Appendix A

Mathematical background

A.1 Multivariable analysis

This section reviews basic concepts of multivariable analysis, including concepts from linear algebra and calculus. We focus on real vector spaces, since those suffice for the purposes of this book, but most concepts extend *mutatis mutandis* to complex spaces. Readers unfamiliar with the topics in this appendix are should get acquainted with the main concepts from one of the many textbook available on these subjects . Good references include the texts of Strang (2009) on linear algebra and Carothers, 2000 on calculus.

More advanced treatments can be found in the books of (Rudin, 1976) or Halmos (1987). The presentation herein roughly follows Mikusiński and M. Taylor (2002).

A.1.1 Vectors and matrices

In the context of this section, vectors refer to elements of a *vector space*, and should not be taken to mean an element of $\mathbb{R}^N$—although $\mathbb{R}^N$ with the usual addition and scalar multiplication is a vector space and its elements are, in that sense, vectors.

Vector spaces

A *vector space* is a set of elements $\mathcal{X}$ endowed with two operations—scalar multiplication and addition—such that

(i) For any $x, y \in \mathcal{X}$, $x + y \in \mathcal{X}$;

(ii) For any $x \in \mathcal{X}$, $\alpha x \in \mathcal{X}$, with α a (real-valued) scalar;

(iii) For any $x, y \in \mathcal{X}$, $x + y = y + x$;

(iv) For any $x, y, z \in \mathcal{X}$, $(x + y) + z = x + (y + z)$.

(v) There is an element $0 \in \mathcal{X}$ such that, for any $x \in \mathcal{X}$, $x + 0 = 0 + x = x$;

(vi) For every $x \in \mathcal{X}$ there is an element $-x \in \mathcal{X}$ such that $x+(-x) = (-x)+x = 0$;

(vii) For every $x \in \mathcal{X}$ and any scalars $\alpha, \beta \in \mathbb{R}$, $\alpha(\beta x) = (\alpha\beta)x$;

(viii) For every $x \in \mathcal{X}$, $1x = x$;

(ix) For any $x, y \in \mathcal{X}$ and any scalar $\alpha \in \mathbb{R}$, $\alpha(x + y) = \alpha x + \alpha y$;

(x) For any $x \in \mathcal{X}$ and any scalars $\alpha, \beta \in \mathbb{R}$, $(\alpha + \beta)x = \alpha x + \beta x$;

We refer to the elements of $\mathcal{X}$ as *vectors*. Notable examples of vector spaces used throughout the book include $\mathbb{R}^N$, the space of $N \times N$ matrices, or the space of square integrable real-valued functions.

Given a vector space $\mathcal{X}$, a *subspace* is any subset of $\mathcal{X}$ (endowed with the same operations) such that properties (i) and (ii) hold. The *span* of a set $U \subset \mathcal{X}$, denoted as span(U), is the intersection of all subspaces that contain U. We say that a subspace $\mathcal{Y} \subset \mathcal{X}$ is *spanned by* U if span$(U) = \mathcal{Y}$.

Given two subspaces $\mathcal{Y}, \mathcal{Z} \subset \mathcal{X}$, the set $\mathcal{Y} \oplus \mathcal{Z}$ is called the *direct sum* of $\mathcal{Y}$ and $\mathcal{Z}$ and is defined as

$$\mathcal{Y} \oplus \mathcal{Z} = \{x \in \mathcal{X} \mid x = y + z, y \in \mathcal{Y}, z \in \mathcal{Z}\}.$$

The direct sum of two subspaces is also a subspace, with dimension $\dim(\mathcal{Y}) + \dim(\mathcal{Z})$.

Linear independence

One of the key properties of vector spaces is that we can combine its elements to get other elements. Given a vector space $\mathcal{X}$ and a set $U = \{u_1, \ldots, u_N\} \subset \mathcal{X}$, a *linear combination* of the vectors in U is any $x \in \mathcal{X}$ of the form

$$x = \sum_{n=1}^{N} \lambda_n u_n, \tag{A.1}$$

with $\lambda_1, \ldots, \lambda_N \in \mathbb{R}$. The scalars $\lambda_1, \ldots, \lambda_N$ are called the *coefficients* of the linear combination. The span of U is, precisely, the set of *all* linear combinations of the vectors in U. The linear combination (A.1) is called a *convex combination* if

$$\sum_{n=1}^{N} \lambda_n = 1.$$

The *convex hull* of U, conv(U), is the set of all convex combinations of the vectors in U.

A second fundamental concept is that of *linear independence*. A set of vectors $\{u_1, \ldots, u_N\} \subset \mathcal{X}$ is *linearly independent* when

$$\sum_{n=1}^{N} \lambda_n u_n = 0$$

if and only if $\lambda_1 = \lambda_2 = \ldots = \lambda_n = 0$. Linear independence allows us to identify the smallest set of vectors that spans a (sub)space. Formally, given a vector space $\mathcal{X}$, a set U is a *basis* for $\mathcal{X}$ if:

- The vectors in U are linearly independent;
- U spans $\mathcal{X}$.

If $\mathcal{X}$ is a vector space and U a basis for $\mathcal{X}$, the elements of U are called *basis vectors*. Any $x \in \mathcal{X}$ can be written as a (unique) linear combination of the basis vectors,

$$x = \sum_{n=1}^{N} x_n u_n.$$

The values $x_1, \ldots, x_N$ are the *coordinates* of x in the basis U. Given a basis for $\mathcal{X}$, any vector is uniquely identified by its coordinates $x_1, \ldots, x_N$. Additionally, if U and V are two bases for $\mathcal{X}$, then both basis have the same number of vectors. This number is called the *dimension* of the vector space and is denoted as $\dim(\mathcal{X})$.

As noted above, given a basis U for $\mathcal{X}$, any element $x \in \mathcal{X}$ can be represented uniquely as a linear combination of the elements of U. Therefore, we can refer to a vector $x \in \mathcal{X}$ using its representation in the basis U as a point in $\mathbb{R}^N$,

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}.$$

We henceforth write x to refer to the element of $\mathcal{X}$ and reserve the boldface symbol $\boldsymbol{x}$ to denote the representation of x in some (implicit or explicit) basis.

Given two basis $U = \{u_1, \ldots, u_N\}$ and $V = \{v_1, \ldots, v_N\}$ for $\mathcal{X}$, a vector $x \in \mathcal{X}$ can be written, equivalently, as

$$x = \sum_{n=1}^{N} {}^{u}x_n u_n \qquad \text{or} \qquad x = \sum_{n=1}^{N} {}^{v}x_n v_n,$$

where ${}^{i}x_j$ denotes the jth component of x in basis i. Let ${}^{u}\boldsymbol{x}$ and ${}^{v}\boldsymbol{x}$ denote the vector representations of x in basis U and V, respectively. Similarly, let ${}^{v}\boldsymbol{u}_1, \ldots, {}^{v}\boldsymbol{u}_N$ denote the vector representation of the vectors in U in the basis V. In particular, let ${}^{v}u_{n,m}$ denote the nth coordinate of ${}^{v}\boldsymbol{u}_m$. Then,

$${}^{v}\boldsymbol{x} = \sum_{m=1}^{N} {}^{u}x_m \, {}^{v}\boldsymbol{u}_m = \sum_{m=1}^{N} {}^{u}x_m \sum_{n=1}^{N} {}^{v}u_{n,m} v_n = \sum_{n=1}^{N} \left[\sum_{m=1}^{N} {}^{u}x_m \, {}^{v}u_{n,m} \right] v_n.$$

Letting ${}^{v}\boldsymbol{M}_u$ denote the matrix with mth column given by ${}^{v}\boldsymbol{u}_m$, we finally get

$${}^{v}\boldsymbol{x} = {}^{v}\boldsymbol{M}_u \, {}^{u}\boldsymbol{x}.$$

The matrix ${}^{v}\boldsymbol{M}_u$, containing the representation of the vectors in U in the basis V can be used to obtain the representation of x in the basis V from the representation of x in the basis U. For this reason, it is called a *change of basis matrix.*

Norms and inner products

Given a vector space $\mathcal{X}$, a *norm* is a mapping $\|\cdot\|$ from $\mathcal{X}$ to the real numbers such that:

- It is *non-negative* and $\|x\| = 0$ if only if $x = 0$;
- It is *homogeneous*, i.e., $\|\alpha x\| = |\alpha| \, \|x\|$, for all $x \in \mathcal{X}$ and all $\alpha \in \mathbb{R}$;
- It verifies the *triangle inequality*, i.e., $\|x + y\| \leq \|x\| + \|y\|$, for all $x, y \in \mathcal{X}$.

Intuitively, one can interpret the norm of a vector as its "length". Commonly used norms include

- The *absolute value norm* in $\mathbb{R}$, given by
$$\|x\| = |x| ,$$
for $x \in \mathbb{R}$.
- The *Euclidean norm* in $\mathbb{R}^N$, given by
$$\|\boldsymbol{x}\|_2 = \sqrt{\sum_{n=1}^{N} x_n^2},$$
for $\boldsymbol{x} \in \mathbb{R}^N$.
- The *supremum norm* in $\mathbb{R}^N$, given by
$$\|\boldsymbol{x}\|_\infty = \max \left\{ |x_n| , n = 1, \ldots, N \right\} ,$$
for $\boldsymbol{x} \in \mathbb{R}^N$.
- More generally, the p-norm in $\mathbb{R}^N$, given by
$$\|\boldsymbol{x}\|_p = \left(\sum_{n=1}^{N} |x_n|^p \right)^{\frac{1}{p}} ,$$
for $\boldsymbol{x} \in \mathbb{R}^N$. The Euclidean norm corresponds to the p-norm with $p = 2$, and the supremum norm can be seen as the limit of the p-norm when $p \to \infty$.
- The p-norm can be extended to spaces of (infinite) sequences, in each case yielding a normed vector space usually denoted ℓ_p. Notable cases include ℓ_1 (the space of absolutely summable sequences), ℓ_2 (the space of square summable sequences) and ℓ_∞ (the space of bounded sequences).
- Similarly, the p-norm can also be extended to spaces of functions, in each case yielding a normed vector space usually denoted L_p. Notable cases include L_2 (the space of square integrable functions) and L_∞ (the space of essentially bounded functions).

A vector space endowed with a norm is called a *normed vector space*. A vector $x \in \mathcal{X}$ such that $\|x\| = 1$ is called a *unit vector* or a *normalized vector*. Any non-zero vector $x \in \mathcal{X}$ can be normalized as

$$\bar{x} = \frac{1}{\|x\|} x.$$

Given a vector space $\mathcal{X}$, an *inner product* is a mapping $\langle \cdot, \cdot \rangle$ from $\mathcal{X} \times \mathcal{X}$ to the real numbers such that

- It is *positive semi-definite*, i.e., $\langle x, y \rangle \geq 0$. Moreover, $\langle x, x \rangle = 0$ if and only if $x = 0$;
- It is *symmetric*, i.e., $\langle x, y \rangle = \langle y, x \rangle$, for all $x, y \in \mathcal{X}$;
- It is *bilinear*, i.e.,

  $$\begin{aligned} \langle \alpha x + \beta y, z \rangle &= \alpha \langle x, z \rangle + \beta \langle y, z \rangle \\ \langle x, \alpha y + \beta z \rangle &= \alpha \langle x, y \rangle + \beta \langle x, z \rangle , \end{aligned}$$

  for all $x, y, z \in \mathcal{X}$ and all $\alpha, \beta \in \mathbb{R}$.

A vector space endowed with an inner product is called a *inner product space*. An inner product space is also a normed vector space, since the inner product induces the norm

$$\|x\| = \langle x, x \rangle^{\frac{1}{2}} ,$$

and the properties of the norm easily follow from those of the inner product. If $U = \{u_1, \ldots, u_N\}$ is a basis for $\mathcal{X}$, then

$$\langle x, y \rangle = \left\langle \sum_{m=1}^{N} x_m u_m, \sum_{n=1}^{N} y_n u_n \right\rangle = \sum_{m=1}^{N} \sum_{n=1}^{N} x_m y_n \langle u_m, u_n \rangle .$$

This can be written in matrix form as

$$\langle x, y \rangle = \begin{bmatrix} x_1 & \ldots & x_N \end{bmatrix} \underbrace{\begin{bmatrix} \langle u_1, u_1 \rangle & \ldots & \langle u_1, u_N \rangle \\ \vdots & \ddots & \vdots \\ \langle u_N, u_1 \rangle & \ldots & \langle u_N, u_N \rangle \end{bmatrix}}_{\boldsymbol{M}} \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

The matrix $\boldsymbol{M}$ represents the inner product in the basis U and is thus called the *inner product matrix*.

Orthogonality

Given an inner product space $\mathcal{X}$, it can be shown that

$$|\langle x, y \rangle| \leq \|x\| \, \|y\| . \tag{A.2}$$

The inequality (A.2) is known as the *Cauchy-Schwartz inequality* and implies that, for any non-zero vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$,

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

We thus define the *angle* between x and y as

$$\theta = \cos^{-1} \frac{\langle x, y \rangle}{\|x\| \, \|y\|},$$

and we say that two vectors x, y are *orthogonal* if

$$\langle x, y \rangle = 0.$$

If U is a basis for $\mathcal{X}$ such that $\langle u_m, u_n \rangle = \mathbb{I}\left[m = n\right]$ for all $u_m, u_n \in U$, the basis is called *orthonormal.* In this case, the inner product matrix reduces to the identity $\boldsymbol{I}$, and the inner product between two vectors x and y reduces to the usual expression

$$\langle x, y \rangle = \sum_{n=1}^{N} x_n y_n,$$

where $x_1, \ldots, x_N$ and $y_1, \ldots, y_N$ are the coordinates of x and y in the basis U.

Let $\mathcal{X}$ denote an inner product space, and $\mathcal{Y} \subset \mathcal{X}$ a subspace. The set

$$\mathcal{Y}^{\perp} \overset{\text{def}}{=} \{x \in \mathcal{X} \mid \langle x, y \rangle = 0, y \in \mathcal{Y}\}$$

is called the *orthogonal complement of* $\mathcal{Y}$ and is also a subspace of $\mathcal{X}$. Every vector in $\mathcal{Y}$ is orthogonal to every vector in $\mathcal{Y}^{\perp}$. Additionally, $\mathcal{X} = \mathcal{Y} \oplus \mathcal{Y}^{\perp}$, i.e., any vector $x \in \mathcal{X}$ can be written as

$$x = \mathbf{Proj}_{\mathcal{Y}}(x) + \mathbf{Proj}_{\mathcal{Y}^{\perp}}(x),$$

where $\mathbf{Proj}_{\mathcal{Y}}(x) \in \mathcal{Y}$ and is called the *orthogonal projection of* x *onto* $\mathcal{Y}$. Similarly, $\mathbf{Proj}_{\mathcal{Y}^{\perp}}(x) \in \mathcal{Y}^{\perp}$ and is called the orthogonal projection of x onto $\mathcal{Y}^{\perp}$. Note that the projection of x onto $\mathcal{Y}$ is orthogonal to all vectors in $\mathcal{Y}^{\perp}$ and, conversely, the projection of x onto $\mathcal{Y}^{\perp}$ is orthogonal to all vectors in $\mathcal{Y}$.

Let U be an orthogonal basis for $\mathcal{Y}$. Since $\mathbf{Proj}_{\mathcal{Y}}(x) \in \mathcal{Y}$, we can write

$$\mathbf{Proj}_{\mathcal{Y}}(x) = \sum_{m=1}^{M} c_m u_m,$$

for some scalars $c_1, \ldots, c_M$. Hence,

$$\langle u_i, x \rangle = c_i \left\|u_i\right\|^2 + \underbrace{\left\langle u_i, \mathbf{Proj}_{\mathcal{Y}^{\perp}}(x) \right\rangle}_{=0}.$$

Solving for c_i yields the expression of the ith coordinate of $\mathbf{Proj}_{\mathcal{Y}}(x)$:

$$c_i = \frac{\langle u_i, x \rangle}{\left\|u_i\right\|^2}.$$

More generally, let ${}^v\boldsymbol{x}$ denote the vector representation of $x \in \mathcal{X}$ in some basis V, and let $\{u_n, n = 1, \ldots, M\}$ be an arbitrary basis for $\mathcal{Y}$, where we write ${}^v\boldsymbol{u}_m$ to denote the vector representation of u_m in V. If $\boldsymbol{M}$ is the matrix with mth column given by ${}^v\boldsymbol{u}_m$, then

$$\mathbf{Proj}_{\mathcal{Y}}(x) = \boldsymbol{M}(\boldsymbol{M}^\top \boldsymbol{M})^{-1}\boldsymbol{M}^\top {}^v\boldsymbol{x}.$$

We conclude with two remarks. First, the projection of x onto $\mathcal{Y}$ is the element of $\mathcal{Y}$ that is closest to x, i.e.,

$$\mathbf{Proj}_{\mathcal{Y}}(x) = \operatorname*{argmin}_{y \in \mathcal{Y}} \|x - y\|.$$

This is particularly useful when approximately solving systems of linear equations of the form

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \tag{A.3}$$

where $\boldsymbol{A}$ is a $M \times N$ matrix and $\boldsymbol{x} \in \mathbb{R}^N$ and $\boldsymbol{b} \in \mathbb{R}^M$. Solving (5.12) consists in determining vector $\boldsymbol{x}^*$ such that

$$\boldsymbol{x}^* = \operatorname*{argmin}_{\boldsymbol{x}} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|$$

where the norm is induced by the usual inner product in $\mathbb{R}^M$. Equivalently, solving (A.3) consists in determining the linear combination of the columns of $\boldsymbol{A}$ that is closest to $\boldsymbol{b}$. That is, precisely, the orthogonal projection of $\boldsymbol{b}$ onto the subspace spanned by the columns of $\boldsymbol{A}$. Therefore,

$$\boldsymbol{A}\boldsymbol{x}^* = \boldsymbol{A}(\boldsymbol{A}^\top \boldsymbol{A})^{-1}\boldsymbol{A}^\top \boldsymbol{b}$$

or, equivalently,

$$\boldsymbol{x}^* = (\boldsymbol{A}^\top \boldsymbol{A})^{-1}\boldsymbol{A}^\top \boldsymbol{b}.$$

For this reason, the matrix $(\boldsymbol{A}^\top \boldsymbol{A})^{-1}\boldsymbol{A}^\top$ is also called the *Moore-Penrose pseudo-inverse of $\boldsymbol{A}$*.

A second remark is that it is possible to generalize the orthogonal projection to convex sets. Let $\mathcal{X}$ denote an inner product space. A set $U \subset \mathcal{X}$ is *convex* if, for any $x, y \in U$, it holds that

$$\lambda x + (1 - \lambda)y \in U,$$

for any $\lambda \in [0, 1]$. In other words, the straight line between any two points in U is itself in U. The orthogonal projection onto a convex set U is defined precisely as

$$\mathbf{Proj}_U(x) = \operatorname*{argmin}_{y \in \mathcal{Y}} \|x - y\|.$$

That such a projection always exists and is unique follows from the Weierstrass extreme value theorem (see Theorem A.2 in Section A.1.2).

Linear transformations

Given two linear spaces $\mathcal{X}$ and $\mathcal{Y}$, a mapping $F : \mathcal{X} \to \mathcal{Y}$ is called a *linear transformation* if

$$F(\alpha x + \beta y) = \alpha F(x) + \beta F(y),$$

for all $x, y \in \mathcal{X}$ and all scalars $\alpha, \beta \in \mathbb{R}$.

Given a linear transformation $F : \mathcal{X} \to \mathcal{Y}$, the set

$$\mathcal{N}(F) = \{x \in \mathcal{X} \mid F(x) = \mathbf{0}\}$$

is called the *null space* or *kernel* of F. The set

$$\mathcal{R}(F) = \{y \in \mathcal{Y} \mid y = F(x), \text{ for some } x \in \mathcal{X}\}$$

is called the *range* or *image* of F. $\mathcal{N}(F)$ is a subspace of $\mathcal{X}$ and $\mathcal{R}(F)$ is a subspace of $\mathcal{Y}$. The dimension of the subspace $\mathcal{R}(F)$ is called the *rank* of F. A linear transformation $F : \mathcal{X} \to \mathcal{Y}$ for which $\mathcal{N}(F) = \mathbf{0}$ is called *invertible*, and the corresponding inverse transformation is represented by F^{-1}. It can be shown that the inverse of a linear transformation, when it exists, is also a linear transformation.

If $U = \{u_n, n = 1, \ldots, N\}$ is a basis for $\mathcal{X}$, then we can write x as a linear combination of the basis vectors, yielding

$$F(x) = F\left(\sum_{n=1}^{N} x_n u_n\right) = \sum_{n=1}^{N} x_n F(u_n).$$

Since each $F(u_n), n = 1, \ldots, N$, is a vector in $\mathcal{Y}$, we can write

$$F(x) = \boldsymbol{F}\boldsymbol{x},$$

where $\boldsymbol{F}$ is a matrix where the nth column is the vector representation of $F(u_n)$ in some basis for $\mathcal{Y}$. Therefore, given basis for $\mathcal{X}$ and $\mathcal{Y}$, there is a one-to-one correspondence between linear transformations and matrices. This means, for example, that the change of basis discussed earlier can be seen as a linear transformation.

Note that we can equate the *composition* of linear transformations with the product of the corresponding matrices. It can also be shown that the inverse of a linear transformation F can be represented by the inverse of the matrix representation of F.

Determinants

Given a vector space $\mathcal{X}$ with dimension N, let $U = \{u_1, \ldots, u_N\}$ be a set of arbitrary vectors in $\mathcal{X}$. The *determinant* of the set U, $\det(U)$, is a real-valued function $\det : \mathcal{X}^N \to \mathbb{R}$ such that

- It is a *multilinear function*, i.e., ,

$$\begin{aligned}&\det(\boldsymbol{u}_1, \ldots, \alpha\boldsymbol{u}_n + \beta\boldsymbol{v}_n, \ldots, \boldsymbol{u}_N)\\ &\qquad = \alpha \det(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n, \ldots, \boldsymbol{u}_N) + \beta \det(\boldsymbol{u}_1, \ldots, \boldsymbol{v}_n, \ldots, \boldsymbol{v}_N).\end{aligned}$$

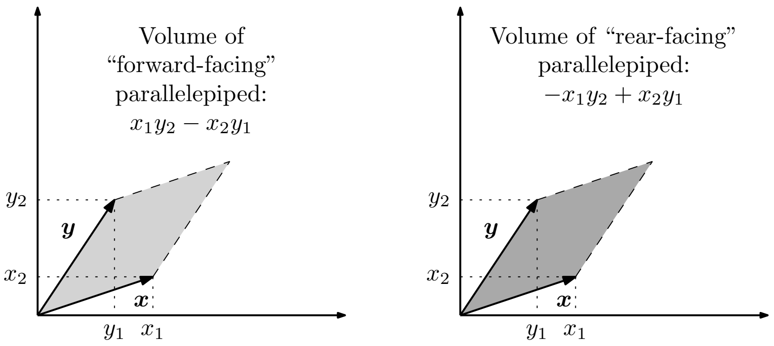

Figure A.1 Illustration of the determinant as an oriented volume. Given a matrix $\boldsymbol{A} = \left[\begin{smallmatrix} x_1 & y_1 \\ x_2 & y_2 \end{smallmatrix}\right]$ the determinant $\det(\boldsymbol{A})$ can be seen as the volume of the parallelepiped with vertices $\boldsymbol{x} = [\ x_1\ x_2\ ]^\top$ and $\boldsymbol{y} = [\ y_1\ y_2\ ]^\top$.

- It is *alternating*, i.e., for any $m, n \in \{1, \ldots, N\}$,

$$\det(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m, \ldots, \boldsymbol{u}_n, \ldots, \boldsymbol{u}_N) = -\det(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n, \ldots, \boldsymbol{u}_m, \ldots, \boldsymbol{u}_N).$$

- If $\boldsymbol{e}_n$ denote the unitary vector with 1 in the nth position and 0 elsewhere,

$$\det(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_N) = 1.$$

Intuitively, the determinant can be seen as the "oriented volume", in N-dimensional space, of the parallelepiped with vertices $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N\}$ (see Fig. A.1 for an illustration). It is multilinear since changing one "side" of the parallelepiped changes its volume accordingly. It is alternating because switching two "sides" flips the orientation, yielding a symmetric "oriented volume". Finally, the last property ensures that the unit square has volume 1.

Now given a linear transformation $F : \mathcal{X} \to \mathcal{X}$, we define the determinant of F as the determinant of the columns of the matrix representation of F. In other words, if U is a basis for $\mathcal{X}$ and $\boldsymbol{f}_n = F(u_n), n = 1, \ldots, N$, we define

$$\det(F) = \det(\boldsymbol{f}_1, \ldots, \boldsymbol{f}_N).$$

In this sense, if $\mathcal{X}$ and $\mathcal{Y}$ are vector spaces and we denote by $\mathcal{L}(\mathcal{X}, \mathcal{Y})$ the set of *all* linear transformations from $\mathcal{X}$ to $\mathcal{Y}$, the determinant is a mapping $\det : \mathcal{L}(\mathcal{X}, \mathcal{X}) \to \mathbb{R}$. An important property of the determinant is that $\det(F \circ G) = \det(F)\det(G)$.

Eigenvalues and eigenvectors

Let $F : \mathcal{X} \to \mathcal{X}$ denote a linear transformation from $\mathcal{X}$ to itself, and $\mathcal{Y} \subset \mathcal{X}$ a subspace of $\mathcal{X}$. The space $\mathcal{Y}$ is *invariant* under F if

$$F(x) \in \mathcal{Y}, \qquad \text{for every } x \in \mathcal{Y}.$$

For example, the orthogonal projection onto a subspace $\mathcal{Y}$ discussed before is a linear transformation under which $\mathcal{Y}$ is invariant. This means, in particular, that the orthogonal projection is *idempotent*.[1]

[1] A linear transformation $F : \mathcal{X} \to \mathcal{X}$ is *idempotent* if $(F \circ F)(x) = F(x)$ for all $x \in \mathcal{X}$.

Given a linear transformation $F : \mathcal{X} \to \mathcal{X}$, a non-zero vector $x \in \mathcal{X}$ is an *eigenvector* of F if there is $\lambda \in \mathbb{R}$ such that

$$F(x) = \lambda x.$$

The value λ is called an *eigenvalue* of F. If x and y are two eigenvectors associated with the same eigenvalue λ, then

$$F(\alpha x + \beta y) = \alpha F(x) + \beta F(y) = \lambda(\alpha x + \beta y),$$

which implies that the set of eigenvectors associated with a given eigenvalue λ is a subspace of $\mathcal{X}$. We refer to this subspace as the *eigenspace* associated with λ. Moreover,

$$(F \circ F)(x) = F(\lambda x) = \lambda^2 x,$$

meaning that eigenspaces are invariant under F. The dimension of the eigenspace associated with an eigenvalue λ is called the *multiplicity of* λ. The set of all eigenvalues of a linear transformation F is called the *spetrum of* F, $\sigma(F)$, and corresponds to the set

$$\sigma(F) = \{\lambda \in \mathbb{R} \mid F - \lambda I \text{ is not invertible}\}.$$

Given a linear transformation $F : \mathcal{X} \to \mathcal{X}$, the direct sum of the eigenspaces of F is a subspace of $\mathcal{X}$. When the dimensions of all eigenspaces of F add to the dimension of $\mathcal{X}$, we have that

$$\mathcal{X} = \bigoplus_{\lambda \in \sigma(F)} \mathcal{E}_\lambda,$$

where $\mathcal{E}_\lambda$ is the eigenspace associated with eigenvalue λ. Equivalently, if U_λ is a basis for $\mathcal{E}_\lambda$, the set

$$U = \bigcup_{\lambda \in \sigma(F)} U_\lambda$$

is a basis for $\mathcal{X}$. This means that every $x \in \mathcal{X}$ can be written as

$$x = \sum_{\lambda \in \sigma(F)} \sum_{n=1}^{N_\lambda} x_n^\lambda u_n^\lambda,$$

where N_λ is the multiplicity of λ. But then,

$$F(x) = F\left(\sum_{\lambda \in \sigma(F)} \sum_{n=1}^{N_\lambda} x_n^\lambda u_n^\lambda\right) = \sum_{\lambda \in \sigma(F)} \sum_{n=1}^{N_\lambda} x_n^\lambda F(u_n^\lambda) = \sum_{\lambda \in \sigma(F)} \sum_{n=1}^{N_\lambda} \lambda x_n^\lambda u_n^\lambda. \quad \text{(A.4)}$$

Given an arbitrary basis U_0 for $\mathcal{X}$, let $\boldsymbol{U}_\Lambda$ denote the matrix containing in its columns the representation of the eigenvectors of F in U_0. Then, $\boldsymbol{U}_\Lambda$ is a change of basis matrix; if $\boldsymbol{x}_\Lambda$ is the representation of $x \in \mathcal{X}$ in the basis U_Λ of the eigenvectors of F, then

$$\boldsymbol{x}_0 = \boldsymbol{U}_\Lambda \boldsymbol{x}_\Lambda,$$

where $\boldsymbol{x}_0$ is the representation of x in U_0. If $\boldsymbol{F}_0$ and $\boldsymbol{F}_\Lambda$ denote the representations of F in U_0 and U^λ, respectively, it holds that

$$\boldsymbol{F}_0\boldsymbol{x}_0 = \boldsymbol{U}_\Lambda\boldsymbol{F}_\Lambda\boldsymbol{x}_\Lambda = \boldsymbol{U}_\Lambda\boldsymbol{F}_\Lambda\boldsymbol{U}_\Lambda^{-1}\boldsymbol{x}_0.$$

On the other hand, by virtue of (A.4),

$$\boldsymbol{F}_\Lambda = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 & 0 \\ 0 & \lambda_2 & \dots & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \dots & 0 & \lambda_N \end{bmatrix},$$

and the matrix $\boldsymbol{F}_0$ is called *diagonalizable*. Diagonalizability of a matrix $\boldsymbol{F}$ is a convenient property, since it allows several algebraic manipulations of $\boldsymbol{F}$ can be performed in terms of its diagonal component, often leading to significant computational savings.

A linear transformation $F : \mathcal{X} \to \mathcal{X}$ is

- *Positive definite*, if all its eigenvalues are strictly positive.
- *Positive semi-definite*, if all its eigenvalues are non-negative.
- *Negative definite*, if all its eigenvalues are strictly negative.
- *Negative semi-definite*, if all its eigenvalues are non-positive.
- *Indefinite*, otherwise.

In light of the one-to-one correspondence between linear transformations and matrices, we define positive (semi-)definite, negative (semi-)definite, and indefinite matrices in the same manner.

We conclude with an example that is used throughout the book. A *stochastic matrix* is a matrix of non-negative entries whose rows add to 1. Formally, an $N \times N$ matrix $\boldsymbol{P}$ is a stochastic matrix if $P_{i,j} \geq 0$ and

$$\sum_{j=1}^{N} P_{i,j} = 1,$$

for all $i = 1, \dots, N$. It follows that

$$\boldsymbol{P}\mathbf{1} = \mathbf{1}$$

and $\mathbf{1}$ is an eigenvector of $\boldsymbol{P}$ associated to the eigenvalue 1. Note also that no eigenvalue can be larger than 1. To see why this is so, suppose that there is an eigenvalue $\lambda > 1$ with an associated eigenvector $\boldsymbol{x}$. Then,

$$\max_n [\boldsymbol{P}\boldsymbol{x}]_n = \lambda \max_n x_n > \max_n x_n.$$

However, since the rows of $\boldsymbol{P}$ add to one,

$$[\boldsymbol{P}\boldsymbol{x}]_m = \sum_{k=1}^{N} P_{m,k} x_k \leq \sum_{k=1}^{N} P_{m,k} \max_n x_n \leq \max_n x_n,$$

which is a contradiction. Therefore, we can conclude that $\sigma(\boldsymbol{P}) \subset [0,1]$, and the matrix $\boldsymbol{P}$ is at least positive semi-definite.

A.1.2 Metric spaces

The study of multivariable analysis relies critically on the idea of *convergence*: continuity, differentiability and integration all rely on convergence in different ways. Convergence, in turn, is rooted on the idea of an *open set*.

Open sets can be introduced as members of a *topology* on a set $\mathcal{X}$, without resorting to the notion of metric. However, for our purposes, such level of abstraction is unnecessary, so we introduce open sets as defined by a *metric*.

Metric

A *metric* on a set $\mathcal{X}$ is any mapping $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that

- It is *non-negative*, i.e., for any $x, y \in \mathcal{X}$, $d(x,y) \geq 0$. Moreover, $d(x,y) = 0$ if and only if $x = y$.
- It is *symmetric*, i.e., for any $x, y \in \mathcal{X}$, $d(x,y) = d(y,x)$.
- It verifies the *triangle inequality*, i.e., for any $x, y, z \in \mathcal{X}$, $d(x,y) \leq d(x,z) + d(z,y)$.

A set $\mathcal{X}$ endowed with a metric d is called a *metric space*.

We denote a metric space as a pair $(\mathcal{X}, d)$ or, when the metric is clear from the context or immaterial for the discussion, simply as $\mathcal{X}$. Common metrics include:

- The *discrete metric*, defined as

$$d(x,y) = 1 - \mathbb{I}\left[x = y\right]. \tag{A.5}$$

  The discrete metric is usually used in discrete sets.

- The *Manhattan metric* in $\mathbb{R}^N$ is derived from the 1-norm discussed in Section A.1.1 and is defined as

$$d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_1 \stackrel{\text{def}}{=} \sum_{n=1}^{N} |x_n - y_n|. \tag{A.6}$$

- Similarly, the *Euclidean metric* in $\mathbb{R}^N$ is derived from the 2-norm and is defined as

$$d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2 \overset{\text{def}}{=} \sqrt{\sum_{n=1}^{N} (x_n - y_n)^2}. \tag{A.7}$$

- More generally, if $\mathcal{X}$ is a normed vector space, the norm $\|\cdot\|$ induces a metric d on $\mathcal{X}$, where

$$d(x, y) = \|x - y\|. \tag{A.8}$$

Given a point $x_0 \in \mathcal{X}$, the *open ball* with center in x_0 and radius $\varepsilon > 0$ is the set

$$B_\varepsilon(x_0) = \{x \in \mathcal{X} \mid d(x, x_0) < \varepsilon\}.$$

An *open set* is any set $U \subset \mathcal{X}$ such that, for any $x \in U$, there is $\varepsilon > 0$ such that $B_\varepsilon(x) \subset U$. A *neighborhood* of a point $x \in \mathcal{X}$ is any open set that contains x.

It is possible to show that:

- $\mathcal{X}$ and $\emptyset$ are open sets.
- For any collection of open sets, $\{U_\alpha\}$, the set $\bigcup_\alpha U_\alpha$ is an open set.
- For any finite collection of open sets, $\{U_k, k = 1, \dots, K\}$, the set $\bigcap_k U_k$ is an open set.

The *interior* of a set $U \subset \mathcal{X}$, $\mathrm{int}(U)$, is the union of all open subsets contained in U. It follows that the interior of a set is an open set—it is the largest open set contained in U. Since the largest open set contained in an open set U is U itself, it can easily be shown that a set U is open if and only if $U = \mathrm{int}(U)$. A point $x \in \mathrm{int}(U)$ is called an *interior point of* U.

A *closed set* U is any set such that $\bar{U}$ is open, where $\bar{U}$ is the complement of U in $\mathcal{X}$, i.e,

$$\bar{U} = \{x \in \mathcal{X} \mid x \notin U\}.$$

The *closure* of a set $U \subset \mathcal{X}$, $\mathrm{cl}(U)$, is the intersection of all closed sets that contain U and is itself a closed set. A set U is closed if and only if $U = \mathrm{cl}(U)$. The *boundary* of a set $U \subset \mathcal{X}$ is the intersection of $\mathrm{cl}(U)$ and $\mathrm{cl}(\bar{U})$ and is also a closed set.

Two other topological concepts play an important role in the study of metric spaces. A set $U \subset \mathcal{X}$ is *bounded* if, for every $x, y \in U$,

$$d(x, y) \leq K,$$

for some $K < \infty$. Intuitively, a set U is bounded if no two points in U are too far from one another.

A set $U \subset \mathcal{X}$ is *compact* if every open cover of U has a finite sub-cover. Recall that an open cover of a set U is any collections of open sets $\{U_\alpha, U_\alpha \subset \mathcal{X}\}$ such that

$$U \subset \bigcup_\alpha U_\alpha.$$

We conclude by noting that every compact set is bounded and closed.

Convergence

A sequence $\{x_n, n \in \mathbb{N}\} \subset \mathcal{X}$ *converges to* $x^* \in \mathcal{X}$ if, for any $\varepsilon > 0$, there is $N \in \mathbb{N}$ such that $x_n \in B_\varepsilon(x_0)$ for all $n \geq N$. We denote such fact as $x_n \to x^*$ or $\lim_{n\to\infty} x_n = x^*$. Intuitively, a sequence converges to a limit x^* if all but a finite number of elements of the sequence are arbitrarily close to x^*. Given $x, x' \in \mathcal{X}$, if a sequence $\{x_n, n \in \mathbb{N}\}$ converges both to x and to x', then $x = x'$. Moreover, if $x_n \to x$ and $\{x_{u_n}, n \in \mathbb{N}\}$ is a subsequence of $\{x_n, n \in \mathbb{N}\}$, then $x_{u_n} \to x$.

We list a number of results that further reinforce the relation between convergence and the topological concepts discussed earlier.

- A set U is open if and only if every sequence converging to a point in U has all but a finite number of its terms in U.
- A set U is closed if and only if every convergent sequence in U converges to a point in U.
- A set U is compact if every sequence in U has a subsequence that converges to an element in U.

We note that two distinct metrics d_1 and d_2 may induce the same topology and, as such, whichever sequence converges in d_1 also converges in d_2 and vice versa. When that is the case, the metrics are said *equivalent*. For example, in $\mathbb{R}^N$ all p-norm induced metrics,

$$d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_p \overset{\text{def}}{=} \left(\sum_{n=1}^{N} |x_n - y_n|^p\right)^{\frac{1}{p}}, \tag{A.9}$$

are equivalent.

A second important concept related to convergence is that of a *Cauchy sequence.* A sequence $\{x_n, n \in \mathbb{N}\} \subset \mathcal{X}$ is a Cauchy sequence if, for any $\varepsilon > 0$, there is $N \in \mathbb{N}$ such that $d(x_n, x_m) < \varepsilon$ for all $m, n \geq N$. Intuitively, in a Cauchy sequence all but a finite number of elements in the sequence are arbitrarily close to one another.

Since, in a convergent sequence, all but a finite number of elements are arbitrarily close to its limit, they are also arbitrarily close to one another. Therefore, every convergent sequence is a Cauchy sequence. Moreover, Cauchy sequences are bounded, and subsequences of a Cauchy sequence are Cauchy sequences themselves.

Even though all convergent sequences are Cauchy sequences, the converse may not be true. We say that a metric space $\mathcal{X}$ is *complete* when all Cauchy sequences are convergent. Therefore, in a complete metric space, a sequence is convergent if and only if it is a Cauchy sequence.

We now introduce a fundamental result, used extensively throughout the book. A mapping $F : \mathcal{X} \to \mathcal{X}$ is called a *contraction* if there is a nonnegative scalar $\gamma < 1$ such that

$$d(F(x), F(y)) \leq \gamma d(x, y),$$

for any $x, y \in \mathcal{X}$.

Theorem A.1 (Banach fixed-point theorem). *Let $(\mathcal{X}, d)$ be a complete metric space, and $F : \mathcal{X} \to \mathcal{X}$ a contraction mapping. Then F admits a unique* fixed-point *in $\mathcal{X}$, i.e., there is a unique point $x^* \in \mathcal{X}$ such that*

$$x^* = F(x^*). \tag{A.10}$$

It is educative to establish the statement in Theorem A.1, since it is constructive and provides an algorithmic approach to computing the fixed-point x^* of a contraction mapping F.

Let $x_0 \in \mathcal{X}$ denote an arbitrary point, and construct a sequence $\{x_n, n \in \mathbb{N}\}$ recursively as follows:

$$x_{n+1} = F(x_n).$$

We show that the sequence thus obtained is a Cauchy sequence. We have that

$$d(x_{n+1}, x_n) \leq \gamma d(x_n, x_{n-1}) \leq \ldots \leq \gamma^n d(x_1, x_0) = K\gamma^n,$$

with $K = d(x_1, x_0)$. Then, by the triangle inequality,

$$d(x_{m+1}, x_n) \leq K \sum_{k=n}^{m} \gamma^k \leq K \frac{\gamma^n}{1-\gamma}.$$

Then, given any $\varepsilon > 0$, we can set

$$N = \left\lceil \log_\gamma \frac{\varepsilon(1-\gamma)}{K} \right\rceil.$$

This shows that the sequence $\{x_n, n \in \mathbb{N}\}$ is a Cauchy sequence and the theorem is proved.

Note, moreover, that the sequence $\{x_n, n \in \mathbb{N}\}$ can be constructed computationally (by successively applying F to $x_n, n = 0, 1, \ldots$) as a way to successively approximate x^*.

Continuity

Let $(\mathcal{X}, d_X)$ and $(\mathcal{Y}, d_Y)$ denote two metric spaces. A function $F : \mathcal{X} \to \mathcal{Y}$ is *continuous at a point* $x \in \mathcal{X}$ if, for every $\delta > 0$ there is $\varepsilon > 0$ such that $F(B_\varepsilon(x)) \subset B_\delta(F(x))$, i.e.,

$$d_X(x, y) < \varepsilon \qquad \Longrightarrow \qquad d_Y(F(x), F(y)) < \varepsilon,$$

for all $y \in \mathcal{X}$. If F is continuous at every point $x \in \mathcal{X}$ we simply say that F is continuous.

Continuous function exhibit the useful property that they transform convergent sequences into convergent sequences. Formally, let $\{x_n, n \in \mathbb{N}\}$ denote a convergent sequence in $\mathcal{X}$ such that $x_n \to x^*$, for some $x^* \in \mathcal{X}$. Let $F : \mathcal{X} \to \mathcal{Y}$ denote a continuous mapping. Then, the sequence $\{f(x_n), n \in \mathbb{N}\}$ is convergent and

$$\lim_{n\to\infty} F(x_n) = F(x^*).$$

Additionally, if $F : \mathcal{X} \to \mathcal{Y}$ and $G : \mathcal{Y} \to \mathcal{Z}$ are continuous functions, the composition of F and F, defined for every $x \in \mathcal{X}$ as

$$(G \circ F)(x) = g(f(x)),$$

is also continuous. From this simple result, one can show, for example, that if F and G are two continuous real-valued functions defined on some metric space $\mathcal{X}$, then the functions $F + G$, FG and F/G, defined for every $x \in \mathcal{X}$ as

$$\begin{aligned}(F + G)(x) &= F(x) + G(x),\\ (FG)(x) &= F(x)G(x),\\ (F/G)(x) &= F(x)/G(x), \quad G(x) \neq 0,\end{aligned}$$

are also continuous. Similarly, if $F_1, \ldots, F_N$ are continuous real-valued functions defined on some metric space $\mathcal{X}$, the function $F : \mathcal{X} \to \mathbb{R}^N$ defined as

$$F(x) = \begin{bmatrix} F_1(x) \\ \vdots \\ F_N(x) \end{bmatrix}$$

is also continuous. Another useful example is that, given two sequences $\{x_n, n \in \mathbb{N}\}$ and $\{y_n, n \in \mathbb{N}\}$ converging to x and y, respectively, then $d(x_n, y_n) \to d(x, y)$, i.e., the metric itself is a continuous mapping from $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

Theorem A.2 (Weierstrass extreme value theorem). *Let F be a continuous real-valued function defined on some metric space $\mathcal{X}$ and U a compact subset of $\mathcal{X}$. Then, the function F attains a maximum and a minimum in U.*

If $(\mathcal{X}, d_X)$ and $(\mathcal{Y}, d_Y)$ are metric spaces, a function $F : \mathcal{X} \to \mathcal{Y}$ is called *Lipschitz continuous* if there is a constant $K < \infty$ such that

$$d_Y(F(x), F(y)) \leq K d_X(x, y)$$

for all $x, y \in \mathcal{X}$. A Lipschitz continuous function is continuous.

A.1.3 Differentiation

Differentiation is concerned with the description of how a function changes as its arguments change. For real-valued functions $F : \mathbb{R} \to \mathbb{R}$, such description is provided by the *derivative*, $\frac{dF}{dx}$, which precisely describes the rate of change of F with (its argument) x.

Differentiation can be extended to more general spaces by bringing together the machinery from the two previous sections. For the purposes of this book, full generality is unnecessary. For this reason, we gradually specialize our presentation to the case of functions defined on $\mathbb{R}^N$. For those interested in a more general treatment of these topics are referred to any book on calculus on manifolds (see, for example, Abraham, Marsden, and Ratiu, 1988).

Let $\mathcal{X}$ be a normed vector space. As discussed in Section A.1.2, the norm $\|\cdot\|$ induces a metric on $\mathcal{X}$ such that

$$d(x, y) = \|x - y\|.$$

Endowed with such metric, we can now discuss convergence of sequences and continuity of functions in the vector space $\mathcal{X}$. We say that the normed vector space $\mathcal{X}$ is a *Banach space* if it is complete—i.e., if all Cauchy sequences in $\mathcal{X}$ are convergent.

Let $\mathcal{X}$ denote a finite dimensional Banach space and U be an open subset of $\mathcal{X}$. A function $F : \mathcal{X} \to \mathbb{R}$ is *differentiable at a point* $x \in U$ if there is a bounded linear transformation DF which depends on x and such that

$$\lim_{y \to 0} \frac{F(x + y) - F(x) - DF(y)}{\|y\|} = 0.$$

for all $y \in \mathcal{X}$. When the linear transformation DF exists, it is unique; we call it *the derivative of* F *at* x. If F is differentiable at every point $x \in \mathcal{X}$, we simply say that F is differentiable in $\mathcal{X}$. The function $DF : \mathcal{X} \to \mathcal{L}(\mathcal{X}, \mathbb{R})$,[2] which maps each $x \in \mathcal{X}$ to the derivative of F at x, is simply called the *derivative of* F and has several important properties.

- It is *linear* in the sense that, for any differentiable functions $F, G : \mathcal{X} \to \mathbb{R}$ and any $\alpha \in \mathbb{R}$,

$$D(\alpha F + G)(x) = \alpha DF(x) + DG(x).$$

- If $F_1, \ldots, F_N$ are differentiable functions, with each $F_n : \mathcal{X} \to \mathbb{R}$, then the function $F : \mathcal{X} \to \mathbb{R}^N$ defined as

$$F(x) = \begin{bmatrix} F_1(x) \\ \vdots \\ F_N(x) \end{bmatrix}$$

is also differentiable and

$$DF(x) = \begin{bmatrix} DF_1(x) \\ \vdots \\ DF_N(x) \end{bmatrix}.$$

In the particular case where F is defined on $\mathbb{R}^N$, the *partial derivatives of* F are defined as

$$\frac{\partial F(\boldsymbol{x})}{\partial x_n} = \lim_{\alpha \to 0} \frac{F(\boldsymbol{x} + \alpha \boldsymbol{e}_n) - F(\boldsymbol{x})}{\alpha},$$

[2] Recall that, given two vector spaces $\mathcal{X}$ and $\mathcal{Y}$, $\mathcal{L}(\mathcal{X}, \mathcal{Y})$ denotes the set of all linear transformations from $\mathcal{X}$ to $\mathcal{Y}$.

for $n = 1, \ldots, N$. For convenience of notation, the partial derivative $\partial F/\partial x_n$ is also denoted $D_n F$. It is, itself, a function $D_n F : \mathbb{R}^N \to \mathbb{R}$ of which we can, in turn, compute the derivative. We write $D_{mn}F$ or $\frac{\partial^2 F}{\partial x_n \partial x_m}$ to denote the *second order derivative* $D_n(D_m F)$, and $D_{n_1 \ldots n_k} F$ to denote the kth order derivative $D_{n_k}(\ldots(D_{n_1}F)\ldots)$.

The existence of partial derivatives is not sufficient to ensure differentiability. A function $F : \mathbb{R}^N \to \mathbb{R}$ is of class C^1 (or *continuously differentiable*) at a point $\boldsymbol{x} \in \mathbb{R}^N$ if all partial derivatives $D_n F(\boldsymbol{x})$ exist and are continuous at $\boldsymbol{x}$. Continuously differentiable functions are differentiable. More generally, a function $F : \mathbb{R}^N \to \mathbb{R}$ is of class C^k at a point $\boldsymbol{x} \in \mathbb{R}^N$ if all partial derivatives $D_{n_1 \ldots n_k} F$ exist and are continuous in $\boldsymbol{x}$. It is also possible to establish differentiability from Lipschitz continuity.

Theorem A.3 (Rademacher Theorem). *Let $U \subset \mathbb{R}^N$ be an open subset and $F : U \to \mathbb{R}^M$ a function that is Lipschitz continuous in U. Then, F is differentiable almost everywhere in U.*

We can extend the notion of partial derivative. The *directional derivative of $F : \mathbb{R}^N \to \mathbb{R}$ at $\boldsymbol{x}$ in the direction $\boldsymbol{v}$* is defined as

$$D_{\boldsymbol{v}} F(\boldsymbol{x}) = \lim_{\alpha \to 0} \frac{F(\boldsymbol{x} + \alpha \boldsymbol{v}) - F(\boldsymbol{x})}{\alpha}.$$

If $F : \mathbb{R}^N \to \mathbb{R}$ is differentiable at $\boldsymbol{x}$, it holds that

$$D_{\boldsymbol{v}} F(\boldsymbol{x}) = \langle \nabla F(\boldsymbol{x}), \boldsymbol{v} \rangle \tag{A.11}$$

where the vector $\nabla F(\boldsymbol{x})$, defined as

$$\nabla F(\boldsymbol{x}) \overset{\text{def}}{=} \begin{bmatrix} D_1 F(\boldsymbol{x}) & \ldots & D_N F(\boldsymbol{x}) \end{bmatrix}^\top,$$

is the *gradient of F at $\boldsymbol{x}$*. Finally, if $F : \mathbb{R}^M \to \mathbb{R}^N$ is differentiable,

$$DF(\boldsymbol{x}) = \begin{bmatrix} D_1 F_1(\boldsymbol{x}) & \ldots & D_M F_1(\boldsymbol{x}) \\ \vdots & \ddots & \vdots \\ D_1 F_N(\boldsymbol{x}) & \ldots & D_N F_N(\boldsymbol{x}) \end{bmatrix}. \tag{A.12}$$

The matrix in (A.12) is called the *Jacobian matrix* of F.

Taylor expansion

Sometimes it is useful to approximate a function $F : \mathbb{R}^N \to \mathbb{R}$ by a low-order polynomial. Towards that goal, we define

$$\langle \boldsymbol{v}, \nabla \rangle F(\boldsymbol{x}) \overset{\text{def}}{=} \langle \boldsymbol{v}, \nabla F(\boldsymbol{x}) \rangle = D_{\boldsymbol{v}} F(\boldsymbol{x}),$$

where the last equality follows from (A.11). More generally, we let

$$\langle \boldsymbol{v}, \nabla \rangle^{k+1} F(\boldsymbol{x}) = \langle \boldsymbol{v}, \nabla \rangle \left(\langle \boldsymbol{v}, \nabla \rangle^k F(\boldsymbol{x}) \right).$$

Note that, although ∇ is a differentiation operator, it can safely be treated algebraically in treating the term $\langle \boldsymbol{v}, \nabla \rangle$ as a "standard" inner product in $\mathbb{R}^N$. To see why this is so, consider for example a function $F : \mathbb{R}^2 \to \mathbb{R}$. We get

$$\langle \boldsymbol{v}, \nabla \rangle F(\boldsymbol{x}) = v_1 D_1 F(\boldsymbol{x}) + v_2 D_2 F(\boldsymbol{x}) = (v_1 D_1 + v_2 D_2) F(\boldsymbol{x})$$

and

$$\begin{aligned} \langle \boldsymbol{v}, \nabla \rangle^2 F(\boldsymbol{x}) &= v_1 D_1 \big(\langle \boldsymbol{v}, \nabla \rangle F(\boldsymbol{x}) \big) + v_2 D_2 \big(\langle \boldsymbol{v}, \nabla \rangle F(\boldsymbol{x}) \big) \\ &= v_1 D_1 (v_1 D_1 F(\boldsymbol{x}) + v_2 D_2 F(\boldsymbol{x})) + v_2 D_2 (v_1 D_1 F(\boldsymbol{x}) + v_2 D_2 F(\boldsymbol{x})) \\ &= (v_1 D_1 + v_2 D_2)^2 F(\boldsymbol{x}). \end{aligned}$$

The following theorem identifies conditions under which a function $F : \mathbb{R}^N \to \mathbb{R}$ can be approximated as a polynomial.

Theorem A.4 (Taylor theorem). *Let $F : \mathbb{R}^N \to \mathbb{R}$ denote a function of class $C^K, K > 1$ in some open convex set $U \subset \mathbb{R}^N$, and let $\boldsymbol{x}, \boldsymbol{y} \in U$. Then, there is $\lambda \in [0, 1]$ such that*

$$F(\boldsymbol{y}) = \sum_{k=0}^{K-1} \frac{1}{k!} \langle \boldsymbol{y} - \boldsymbol{x}, \nabla \rangle^k F(\boldsymbol{x}) + \frac{1}{K!} \langle \boldsymbol{y} - \boldsymbol{x}, \nabla \rangle^K F(\boldsymbol{z}),$$

for $\boldsymbol{z} = \lambda \boldsymbol{x} + (1 - \lambda) \boldsymbol{y}$.

Note that, for $F : \mathbb{R} \to \mathbb{R}$, the above expression reduces to the standard Taylor expression

$$F(y) = \sum_{k=0}^{K-1} (y - x)^k \; \frac{F^{(k)}(x)}{k!} + (y - x)^K \; \frac{F^{(K)}(z)}{K!},$$

for some $z \in [x, y]$, where $F^{(k)}$ denotes the kth order derivative of F. Moreover, if F is a class C^∞ function, we can represent F as the power series

$$F(y) = \sum_{k=0}^{\infty} (y - x)^k \; \frac{F^{(k)}(x)}{k!}.$$

The Taylor/power series expansion can be used to derive useful approximations for common functions.

- The Taylor expansion of $\log(1 - x)$ around the origin is given by

  $$\log(1 - x) = -\sum_{k=1}^{\infty} \frac{x^k}{k!}, \quad \text{for } |x| < 1.$$

  Therefore,

  $$\log(1 - x) \le -x, \quad \text{for } x \in [0, 1], \tag{A.13}$$

  and

  $$-\log(1 - x) \le x(1 + x), \quad \text{for } x \in [0, \tfrac{1}{2}]. \tag{A.14}$$

- The Taylor expansion of $\log(x)$ around 1 is given by

$$\log(x) = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{(x-1)^k}{k}, \quad \text{for } |x-1| \leq 1, x \neq 0,$$

and we immediately get that

$$\log(x) \leq x - 1. \tag{A.15}$$

- The Taylor expansion for the exponential is given by

$$\exp(x) = \sum_{k=1}^{\infty} \frac{x^k}{k!}, \quad \text{for all } x.$$

Hence,

$$e^x \leq 1 + x + x^2(e-2), \quad \text{for } x \leq 1. \tag{A.16}$$

We conclude this section with a brief overview of optimization.

A.1.4 Optimization

We overview several methods for minimizing a function $F : \mathbb{R}^N \to \mathbb{R}$. The function F is usually known as the *objective function*, and $F(\boldsymbol{x})$ roughly translates the "cost" of $\boldsymbol{x}$ and the goal is, of course, to find the $\boldsymbol{x} \in \mathbb{R}^N$ with minimum cost. Of course, the same exact approach can be used to address *maximization problems*, simply by changing the sign of the objective function.

Optimization problems can be categorized according to several criteria, mostly related with the objective function F.

Convexity When the objective function is convex, local minima are global minima and are unique, and many methods exist to find the optimal solution. Nonconvex problems, in contrast, are generally hard to find, and most methods are plagued by local minima. The topic of convex optimization is a mature one, and many excellent references exist—the textbooks of Boyd and Vandenberghe (2004) and Bertsekas (2015) are just two examples.

Differentiability When the objective function is differentiable, we can use information from the derivatives to search for (local) minima. Recall that the directional derivative $D_{\boldsymbol{v}}F(\boldsymbol{x})$ of a differentiable function $F : \mathbb{R}^N \to \mathbb{R}$ is given by

$$\langle \boldsymbol{v}, \nabla F(\boldsymbol{x}) \rangle = \|\boldsymbol{v}\| \, \|\nabla F(\boldsymbol{x})\| \cos\theta,$$

where θ is the angle between $\boldsymbol{v}$ and $\nabla F(\boldsymbol{x})$. The function F grows the most in the direction of the gradient (where $\theta = 0$) and decreases the most in the opposite direction.

In the remainder of this section, we briefly survey standard methods for both unconstrained and constrained optimization problems.

Unconstrained optimization problems

We start with the optimization problem

$$\text{minimize} \quad F(\boldsymbol{x}), \tag{A.17}$$

where $F : \mathbb{R}^N \to \mathbb{R}$ and $\boldsymbol{x}$ is the *optimization variable*. In the simplest cases, $\boldsymbol{x}$ can take any value in $\mathbb{R}^N$, and we say that a point $\boldsymbol{x} \in \mathbb{R}^N$ is a *local minimum of* F if there is $\varepsilon > 0$ such that

$$F(\boldsymbol{y}) \geq F(\boldsymbol{x}), \tag{A.18a}$$

for any $\boldsymbol{y}$ such that $\|\boldsymbol{x} - \boldsymbol{y}\| < \varepsilon$. If the objective function F is continuously differentiable and $\boldsymbol{x}^*$ is a local minimum of F, then

$$\nabla F(\boldsymbol{x}^*) = \mathbf{0} \tag{A.18b}$$

and the *Hessian matrix* of F at $\boldsymbol{x}^*$, defined as

$$\nabla^2 F(\boldsymbol{x}^*)) \overset{\text{def}}{=} \nabla(\nabla F(\boldsymbol{x}^*)^\top),$$

is positive semi-definite. In the particular case where F is a convex function, local minima are also global minima, so solving the problem (A.17) amounts to solving (A.18a).

The two conditions (A.18) are *necessary conditions for optimality*, but not sufficient. However, if we replace (A.18b) by the stronger requirement that $\nabla^2 F(\boldsymbol{x}^*))$ must be positive definite, we get a set of *sufficient conditions for optimality*:

$$\nabla F(\boldsymbol{x}^*) = \mathbf{0} \tag{A.19a}$$

$$\nabla^2 F(\boldsymbol{x}^*) \quad \text{is positive definite.} \tag{A.19b}$$

◇

We survey a standard class of methods to solve unconstrained minimization problems which iteratively build a sequence of estimates, $\{\boldsymbol{x}^{(t)}, t \in \mathbb{N}\}$, such that $F(\boldsymbol{x}^{(t+1)}) < F(\boldsymbol{x}^{(t)})$. Such methods fall under the general designation of *iterative descent methods*, the most common of which is, perhaps, the *gradient descent method*.

Let us suppose that the objective function $F : \mathbb{R}^N \to \mathbb{R}$ is continuously differentiable. The gradient descent method departs from an estimate $\boldsymbol{x}^{(0)}$ and builds the sequence $\{\boldsymbol{x}^{(t)}, t \in \mathbb{N}\}$ recursively as

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - \alpha_t \nabla F(\boldsymbol{x}^{(t)}), \tag{A.20}$$

where α_t is a small positive scalar. From the Taylor expansion of F,

$$\begin{aligned} F(\boldsymbol{x}^{(t+1)}) &\approx F(\boldsymbol{x}^{(t)}) + (\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)})^\top \nabla F(\boldsymbol{x}^{(t)}) + \text{small term} \\ &= F(\boldsymbol{x}^{(t)}) - \alpha_t \left\|\nabla F(\boldsymbol{x}^{(t)})\right\|^2 + \text{small term} \end{aligned}$$

and, for small α_t, it follows that $F(\boldsymbol{x}^{(t+1)}) < F(\boldsymbol{x}^{(t)})$.

There are numerous variations of the gradient descent approach just described, in which (A.20) is modified as

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - \alpha_t \boldsymbol{D}_t \nabla F(\boldsymbol{x}^{(t)}), \tag{A.21}$$

where $\boldsymbol{D}_t$ is some positive definite matrix. The analysis above remains valid, and the simple inclusion of $\boldsymbol{D}_t$ often boosts significantly the performance of the algorithm. In any case, when $\nabla F(\boldsymbol{x}^{(t)}) = 0$ the method will surely have converged to a local minimum and can, therefore, stop. In practice, the method is usually stopped when $\left\|\nabla F(\boldsymbol{x}^{(t)})\right\| < \varepsilon$, for some adequately selected $\varepsilon > 0$.

The standard gradient descent method is obtained when $\boldsymbol{D}_n = \boldsymbol{I}$. Another common variation can be obtained when the objective function is of class C^2. In this case, whenever $\nabla^2 F(\boldsymbol{x})$ is positive semi-definite, we can set

$$\boldsymbol{D}_t = \left(\nabla^2 F(\boldsymbol{x}^{(t)})\right)^{-1}$$

to get the update for *Newton's method*,

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - \alpha_t \left(\nabla^2 F(\boldsymbol{x}^(t))\right)^{-1} \nabla F(\boldsymbol{x}^{(t)}). \tag{A.22}$$

When applicable, Newton's method is significantly faster than standard gradient descent. However, the requirement that $\nabla^2 F(\boldsymbol{x})$ is twice differentiable is often too restrictive, which restricts the applicability of the method.

Another common variation—useful when $F(\boldsymbol{x})$ takes the form $F(\boldsymbol{x}) = \mathbb{E}\left[f(\boldsymbol{x}, \mathrm{z})\right]$ for some r.v z—is known as *stochastic gradient descent* and is discussed in greater detail in Appendix B.

We conclude by noting that the performance of the generalized gradient descent algorithm in (A.21) depends critically on (i) the direction of the update, governed by $\boldsymbol{D}_n$; (ii) the size of the update, governed by the step-size α_n; and (iii) the stopping condition of the algorithm.

Constrained optimization problems

We focus on the general problem

$$\begin{array}{ll} \text{minimize} & F(\boldsymbol{x}) \\ \text{subject to} & G_m(\boldsymbol{x}) \leq 0, m = 1, \ldots, M \end{array} \tag{A.23}$$

where $F : \mathbb{R}^N \to \mathbb{R}$ and each $G_m : \mathbb{R}^N \to \mathbb{R}, m = 1, \ldots, M$. The expressions $G_m(\boldsymbol{x}) \leq 0$ describe the *constraints* imposed on the solution. A point $\boldsymbol{x} \in \mathbb{R}^N$ that verifies the constraints is called *feasible.* If it exists, we denote the optimal solution as $\boldsymbol{x}^*$.

The formulation in (A.23) is used to translate an optimization problem that we wish to solve, and is referred as the *primal.* However, it is possible to derive a second optimization problem, known as the *dual*, which takes a form that, in a sense, is complementary to that of (A.23). The dual problem for (A.23) is

$$\begin{aligned} &\text{maximize} \quad Q(\boldsymbol{\lambda}) \\ &\text{subject to} \quad \lambda_m \geq 0, m = 1, \ldots, M, \end{aligned} \tag{A.24}$$

where the *dual function* Q is given by

$$Q(\boldsymbol{\lambda}) \overset{\text{def}}{=} \inf_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\lambda})$$

and

$$L(\boldsymbol{x}, \boldsymbol{\lambda}) = F(\boldsymbol{x}) + \sum_{m=1}^{M} \lambda_m G_m(\boldsymbol{x}). \tag{A.25}$$

The function L in (A.25) is called the *Lagrangian* for the problem; the scalars $\lambda_m, m = 1 \ldots, M$ correspond to the variables for the dual problem and are known as *Lagrange multipliers*. We denote by $\boldsymbol{\lambda}^*$ the solution to the dual problem.

An important result—known as *weak duality*—is that $Q(\boldsymbol{\lambda}^*) \leq F(\boldsymbol{x}^*)$, since by definition

$$Q(\boldsymbol{\lambda}) \leq L(\boldsymbol{x}, \boldsymbol{\lambda}) = F(\boldsymbol{x}) + \sum_{m=1}^{M} \lambda^\top G(\boldsymbol{x}) \leq F(x)$$

for all feasible $\boldsymbol{\lambda}$ and all feasible $\boldsymbol{x}$. The difference $F(\boldsymbol{x}^*) - Q(\boldsymbol{\lambda}^*)$ is known as the *optimal duality gap*. When the duality gap is 0, the problem is said to exhibit *strong duality*. The following result identifies conditions under which strong duality holds.

Proposition A.5. *Problems* (A.23) *and* (A.24) *have optimal solutions* $\boldsymbol{x}^*$ *and* $\boldsymbol{\lambda}^*$ *and* $F(\boldsymbol{x}^*) = H(\boldsymbol{\lambda}^*)$ *if and only if*

- *The solution vector* $\boldsymbol{x}^*$ *is feasible, i.e.,* $G_m(\boldsymbol{x}) \leq 0, m = 1, \ldots, M$*;*
- *The solution vector* $\boldsymbol{\lambda}^*$ *is feasible, i.e.,* $\lambda_m \geq 0, m = 1, \ldots, M$*;*
- *The solution vector* $\boldsymbol{x}^*$ *verifies*

$$\boldsymbol{x}^* = \operatorname*{argmin}_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\lambda}^*);$$

- *The solution vector* $\boldsymbol{\lambda}^*$ *verifies*

$$\sum_{m=1}^{M} \lambda_m^* G_m(\boldsymbol{x}^*) = 0.$$

To better understand the last condition, we start with the following concept. An inequality constraint $G_m(\boldsymbol{x}) \leq 0$ is *active at* $\boldsymbol{x}$ if

$$G_m(\boldsymbol{x}) = 0,$$

and *inactive* otherwise. Let $\mathrm{Act}(\boldsymbol{x})$ denote the set of all active constraints at $\boldsymbol{x}$, i.e.,

$$\mathrm{Act}(\boldsymbol{x}) = \{m \mid G_m(\boldsymbol{x}) = 0\}\,.$$

If $\boldsymbol{x}^*$ is an optimal solution to (A.26), then it is also a solution to the problem where the active inequality constraints at $\boldsymbol{x}$ are replaced by equality constraints, and the inactive inequality constraints at $\boldsymbol{x}^*$ are removed. For this reason, if the mth inequality constraint is inactive at the solution $\boldsymbol{x}^*$, $\lambda_m^* = 0$. If, in contrast, the mth inequality constraint is active, then $G_m(\boldsymbol{x}^*) = 0$. In both cases, the last condition in Proposition A.5 holds.

The above framework applies unchanged to the more general problem

$$\begin{aligned} &\text{minimize} && F(\boldsymbol{x}) \\ &\text{subject to} && G_m(\boldsymbol{x}) \le 0, m = 1, \dots, M \\ &&& H_k(\boldsymbol{x}) = 0, k = 1, \dots, K. \end{aligned} \tag{A.26}$$

To see how, note that the points $\boldsymbol{x} \in \mathbb{R}^N$ such that $H_k(\boldsymbol{x}) = 0$ verify, simultaneously $H_k(\boldsymbol{x}) \le 0$ and $-H_k(\boldsymbol{x}) \le 0, k = 1, \dots, K$. The resulting Lagrangian is

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = F(\boldsymbol{x}) + \sum_{m=1}^{M} \lambda_m G_m(\boldsymbol{x}) + \sum_{k=1}^{K} \mu_k H_k(\boldsymbol{x}),$$

where $\lambda_m \ge 0, m = 1, \dots, M$, and the $\mu_k \in \mathbb{R}, k = 1, \dots, K$.

We can now enumerate the necessary conditions for optimality when F, $G_m, m = 1, \dots, M$, and $H_k, k = 1, \dots, K$, known as *Karush-Kuhn-Tucker necessary conditions.*

Theorem A.6 (Karush-Kuhn-Tucker conditions). *Let* $(\boldsymbol{x}^*)$ *denote an optimal solution to the primal problem, and* $(\boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ *an optimal solution for the dual problem. Then,*

$$\begin{aligned} &\nabla L(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = 0, && \\ &G_m(\boldsymbol{x}^*) \le 0, && m = 1, \dots, M, \\ &\lambda_m^* \ge 0, && m = 1, \dots, M, \\ &\lambda_m^* G_m(\boldsymbol{x}^*) = 0, && m = 1, \dots, M, \\ &H_k(\boldsymbol{x}^*) = 0, && k = 1, \dots, K. \end{aligned} \tag{A.27}$$

Moreover, if F*,* $G_m, m = 1, \dots, M$*, and* $H_k, k = 1, \dots, K$*, are convex, the conditions* (A.27) *are also sufficient.*

The appealing aspect of the above theory is that, when the optimal duality gap is 0, we can solve the primal problem by minimizing the Lagrangian. For problems with only equality constraints, unconstrained minimization methods such as the ones discussed before can be applied directly. For problems involving also inequality constraints, one possible approach is to turn the inequality constraints as penalty terms and then solve the resulting problem.

Linear programming

A particular situation in which exact methods are available occurs when the objective function F and the constraints $G_m, m = 1, \ldots, M$ and $H_k, k = 1, \ldots, K$, are linear. Such optimization problems are known as *linear programs*, and are usually take the following standard form

$$\begin{aligned} \text{minimize} \quad & \boldsymbol{c}^\top \boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}, \\ & \boldsymbol{x} \geq 0. \end{aligned} \tag{A.28}$$

When the constraints take the form of inequalities $\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$, the problem can be converted to the form (A.28) by adding *slack variables* $y_m, m = 1, \ldots, m$ to yield

$$\begin{aligned} \text{minimize} \quad & \boldsymbol{c}^\top \boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} + \boldsymbol{y} = \boldsymbol{b}, \\ & \boldsymbol{x} \geq 0, \boldsymbol{y} \geq 0. \end{aligned}$$

Similarly, if the constraints take the form of inequalities $\boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{b}$, the problem can, once again, be converted to the form (A.28) by adding *surplus variables* $z_m, m = 1, \ldots, m$ to yield

$$\begin{aligned} \text{minimize} \quad & \boldsymbol{c}^\top \boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} - \boldsymbol{z} = \boldsymbol{b}, \\ & \boldsymbol{x} \geq 0, \boldsymbol{z} \geq 0. \end{aligned}$$

Linear programs can be interpret geometrically, as depicted in Fig. A.2: the set of feasible solutions is the intersection of a set of hyperplanes (a polyhedron) and if there is any optimal solution, then there is at least one optimal solution that corresponds to a vertex of the feasible polyhedron. Vertices of the polyhedron are called *basic solutions*.

Solution methods can take advantage of such structure to determine the optimal solution, for example by traversing all vertices of the polyhedron until optimality is attained. The *simplex method* is built precisely on such intuition, it moves from one vertex to another in such a way as to always decrease the value of the objective function. We do not detail the algorithm here, and instead refer to a specialized book (see, for example, Luenberger, 1979).

We conclude by noting that duality is also fundamental in linear programming. We start by writing the linear program in the form

$$\begin{aligned} \text{minimize} \quad & \boldsymbol{c}^\top \boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{b}, \\ & \boldsymbol{x} \geq \boldsymbol{0}. \end{aligned} \tag{A.29}$$

The corresponding Lagrangian is

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \boldsymbol{c}^\top \boldsymbol{x} - \boldsymbol{\lambda}^\top (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) - \boldsymbol{\mu}^\top \boldsymbol{x},$$

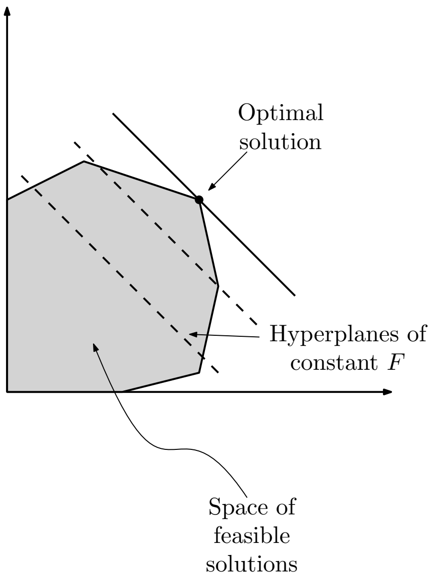

Figure A.2 Illustration of a linear program. The dashed lines correspond to *level sets* of the objective function F—i.e., hyperplanes where the function F is constant. The gray area corresponds to the set of feasible solutions, and the optimal solution is the vertex of the polyhedron closest to the origin.

and minimizing with respect to $\boldsymbol{x}$ yields a finite solution only when $\boldsymbol{\mu} = \boldsymbol{c} - \boldsymbol{A}\boldsymbol{\lambda}$. Therefore, we get

$$\begin{aligned} Q(\boldsymbol{\lambda}) &= \boldsymbol{c}^\top \boldsymbol{x}^* - \boldsymbol{\lambda}^\top (\boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{b}) - (\boldsymbol{c} - \boldsymbol{A}\boldsymbol{\lambda})^\top \boldsymbol{x}^* \\ &= \boldsymbol{\lambda}^\top \boldsymbol{b}. \end{aligned}$$

The associated constraints are $\boldsymbol{\lambda} \geq \boldsymbol{0}$ and $\boldsymbol{\mu} \geq \boldsymbol{0}$ or, equivalently, $\boldsymbol{A}\boldsymbol{\lambda} \leq \boldsymbol{c}$, finally yielding the program

$$\begin{aligned} \text{maximize} \quad & \boldsymbol{\lambda}^\top \boldsymbol{b} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{\lambda} \leq \boldsymbol{c} \\ & \boldsymbol{\lambda} \geq \boldsymbol{0}. \end{aligned} \tag{A.30}$$

A.2 Probability theory

In this section reviews basic concepts from probability theory. We aim at refreshing the notation and reviewing some useful results used in the main text. Readers unfamiliar with probabilities are strongly encouraged to read any introductory text on the subject (such as Bertsekas and Tsitsiklis, 2008).

◇

Let Ω denote an arbitrary set, and A an arbitrary subset of A. We write $\bar{A}$ to denote the complement of A on Ω, defined as

$$\bar{A} \overset{\text{def}}{=} \{\omega \in \Omega : \omega \notin A\}.$$

Let $\mathcal{F}$ denote any family of subsets of Ω such that

- $\Omega \in \mathcal{F}$;
- $\mathcal{F}$ is *closed under complements*, i.e., if $A \in \mathcal{F}$ then $\bar{A} \in \mathcal{F}$;
- $\mathcal{F}$ is *closed under countable unions*, i.e., given any countable collection $\{A_n\}$ such that $A_n \in \mathcal{F}$ for all n, then $A \in \mathcal{F}$, with

$$A = \bigcup_n A_n.$$

The family $\mathcal{F}$ is called a *σ-algebra*, and the elements of $\mathcal{F}$ are called *events*. Intuitively, the sets in $\mathcal{F}$ are those that we can measure. If $\mathcal{X}$ is a family of subsets of Ω, the smallest σ-algebra containing $\mathcal{X}$ is called the σ-algebra generated by $\mathcal{X}$ and is denoted as $\mathcal{F} = \sigma(\mathcal{X})$.

A *probability measure* on Ω is a function $\mathbb{P} : \mathcal{F} \to [0, 1]$, satisfying the following properties

- $\mathbb{P}[\emptyset] = 0$;
- $\mathbb{P}\left[\bar{A}\right] = 1 - \mathbb{P}[A]$ for every $A \in \mathcal{F}$;
- Given any two disjoint events A and B, $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B]$.

It follows that $\mathbb{P}[\Omega] = 1$. The properties above also yield the *union bound* (also known as *Boole's inequality*), which states that, given a finite collection of events $A_1, \ldots, A_N \in \mathcal{F}$, not necessarily disjoint, then

$$\mathbb{P}[A_1 \cup A_2 \cup \ldots \cup A_N] \leq \sum_{n=1}^{N} \mathbb{P}[A_n]. \tag{A.31}$$

We refer to the triplet $(\Omega, \mathcal{F}, \mathbb{P})$ as a *probability space*.

Conditional probabilities

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and two events $A, B \in \mathcal{F}$ such that $\mathbb{P}[B] > 0$, the *conditional probability of A given B* is defined as

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$$

Intuitively, if we interpret the probability $\mathbb{P}[A]$ as indicating how likely event A is to occur, the conditional probability can be seen as a "revised likelihood" regarding the occurrence of A if we know that B occurred. If knowing B does not change the likelihood of occurrence of A, then we say that A is *independent of* B. Formally,

$$\mathbb{P}[A \mid B] = \mathbb{P}[A]$$

or, equivalently,

$$\mathbb{P}[A \cap B] = \mathbb{P}[A]\,\mathbb{P}[B]\,. \tag{A.32}$$

Similarly, two events $A, B \in \mathcal{F}$ are *conditionally independent given event* C if

$$\mathbb{P}[A \cap B \mid C] = \mathbb{P}[A \mid C]\,\mathbb{P}[B \mid C]\,. \tag{A.33}$$

The relations in (A.32) and (A.33) extend to more than two events. For example, a collection of events $\mathcal{A} = \{A_1, \ldots, A_N\}$ is independent if, given any $\mathcal{A}' \subset \mathcal{A}$,

$$\mathbb{P}\left[\bigcap_{A \in \mathcal{A}'} A\right] = \prod_{A \in \mathcal{A}'} \mathbb{P}[A]\,.$$

The conditional probability is useful to describe the probability of an event A in terms of the probability of other events. For example, if $\{A_1, \ldots, A_N\}$ is a collection of disjoint events such that

$$\bigcup_{n=1}^{N} A_n = \Omega,$$

it follows that, for any event $A \in \mathcal{F}$,

$$\mathbb{P}[A] = \mathbb{P}\left[\bigcup_{n=1}^{N} (A \cap A_n)\right] = \sum_{n=1}^{N} \mathbb{P}[A \mid A_n]\,\mathbb{P}[A_n]\,. \tag{A.34}$$

Equation A.34 is known as the *total probability law* and is useful to decompose the probability of an event A into the sum of conditional probabilities. Another useful result is *Bayes rule*, which follows directly from the definition of conditional probability and states that

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[B \mid A]\,\mathbb{P}[A]}{\mathbb{P}[B]}. \tag{A.35}$$

Random variables

Probability spaces are usually used to describe random phenomena. The set Ω is known as the set of possible *outcomes* and $\mathcal{F}$ corresponds to the set of events to which we can assign a probability. In many situations, however, it is unpractical to work directly with probability spaces. *Random variables* (r.v.s) provide an alternative representation for the outcomes of the phenomenon of interest, usually in terms of some numerical quantity.

Formally, a r.v. x is a mapping $\mathrm{x} : \Omega \to \mathcal{X}$, where $\mathcal{X}$ is the set of values that the r.v. x can take.[3] If $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and x a r.v. defined on

[3] There are some technical requirements in order for x to be a proper r.v.—essentially, that statements about the events in $\mathcal{F}$ can be expressed in terms of the values of x. We refer to any book on probability for further details (see, for example, Bertsekas and Tsitsiklis, 2008.

$\mathcal{X}$, we say that x follows the *(probability) distribution* p_{x},[4] or that x is distributed according to p_{x} if

$$p_{\mathrm{x}}(x) = \mathbb{P}\left[\mathrm{x} = x\right], \qquad \text{for } x \in \mathcal{X},$$

and denote it as $\mathrm{x} \sim p_{\mathrm{x}}$. Given a r.v. x, taking values in some set $\mathcal{X}$ and following some distribution p_{x}, and a function $F : \mathcal{X} \to \mathbb{R}$, the *expected value of F according to p_{x}* is defined as

$$\mathbb{E}\left[F(\mathrm{x})\right] \overset{\text{def}}{=} \sum_{x} F(x)\, p_{\mathrm{x}}(x).^{5}$$

The expected value is also known as the *mean*, and is a linear operator in the sense that

$$\mathbb{E}\left[aF(\mathrm{x}) + b\right] = a\mathbb{E}\left[F(\mathrm{x})\right] + b,$$

for any scalars a, b.

Similarly, the *variance of F according to p_{x}* is defined as

$$\begin{aligned} \operatorname{var}\left[F(\mathrm{x})\right] &\overset{\text{def}}{=} \mathbb{E}\left[\left(F(\mathrm{x}) - \mathbb{E}_{\mathrm{x}\sim p}\left[F(\mathrm{x})\right]\right)^2\right] \\ &= \mathbb{E}\left[F(\mathrm{x})^2\right] - \mathbb{E}\left[F(\mathrm{x})\right]^2. \end{aligned}$$

Unlike the mean, the variance is not a linear operator. Instead, we have that

$$\operatorname{var}\left[aF(\mathrm{x}) + b\right] = a^2 \operatorname{var}\left[F(\mathrm{x})\right],$$

for any scalars a, b.

Joint and conditional distributions

Consider now two r.v.s x, y taking values in the sets $\mathcal{X}$ and $\mathcal{Y}$, respectively, and defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The event $\mathrm{x} = x, \mathrm{y} = y$, with $x \in \mathcal{X}, y \in \mathcal{Y}$, corresponds to

$$(\mathrm{x} = x, \mathrm{y} = y) \overset{\text{def}}{=} \left\{\omega \in \Omega : \mathrm{x}(\omega) = x \wedge \mathrm{y}(\omega) = y\right\},$$

and we can define a *joint distribution of* x *and* y as

$$p_{\mathrm{x},\mathrm{y}}(x, y) \overset{\text{def}}{=} \mathbb{P}\left[\mathrm{x} = x, \mathrm{y} = y\right].$$

Each of the two r.v.s x and y is distributed according to the corresponding *marginal distribution*,

$$p_{\mathrm{x}}(x) = \sum_{y \in \mathcal{Y}} p_{\mathrm{x},\mathrm{y}}(x, y) = \mathbb{P}\left[\mathrm{x} = x\right]; \qquad p_{\mathrm{y}}(y) = \sum_{x \in \mathcal{X}} p_{\mathrm{x},\mathrm{y}}(x, y) = \mathbb{P}\left[\mathrm{y} = y\right].$$

[4] In the case of discrete variables, this function is often designated as *probability mass function.* For continuous variables, the corresponding function is designated *probability density function.* However, since most r.v.s in the book are discrete in nature, we adopt the rather generic designation of “distribution function”.

[5] Sometimes we write $\mathbb{E}_{\mathrm{x}\sim p}\left[F(\mathrm{x})\right]$ to make explicit that the expectation is taken w.r.t. the distribution p.

The concept of joint distribution can naturally be extended to more than two r.v.s. It can also be used to introduce the concept of *independence* and *conditional distribution* of r.v.s. We say that the r.v.s x, y are *independent* if

$$p_{\mathrm{x},\mathrm{y}}(x,y) = p_{\mathrm{x}}(x)p_{\mathrm{y}}(y),$$

for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. The extension of the notion of independence to more than two r.v.s is similar to that described for events.

The *conditional distribution of* x *given* y is defined as

$$p_{\mathrm{x}|\mathrm{y}}(x \mid y) \overset{\text{def}}{=} \mathbb{P}\left[\mathrm{x} = x \mid \mathrm{y} = y\right] = \frac{p_{\mathrm{x},\mathrm{y}}(x,y)}{p_{\mathrm{y}}(y)}.$$

Then, given a function $F : \mathcal{X} \to \mathbb{R}$, the *conditional expectation of F given* y is defined as

$$\mathbb{E}\left[F(\mathrm{x}) \mid \mathrm{y} = y\right] = \sum_{x \in \mathcal{X}} F(x) p_{\mathrm{x}|\mathrm{y}}(x \mid y).$$

Common distributions

We now briefly go over common probability distributions that are used in the main text, as well as some of their main properties.

The *Bernoulli distribution* is, perhaps, the simplest distribution, used to describe binary r.v.s. A r.v. x follows a Bernoulli distribution with parameter $p \in [0,1]$—writen $\mathrm{x} \sim \mathsf{Bernoulli}(p)$—if

$$\mathbb{P}\left[\mathrm{x} = 1\right] = 1 - \mathbb{P}\left[\mathrm{x} = 0\right] = p$$

or, equivalently,

$$\mathbb{P}\left[\mathrm{x} = x\right] = p^x(1-p)^{1-x},$$

with $x \in \{0,1\}$. We have that

$$\begin{aligned}\mathbb{E}\left[\mathrm{x}\right] &= 1 \times p + 0 \times (1-p) = p \\ \operatorname{var}\left[\mathrm{x}\right] &= \mathbb{E}\left[\mathrm{x}^2\right] - \mathbb{E}\left[\mathrm{x}\right]^2 = p(1-p).\end{aligned}$$

The Bernoulli is a particular case of another common distribution, the *binomial distribution.* A r.v. x taking values in $\{0, \ldots, N\}$ follows a Binomial distribution with parameters N and $p > 0$—denoted as $\mathrm{x} \sim \mathsf{Bin}(N, p)$—if

$$\mathbb{P}\left[\mathrm{x} = n\right] = \binom{N}{n} p^n (1-p)^{N-n},$$

with $N > 0$ and $p \in [0,1]$. The Bernoulli distribution corresponds to the case where $N = 1$. Alternatively, the Binomial distribution can be seen as describing the sum of N independent Bernoulli r.v.s, i.e., if $\mathrm{x}_1, \ldots, \mathrm{x}_N$ are independent r.v.s following a Bernoulli distribution with parameter p, then the r.v.

$$\mathrm{x} = \sum_{n=1}^{N} \mathrm{x}_n$$

follows a binomial distribution with parameters N and p. It immediately follows that

$$\mathbb{E}\left[\mathrm{x}\right] = np \qquad \text{and} \qquad \mathrm{var}\left[\mathrm{x}\right] = np(1-p).$$

We can also generalize the Bernoulli distribution to take values in an arbitrary discrete set $\{x_1, \ldots, x_K\}$. We say that a r.v. x follows a *generalized Bernoulli distribution* (also known as *categorical distribution* or *multinoulli distribution*) with parameters $p_1, \ldots, p_K$ if

$$\mathbb{P}\left[\mathrm{x} = x_k\right] = p_k,$$

and we write $\mathrm{x} \sim \mathsf{Multinoulli}(p_1, \ldots, p_k)$ to denote such fact. If $p_1 = \ldots = p_K$, the distribution is called *uniform*, and we write $\mathrm{x} \sim \mathsf{Uniform}(\mathcal{X})$ to denote a uniformly-distributed r.v. taking values in the set $\mathcal{X}$.

A r.v. x taking values in $\mathbb{R}$ follows a *normal* or *Gaussian distribution* with parameters μ and σ^2—denoted $\mathrm{x} \sim \mathsf{Normal}(\mu, \sigma^2)$—if

$$\mathbb{P}\left[\mathrm{x} \in U\right] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_U \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx,$$

with $U \subset \mathbb{R}$, $\mu \in \mathbb{R}$ and $\sigma > 0$. We have that

$$\mathbb{E}\left[\mathrm{x}\right] = \mu \qquad \text{and} \qquad \mathrm{var}\left[\mathrm{x}\right] = \sigma^2.$$

The normal distribution plays a central role in much of statistics, mostly due to its many useful properties. In particualr,

- The mean, mode and median of a normal distribution all coincide with μ; the distribution is *symmetric* around μ.
- If x and y are two independent and normally distributed r.v.s, then $\mathrm{z} = a\mathrm{x} + \mathrm{y}$ is also normally distributed and

$$\mathbb{E}\left[\mathrm{z}\right] = a\mathbb{E}\left[\mathrm{x}\right] + \mathbb{E}\left[\mathrm{y}\right] \qquad \text{and} \qquad \mathrm{var}\left[\mathrm{z}\right] = a^2\mathrm{var}\left[\mathrm{x}\right] + \mathrm{var}\left[\mathrm{y}\right].$$

- The product of two normal distributions $\mathsf{Normal}(\mu_1, \sigma_1^2)$ and $\mathsf{Normal}(\mu_2, \sigma_2^2)$ is *proportional* to a normal distribution $\mathsf{Normal}(\mu_{\mathrm{prod}}, \sigma^2_{\mathrm{prod}})$, where

$$\begin{aligned}\mu_{\mathrm{prod}} &= \frac{\sigma_1^2\mu_2 + \sigma_2^2\mu_1}{\sigma^1 + \sigma^2} \\ \sigma^2_{\mathrm{prod}} &= \frac{\sigma_1^2\sigma_2^2}{\sigma^1 + \sigma^2}.\end{aligned}$$

- Finally, the normal distribution is the fundamental element in the many variations of the *central limit theorem*. In its simplest form, let $\mathrm{x}_1, \ldots, \mathrm{x}_N$ denote a set of *independent and identically distributed*[6] r.v.s taking values in $\mathcal{X} \subset \mathbb{R}$, having mean μ and variance $\sigma^2 < \infty$. Let s_N denote the r.v.

$$\mathrm{s}_N = \frac{1}{N} \sum_{n=1}^{N} \mathrm{x}_n.$$

[6] "Independent and identically distributed" is abbreviated to i.i.d.

Then, the *central limit theorem* ensures that

$$\frac{1}{\sqrt{N}}\sum_{n=1}^{N} \mathrm{x}_n \xrightarrow{d} \mathsf{Normal}(\mu, \sigma^2),$$

where $\xrightarrow{d}$ denotes *convergence in distribution*. In other words, as $N \to \infty$, the distribution of the r.v. s_n converges to $\mathsf{Normal}(\mu, \sigma^2/N)$.

The generalization of the normal distribution to $\mathbb{R}^p$ is relatively straightforward. The r.v. $\mathbf{x}$ taking values in $\mathbb{R}^p$ follows a *multivariate normal distribution* with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ if

$$\mathbb{P}\left[\mathbf{x} \in U\right] = \frac{1}{\sqrt{(2\pi)^p \det(\boldsymbol{\Sigma})}} \int_U \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right] d\boldsymbol{x},$$

with $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\boldsymbol{\Sigma}$ is a positive definite matrix. The parameter $\boldsymbol{\mu}$ corresponds to the mean of the distribution, while $\boldsymbol{\Sigma}$ corresponds to the corresponding covariance matrix.

We conclude by mentioning the *Dirac distribution* δ_0, which can be obtained from the normal distribution as

$$\delta(x) = \lim_{\sigma \to 0} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{x^2}{2\sigma^2}\right].$$

Intuitively, one can think of the Dirac distribution as an infinite spike at the origin:

$$\delta_0(x) = \begin{cases} +\infty & \text{if } x = 0 \\ 0 & \text{otherwise.} \end{cases}$$

The Dirac distribution is such that

$$\int_{\mathbb{R}} \delta_0(x) dx = 1$$

and, for any function $F : \mathbb{R} \to \mathbb{R}$,

$$\int_{\mathbb{R}} F(x)\delta_0(x) dx = f(0).$$

More generally, we write δ_y to denote the Dirac distribution centered in y, given by

$$\delta_y(x) = \delta_0(x - y).$$

Finally, the Dirac distribution can be generalized to arbitrary sets $\mathcal{X}$. For any $y \in \mathcal{X}$ and set $U \subset \mathcal{X}$, the Dirac distribution centered in y is such that

$$\int_U \delta_y(x) dx = \begin{cases} 1 & \text{if } y \in U \\ 0 & \text{otherwise.} \end{cases}$$

Comparing distributions

It is often convenient to measure whether two distributions are close of far apart. Several criteria are proposed in the literature for that purpose, with different properties.

The *Kulback-Liebler (KL) divergence* offers a criterion to compare two distributions, from an information-theoretic perspective. The KL-divergence between distributions p_1 and p_2 over some set $\mathcal{X}$ is defined as

$$\mathrm{KL}(p_1 \parallel p_2) = \sum_{x \in \mathcal{X}} p_1(x) \log \frac{p_1(x)}{p_2(x)}.$$

Intuitively, the KL-divergence measures the difference—in terms of expected message size—between an encoding optimized for p_2 and an encoding optimized for p_1, when the messages actually follows p_1. An important property of the KL divergence is that it is always positive. However, it is not symmetric and, as such, it is not a proper “distance”. However, its information-theoretic interpretation makes it an often used comparison criteria.

Another commonly used metric is the *total variation distance*. Given a set $\mathcal{X}$ and two distributions p_1 and p_2 over $\mathcal{X}$, the total variation distance between p_1 and p_2 is defined as

$$d(p_1, p_2) = \frac{1}{2} \sum_{x \in \mathcal{X}} \left| p_1(x) - p_2(x) \right|.$$

It is possible to assign a probabilistic meaning to the total variation distance. In particular, it is possible to show that

$$d(p_1, p_2) = \max_{U \subset \mathcal{X}} \left| \mathbb{P}_{\mathrm{x} \sim p_1} \left[\mathrm{x} \in U \right] - \mathbb{P}_{\mathrm{x} \sim p_2} \left[\mathrm{x} \in U \right] \right|.$$

Useful inequalities

Often, when performing probabilistic inference, it is possible to provide some form of guarantees regarding the outcome of such inference. Such guarantees often take the form of *error bounds*, which bound the probability of making an error of large magnitude. The derivation of such bounds often relies on a number of well-established inequalities from the literature. We summarize below some of the most useful inequalities.

Let x be a real-valued r.v. and a a positive scalar. The *Markov inequality* states that

$$\mathbb{P} \left[|\mathrm{x}| > a \right] \leq \frac{\mathbb{E} \left[|\mathrm{x}| \right]}{a}. \tag{A.36}$$

The *Chebyshev inequality* follows from (A.36), and states that

$$\mathbb{P} \left[|\mathrm{x} - \mathbb{E} \left[\mathrm{x} \right]| \geq a \right] \leq \frac{\mathrm{var} \left[\mathrm{x} \right]}{a^2}. \tag{A.37}$$

Establishing (A.37) usually resorts to another important inequality, called *Jensen's inequality*. Given a r.v. x taking values in some vector space $\mathcal{X}$ and a convex function $F : \mathcal{X} \to \mathbb{R}$,

$$F(\mathbb{E} \left[\mathrm{x} \right]) \leq \mathbb{E} \left[F(\mathrm{x}) \right]. \tag{A.38}$$

Hoeffding's lemma offers a bound for the Laplace transform of a random variable. Let x denote a r.v. taking values in $[a, b]$ and η some positive scalar. Then,

$$\mathbb{E}\left[e^{\eta \mathrm{x}}\right] \leq e^{\eta \mathbb{E}[\mathrm{x}]} e^{\eta^2 (b-a)^2/8}. \tag{A.39}$$

Finally, if $\mathrm{x}_1, \ldots, \mathrm{x}_N$ is a collection of N i.i.d. r.v.s taking values in $[0, 1]$, let s_n denote the r.v. defined as

$$\mathrm{s}_n = \mathrm{x}_1 + \ldots + \mathrm{x}_n.$$

If $\mathbb{E}\left[\mathrm{x}_i\right] = \mu, i = 1, \ldots, n$, then *Hoeffding's inequality* states that

$$\mathbb{P}\left[\mathrm{s}_n - n\mu \geq \varepsilon\right] \leq e^{-2\frac{\varepsilon^2}{n}}. \tag{A.40}$$