x}_0).$$

If $f\colon \mathbb{R}^n \to \mathbb{R}$ is differentiable with continuous derivative near $\boldsymbol{x}_0$, then

$$D_{\boldsymbol{v}} f(\boldsymbol{x}_0) = \nabla f(\boldsymbol{x}_0)^\top \boldsymbol{v} = \langle \nabla f(\boldsymbol{x}_0), \boldsymbol{v} \rangle.$$

If $f\colon \mathbb{R}^n \to \mathbb{R}^m$ and $g\colon \mathbb{R}^m \to \mathbb{R}^p$ are differentiable in a neighbourhood of $\boldsymbol{x}_0 \in \mathbb{R}^n$ and $f(\boldsymbol{x}_0) \in \mathbb{R}^m$, respectively, then the composition $h = g \circ f\colon \mathbb{R}^n \to \mathbb{R}^p$ is continuously differentiable in a neighbourhood of $\boldsymbol{x}_0$, and the Jacobian matrix is defined by the **chain rule**:

chain rule

$$Jh(\boldsymbol{x}_0) = Jg(f(\boldsymbol{x}_0))Jf(\boldsymbol{x}_0).$$

If $n = 1$, then $f\colon \mathbb{R} \to \mathbb{R}^n$ is called a **curve**, and we write

curve

$$Jf = \frac{\mathrm{d}f}{\mathrm{d}t} = \dot{f} = (\dot{f}_1, \dots, \dot{f}_n)^\top \in \mathbb{R}^n$$

for the derivative of the curve. If $\dot{f}(t_0) = \boldsymbol{v} \in \mathbb{R}^n$ and $g\colon \mathbb{R}^n \to \mathbb{R}$, then by the chain rule, the derivative of $g \circ f\colon \mathbb{R} \to \mathbb{R}$ is the directional derivative of g in the direction $\boldsymbol{v}$,

$$\frac{\mathrm{d}g \circ f(t_0)}{\mathrm{d}t} = \langle \nabla g(f(t_0)), \boldsymbol{v} \rangle.$$

Before going on to deal with higher derivative, we state the generalisation of the Mean Value Theorem to higher dimensions.

Multivariate MVT

Theorem (Multivariate Mean Value Theorem)**.** Let $f \in C^1(U)$ for an open set U with $\boldsymbol{x}_0, \boldsymbol{x} \in U$, $\boldsymbol{x} \neq \boldsymbol{x}_0$. Then there exists $t \in (0,1)$ such that

$$f(\boldsymbol{x}) - f(\boldsymbol{x}_0) = \langle \nabla f(t\boldsymbol{x} + (1-t)\boldsymbol{x}_0), \boldsymbol{x} - \boldsymbol{x}_0 \rangle.$$

Note that $t\boldsymbol{x} + (1-t)\boldsymbol{x}_0$ parametrises the line segment connecting $\boldsymbol{x}$ and $\boldsymbol{x}_0$.

For a tuple of natural number $\alpha = (\alpha_1, \dots, \alpha_n)$, set $|\alpha| = \sum_{i=1}^n \alpha_i$, and define the higher order partial derivative

$$D^\alpha f(\boldsymbol{x}) = \frac{\partial^{|\alpha|} f(\boldsymbol{x})}{\partial^{\alpha_1} x_1 \cdots \partial^{\alpha_n} x_n}.$$

For a set $S \subseteq \mathbb{R}^n$, denote by $C^k(S, \mathbb{R}^m)$ the set of functions f such that all partial derivatives $D^\alpha f$ with $|\alpha| \leq k$ exists and are continuous on $\mathrm{int} S$. If $m = 1$, we write $C^k(S) := C^k(S, \mathbb{R})$.

Define, for a vector $\boldsymbol{x}$ and multi-index α,

$$\boldsymbol{x}^\alpha := x_1^{\alpha_1} \cdots x_n^{\alpha_n}.$$

Taylor series

We then have the **Taylor expansion** around a point $\boldsymbol{x}_0$,

$$f(\boldsymbol{x}) = \sum_{|\alpha| \leq k} D^\alpha f(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0)^\alpha + \sum_{|\alpha| = k} r_\alpha(\boldsymbol{x})(\boldsymbol{x} - \boldsymbol{x}_0)^\alpha,$$

with $r_\alpha(\boldsymbol{x}) \to 0$ as $\boldsymbol{x} \to \boldsymbol{x}_0$.

Lagrange multipliers

If a differentiable function $f(\boldsymbol{x})$ has a local minimum or maximum at a point $\boldsymbol{x}$, then this point satisfies $\nabla f(\boldsymbol{x}) = 0$, that is, it is a critical point. The **Lagrange multiplier theorem** says something about local extrema under certain constraints.

Theorem (Lagrange multipliers). Let $\boldsymbol{x}^*$ be maximum of $f(\boldsymbol{x})$ under the constraint $g(\boldsymbol{x}) = c$ (that is, a maximum among all points $\boldsymbol{x}$ such that $g(\boldsymbol{x}) = c$). Then there exist a $\lambda \in \mathbb{R}$ such that

$$\nabla f(\boldsymbol{x}^*) = \lambda \nabla g(\boldsymbol{x}^*).$$

Lagrangian

The **Lagrangian** of a function $f(\boldsymbol{x})$ with constraint $g(\boldsymbol{x}) = c$ is the function $\Lambda\colon \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ defined by

$$\Lambda(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) - \lambda(g(\boldsymbol{x}) - c).$$

The Lagrange multiplier theorem then says that if $\boldsymbol{x}^*$ is a maximum point of $f(\boldsymbol{x})$ under the constraint $g(\boldsymbol{x}) = c$, then there exists $\lambda \in \mathbb{R}$ such that the pair $(\boldsymbol{x}^*, \lambda)$ is a critical point of the Lagrangian $\Lambda(\boldsymbol{x}, \lambda)$.

Implicit Function Theorem

The **Implicit Function Theorem** is one of the most important results in analysis, and underlies much of differential geometry and physics. Let $F\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ by a function that is continuously differentiable in a neighbourhood of a point $(\boldsymbol{x}_0, \boldsymbol{y}_0)$, with $\boldsymbol{x}_0 \in \mathbb{R}^n$ and $\boldsymbol{y}_0 \in \mathbb{R}^m$. The Jacobian $J_{\boldsymbol{x}}(\boldsymbol{x}_0, \boldsymbol{y}_0)$ with respect to the first set of n coordinates consists of the first n columns of the Jacobian matrix,

$$J_{\boldsymbol{x}}(\boldsymbol{x}_0, \boldsymbol{y}_0) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\boldsymbol{x}_0, \boldsymbol{y}_0) & \cdots & \frac{\partial f_1}{\partial x_n}(\boldsymbol{x}_0, \boldsymbol{y}_0) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\boldsymbol{x}_0, \boldsymbol{y}_0) & \cdots & \frac{\partial f_m}{\partial x_n}(\boldsymbol{x}_0, \boldsymbol{y}_0) \end{pmatrix}$$

The interpretation is that we consider f as a function in only the first set of coordinates, with the remaining ones (denoted by $\boldsymbol{y}$) considered as parameters.

Theorem (Implicit Function Theorem). Let $f\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ be k times continuously differentiable in an open neighbourhood of $(\boldsymbol{x}_0, \boldsymbol{y}_0) \in \mathbb{R}^n \times \mathbb{R}^m$, and assume that $f(\boldsymbol{x}_0, \boldsymbol{y}_0) = \boldsymbol{0}$. Assume further that the Jacobian $J_{\boldsymbol{x}} f(\boldsymbol{x}_0, \boldsymbol{y}_0) \in \mathbb{R}^{n \times n}$ in the first n coordinates is *non-singular* at $(\boldsymbol{x}_0, \boldsymbol{y}_0)$. Then there exists an open neighbourhood $\boldsymbol{y}_0 \in U_{\boldsymbol{y}} \subseteq \mathbb{R}^m$, and a function $h \in C^k(U_{\boldsymbol{y}}, \mathbb{R}^n)$ such that

- $h(\boldsymbol{y}_0) = \boldsymbol{x}_0$,
- $f(h(\boldsymbol{y}), \boldsymbol{y}) = 0$ for $\boldsymbol{y} \in U_{\boldsymbol{y}}$.

Moreover, the Jacobian of h is given by

$$Jh(\boldsymbol{y}) = -Jf_{\boldsymbol{y}}(h(\boldsymbol{y}), \boldsymbol{y})(J_{\boldsymbol{x}} f(h(\boldsymbol{y}), \boldsymbol{y}))^{-1}$$

for all $\boldsymbol{y} \in U_{\boldsymbol{y}}$.

Example 3.1. Let $f(x, y) = x^2 + y^2 - 1$ and $(x_0, y_0) = (1, 0)$. The Jacobian of f in the first coordinate is

$$\frac{\partial f}{\partial x}(1, 0) = 2 \neq 0,$$

which is non-singular. Choosing the neighbourhood $U_y = (-1, 1)$, the open interval between -1 and 1, we get the function $h \colon U_y \to \mathbb{R}$ as

$$h(y) = \sqrt{1 - y^2}.$$

This function is defined and differentiable on $(-1, 1)$, and satisfies

$$f(h(y), y) = h(y)^2 + y^2 - 1 = (1 - y^2) + y^2 - 1 = 0, \quad y \in U_y.$$

The derivative of h can be computed using the chain rule:

$$\frac{\mathrm{d}f(h(y), y)}{\mathrm{d}y} = \frac{\partial f}{\partial x}(h(y), y)\frac{\mathrm{d}h}{\mathrm{d}y}(y) + \frac{\partial f}{\partial y}(h(y), y)\frac{\mathrm{d}y}{\mathrm{d}y}$$

from which we get

$$\frac{\mathrm{d}h}{\mathrm{d}y}(y) = -\frac{\partial f}{\partial y}(h(y), y)\left(\frac{\partial f}{\partial x}(h(y), y)\right)^{-1} = -y(1 - y^2)^{-3/2}.$$

In this example, the implicit function theorem just gives the usual way of parametrising part of the circle.

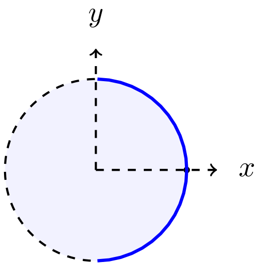


Figure 3.6: The blue arc is parametrised by $h(y) = \sqrt{1 - y^2}$ for $y \in (-1, 1)$.

4 Finite precision arithmetic

In practical applications, one often cannot simply plug numbers into formulae and get all the exact results. Most numerical data also requires an infinite amount of storage (just try to store π on a computer!), but a piece of paper or a computer only has limited space. These are some of the reasons that lead us to work with **approximations**.

approximation

Measuring errors

To measure the quality of approximations, we use the concept of **relative error**. Given a quantity x and a computed approximation $\hat{x}$, the **absolute error** is given by

absolute and relative error

$$E_{\text{abs}}(\hat{x}) = |x - \hat{x}|,$$

while the *relative error* is given as

$$E_{\text{rel}}(\hat{x}) = \frac{|x - \hat{x}|}{|x|}.$$

The benefit of working with relative errors is that they are scale invariant. Absolute errors can be meaningless at times: for example, an error of one hour is irrelevant when estimating the age of Stan the Tyrannosaurus Rex at Manchester Museum, but it is crucial when determining the time of a lecture. That is because in the former one hour corresponds to a relative error is of the order 10^{-11}, while in the latter it is of the order 1.

Floating point and significant figures

floating-point

The established way of representing real numbers on computers is using **floating-point arithmetic**. In the double precision version of the IEEE standard for floating-point arithmetic, a number is represented using 64 bits, where a bit is either 1 or 0. A number is written

$$x = \pm f \times 2^{e},$$

where f is a fraction in $[0, 1]$, represented using 52 bits, and e is the exponent, using 11 bits, and one bit is for the sign. There are largest possible numbers, and there are gaps between representable numbers. The largest and smallest numbers representable in this form are of the order of $\pm 10^{308}$, enough for most practical purposes. A bigger concern are the gaps, which means that the results of many computations almost always have to be rounded to the closest floating-point number.

significant figure

When going through calculations without using a computer, we usually use the terminology of **significant figures** (s.f.) and work with 4 significant figures in base 10. For example, in base 10, $\sqrt{3}$ equals 1.732 to 4 significant figures. To count the number of significant figures in a given number, start with the first non-zero digit from the left and, moving to the right, count all the digits thereafter, counting final zeros if they are to the right of the decimal point. For example, 1.2048, 12.040, 0.012048, 0.0012040 and 1204.0 all have 5 significant figures (s.f.). In rounding or truncation of a number to n s.f., the original is replaced by the closest number with n s.f. An approximation $\hat{x}$ of a number x is said to be **correct to n significant figures** if both $\hat{x}$ and x round to the same n s.f. number.

Remark 4.1. Note that final zeros to left of the decimal point may or may not be significant: the number 1204000 has a least 4 significant figures, but without any more information there is no way of knowing whether or not any more figures are significant. When 1203970 is rounded to 5 significant figures to give 1204000, an explanation that this has 5 significant figures is required. This could be made clear by writing it in scientific notation: 1.2040×10^6. In some cases we also have to agree whether to round up or round down: for example, 1.25 could equal 1.2 or 1.3 to two significant figures. If we agree on rounding up, then to say that $a = 1.2048$ to 5 s.f. means that the exact value of a satisfies $1.20475 \leq a < 1.40485$.

Example 4.2. Suppose we want to find the solution to the quadratic equation

$$ax^2 + bx + c = 0.$$

The two solutions to this problem are given by

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}, \quad x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}. \tag{4.1}$$

In principle, to find x_1 and x_2 one only needs to evaluate the expressions for given a, b, c. Assume, however, that we are only allowed to compute to four significant figures, and consider the particular equation

$$x^2 + 39.7x + 0.13 = 0.$$

Using the formula 4.1, we have, always rounding to four significant figures,

$$a = 1, b = 39.7, c = 0.13,$$

$$b^2 = 1576.09 = 1576 \text{ (to 4 s.f.)}\ , 4ac = 0.52 \text{ (to 4 s.f.)},$$

$$b^2 - 4ac = 1575.48 = 1575 \text{ (to 4 s.f.)}\ , \sqrt{b^2 - 4ac} = 39.69.$$

Hence, the computed solutions (to 4 significant figures) are given by

$$\overline{x}_1 = -0.005,\ \overline{x}_2 = -39.69$$

The exact solutions, however, are

$$x_1 = -0.0032748...,\ x_2 = -39.6907...$$

The solution x_1 is completely wrong, at least if we look at the relative error:

$$\frac{|\overline{x}_1 - x_1|}{|x_1|} = 0.5268.$$

While the accuracy can be increased by increasing the number of significant figures during the calculation, such effects happen all the time in scientific computing and the possibility of such effects has to be taken into account when designing numerical algorithms.

By analysing what causes the error it is sometimes possible to modify the method of calculation in order to improve the result. In the present example, the problems are being caused by the fact that $b \approx \sqrt{b^2 - 4ac}$, and therefore

$$\frac{-b + \sqrt{b^2 - 4ac}}{2a} = \frac{-39.7 + 39.69}{2}$$

causes what is called "catastrophic cancellation". A way out is provided by the observation that the two solutions are related by

$$x_1 \cdot x_2 = \frac{c}{a}. \tag{4.2}$$

When $b > 0$, the calculation of x_2 according to (4.1) shouldn't cause any problems, in our case we get -39.69 to four significant figures. We can then use (4.2) to derive $\overline{x}_1 = c/(a\overline{x}_2) = -0.00327$.

There are other potential sources of error besides those introduced by rounding operations.

1. Overflow
2. Errors in the model
3. Human or measurements errors
4. Truncation or approximation errors

The first is rarely an issue, as we can represent numbers of order 10^{308} on a computer. The second two are important factors that need to be addressed when working on real-world problems. The fourth has to do with the fact that many computations are done approximately rather than exactly. For computing the exponential, for example, we might use a method that gives the approximation

$$e^x \approx 1 + x + \frac{x^2}{2}.$$

As it turns out, many optimization problems work with approximations of the functions of interest, and the solution found is only an approximation to the "true" solution of the problem. Quantifying the quality of such an approximation is an important aspect in the design and analysis of optimization algorithms.