Carl de Boor

Linear Algebra

draft 25jan13

Contents

9 Duality

10 The powers of a linear map and its spectrum

11 Convergence of the power sequence

12 Canonical forms

13 Localization of eigenvalues

14 Optimization and quadratic forms

15 More on determinants

Preface

This book is motivated by the following realizations:

(1) The linear maps between a vector space X over the scalar field $\mathbb{F}$ and the associated coordinate spaces $\mathbb{F}^n$ are efficient tools for work on theoretical and practical problems involving X. Those from $\mathbb{F}^n$ to X share with matrices the feature of columns, hence are called column maps, while those from X to $\mathbb{F}^n$ share with matrices the feature of rows, hence are called row maps. Work with a linear map A usually requires its factorization into a column map and a row map. Such factorization is most efficient for the task if the particular column map is invertible as a map to the range of A, i.e., if it is a basis for that range.

(2) Gauss elimination is applied to matrices for the purpose of obtaining bases for their nullspace and for their range. It results in a sequence of matrices all with the same nullspace, with the last matrix making the nullspace quite evident.

(3) A change of basis amounts to interpolation and *vice versa.*

(4) Since the eigenstructure of a linear map A on a vector space X over the scalar field $\mathbb{F}$ is of interest in the study of the sequence $A^0 = \mathrm{id}, A^1 = A, A^2, \ldots$ of the powers of A, its derivation and discussion is best handled in terms of polynomials $p(A)$ in that linear map with coefficients in $\mathbb{F}$. While determinants are indispensible and powerful tools in certain situations, they do not provide the best path to understanding eigenstructure.

(5) In applications, vector spaces are, by and large, spaces of maps with the vector operations defined pointwise, or derived from such spaces in a straightforward manner. The coordinate spaces $\mathbb{F}^n$ are merely the simplest examples of such vector spaces.

These realizations led me to volunteer to teach the follow-up linear algebra course offered in the Mathematics Department of the University of Wisconsin-Madison and taken by undergraduate Math majors and graduate students from science and engineering departments. Each time, I produced lecture notes, and this book derives from them.

The numbered problems, posed throughout the book and typeset in the smaller font of this paragraph, are meant to deepen understanding of the material. While there are answers available to all the problems, the book contains only the answers to those problems that are referred to in the text for some missing proof detail or as an illustration of a point being made. The latter problems are starred.

I adhere to certain notational conventions:

(1) I distinguish between equalities being asserted or derived and equalities that hold by definition. For the latter, I use $:=$ or $=:$ depending on which side is being defined.

(2) I distinguish between terms or phrases being defined and those being emphasized. The former are set in **boldface** (to make them easy to find), the latter in *italic*.

(3) I use only one numerical sequence for labeling all the equations, figures and formal statements in each chapter as that seems to me more helpful for finding any particular labeled item than the more standard separate enumeration of various classes of items. I do, however, number separately the problems given throughout.

(4) I use standard symbols, like $\forall$ ('for all') and $\exists$ ('there exist(s)') with the subsequent 'such that' not written out, and standard abbreviations, like 'iff' for 'if and only if', and use braces, $\{...\}$, only to delimit the description of a set.

(5) A question mark in an equation indicates the unknown item, the item for which to solve the equation.

(6) n-vectors, i.e., elements of $\mathbb{R}^n$ or, more generally, of $\mathbb{F}^n$, are written in boldface, like $\mathbf{x} \in \mathbb{R}^n$, with their entries written in subscripted italics, e.g., $\mathbf{x} = (x_1, \ldots, x_n)$. In particular, n-vectors are not written as 1-column matrices.

The study of Linear Algebra is incomplete without some numerical experimentation. I carry out such experimentation with the help of `MATLAB`, a program that has grown well beyond its initial purpose of being a "`Matrix laboratory`" into a very handy tool for experimentation in general scientific computing. Throughout this book, there are paragraphs, typeset like this one, that provide information about `MATLAB` essential for experimenting with the material under discussion. Some of the problems also require `MATLAB`, but most of these are easily adapted to other programming languages. With that proviso, any reader not interested in numerical experimentation or well familiar with `MATLAB` can safely skip all such paragraphs.

Overview

Here is a quick run-down on this book, with various terms to be learned by studying this book printed in **boldface**.

Much of scientific work involves relationships called **map**s, specified in this book by the notational template

$$f : X \to Y : x \mapsto y$$

that is read: f is a map from the set X to the set Y, and maps $x \in X$ to $y \in Y$, with $\mapsto$ read 'maps to' and with "y" usually replaced by some expression or formula involving x. If these latter details do not matter, then the shorthand

$$f : X \to Y$$

indicates that f is a map with **domain** $X =: \operatorname{dom} f$ and **target** $Y =: \operatorname{tar} f$. When X and Y are understood from the context, the shorter form

$$f : x \mapsto y$$

is often used, at times even without mention of the name "f".

For example,

- ○ time ↦ the population of the US;
- ○ temperature ↦ pressure in a bottle;
- ○ location (longitude, latitude, altitude) ↦ (barometric pressure, humidity, temperature);
- ○ mother's age ↦ frequency of newborn with Down syndrom;
- ○ available resources (capital, raw materials, labor pool, etc) ↦ output of the US economy;
- ○ etc. .

All this is part of our hope to understand effects in terms of causes.

Once we feel we understand such a relationship, we are eager to put it to use in order to find out how to cause certain effects. Mathematically, we are trying to solve the equation

$$f(?) = y$$

for given $f : X \to Y$ and given $y \in Y$ where, here and throughout the book, the question mark indicates exactly what quantity we are hoping to determine.

In this generality and vagueness, nothing much can be said other than to urge familiarity with basic map terms, such as, **domain**, **target** and **range** of a map, the map properties **1-1** (equivalent to **uniqueness** of solutions), **onto** (equivalent to **existence** of a solution for any y), **invertible** (equivalent to having exactly one solution for any $y \in Y$, the best-possible situation), and the notions of **left inverse**, **right inverse** and **inverse** related to the earlier notions by the concept of **map composition**.

Often, though, the map f is a **smooth** map, from some subset X of **real n-dimensional coordinate space**, $\mathbb{R}^n$, to $\mathbb{R}^m$, say. With the list $\mathbf{x} = (x_1, \ldots, x_n)$ our notation for $\mathbf{x} \in \mathbb{R}^n$, this means that, first of all,

$$f(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x})) \ \in \mathbb{R}^m$$

with each f_j a real-valued function, and, secondly, at any point $\mathbf{p} \in X$, we can expand each f_j into a **Taylor series**:

$$f_j(\mathbf{p} + \mathbf{h}) = f_j(\mathbf{p}) + Df_j(\mathbf{p})^{\mathrm{t}}\mathbf{h} + o(\mathbf{h}), \quad j = 1, \ldots, m,$$

with

$$Df_j(\mathbf{p}) := (D_1 f_j(\mathbf{p}), \ldots, D_n f_j(\mathbf{p})) \ \in \mathbb{R}^n$$

the **gradient** of f_j at $\mathbf{p}$, and $\mathbf{x}^{\mathrm{t}}\mathbf{y} := x_1 y_1 + \cdots + x_n y_n$ the **scalar product** of the **n-vectors** $\mathbf{x}$ and $\mathbf{y}$, and the $o(\mathbf{h})$ denoting 'higher-order' terms that we eventually are going to ignore in best scientific fashion.

This implies that

$$f(\mathbf{p} + \mathbf{h}) = f(\mathbf{p}) + Df(\mathbf{p})\mathbf{h} + o(\mathbf{h}),$$

with

$$Df(\mathbf{p}) := \begin{bmatrix} D_1 f_1(\mathbf{p}) & \cdots & D_n f_1(\mathbf{p}) \\ \vdots & \cdots & \vdots \\ D_1 f_m(\mathbf{p}) & \cdots & D_n f_m(\mathbf{p}) \end{bmatrix}$$

the **Jacobian** matrix of f at $\mathbf{p}$.

With this, a standard approach to finding a solution to the equation

$$f(?) = \mathbf{y}$$

is **Newton's method**: We are looking for a **correction $\mathbf{h}$** to our **current guess $\mathbf{x}$** for the solution for which

$$\mathbf{y} = f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + Df(\mathbf{x})\mathbf{h} + o(\mathbf{h});$$

we ignore the 'higher-order' terms that hide behind the expression $o(\mathbf{h})$, and so get a *linear equation* for $\mathbf{h}$:

$$\mathbf{y} - f(\mathbf{x}) = Df(\mathbf{x})?,$$

which we solve for $\mathbf{h}$, add this correction to our current $\mathbf{x}$ to get a new guess

$$\mathbf{x} \leftarrow \mathbf{x}+\mathbf{h} = \mathbf{x} + Df(\mathbf{x})^{-1}(\mathbf{y} - f(\mathbf{x}))$$

and repeat. Under suitable circumstances, the process converges, to a solution.

The *key idea* here is the reduction, from solving a general equation $f(?) = \mathbf{y}$ to solving a sequence of **linear** equations, $Df(\mathbf{x})? = \mathbf{z}$. This works since, in principle, we can always solve a linear system.

Most equations $f(?) = \mathbf{y}$ that can be solved are actually solved by this process or some variant thereof, hence the importance of knowing how to solve *linear* equations.

For this reason, our first task will be to introduce **linear map**s and **vector space**s over the scalar field $\mathbb{F}$, with $\mathbb{F}$ either the real ($\mathbb{R}$) or the complex ($\mathbb{C}$) numbers (though it pays to think through the material for the case of a more general commutative field $\mathbb{F}$). We focus on **linear spaces of functions**, i.e., vector spaces in which the basic **vector operations**, namely **vector addition** and **multiplication by a scalar**, are defined **pointwise**. These provide the proper means for expressing the concept of **linearity**. We recognize that, for a linear map A from the vector space X to the vector space Y, the linear equation $A? = y$ has a solution, x_0 say, if and only if y is an element of $\operatorname{ran} A$, the **range** of A, in which case the general solution to the linear equation $A? = y$ is of the form $x_0 + \operatorname{null} A$, with $\operatorname{null} A$ the **nullspace** of A, i.e., the set of solutions to the **homogeneous** equation $A? = 0$.

Both $\operatorname{ran} A$ and $\operatorname{null} A$ are linear subspaces, of Y and X respectively, and efficient descriptions for them are in terms of a **basis**, i.e., in terms of an invertible linear map V from some **coordinate space** $\mathbb{F}^n$ to the linear subspace in question. This identifies bases as particular **column map**s, i.e., linear maps from some coordinate space, i.e., maps of the form

$$\mathbb{F}^n \to X : \mathbf{a} \mapsto a_1 v_1 + \cdots + a_n v_n =: [v_1, \ldots, v_n]\mathbf{a}$$

for some sequence $v_1, \ldots, v_n$ in the linear space X in question.

We will spend some time recalling various details about bases, how to construct them (using the concept of **bound** and **free** columns of column map), how to use them, and will also mention their generalization, **direct sum**s and their associated **linear projector**s or **idempotent**s. We stress the notion of **dimension** (:= number of columns or elements in a basis), in particular the **Dimension Formula**

$$\dim \operatorname{dom} A = \dim \operatorname{ran} A + \dim \operatorname{null} A,$$

valid for any linear map A, which summarizes much of what is important about dimension.

Then we recall **elimination** as the method for solving a **homogeneous** linear system

$$A? = \mathbf{0}$$

with $A \in \mathbb{F}^{m\times n}$. Specifically, we show that elimination identifies the bound and free columns of a matrix A, and this leads to **row echelon forms** for the matrix A, in particular the **rrref** or **really reduced row echelon form**, from which we can obtain a complete description of the solution set of $A? = \mathbf{0}$, i.e., for $\operatorname{null} A$ as well as an efficient description of $\operatorname{ran} A$. Thus equipped, we deal with the general linear system $A? = \mathbf{y}$ via the homogeneous linear system $[A, \mathbf{y}]? = \mathbf{0}$.

We will also worry about how to determine the **coordinates** of a given $x \in X$ with respect to a given basis V for X, i.e., how to solve the equation

$$V? = x.$$

This will lead us to **row map**s, i.e., linear maps from some vector space to coordinate space, i.e., maps of the form

$$X \to \mathbb{F}^n : x \mapsto (\lambda_1 x, \ldots, \lambda_n x) =: [\lambda_1, \ldots, \lambda_n]^{\mathrm{t}} x$$

for some sequence $\lambda_1, \ldots, \lambda_n$ of **linear functional**s, i.e., scalar-valued linear maps, on the vector space X in question. It will also lead us to **interpolation** aka **change of basis**, and will make us single out **inner product space**s as spaces with a ready supply of suitable row maps, and thence to **least-squares**, to particularly good bases, namely **o.n.** (:= **orthonormal**) bases (which are the **isometries** for the standard **norm**, the **Euclidean norm** $\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^{\mathrm{t}}\mathbf{x}}$ associated with the standard **inner product**, and which can be constructed from an arbitrary basis by **Gram-Schmidt**).

We will find that bases also show up naturally when we try to **factor** a given linear map $A \in L(X, Y)$ in the most efficient way, as a product

$$A = V\Lambda^{\mathrm{t}}$$

with $\Lambda^{\mathrm{t}} \in L(X, \mathbb{F}^r)$ and $V \in L(\mathbb{F}^r, Y)$ and r as small as possible. It will be one of my tasks to convince you that you have actually carried out such factorizations, in fact had to do this in order to do certain standard operations, like differentiating or integrating polynomials and other functions. Such factorizations are intimately connected with the **rank** of A (since the smallest possible r is the rank of A) and lead, for a matrix A, to the **SVD**, or **Singular Value Decomposition**,

$$A = V\Sigma W^{\mathrm{c}}$$

with V, W **orthonormal** bases, W^{c} the **conjugate transpose** of W, and Σ diagonal, a factorization that is, in a certain sense, a best way of describing the action of the linear map A. Other common factorizations for matrices are the PLU **factorization** with P a **permutation matrix**, L **unit lower triangular**, and U **upper triangular** (generated during elimination); and the (more stable) QR **factorization**, with Q **unitary** (i.e., an orthonormal basis) and R **upper**, or, **right triangular**, obtained by elimination with the aid of specific **elementary matrices** called **Householder reflection**s.

For *square* matrices, one hopes to (but does not always) get factorizations of the form $A = V\Sigma V^{-1}$ with Σ diagonal (the simplest example of a matrix without such a factorization is the **nilpotent** matrix $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$), but often must be (and is) content to get the **Schur form**, which is available for any square matrix and is of the form $A = VUV^{\mathrm{c}}$ with V an o.n. basis and U upper triangular. In either case, A is then said to be **similar** to Σ and U, respectively. These latter factorizations, or **similarities**, are essential for an understanding of the **power sequence**

$$A^0 = \mathrm{id}, A^1 = A, A^2 = AA, A^3 = AAA, ...$$

of the square matrix A and, more generally, for an understanding of the **matrix polynomial** $p(A)$, since, e.g.,

$$A = V\,\mathrm{diag}(\mu_1, \ldots, \mu_n)V^{-1} \quad \Longrightarrow \quad p(A) = V\,\mathrm{diag}(p(\mu_1), \ldots, p(\mu_n))V^{-1},$$

for any polynomial p and even for some well-behaved functions p like the **exponential** $p : t \mapsto \exp(t)$. In particular, then

$$A^k = V\,\mathrm{diag}(\mu_1^k, \ldots, \mu_n^k)V^{-1}, \quad k = 0, 1, 2, \ldots,$$

therefore we can describe the behavior of the *matrix* sequence $(A^k : k = 0, 1, 2, \ldots)$ entirely in terms of the *scalar* sequences $(\mu_j^k : k = 0, 1, 2, \ldots)$. Specifically, we can characterize **power-boundedness**, **convergence**, and **convergence to** 0.

There are many reasons for wanting to understand the power sequence of a matrix; here is one. Often, elimination is not the most efficient way to solve a linear system. Rather, the linear system

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

itself is solved by iteration, by splitting $A =: M - N$ with M 'easily' invertible, and looking at the **equivalent equation**

$$M? = N? + \mathbf{y}$$

which leads to the **iteration**

$$\mathbf{x} \leftarrow M^{-1}(N\mathbf{x} + \mathbf{y}) =: B\mathbf{x} + \mathbf{c}.$$

Convergence of this process depends crucially on the behavior of the power sequence for B (and does not at all depend on the particular **vector norm** or **map norm** used).

The factorization

$$A = V \operatorname{diag}(\mu_1, \ldots, \mu_n) V^{-1}$$

is equivalent to having $AV = V \operatorname{diag}(\mu_1, \ldots, \mu_n)$, i.e.,

$$[Av_1, \ldots, Av_n] = [\mu_1 v_1, \ldots, \mu_n v_n]$$

for some invertible $V = [v_1, \ldots, v_n] : \mathbb{F}^n \to \operatorname{dom} A$, i.e., to having a basis V consisting of **eigenvector**s for A, with the μ_j the corresponding **eigenvalue**s. For this reason, we will study the **eigenstructure** of A and the **spectrum** of A, as well as **similarity**, i.e., the **equivalence relation**

$$A \sim C \ := \ \exists V \ \ A = VCV^{-1}.$$

In this study, we make use of polynomials, particular the **annihilating polynomial**s (which are the nontrivial polynomials p for which $p(A) = 0$) and their cousins, the nontrivial polynomials p for which $p(A)x = 0$ for some $x \neq 0$, and the unique **monic** annihilating polynomial of minimal degree, called the **minimal polynomial** for A, as well as the **Krylov sequence** $x, Ax, A^2x, \ldots$.

We will discuss the most important classification of eigenvalues, into **defective** and **non-defective** eigenvalues, and give a complete description of the asymptotic behavior of the power sequence $A^0, A^1, A^2, \ldots$ in terms of the eigenstructure of A, even when A is not **diagonalizable**, i.e., is not similar to a diagonal matrix (which is equivalent to some eigenvalue of A being defective).

We will also discuss standard means for locating the **spectrum**, i.e., the collection of all eigenvalues, of a matrix, such as **Gershgorin's disks** and the **characteristic polynomial** of a matrix, and give the Perron-Frobenius theory concerning the dominant eigenvalue of a positive matrix.

From the Schur form (mentioned earlier), we derive the basic facts about the eigenstructure of **hermitian** and of **normal** matrices. We give the **Jordan form** only because of its mathematical elegance since, in contrast to the

Schur form, it cannot be constructed reliably numerically. We briefly discuss a related form, the **Weyr form**.

Further, we also consider briefly **minimization** of a real-valued map

$$f : K \to \mathbb{R}$$

with $K \subset \mathbb{R}^n$. Returning to our Taylor expansion

$$f(\mathbf{p}+\mathbf{h}) = f(\mathbf{p}) + Df(\mathbf{p})^{\mathrm{t}}\mathbf{h} + o(\mathbf{h}),$$

we notice that, usually, $\mathbf{p}$ cannot be a minimum point for f unless it is a **critical point**, i.e., unless the gradient, $Df(\mathbf{p})$, of f at $\mathbf{p}$ is the zero vector. However, even with $Df(\mathbf{p}) = \mathbf{0}$, we only know that f is 'flat' at $\mathbf{p}$. In particular, a critical point could also be a (local) maximum point, or a saddle point, etc. . To distinguish between the various possibilities, we must look at the **second-order** terms, i.e., we must write and know, more explicitly, that

$$f(\mathbf{p}+\mathbf{h}) = f(\mathbf{p}) + Df(\mathbf{p})^{\mathrm{t}}\mathbf{h} + \mathbf{h}^{\mathrm{t}}D^2f(\mathbf{p})\mathbf{h}/2 + o(\mathbf{h}^{\mathrm{t}}\mathbf{h}),$$

with

$$H := D^2 f := \begin{bmatrix} D_1D_1f & \cdots & D_1D_nf \\ \vdots & \cdots & \vdots \\ D_nD_1f & \cdots & D_nD_nf \end{bmatrix}$$

the **Hessian** for f, hence

$$\mathbf{h} \mapsto \mathbf{h}^{\mathrm{t}}D^2f(\mathbf{p})\mathbf{h} = \sum_{i,j} D_iD_jf(\mathbf{p})h_ih_j$$

the associated **quadratic form**.

We will learn to distinguish between maxima, minima, and saddle points by the signs of the eigenvalues of the Hessian, mention **Sylvester's Law of Inertia**, and show how to estimate the effect of **perturbations** of H on the spectrum of H, using ideas connected with the **Rayleigh quotient**.

At this point, you will realize that this book is strongly influenced by the use of Linear Algebra in Analysis, with important applications, e.g., in Graph Theory, ???, or ???, being ignored (partly through ignorance).

Finally, although **determinant**s have little to contribute to Linear Algebra at the level of this book, we will give a complete introduction to this very important Linear Algebra tool, and then discuss the **Schur complement**, **Sylvester's determinant identity**, and the **Binet-Cauchy formula**.

As a taste of the many different applications of Linear Algebra, we discuss briefly: the solution of a system of constant-coefficient ODEs, Markov processes, subdivision in CAGD, Linear Programming, the Discrete Fourier Transform, approximation by broken lines, the B-spline basis of a spline space, multivariate polynomial interpolation, the reduced monic Gröbner basis for a zero-dimensional polynomial ideal, and the use of flats in analysis and CAGD.

Throughout, we will rely on needed material from ancillary courses as collected in an appendix called **Background**.

1 Sets, assignments, lists, and maps

The basic objects of Mathematics are sets and maps. Linear Algebra is perhaps the first course where this fact becomes evident and where it can be illustrated in a relatively straightforward context. Since a complete understanding of the course material requires a thorough appreciation of the basic facts about maps, we begin with these and their simpler cousins, lists and assignments, after a brief review of standard language and notation concerning sets.

Sets

Sets of interest in this book include

- the **natural numbers** : $\mathbb{N} := \{1, 2, \ldots\}$;
- the **integers** : $\mathbb{Z} := \{\ldots, -1, 0, 1, \ldots\} = (-\mathbb{N}) \cup \{0\} \cup \mathbb{N}$;
- the **integer interval** $m{:}n := \{m, m+1, \ldots, n\}$ with $m, n \in \mathbb{Z}$ which is empty if $m > n$;
- the nonnegative integers : $\mathbb{Z}_+ := \{p \in \mathbb{Z} : p \geq 0\} = \mathbb{N} \cup \{0\}$;
- the **rational numbers** : $\mathbb{Q} := \mathbb{Z} \div \mathbb{N} := \{p/q : p \in \mathbb{Z}, q \in \mathbb{N}\}$;
- the **real numbers**: $\mathbb{R}$;
- the nonnegative reals: $\mathbb{R}_+ := \{x \in \mathbb{R} : x \geq 0\}$;
- the **open interval** $(a \,..\, b) := \{r \in \mathbb{R} : a < r < b\}$ with **endpoint**s $a, b \in \mathbb{R}$;
- the **closed interval** $[a \,..\, b] := \{r \in \mathbb{R} : a \leq r \leq b\}$;
- the **half-open interval** $[a \,..\, b) := \{r \in \mathbb{R} : a \leq r < b\}$;
- the **complex numbers** : $\mathbb{C} := \mathbb{R} + \mathrm{i}\mathbb{R} = \{x + \mathrm{i}y : x, y \in \mathbb{R}\}$, $\quad \mathrm{i} := \sqrt{-1}$.

As these examples show, a set is often specified in the form $\{x : P(x)\}$ which is read 'the set of all x that have the property $P(x)$'. Note the use of the colon, ':', (rather than a vertical bar, '|') to separate the initial, provisional, description of the typical element of the set from the conditions imposed on it for membership in the set, with that provisional description often only clear from the context. In this book, braces, '{', '}', are used solely in the description of sets.

Standard notation concerning sets includes:

- $\#S$ denotes the **cardinality** of the set S, i.e., the count of its elements.
- $x \in S$ and $S \ni x$ both mean that x is an element of S.
- $S \subset T$, $T \supset S$ both mean that S is a **subset** of T, i.e., all the elements of S are also elements of T; if we want to convey that S is a **proper subset** of T, meaning that $S \subset T$ but $S \neq T$, we write $S \subsetneq T$.
- $\{\}$ denotes the **empty set**, the set with no elements.
- $S \cap T := \{x : x \in S \text{ and } x \in T\}$ is the **intersection** of S and T.
- $S \cup T := \{x : x \in S \text{ or } x \in T\}$ is the **union** of S and T.
- $S \backslash T := \{x : x \in S \text{ but not } x \in T\}$ is the **difference** of S from T and is often read 'S take away T'. In this book, this difference is *never* written $S - T$, as the latter is reserved for the set $\{s - t : s \in S, t \in T\}$ formable when both S and T are subsets of the same vector space.

1.1 What is the standard name for the elements of $\mathbb{R}\backslash\mathbb{Q}$?

1.2 What is the standard name for the elements of $\mathrm{i}\mathbb{R}$?

1.3 Work out each of the following sets. (a) $(\{-1, 0, 1\} \cap \mathbb{N}) \cup \{-2\}$; (b) $(\{-1, 0, 1\} \cup \{-2\}) \cap \mathbb{N}$; (c) $\mathbb{Z}\backslash(2\mathbb{Z})$; (d) $\{z^2 : z \in \mathrm{i}\mathbb{R}\}$.

1.4 Determine $\#((\mathbb{R}_+\backslash\{x \in \mathbb{R} : x^2 > 16\}) \cap \mathbb{N})$.

Assignments, lists

(1.1) Definition: An **assignment** or, more precisely, an **assignment on I** or **I-assignment**

$$f = (f_i : i \in I)$$

associates with each element i in its **domain** (or, **index set**)

$$\operatorname{dom} f := I$$

some **term** or **item** or **entry** or **value** f_i. In symbols:

$$f : \operatorname{dom} f : i \mapsto f_i.$$

The set

$$\operatorname{ran} f := \{f_i : i \in \operatorname{dom} f\}$$

of all items appearing in the assignment f is called the **range** of the assignment.

If also g is an assignment, then $f = g$ exactly when $f_i = g_i$ for all $i \in \operatorname{dom} f = \operatorname{dom} g$.

Very confusingly, many mathematicians call an assignment an *indexed set*, even though it is not a set whose elements have been indexed. The term **family** is also used; however it, too, smacks too much of a set or collection.

We call the assignment f **1-1** if $f_i = f_j$ implies $i = j$.

The simplest assignment is the **empty assignment**, (), i.e., the unique assignment whose domain is the empty set. Note that the empty assignment is 1-1 (why?? See Problem 1.5).

An assignment with domain the set

$$\underline{n} := \{1, 2, \ldots, n\}$$

of the first n natural numbers is called a **list**, or, more explicitly, an **n-list**. Note that the empty assignment, (), is the only **0-list**. We use the notation

$$\#f := n$$

for the number of entries of the n-list f.

To specify an n-list f, it is sufficient to write down the **sequence** $f_1, f_2, \ldots, f_n$ of its terms or values:

$$f = (f_1, f_2, \ldots, f_n).$$

For example, the **cartesian product**

$$\times_{i=1}^{n} X_i := X_1 \times X_2 \times \cdots \times X_n := \{(x_1, x_2, \ldots, x_n) : x_i \in X_i, i \in \underline{n}\}$$

of the set sequence $X_1, \ldots, X_n$ is, by definition, the collection of all n-lists with the ith item or **coordinate** taken from X_i, all i.

In this book, we deal with *n-vectors*, i.e., n-lists of *numbers*, such as the 3-lists $(1, 3.14, -14)$ or $(3, 3, 3)$. (Note that the *list* $(3, 3, 3)$ is quite different from the *set* $\{3, 3, 3\}$. The list $(3, 3, 3)$ has three terms, while the set $\{3, 3, 3\}$ has exactly one element.)

(1.2) Definition: An **n-vector**

$$\mathbf{x} = (x_1, \ldots, x_n)$$

is a list of n scalars (numbers). The collection of all **real** (**complex**) n-vectors is denoted by $\mathbb{R}^n$ ($\mathbb{C}^n$).

In this book, a single boldface roman letter always denotes an n-vector, with the n clear from the context. However, the ith entry of the n-vector $\mathbf{x}$ is not denoted $\mathbf{x}_i$ but x_i.

In MATLAB, there are (at least) two ways to specify an n-vector, namely as a one-row matrix (colloquially known as a **row vector**), or as a one-column matrix (colloquially known as a **column vector**). For example, one can record the 3-vector $x = (1.3, 3.14, -15)$ as the one-row matrix

```
x_as_row = [1.3,3.14,-15]; % with the commas optional
```

or as the one-column matrix

```
x_as_col = [1.3;3.14;-15]; % with semicolons separating the rows
```

One can also write a one-column matrix as a column, without the need for the semicolons, e.g.,

```
x_as_col = [1.3
            3.14
            -15 ];
```

Back to general assignments. If $\operatorname{dom} f$ is finite, say $\#\operatorname{dom} f = n$, then we could always describe f by listing the n pairs (i, f_i), $i \in \operatorname{dom} f$, in some fashion. However, that may not always be the most helpful thing to do. Here is a famous example.

During the Cholera outbreak in 1854 in London, Dr. John Snow recorded the deaths by address, thus setting up an assignment whose domain consisted of all the houses in London. But he did not simply make a list of all the addresses and then record the deaths in that list. Rather, he took a map of London and marked the number of deaths at each address as a corresponding number of black dots right on the map. He found that the deaths clustered around one particular public water pump, jumped to a conclusion (remember that this was well before Pasteur's discoveries), had the handle of that pump removed, and had the satisfaction of seeing the epidemic fade.

Thus, one way to think of an assignment is to visualize its domain in some convenient fashion, and, 'at' each element of the domain, its assigned item or value.

This is routinely done for matrices, another basic object in this book.

1.5* Why is any assignment 1-1 whose domain contains fewer than 2 elements?

1.6 In some courses, students are assigned to specific seats in the class room. (a) If you were the instructor in such a class, how would you record this seating assignment? (b) What are the range and domain of this assignment?

1.7* A **relation** between the sets X and Y is any subset of $X \times Y$. Each such relation relates or associates with some elements of X one or more elements of Y. For each of the following relations, determine whether or not it provides an assignment on the set $X := \underline{3} =: Y$. (i) $R = X \times Y$; (ii) $R = \{(x,x) : x \in X\}$; (iii) $R = \{(1,2),(2,2)\}$; (iv) $R = \{(1,2),(2,1)\}$; (v) $R = \{(1,2),(3,1),(2,1)\}$; (vi) $R = \{(1,2),(2,2),(3,1),(2,1)\}$.

Matrices

(1.3) Definition: A **matrix**, or, more precisely, an $m \times n$**-matrix**, is any assignment with domain the cartesian product

$$\underline{m} \times \underline{n} \;=\; \{(i,j) : i \in \underline{m}, j \in \underline{n}\}$$

of $\underline{m}$ with $\underline{n}$, for some nonnegative m and n.

A matrix is called **real** resp. **complex** if all its entries are real, resp. complex numbers.

The collection of all **real**, resp. **complex** $m \times n$-matrices is denoted by $\mathbb{R}^{m\times n}$, resp. $\mathbb{C}^{m\times n}$.

It is customary to display an $m \times n$-matrix A as a rectangle of items:

$$A \;=\; \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,n} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m,1} & A_{m,2} & \cdots & A_{m,n} \end{bmatrix}.$$

This means that we must think of its domain rotated clockwise 90° when compared to the ordinary (x,y)-plane, i.e., the domain of many other bivariate assignments (or maps).

This way of displaying a matrix has led to the following language.

(1.4) Let A be an $m \times n$-matrix. The item

$$A_{ij} := A_{i,j}$$

corresponding to the index (i,j) is also called the **(i,j)-entry** of A. The list $A_{i:} := A_{i,:} := (A_{i,j} : j \in \underline{n})$ is called the **ith row of** A, the list $A_{:j} := A_{:,j} := (A_{i,j} : i \in \underline{m})$ is called the **jth column of** A, and the list $(A_{ii} = A_{i,i} : 1 \le i \le \min\{m,n\})$ is called the **(main) diagonal of** A.

A matrix with nonzero entries only on or above (below) the diagonal is called **upper (lower) triangular**. A **diagonal matrix** is one that is both upper and lower triangular.

By definition, A^{t} denotes the **transpose** of the matrix A, i.e., the $n \times m$-matrix whose (i,j)-entry is A_{ji}, all i,j. A is **symmetric** if $A^{\mathrm{t}} = A$. Because of its importance in the later parts of this book, we usually use the **conjugate transpose** $A^{\mathrm{c}} := \overline{A}^{\mathrm{t}}$ whose (i,j)-entry is the scalar $\overline{A_{ji}}$, with $\overline{\alpha}$ the complex conjugate of the scalar α. A is **Hermitian** if $A^{\mathrm{c}} = A$.

When $m = n$, A is called a **square matrix** of **order** n.

The notation $A_{i:}$ for the ith row and $A_{:j}$ for the jth column of the matrix A is taken from MATLAB, where, however, `A(i,:)` is a one-row matrix and `A(:,j)` is a one-column matrix (rather than just a vector). The (main) diagonal of a matrix `A` is obtained in MATLAB by the command `diag(A)`, which returns, in a one-column matrix, the list of the diagonal elements. The upper (lower) triangular part of a matrix `A` is the matrix obtained from `A` by setting to zero all entries below (above) the diagonal; it is provided by the command `triu(A)` (`tril(A)`). The conjugate transpose of a matrix `A` is obtained by `A'`. This is the same as the transpose if `A` is real. To get the mere *transpose* $\mathtt{A}^{\mathrm{t}}$ in the contrary case, you must use the notation `A.'` which is strange since there is nothing *pointwise* about this operation.

MATLAB's command `mesh(A)` plots the matrix `A`, treating its entries as the values of a function on a corresponding rectangular mesh but not in the way suggested earlier but rather by taking `A(i,j)` as the value at the point `(j,i)`. Here, for example, is the 'picture' of the 8×16-matrix $A :=$ `eye(8,16)` as generated by the command `mesh(eye(8,16))`. This matrix has all its diagonal entries equal to 1 and all other entries equal to 0. But note that a careless interpretation of this figure would lead one to see a matrix with 16 rows and only 8 columns.

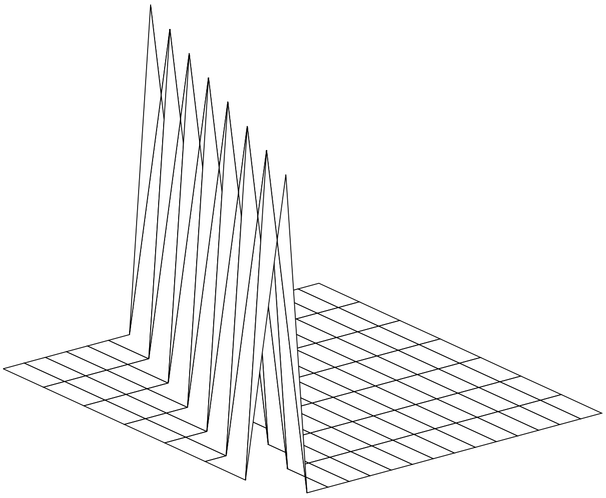


Figure. The rectangular identity matrix `eye(8,16)` as plotted in `MATLAB` via `mesh`

While lists can be concatenated in just one way, by letting one follow the other, matrices can be 'concatenated' by laying them next to each other and/or one underneath the other. The only requirement is that the result be again a matrix. If, for example,

$$A := [1 \quad 2], \qquad B := \begin{bmatrix} 3 \\ 6 \\ 9 \end{bmatrix}, \qquad C := \begin{bmatrix} 4 & 5 \\ 7 & 8 \end{bmatrix},$$

then there are four different ways to 'concatenate' these three matrices, namely

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}, \begin{bmatrix} 4 & 5 & 3 \\ 7 & 8 & 6 \\ 1 & 2 & 9 \end{bmatrix}, \begin{bmatrix} 3 & 1 & 2 \\ 6 & 4 & 5 \\ 9 & 7 & 8 \end{bmatrix}, \begin{bmatrix} 3 & 4 & 5 \\ 6 & 7 & 8 \\ 9 & 1 & 2 \end{bmatrix}.$$

In `MATLAB`, one would write the three matrices

```
A = [1 2];   B = [3;6;9];   C = [4 5; 7 8];
```

and would describe the four possible 'concatenations' as follows:

```
[[A;C],B];   [[C;A],B];   [B,[A;C]];   [B,[C;A]];
```

We saw earlier that even vectors are described in `MATLAB` by matrices since plain `MATLAB` only knows matrices.

The MATLAB use of semicolons and commas in the preceding MATLAB discussion is so handy that I will use it throughout the book in the description of matrices composed of submatrices.

1.8 For the matrix A given by `[[0 0 0 0];eye(2,4)]`, determine the following items: (a) the main diagonal; (b) the second column; (c) the third row; (d) A_{32}; (e) A^{t}; (f) A^{c}. (g) Is A lower or upper triangular?

Lists of lists

Matrices are often used to record or represent a list $f = (f_1, f_2, \ldots, f_n)$ in which all the items f_j are themselves lists. This can always be done if all the items f_j in that list have the same length, i.e., for some m and all j, $\#f_j = m$. Further, it can be done in two ways, by columns or by rows.

Offhand, it seems more natural to think of a matrix as a list of its rows, particularly since we are used to writing things in rows from left to right, each new row underneath the previous row. Nevertheless, in this book, it will always be done by column (the Chinese and Japanese way or the way the ancient Law Code of Hammurabi was written), i.e., the list $(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n)$ of m-vectors will be associated with the $m \times n$-matrix A whose jth column is $\mathbf{a}_j$, all j. We write this fact in this way:

$$A = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n]; \qquad \text{i.e., } A_{:j} = \mathbf{a}_j,\ j \in \underline{n}.$$

This makes it acceptable to denote by

$$\#A$$

the number of columns of the matrix A. If I need to refer to the number of rows of A, I will simply count the number of columns of its transpose, A^{t}, or its conjugate transpose, A^{c}, i.e., write

$$\#A^{\mathrm{t}} \quad \text{or} \quad \#A^{\mathrm{c}},$$

rather than introduce yet another notation.

Maps

(1.5) Definition: A **map**

$$f : X \to Y : x \mapsto f(x)$$

associates with each element x of its **domain** $\operatorname{dom} f := X$ a unique element $y = f(x)$, called the **value of** f **at** x, from its **target** $\operatorname{tar} f := Y$. If g is also a map, then $f = g$ means that $\operatorname{dom} f = \operatorname{dom} g$, $\operatorname{tar} f = \operatorname{tar} g$, and $f(x) = g(x)$ for all $x \in \operatorname{dom} f$.

The collection

$$\operatorname{ran} f := \{f(x) : x \in X\}$$

of all values taken by f is called the **range of** f. More generally, for any subset Z of X,

$$fZ := f(Z) := \{f(z) : z \in Z\}$$

is called the **image of** Z **under** f. In these terms,

$$\operatorname{ran} f = f(\operatorname{dom} f).$$

Also, for any $U \subset Y$, the set

$$f^{-1}U := \{x \in X : f(x) \in U\}$$

is called the **pre-image of** U **under** f. The collection of all maps from X to Y is denoted by

$$Y^X \quad .$$

Names other than **map** are in use, such as **function**, **transformation**, **mapping**, **morphism**, **operator**, etc., all longer than 'map'. I have acknowledged the standard name, 'function', by using the letter f for the map appearing in the above definition. However, in this book, I reserve **function** for a scalar-valued map. A map with domain $\mathbb{N}$ is a called a **sequence**.

Somewhat confusingly, many mathematicians use the term 'range' for what I have called here 'target'; the same mathematicians use the term **image** for what I have called here 'range'.

Every map $f : X \to Y$ gives rise to an assignment on X, namely the assignment $(f(x) : x \in X)$. On the other hand, an assignment f on X gives rise to *many* maps, one for each Y that contains $\operatorname{ran} f$, by the prescription

$$f^{|Y} : X \to Y : x \mapsto f_x.$$

We call this the **map into** Y **given by the assignment** f.

If X is empty, then Y^X consists of exactly one element, namely the map given by the empty assignment, and this even holds if Y is empty.

However, if Y is empty and X is not, then there can be no map from X to Y, since any such map would have to associate with each $x \in X$ some $y \in Y$, yet there are no $y \in Y$ to associate with.

"Wait a minute!", you now say, "How did we manage when X was empty?" Well, if X is empty, then there is no $x \in X$, hence the question of what element of Y to associate with never comes up. Isn't Mathematics slick?

1.9 Which of the following lists of pairs describes a map from {o,u,i,a} to {t,h,s}? A: ((u,s), (i,s), (a,t), (o,h), (i,s)); B: ((i,t), (a,s), (o,h), (i,s), (u,s)); C: ((a,s), (i,t), (u,h), (a,s), (i,t)).

1.10 For each of the following `MATLAB` maps, determine their range, as maps on real 2-by-3 matrices: (a) `A` ↦ `max(A)`; (b) `A` ↦ `A_{:2}`; (c) `A` ↦ `diag(A)`; (d) `A` ↦ `size(A)`; (e) `A` ↦ `length(A)`; (f) `A` ↦ `cos(A)`; (g) `A` ↦ `ones(A)`; (h) `A` ↦ `sum(A)`.

1.11 The **characteristic function** χ_S of the subset S of the set T is, by definition, the function on T that is 1 on S and 0 otherwise:

$$\chi_S : T \to \{0,1\} : t \mapsto \begin{cases} 1, & \text{if } t \in S; \\ 0, & \text{otherwise.} \end{cases}$$

Let R and S be subsets of T. Prove that (a) $\chi_{R\cup S} = \max(\chi_R, \chi_S)$; (b) $\chi_{R\cap S} = \min(\chi_R, \chi_S) = \chi_R\chi_S$; (c) $\chi_{S\backslash R} = (1-\chi_R)\chi_S$; (d) $R \subset S$ iff $\chi_R \le \chi_S$.

1.12 Let $f : T \to U$, and consider the map from subsets of U to subsets of T given by the rule

$$R \mapsto f^{-1}R := \{t \in T : f(t) \in R\}.$$

Prove that *this map commutes with the set operations of union, intersection and 'take away'*, i.e., for any subsets R and S of U, (a) $f^{-1}(R\cup S) = (f^{-1}R)\cup(f^{-1}S)$; (b) $f^{-1}(R\cap S) = (f^{-1}R)\cap(f^{-1}S)$; (c) $f^{-1}(R\backslash S) = (f^{-1}R)\backslash(f^{-1}S)$.

1-1 and onto

In effect, a map is an assignment together with a target, with the target necessarily containing the range of the assignment. A major reason for introducing the concept of *map* (as distinct from the notion of *assignment*) is in order to raise the following basic question:

Given the map $f : X \to Y$ and $y \in Y$, find $x \in X$ for which $f(x) = y$, i.e., solve the equation

$$f(?) = y. \tag{1.6}$$

Existence occurs if this equation has a solution for every $y \in Y$, i.e., if $\operatorname{ran} f = \operatorname{tar} f$. **Uniqueness** occurs if there is at most one solution for every $y \in Y$, i.e., if $f(x) = f(z)$ implies that $x = z$, i.e., the assignment $(f(x) : x \in X)$ is 1-1.

Here are the corresponding map properties:

(1.7) Definition: The map $f : X \to Y$ is **onto** in case $\operatorname{ran} f = Y$.

(1.8) Definition: The map $f : X \to Y$ is **1-1** in case $f(x) = f(y)$ implies $x = y$.

Not surprisingly, these two map properties play a major role throughout this book. (At last count, '1-1' appears over 400 times in this book, and 'onto' over 270 times.) – There are other names in use for these properties: An onto map is also called **surjective** or **epimorph(ic)**, while a 1-1 map is also called **injective** or **monomorph(ic)** or **faithful**.

You are, of course, familiar with maps in an atlas, or maps used for travel. These endeavor to associate with each point in their domain (usually a rectangle) some point on the earth's surface in a "continuous" 1-1 manner.

Cardinality and the pigeonhole principle

Perhaps the simplest useful examples of maps are those derived from lists, i.e., maps from some $\underline{n}$ into some set Y. Here is the basic observation concerning such maps being 1-1 or onto.

(1.9) Lemma If $g : \underline{n} \to Y$ is 1-1 and $f : \underline{m} \to Y$ is onto, then $n \le m$, with equality if and only if g is also onto and f is also 1-1.

Proof: The list $(f(1), \ldots, f(m))$ contains every element of Y, but may also contain duplicates of some. Throw out all duplicates to arrive at the list $(h(1), \ldots, h(q))$ which still contains all elements of Y but each one only once. In effect, we have 'thinned' f to a map $h : \underline{q} \to Y$ that is still onto but is also 1-1. In particular, $q \le m$, with equality if and only if there were no duplicates, i.e., f is also 1-1.

Now remove from the list $(h(1), \ldots, h(q))$ every entry of the list $(g(1), \ldots, g(n))$. Since h is onto and 1-1, each of the n distinct entries $g(j)$ does appear in h's list exactly once, hence the remaining list $(k(1), \ldots, k(r))$ has length $r = q - n$. Thus, $n \le q$, with equality, i.e., with $r = 0$, if and only if g is onto. In any case, the concatenation $(g(1), \ldots, g(n), k(1), \ldots, k(r))$ provides an 'extension' of the 1-1 map g to a map to Y that is still 1-1 but is also onto.

Put the two arguments together to get that $n \le q \le m$, with equality if and only if f is also 1-1 and g is also onto. □

Note the particular conclusion that if both $g : \underline{n} \to Y$ and $f : \underline{m} \to Y$ are 1-1 and onto, then necessarily $n = m$. This number is called the **cardinality** of Y and, as noted earlier, is denoted

$$\#Y.$$

Hence, if we know that $\#Y = n$, i.e., that there is some invertible map from $\underline{n}$ to Y, then we know that any map $f : \underline{n} \to Y$ is onto if and only if it is 1-1. This is the

(1.10) Pigeonhole Principle: If $f : \underline{n} \to Y$ with $\#Y = n$, then f is 1-1 if and only if f is onto.

Any map from $\underline{n}$ to $\underline{n}$ that is 1-1 hence onto is called a **permutation of degree** n since its list is a reordering of the first n integers. Thus $(3, 2, 1)$ or $(3, 1, 2)$ are permutations of degree 3 while the map into $\underline{3}$ given by the 3-vector $(3, 3, 1)$ is not a permutation, as it is neither 1-1 nor onto.

By the pigeonhole principle, in order to check whether an n-vector represents a permutation, we only have to check whether its range is $\underline{n}$ (which would mean that it is onto, as a map into $\underline{n}$), or we only have to check whether all its values are different and in $\underline{n}$ (which would mean that it is a 1-1 map into its domain, $\underline{n}$).

The finiteness of $\underline{n}$ is essential here. For example, consider the **right shift**

$$r : \mathbb{N} \to \mathbb{N} : n \mapsto n + 1. \tag{1.11}$$

This maps different numbers to different numbers, i.e., is 1-1, but fails to be onto since the number 1 is not in its range. On the other hand, the **left shift**

$$l : \mathbb{N} \to \mathbb{N} : n \mapsto \max\{n - 1, 1\} \tag{1.12}$$

is onto, but fails to be 1-1 since it maps both 1 and 2 to 1.

In light of this example, it is all the more impressive that such a pigeonhole principle continues to hold for certain special maps $f : X \to Y$ with both X and Y infinite. Specifically, according to (3.24)Corollary, if X and Y are *vector spaces* of the same finite *dimension* and $f : X \to Y$ is a *linear* map, then f is 1-1 if and only f is onto. This result is one of the high points of basic linear algebra. A more down-to-earth formulation of it, as in (3.26)Theorem, is the following: *A linear system with as many equations as unknowns has a solution for every right-hand side if and only if it has only the trivial solution when the right-hand side is* $\mathbf{0}$.

1.13 Prove: *If X and Y are finite sets, then* $\#(Y^X) = (\#Y)^{\#X}$.

1.14 Prove: *Any* $g : \underline{n} \to Y$ *with* $n > \#Y$ *cannot be 1-1.*

1.15 Prove: *Any* $f : \underline{m} \to Y$ *with* $m < \#Y$ *cannot be onto.*

1.16* Let $g : \underline{n} \to Y$ be 1-1, and $f : \underline{m} \to Y$ be onto. Prove that

(i) *for some* $k \ge n$, g *can be 'extended' to a map* $h : \underline{k} \to Y$ *that is 1-1 and onto*;

(ii) *for some* $k \le m$, f *can be 'thinned' to a map* $h : \underline{k} \to Y$ *that is onto and 1-1.*

1.17* Prove: *If T is finite and $S \subset T$, then S is finite, too.* (Hint: consider the set N of all $n \in \mathbb{N} \cup \{0\}$ for which there is a 1-1 map $g : \underline{n} \to S$.)

1.18 Prove that *$S \subset T$ and $\#T < \infty$ implies that $\#S \le \#T$, with equality if and only if $S = T$.*

1.19* Prove the **Chinese Remainder Theorem**: *For $p \in \mathbb{Z}$ and $q \in \mathbb{N}$, let* $\mathrm{rem}(p, q)$ *be the remainder of the division of p by q, i.e., the unique integer* $r \in R_q := \{0, 1, \ldots, q-1\}$ *for which $p - r$ is divisible by q. Then, for any $k \in \mathbb{N}$, and any $\mathbf{q} \in \mathbb{N}^k$, $\mathbf{r} \in \mathbb{Z}^k$ with $r_i \in R_{q_i}$, $i = 1{:}k$, and any two entries of $\mathbf{q}$ relatively prime, there exists a unique $m \in R_q$ with $q := \prod_i q_i$ for which* $\mathrm{rem}(m, q_i) = r_i$, *all i.* (Hint: Prove that the map $f : R_q \to R_{q_1} \times \cdots \times R_{q_k} : m \mapsto (\mathrm{rem}(m, q_i) : i = 1{:}k)$ is 1-1, hence onto.)

Some examples

The next simplest maps after those given by lists are probably those that come to you in the form of a *list of pairs.* For example, at the end of the semester, I am forced to make up a grade map. The authorities send me the domain of that map, namely the students in this class, in the form of a list, and ask me to assign, to each student, a grade, thus making up a list of pairs of the form

name | grade

At my university, the target of the grade map is the set

$$\{\text{A, AB, B, BC, C, D, F, I}\},$$

but there is no requirement to make this map onto. In fact, I could not meet that requirement if there were fewer than 8 students in the class. Neither is it required to make the grade map 1-1. In fact, it is not possible to make the grade map 1-1 if the class has more than 8 students in it. But if the class has exactly 8 students in it, then a grade map that is onto is automatically also 1-1, and a grade map that is 1-1 is automatically also onto.

There are many maps in your life that are given as a list of pairs, such as the list of dorm-room assignments or the price list in the cafeteria. The dorm-room assignment list usually has the set of students wanting a dorm room as its domain and the set of available dorm rooms as its target, is typically not 1-1, but the authorities would like it to be onto. The price list at the cafeteria has all the items for sale as its domain, and the set $\mathbb{N}/100 := \{m/100 : m \in \mathbb{N}\}$ of all positive reals with at most two digits after the decimal point as its target. There is little sense in wondering whether this map is 1-1 or onto.

1.20 Describe an interesting map (not already discussed in class) that you have made use of in the last month or so (or, if nothing comes to mind, a map that someone like you might have used recently). Be sure to include domain and target of your map in your description and state whether or not it is 1-1, onto.

Maps and their graphs

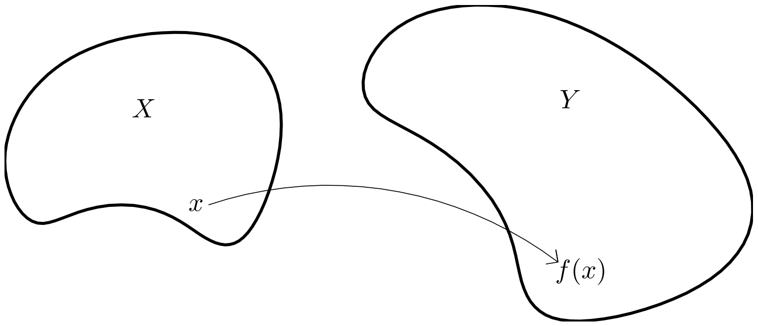


Figure. One way to visualize the map $f : X \to Y : x \mapsto f(x)$.

One successful mental image of a 'map' is to imagine both domain and target as sets of some possibly indistinct shape, with curved arrows indicating with which particular element in the target the map f associates a particular element in the domain. Another successful mental (and more successful mathematical) image of a map $f : X \to Y$ is in terms of its **graph**, i.e., in terms of the set of pairs

$$\{(x, f(x)) : x \in X\}.$$

In fact, the mathematically most satisfying definition of 'map from X to Y' is: *a subset of $X \times Y$ that, for each $x \in X$, contains exactly one pair (x, y)*. In this view, *a map is its graph*.

Here, for example, is the (graph of the) grade map G for a graduate course I taught recently. I abbreviated the students' names, to protect the innocent.

A				●	●	●		●		●	●
AB							●		●		
B	●										
BC		●									
C											
D											
F											
I			●								
	NA	SC	SG	AK	TK	AM	JP	DS	ST	TW	ZH

Figure. The graph of the grade map

You may be more familiar with the graphs of real functions, such as the 'squaring' map

$$()^2 : [0 \mathinner{..} 2] \to [0 \mathinner{..} 4] : x \mapsto x^2,$$

whose graph is shown in the next figure. The arrows indicate the solution process for the equation $y = f(?)$. Reflection across the bisector drawn interchanges X and Y, and changes the graph of f to the graph of f^{-1}.

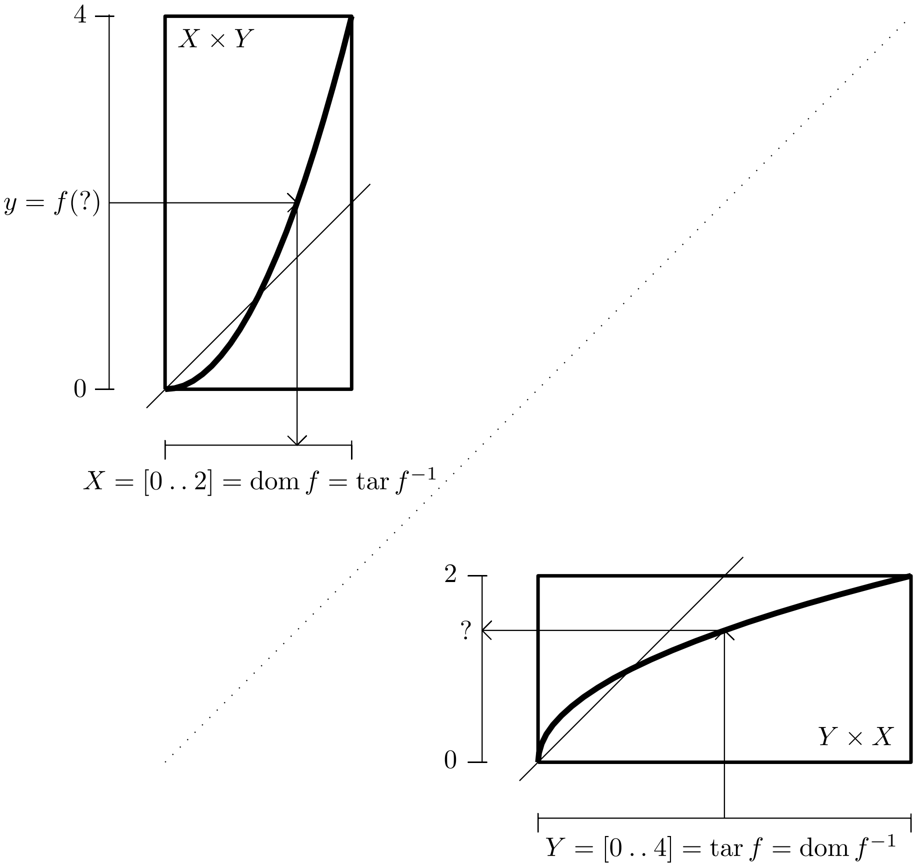


Figure. The graph of the squaring map $f := ()^2 : [0 \mathinner{..} 2] \to [0 \mathinner{..} 4] : x \mapsto x^2$ and of its inverse $f^{-1} = \sqrt{\ } : [0 \mathinner{..} 4] \to [0 \mathinner{..} 2] : x \mapsto \sqrt{x}$.

1.21 For each of the following subsets R of the cartesian product $X \times Y$ with $X = [0 \mathinner{..} 2]$ and $Y = [0 \mathinner{..} 4]$, determine whether it is the graph of a map from X to Y and, if it is, whether that map is 1-1 and/or onto or neither.

(a) $R = \{(x, y) : y = (x - 1/2)^2\}$; (b) $R = \{(x, y) : x \geq 1, y = (2x - 2)^2\}$; (c) $R = \{(x, y) : y = (2x - 2)^2\}$; (d) $R = \{(x, y) : x = y\}$.

1.22 Same as previous problem, but with X and Y interchanged and, correspondingly, R replaced by $R^{-1} := \{(y, x) \in Y \times X : (x, y) \in R\}$. Also, discuss any connections you see between the answers in these two problems.

Invertibility

The graph of a map f helps us solve the standard 'computational' problem involving maps, namely the problem of finding an $x \in X$ that solves the equation

$$f(?) = y$$

for given $f : X \to Y$ and $y \in Y$. The solution set is the pre-image of $\{y\}$ under f, i.e., the set

$$f^{-1}\{y\} = \{x \in X : f(x) = y\}.$$

For example, when looking at the graph of the above grade map G, we see that $G^{-1}\{\mathrm{AB}\} = \{\mathrm{JP},\ \mathrm{ST}\}$, while $G^{-1}\{\mathrm{D}\} = \{\}$ (the empty set). In the first case, we have two solutions, in the second case, we have none.

In effect, when looking for solutions to the equation $f(?) = y$, we are looking at the graph of f with the roles of domain and target interchanged: We are trying to associate with each $y \in Y$ some $x \in X$ in such a way that $f(x) = y$. If f is onto, then there is *at least* one solution for every $y \in Y$, and conversely (**existence**). If f is 1-1, then there is *at most* one solution for any $y \in Y$, and conversely (**uniqueness**). Ideally, there is, for each $y \in Y$, exactly one $x \in X$ for which $f(x) = y$.

(1.13) Definition: The map $f : X \to Y$ is **invertible** := for every $y \in Y$ there exists exactly one $x \in X$ for which $f(x) = y$.

(1.14) Let $f : X \to Y$.

f is invertible if and only if f is 1-1 and onto.

f is invertible if and only if the **inverse of its graph**, i.e., the set

$$\{(f(x), x) : x \in X\} \subset Y \times X,$$

is the graph of a map from Y to X. This latter map is called the **inverse of** f and is denoted by f^{-1}.

Any 1-1 assignment f, taken as a map into its range, is invertible, since it is both 1-1 and onto. The above grade map G fails on both counts to be invertible, it is neither 1-1 nor onto. The squaring map $()^2 : [0\,..\,2] \to [0\,..\,4] : x \mapsto x^2$, on the other hand, is invertible since it is both 1-1 and onto. The figure on page 15 shows the graph of its inverse, obtained from the graph of the squaring map by reversing the roles of domain and target. In effect, we obtain the inverse of the graph of f by looking at the graph of f sideways and can often tell at a glance whether or not it is the graph of a map, i.e., whether f is 1-1 and onto.

A map may be 'half' invertible, i.e., it may be either 1-1 or onto, without being both. For example, the right shift (1.11) is 1-1, but not onto, while the left shift (1.12) is onto, but not 1-1. Only if domain and target happen to have the same *finite* number of elements, then being 1-1 is guaranteed to be the same as being onto, by the pigeonhole principle (see Problem 1.36).

(1.15) If $f : X \to Y$, with $\#X = \#Y < \infty$, then f 1-1 *or* onto implies f 1-1 *and* onto, i.e., invertible.

In particular, for any *finite* X, any map $f : X \to X$ that is 1-1 *or* onto is automatically invertible.

Map composition; left and right inverse

The notion of f being 'half' invertible is made precise by the notions of left and right inverse. Their definition requires the **identity map**, often written

$$\mathrm{id}$$

if its domain (which is also its target) is clear from the context. The full definition is:

$$\mathrm{id}_X : X \to X : x \mapsto x.$$

In other words, the identity map is a particularly boring map, it leaves everything unchanged.

We also need **map composition**:

(1.16) Definition: The **composition** $f \circ g$ of two maps $f : X \to Y$ and $g : U \to W \subset X$ is the map

$$f \circ g : U \to Y : u \mapsto f(g(u)).$$

We write fg instead of $f \circ g$ whenever there is no danger of confusion. Map composition is **associative**, i.e., whenever fg and gh are defined, then

$$(fg)h = f \circ (gh).$$

There is a corresponding definition for the composition $x \circ y$ of two assignments, x and y, under the assumption that $\operatorname{ran} y \subset \operatorname{dom} x$. Thus,

$$x_y := x \circ y = (x_{y_i} : i \in \operatorname{dom} y)$$

is an assignment whose domain is $\operatorname{dom} y$ and whose range is contained in $\operatorname{ran} x$.

As a simple *example*, if $\mathbf{x}$ is an n-vector and $\mathbf{y}$ is an m-vector with $\operatorname{ran}\mathbf{y} \subset \underline{n} = \{1, \ldots, n\}$, then

$$x_{\mathbf{y}} := \mathbf{x} \circ \mathbf{y} = (x_{y_1}, \ldots, x_{y_m}). \tag{1.17}$$

In MATLAB, if x describes the n-vector $\mathbf{x}$ and y describes the m-vector $\mathbf{y}$ with entries in $\underline{n} = \{1, \ldots, n\}$, then z=x(y) describes the m-vector $\mathbf{z} = x_{\mathbf{y}} = \mathbf{x} \circ \mathbf{y}$.

In the same way, if $\mathtt{A} \in \mathbb{F}^{m \times n}$, and b is a k-list with entries from $\underline{m} = \{1, \ldots, m\}$, and c is an l-list with entries from $\underline{n} = \{1, \ldots, n\}$, then $\mathtt{A}(\mathtt{b}, \mathtt{c})$ is a $k \times l$-matrix, namely the matrix $\mathtt{D} := \mathtt{A}(\mathtt{b}, \mathtt{c}) \in \mathbb{F}^{k \times l}$ with

$$\mathtt{D}(i, j) = \mathtt{A}(\mathtt{b}(i), \mathtt{c}(j)), \quad i \in \underline{k}, j \in \underline{l}.$$

In effect, the matrix $\mathtt{D} = \mathtt{A}(\mathtt{b}, \mathtt{c})$ is obtained by choosing from the matrix A the rows $\mathtt{b}(1), \mathtt{b}(2), \ldots, \mathtt{b}(k)$ and columns $\mathtt{c}(1), \mathtt{c}(2), \ldots, \mathtt{c}(l)$ of A, in that order.

If *all* rows, in their natural order, are to be chosen, then use A(:,c). If *all* columns, in their natural order, are to be chosen, then use A(b,:).

In particular, A(1,:) is the matrix having the first row of A as its sole row, and A(:,end) is the matrix having the last column of A as its sole column. The matrix A(1:2:end,:) is made up from all the odd rows of A. A(end:-1:1,:) is the matrix obtained from A by reversing the order of the rows (as could also be obtained by the command flipud(A)). A(:,2:2:end) is obtained by removing from A all odd-numbered columns. If x is a one-row matrix, then x(ones(1,m),:) and x(ones(m,1),:) both give the matrix having all its m rows equal to the single row in x (as would the expression repmat(x,m,1)).

MATLAB permits the expression A(b,c) to appear on the *left* of the equality sign: If A(b,c) and D are matrices of the same size, then the statement

```
A(b,c) = D;
```

changes, for each $(\mathtt{i,j}) \in \operatorname{dom} \mathtt{D}$, the entry A(b(i),c(j)) of A to the value of D(i,j). What if, e.g., b is not 1-1? MATLAB does the replacement for each entry of b, from the first to the last. Hence, the last time is the one that sticks. For example, if a=1:4, then the statement a([2,2,2])=[1,2,3] changes a to [1,3,3,4]. On the other hand, if A appears on both sides of such an assignment, then the one on the right is taken to be as it is at the outset of that assignment. For example,

```
A([i,j],:) = A([j,i],:);
```

is a slick way to interchange the ith row of A with its jth.

As a first use of map composition, here are surprisingly useful sufficient conditions for a map f being onto or a map g being 1-1.

(1.18) If fg is onto, then f is onto; if fg is 1-1, then g is 1-1.

Proof: Since $\operatorname{ran}(fg) \subset \operatorname{ran} f \subset \operatorname{tar} f = \operatorname{tar} fg$, fg onto implies f onto. Also, if $g(y) = g(z)$, then $(fg)(y) = (fg)(z)$, hence fg 1-1 implies $y = z$, i.e., g is 1-1. □

For example, the composition lr of the left shift (1.12) with the right shift (1.11) is the identity, hence l is onto and r is 1-1 (as observed earlier).

(1.19) Definition: If $f \in Y^X$ and $g \in X^Y$ and $fg = \text{id}$, then f (being to the left of g) is a **left inverse** of g, and g is a **right inverse** of f. In particular, any left inverse is onto and any right inverse is 1-1.

To help you remember which of f and g is onto and which is 1-1 in case $fg = \text{id}$, keep in mind that being onto provides conclusions about elements of the target of the map while being 1-1 provides conclusions about elements in the domain of the map.

Now we consider the converse statements.

(1.20) If $f : X \to Y$ is 1-1, then f has a left inverse.

Proof: If f is 1-1 and $x \in X$ is some element, then

$$g : Y \to X : y \mapsto \begin{cases} f^{-1}\{y\} & \text{if } y \in \operatorname{ran} f; \\ x & \text{otherwise}, \end{cases}$$

is well-defined since each $y \in \operatorname{ran} f$ is the image of exactly one element of X. With g so defined, $gf = \text{id}$ follows. (What if Y is empty?) □

The corresponding statement: *If $f : X \to Y$ is onto, then f has a right inverse* would have the following ‘proof’: Since f is onto, we can define $g : Y \to X : y \mapsto$ some point in $f^{-1}\{y\}$. Regardless of how we pick that point $g(y) \in f^{-1}\{y\}$, the resulting map is a right inverse for f. – Some object to this argument since it requires us to pick, for each y, a particular element from that set $f^{-1}\{y\}$. The **belief** that this can *always* be done is known as “The Axiom of Choice”.

(1.21) If f is an invertible map, then f^{-1} is both a right inverse and a left inverse for f. Conversely, if g is a right inverse for f and h is a left inverse for f, then f is invertible and $h = f^{-1} = g$.
Consequently, if f is invertible, then: (i) f^{-1} is also invertible, and $(f^{-1})^{-1} = f$; and, (ii) if also g is an invertible map, with $\operatorname{tar} g = \operatorname{dom} f$, then fg is invertible, and $(fg)^{-1} = g^{-1}f^{-1}$ (note the order reversal).

Proof: Let $f : X \to Y$ be invertible. Since, for every $y \in Y$, $f^{-1}(y)$ solves the equation $f(?) = y$, we have $ff^{-1} = \mathrm{id}_Y$, while, for any $x \in X$, x is a solution of the equation $f(?) = f(x)$, hence necessarily $x = f^{-1}(f(x))$, thus also $f^{-1}f = \mathrm{id}_X$.

As to the converse, if f has both a left inverse h and a right inverse g, then it must be both 1-1 and onto, hence invertible. Further, since $hf = \mathrm{id}_X$ and $fg = \mathrm{id}_Y$, then (using the associativity of map composition),

$$h = h\,\mathrm{id}_Y = h \circ (fg) = (hf)g = \mathrm{id}_X g = g,$$

showing that $h = g$, hence $h = f^{-1} = g$.

As to the consequences, the identities $ff^{-1} = \mathrm{id}_Y$ and $f^{-1}f = \mathrm{id}_X$ explicitly identify f as a right and left inverse for f^{-1}, hence f must be the inverse of f^{-1}. Also, by map associativity, $(fg)g^{-1}f^{-1} = f\,\mathrm{id}_X f^{-1} = ff^{-1} = \mathrm{id}_Y$, etc. . □

While $fg = \mathrm{id}$ implies $gf = \mathrm{id}$ in general only in case $\#\operatorname{dom} f = \#\operatorname{tar} f < \infty$, it does imply that gf is as much of an identity map as it can be: Indeed, if $fg = \mathrm{id}$, then $(gf)g = g \circ (fg) = g\,\mathrm{id} = g$, showing that $(gf)(x) = x$ for every $x \in \operatorname{ran} g$. There is no such hope for $x \notin \operatorname{ran} g$, since such x cannot possibly be in $\operatorname{ran} gf = g(\operatorname{ran} f) \subset \operatorname{ran} g$. However, since $(gf)(x) = x$ for all $x \in \operatorname{ran} g$, we conclude that $\operatorname{ran} gf = \operatorname{ran} g$. This makes gf the identity on its range, $\operatorname{ran} g$. In particular, $(gf) \circ (gf) = gf$, i.e., gf is **idempotent** or, a **projector**.

(1.22) Proposition: If $f : X \to Y$ and $fg = \mathrm{id}_Y$, then gf is a projector, i.e., the identity on its range, and that range equals $\operatorname{ran} g$.

For example, the composition lr of the left shift (1.12) with the right shift (1.11) is the identity, hence rl must be the identity on $\operatorname{ran} r = \{2, 3, \ldots\}$ and, indeed, it is.

If the n-vector c in MATLAB describes a permutation, i.e., if the map c: $\underline{n} \to \underline{n} : j \mapsto$ c(j) is 1-1 or onto, hence invertible, then the n-vector cinv giving its inverse can be obtained with the commands

```
cinv = c; cinv(c) = 1:length(c);
```

The first command makes sure that cinv starts out as a vector of the same size as c. With that, the second command changes cinv into one for which cinv(c) = [1,2,...,length(c)]. In other words, cinv describes a *left* inverse for (the map given by) c, hence the inverse (by the pigeonhole principle).

A second, more expensive, way to construct cinv is with the help of the command sort, as follows:

```
[d, cinv] = sort(c);
```

For, whether or not c describes a permutation, this command produces, in the n-vector d, the list of the items in c in nondecreasing order, and provides, in cinv, the recipe for this re-ordering:

$$\texttt{d(i)=c(cinv(i))}, \quad \texttt{i}= 1{:}n.$$

In particular, if c describes a permutation, then, necessarily, d = [1,2,3,...], therefore c(cinv) = [1,2,...,length(c)], showing that cinv describes a *right* inverse for (the map given by) c, hence the inverse (by the pigeonhole principle).

Both of these methods extend, to the construction of a left, respectively a right, inverse, in case the map given by c has only a left, respectively a right, inverse.

1.23 Let $f : \underline{2} \to \underline{3}$ be given by the list $(2,3)$, and let $g : \underline{3} \to \underline{2}$ be the map given by the list $(2,1,2)$.

(a) Describe fg and gf (e.g., by giving their lists).

(b) Verify that fg is a projector, i.e., is the identity on its range.

1.24 For each of the following maps, state whether or not it is 1-1, onto, invertible. Also, describe a right inverse or a left inverse or an inverse for it or else state why such right inverse or left inverse or inverse does not exist.

The maps are specified in various ways, e.g., by giving their list and their target or by giving both domain and target and a rule for constructing their values.

(a) a is the map to $\{1,2,3\}$ given by the list $(1,2,3)$.

(b) b is the map to $\{1,2,3,4\}$ given by the list $(1,2,3)$.

(c) c is the map to $\{1,2\}$ given by the list $(1,2,1)$.

(d) $d : \mathbb{R}^2 \to \mathbb{R} : \mathbf{x} \mapsto 2x_1 - 3x_2$.

(e) $f : \mathbb{R}^2 \to \mathbb{R}^2 : \mathbf{x} \mapsto (-x_2, x_1)$.

(f) $g : \mathbb{R}^2 \to \mathbb{R}^2 : \mathbf{x} \mapsto (x_1 + 2, x_2 - 3)$.

(g) $h : \mathbb{R} \to \mathbb{R}^2 : y \mapsto (y/2, 0)$.

1.25 Verify that, in the preceding problem, $dh = \mathrm{id}$, and explain geometrically why one would call hd a projector.

1.26 Prove: *If $fg = fh$ for $g, h : S \to T$ and with $f : T \to U$ 1-1, then $g = h$.*

1.27 Prove: *If $fh = gh$ for $f, g : T \to U$ and with $h : S \to T$ onto, then $f = g$.*

1.28* Use the preceding two problems to prove the following converse of (1.22)Proposition: *If $f : X \to Y$ and gf is a projector, then f is onto and g is 1-1 iff $fg = \mathrm{id}_Y$.*

1.29 If both f and g are maps from $\underline{n}$ to $\underline{n}$, then so are both fg and gf. In particular, for any $f \in \underline{n}^{\underline{n}}$, its power sequence

$$f^0 := \mathrm{id}_{\underline{n}}, f^1 := f, f^2 := f \circ f, f^3 := f \circ f^2, \ldots$$

is well defined. Further, since $\underline{n}^{\underline{n}}$ is finite, the sequence $f^0, f^1, f^2, \ldots$ of powers must eventually repeat itself. In other words, there must be a first r such that $f^r = f^j$ for some $j < r$. Let's call the difference $d := r - j$ between these two exponents the **cycle length** of f.

(a) Find the cycle length for the map given by the list $(2, 3, 4, 1, 1)$. (Feel free to use `MATLAB`.)

(b) Also determine the cycle lengths for the following maps:

A:=(2,3,4,5,1);	B:=(2,3,1,5,4);	C:=(1,2,3,4,5);
D:=(2,5,2,2,1);	E:=(2,5,2,5,2);	F:=(2,5,2,2,5).

(c) Given all these examples (and any others you care to try), what is your *guess* as to the special nature of the map f^d in case the cycle length of f is d and f is invertible? What if f is not invertible?

1.30 Finish appropriately the following `MATLAB` function

```
function b = ii(a)
% If  ran(a) = N := {1,2,...,length(a)} , hence  a  describes
% the invertible map
%             f:N --> N : j |--> a(j)
% then  b  describes the inverse of  f , i.e., the map  g:N --> N  for which
%  fg = id_N  and  gf = id_N .
% Otherwise, the message
%  The input doesn't describe an invertible map
% is printed and an empty  b  is returned.
```

1.31 Let $f_i : X \to X$ for $i \in \underline{n}$, hence $g := f_1 \cdots f_n$ is also a map from X to X. Prove that *g is invertible if, but not only if, each f_i is invertible, and, in that case, $g^{-1} = f_n^{-1} \cdots f_1^{-1}$.* (Note the order reversal!)

1.32 Prove: *If $f : S \to T$ is invertible, then f has exactly one left inverse.* Is the converse true?

1.33 Let g be a left inverse for $f : S \to T$, and assume that $\#S > 1$. Prove that *g is the unique left inverse for f iff g is 1-1.* (Is the assumption that $\#S > 1$ really needed?)

1.34 Let g be a right inverse for f. Prove that *g is the unique right inverse for f iff g is onto.*

1.35 Prove: *If $f : S \to T$ is invertible, then f has exactly one right inverse.* Is the converse true?

1.36*

(i) Prove: *If $g : Z \to X$ is invertible, then, for any $f : X \to Y$, f is 1-1 (onto) if and only if the map fg is 1-1 (onto).*

(ii) Derive (1.15) from (1.10).

The inversion of maps

The notions of 1-1 and onto, and the corresponding notions of right and left inverse, are basic to the discussion of the standard 'computational' problem already mentioned earlier: for $f : X \to Y$ and $y \in Y$, solve

$$f(?) = y. \tag{1.6}$$

When we try to solve (1.6), we are really trying to find, for each $y \in Y$, some $x \in X$ for which $f(x) = y$, i.e., we are trying to come up with a right inverse for f. *Existence* of a solution for every right side is the same as having f *onto*, and is ensured by the existence of a right inverse for f. Existence of a left inverse for f ensures *uniqueness*: If $hf = \mathrm{id}$, then $f(x) = f(y)$ implies that $x = h(f(x)) = h(f(y)) = y$. Thus existence of a left inverse implies that f is *1-1*. But existence of a left inverse does *not*, in general, provide a solution.

When f has its domain in $\mathbb{R}^n$ and and its target in $\mathbb{R}^m$, then we can think of solving (1.6) *numerically*. Under the best of circumstances, this still means that we must proceed by *approximation*. The solution is found as the limit of a sequence of solutions to *linear* equations, i.e., equations of the form $A? = \mathbf{y}$, with A a *linear* map. This is so because linear (algebraic) equations are the only kind of equations we can actually solve exactly (ignoring roundoff). This is one reason why Linear Algebra is so important. It provides the mathematical structures, namely vector spaces and linear maps, needed to deal efficiently with linear equations and, thereby, with other equations.

1.37 T/F

(a) 0 is a natural number.

(b) $\#\{3, 3, 3\} = 1$.

(c) $\#(3, 3, 3) = 3$.

(d) $(\{3, 1, 3, 2, 4\} \cap \{3, 5, 4\}) \cup \{3, 3\} = \{4, 3, 3, 3, 3\}$.

(e) If A, B are finite sets, then $\#(A \cup B) = \#A + \#B - \#(A \cap B)$.

(f) $\#\{\} = 1$.

(g) $\{3, 3, 1, 6\}\backslash\{3, 1\} = \{3, 6\}$.

(h) If $f : X \to X$ for some finite X, then f is 1-1 if and only if f is onto.

(i) The map $f : \underline{3} \to \underline{3}$ given by the list $(3, 1, 2)$ is invertible, and its inverse is given by the list $(2, 3, 1)$.

(j) The map $f : \underline{3} \to \underline{2}$ given by the list $(1, 2, 1)$ has a right inverse.

(k) If $U \subset \operatorname{tar} f$, then f maps $f^{-1}U$ onto U.

(l) The map f is invertible if and only if $\{(f(x), x) : x \in \operatorname{dom} f\}$ is the graph of a map.

(m) If $f, g \in X^X$ and $h := fg$ is invertible, then both f and g are invertible.

(n) The matrix $\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$ is diagonal.

(o) The matrix $\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ is upper triangular.

2 Vector spaces and linear maps

Vector spaces, especially spaces of functions

Linear algebra is concerned with vector spaces. These are sets on which two operations, *vector addition* and *multiplication by a scalar*, are defined in such a way that they satisfy various laws. Here they are, for the record:

(2.1) Definition: To say that X is a **linear space** (of **vector**s), or a **vector space**, over the commutative field (see (17.9)) $\mathbb{F}$ (of **scalar**s) means that there are two maps, (i) $X \times X \to X : (x, y) \mapsto x + y$ called **(vector) addition**; and (ii) $\mathbb{F} \times X \to X : (\alpha, x) \mapsto \alpha x =: x\alpha$ called **scalar multiplication**, which satisfy the following rules.

(a) X is a **commutative group with respect to addition**; i.e. (see (17.8)), addition

(a.1) is **associative**: $x + (y + z) = (x + y) + z$;

(a.2) is **commutative**: $x + y = y + x$;

(a.3) has **neutral** element: $\exists 0 \ \forall x, \quad x + 0 = x$;

(a.4) has **inverse**: $\forall x \ \exists y, \quad x + y = 0$.

(s) scalar multiplication is

(s.1) **associative**: $\alpha(\beta x) = (\alpha\beta)x$;

(s.2) **field-addition distributive**: $(\alpha + \beta)x = \alpha x + \beta x$;

(s.3) **vector-addition distributive**: $\alpha(x + y) = \alpha x + \alpha y$;

(s.4) **unitary**: $1x = x$.

Note that *a vector space cannot be empty* since, by condition (a.3), it must contain the neutral element or **zero vector**, 0. For simplicity, I denote the zero vector with the same symbol used for the zero scalar. In the same spirit, vector addition and scalar addition are both denoted by the plus sign. There is, offhand, no sense in adding a scalar to a vector. It is standard to denote the element $y \in X$ for which $x+y=0$ by $-x$ since such y is uniquely determined by the requirement that $x+y=0$. Also, the short-hand $y-x := y+(-x)$ is standard. For reasons to become clear, I often write $x\alpha$ for αx.

While the **scalar**s can come from some abstract field, we will only be interested in the real scalars $\mathbb{R}$ and the complex scalars $\mathbb{C}$. Also, from a practical point of view, the most important linear spaces consist of **function**s, i.e., of scalar-valued maps all on some common domain. This means that the typical vector space we will deal with is (a subset of) the collection of all maps $\mathbb{F}^T$ from some fixed domain T into the **scalar field** $\mathbb{F}$ (either $\mathbb{F}=\mathbb{R}$ or $\mathbb{F}=\mathbb{C}$), with **pointwise** addition and multiplication by scalars. Here is the definition:

(2.2) Definition of pointwise vector operations:

(a) The **sum** $f+g$ of $f, g \in \mathbb{F}^T$ is the function

$$f+g : T \to \mathbb{F} : t \mapsto f(t)+g(t).$$

(s) The **product** αf of the **scalar** $\alpha \in \mathbb{F}$ with the function $f \in \mathbb{F}^T$ is the function

$$\alpha f : T \to \mathbb{F} : t \mapsto \alpha f(t).$$

With respect to these operations, $\mathbb{F}^T$ is a vector space (over $\mathbb{F}$). In particular, the function

$$0 : T \to \mathbb{F} : t \mapsto 0$$

is the neutral element, or zero vector, and, for $f \in \mathbb{F}^T$,

$$-f : T \to \mathbb{F} : t \mapsto -f(t)$$

is the additive inverse for f.

Note that it is not possible to add two functions unless they have the same domain!

Standard examples include:

(i) $T = \underline{n}$, in which case we get n**-dimensional coordinate space**, customarily denoted $\mathbb{F}^n$, whose elements (vectors) we call n-vectors.

(ii) $T = \underline{m} \times \underline{n}$, in which case we get the space of m-by-n matrices, customarily denoted $\mathbb{F}^{m\times n}$.

(iii) $T = \mathbb{R}$, $\mathrm{I\!F} = \mathbb{R}$, in which case we get the space of all real-valued functions on the real line.

(iv) $T = \mathbb{R}^n$, $\mathrm{I\!F} = \mathbb{R}$, in which case we get the space of all real-valued functions of n real variables.

The most common way to get a vector space is as a *linear subspace*:

(2.3) Definition: A nonempty subset Y of a vector space X is a **linear subspace** (of X) in case it is closed under addition and multiplication by a scalar. This means that the two sets

$$Y + Y := \{y + z : y, z \in Y\} \quad\text{and}\quad \mathrm{I\!F}Y := \{\alpha y : \alpha \in \mathrm{I\!F}, y \in Y\}$$

are subsets of Y. Equivalently, $\mathrm{I\!F}Y + \mathrm{I\!F}Y \subset Y$.

(2.4) Proposition: A subset Y of a vector space X is a vector space (with respect to the same vector addition and multiplication by scalars) if and only if Y is a linear subspace (of X).

For a proof, see the answer to Problem 2.3.

Standard examples of vector spaces obtained as linear subspaces include:

(i) The **trivial space** $\{0\}$, consisting of the zero vector alone; it's a great space for testing one's understanding.

(ii) $\Pi_{\le k}$:= the set of all **polynomials of degree** $\le k$ as a subset of $\mathrm{I\!F}^{\mathbb{F}}$ (see page 42).

(iii) The set $C([a\,..\,b])$ of all **continuous functions** on the interval $[a\,..\,b]$ as a subset of $\mathrm{I\!F}^{[a..b]}$.

(iv) The set of all real symmetric matrices of order n as a subset of $\mathbb{R}^{n\times n}$.

(v) The set of all real-valued functions on $\mathbb{R}$ that vanish on some fixed set S.

(vi) The set $\mathrm{BL}_{\boldsymbol{\xi}} \subset C([\xi_1\,..\,\xi_{\ell+1}])$ of all **broken lines** on $[\xi_1\,..\,\xi_{\ell+1}]$ with **(interior) breaks** at $\xi_2 < \cdots < \xi_\ell$, meaning that each $f \in \mathrm{BL}_{\boldsymbol{\xi}}$ is continuous and agrees, on each interval $[\xi_i\,..\,\xi_{i+1}]$, with a straight line.

It is a good exercise to check that, according to the abstract definition of a vector space, any linear subspace of a vector space is again a vector space. Conversely, if a subset of a vector space is *not* closed under vector addition or under multiplication by scalars, then it cannot be a vector space (with respect to the given operations) since it violates the basic assumption that the sum of any two elements and the product of any scalar with any element is again an element of the space. (To be sure, the empty subset $\{\}$ of a vector space is vacuously closed under the two vector operations but fails to be a linear subspace since it fails to be nonempty.)

(2.5) Proposition: The sum, $Y+Z:=\{y+z : y\in Y, z\in Z\}$, and the intersection, $Y\cap Z$, of two linear subspaces, Y and Z, of a vector space are linear subspaces.

For a proof, see the answer to Problem 2.4.

We saw that pointwise addition and multiplication by a scalar makes the collection $\mathbb{F}^T$ of all maps from some set T to the scalars a vector space. The same argument shows that the collection X^T of all maps from some set T into a *vector space* X (over the scalar field $\mathbb{F}$) is a vector space under pointwise addition and multiplication by scalars. This means, explicitly, that we define the sum $f+g$ of $f,g\in X^T$ by

$$f+g : T\to X : t\mapsto f(t)+g(t)$$

and define the product αf of $f\in X^T$ with the scalar $\alpha\in\mathbb{F}$ by

$$\alpha f : T\to X : t\mapsto \alpha f(t).$$

Thus, we can generate from one vector space X many different vector spaces, namely all the linear subspaces of the vector space X^T, with T an arbitrary set.

2.1 For each of the following sets of real-valued assignments or maps, determine whether or not they form a vector space (with respect to pointwise addition and multiplication by scalars), and give a reason for your answer. (a) $\{\mathbf{x}\in\mathbb{R}^3 : x_1=4\}$; (b) $\{\mathbf{x}\in\mathbb{R}^3 : x_1=x_2\}$; (c) $\{\mathbf{x}\in\mathbb{R}^3 : 0\le x_j, j=1,2,3\}$; (d) $\{(0,0,0)\}$; (e) $\{\mathbf{x}\in\mathbb{R}^3 : \mathbf{x}\not\in\mathbb{R}^3\}$; (f) $C([0\,..\,2])$; (g) The collection of all 3×3 matrices with all diagonal entries equal to zero; (h) $\{(x,0) : x\in\mathbb{R}\}\cup\{(0,y) : y\in\mathbb{R}\}$.

2.2* Prove that, *for every x in the vector space X, $0x=0$ and $(-1)x=-x$.*

2.3* Provide a proof for (2.4)Proposition.

2.4* Provide a proof for (2.5)Proposition.

2.5 Prove that *the intersection of any collection of linear subspaces of a vector space is a linear subspace.*

2.6 Prove: *The union of two linear subspaces is a linear subspace if and only if one of them contains the other.*

2.7 Prove: *The finite union of linear subspaces is a linear subspace if and only if one of them contains all the others.* (Hint: reduce to the situation that no subspace is contained in the union of the other subspaces and, assuming this leaves you with at least two subspaces, take from each a point that is in none of the others and consider the straight line through these two points.)

Linear maps

(2.6) Definition: Let X, Y be vector spaces (over the same scalar field $\mathbb{F}$). The map $f : X \to Y$ is called **linear** if it is

(a) **additive**, i.e.,

$$\forall x, z \in X, \quad f(x+z) = f(x) + f(z);$$

and

(s) **homogeneous**, i.e.,

$$\forall (\alpha, x) \in \mathbb{F} \times X, \quad f(\alpha x) = \alpha f(x).$$

We denote the collection of all linear maps from X to Y by

$$L(X, Y), \quad \text{and set} \quad L(X) := L(X, X).$$

Many books call a linear map a **linear transformation** or a **linear operator**. It is customary to denote linear maps by capital letters. Further, if A is a linear map and $x \in \operatorname{dom} A$, then it is customary to write Ax instead of $A(x)$.

Examples: (i) If X is a linear subspace of $\mathbb{F}^T$, then, for every $t \in T$, the map

$$\delta_t : X \to \mathbb{F} : f \mapsto f(t)$$

of evaluation at t is linear since the vector operations are pointwise.

(ii) The map $D : C^{(1)}(\mathbb{R}) \to C(\mathbb{R}) : g \mapsto Dg$ that associates with each continuously differentiable function g its first derivative Dg is a linear map.

(iii) The map $C([a \mathinner{\ldotp\ldotp} b]) \to \mathbb{R} : g \mapsto \int_a^b g(t)\,\mathrm{d}t$ is linear.

(iv) Let $c := \{a \in \mathbb{F}^{\mathbb{N}} : \lim_{n\to\infty} a_n \text{ exists}\}$, i.e., c is the vector space of all convergent sequences. Then the map $c \to \mathbb{F} : a \mapsto \lim_{n\to\infty} a_n$ is linear.

These examples show that the basic operations in Calculus are linear. This is the reason why so many people outside Algebra, such as Analysts and Applied Mathematicians, are so interested in Linear Algebra.

The simplest linear map on a vector space X to a vector space Y is the so-called **trivial map**. It is the linear map that maps every element of X to 0; it is, itself, denoted by

$$0.$$

It is surprising how often this map serves as a suitable illustration or counterexample.

(2.7) For any $A \in L(X, Y)$, $\operatorname{ran} A$ is a linear subspace of Y, and $A0 = 0$.

Indeed, since A is linear, $Ax + Ay = A(x + y) \in \operatorname{ran} A$, and $\alpha Ax = A(\alpha x) \in \operatorname{ran} A$, and $A0 = A(\alpha 0) = \alpha A0$ for any $\alpha \in \mathbb{F}$, hence $A0 = 0$. (The preceding has been an example for how simplicity of notation can come back to bite you; if, instead of simplicity, I had chosen to denote the zero vector of X by 0_X and the zero scalar in $\mathbb{F}$ by $0_{\mathbb{F}}$, I could have written, more concisely, $A(0_X) = A(0_{\mathbb{F}} 0_X) = 0_{\mathbb{F}} A(0_X) = 0_Y$ to prove that A maps $0 \in X$ to $0 \in Y$.)

Example: If $\mathbf{a} \in \mathbb{R}^n$, then[†]

$$\mathbf{a}^{\mathrm{t}} : \mathbb{R}^n \to \mathbb{R} : \mathbf{x} \mapsto \mathbf{a}^{\mathrm{t}}\mathbf{x} := a_1 x_1 + a_2 x_2 + \cdots + a_n x_n \tag{2.8}$$

is a linear map of great practical importance. Indeed, any (real) linear algebraic equation in n unknowns has the form

$$\mathbf{a}^{\mathrm{t}}? = y$$

for some **coefficient vector** $\mathbf{a} \in \mathbb{R}^n$ and some **right side** $y \in \mathbb{R}$. Such an equation has solutions for arbitrary y if and only if $\mathbf{a} \neq \mathbf{0}$. You may have already learned that the general solution can always be written as the sum of a particular solution and an arbitrary solution of the corresponding **homogeneous** equation

$$\mathbf{a}^{\mathrm{t}}? = 0.$$

In particular, the map $\mathbf{a}^{\mathrm{t}}$ cannot be 1-1 unless $n = 1$.

Assume that $\mathbf{a} \neq 0$. For $n = 2$, it is instructive to visualize the solution set as a straight line, parallel to the straight line

$$\operatorname{null} \mathbf{a}^{\mathrm{t}} := \{\mathbf{x} \in \mathbb{R}^2 : \mathbf{a}^{\mathrm{t}}\mathbf{x} = 0\}$$

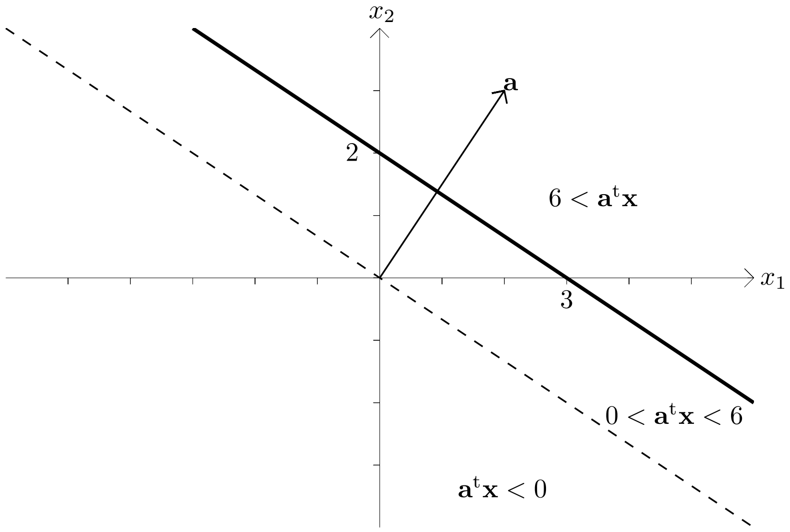
through the origin formed by all the solutions to the corresponding homogeneous problem, hence perpendicular to the coefficient vector $\mathbf{a}$. Note that the ‘nullspace’ $\operatorname{null} \mathbf{a}^{\mathrm{t}}$ splits $\mathbb{R}^2$ into the two **half-spaces**

$$\{\mathbf{x} \in \mathbb{R}^2 : \mathbf{a}^{\mathrm{t}}\mathbf{x} > 0\}, \qquad \{\mathbf{x} \in \mathbb{R}^2 : \mathbf{a}^{\mathrm{t}}\mathbf{x} < 0\},$$

one of which contains $\mathbf{a}$. This is shown in (2.9)Figure, for the particular equation

$$2x_1 + 3x_2 = 6.$$

[†] The use of the superscript $^{\mathrm{t}}$ in (2.8) is explained on page 37.


x_2
$\mathbf{a}$
2
$6 < \mathbf{a}^{\mathrm{t}}\mathbf{x}$
x_1
3
$0 < \mathbf{a}^{\mathrm{t}}\mathbf{x} < 6$
$\mathbf{a}^{\mathrm{t}}\mathbf{x} < 0$

(2.9) Figure. One way to visualize all the parts of the equation $\mathbf{a}^{\mathrm{t}}\mathbf{x} = 6$ with $\mathbf{a} = (2, 3)$. □

By adding or composing two linear maps (if appropriate) or by multiplying a linear map by a scalar, we obtain further linear maps. Here are the details.

The (pointwise) sum $A + B$ of $A, B \in L(X, Y)$ and the product αA of $\alpha \in \mathbb{F}$ with $A \in L(X, Y)$ are again in $L(X, Y)$, hence $L(X, Y)$ is closed under (pointwise) addition and multiplication by a scalar, therefore a linear subspace of the vector space Y^X of all maps from X into the vector space Y.

> **(2.10)** $L(X, Y)$ is a vector space under pointwise addition and multiplication by a scalar.

Linearity is preserved not only under (pointwise) addition and multiplication by a scalar, but also under map *composition*.

> **(2.11)** The composition of two linear maps is again linear (if it is defined).

Indeed, if $A \in L(X, Y)$ and $B \in L(Y, Z)$, then BA maps X to Z and, for any $x, y \in X$,

$$\begin{aligned}(BA)(x + y) = B(A(x + y)) &= B(Ax + Ay)\\ &= B(Ax) + B(Ay) = (BA)(x) + (BA)(y).\end{aligned}$$

Also, for any $x \in X$ and any scalar α,

$$(BA)(\alpha x) = B(A(\alpha x)) = B(\alpha Ax) = \alpha B(Ax) = \alpha(BA)(x).$$

2.8 For each of the following maps, determine whether or not it is linear (give a reason for your answer).

(a) $\Pi_{<k} \to \mathbb{Z}_+ : p \mapsto \#\{x : p(x) = 0\}$ (i.e., the map that associates with each polynomial of degree $< k$ the number of its zeros).

(b) $C([a \,..\, b]) \to \mathbb{R} : f \mapsto \max_{a \le x \le b} f(x)$.

(c) $\mathbb{F}^{3\times 4} \to \mathbb{F} : A \mapsto A_{2,2}$.

(d) $L(X, Y) \to Y : A \mapsto Ax$, with x a fixed element of X (and, of course, X and Y vector spaces).

(e) $\mathbb{F}^{[0..1]} \to \mathbb{F}^{[a..b]} : f \mapsto f \circ \phi$ with $\phi : [a \,..\, b] \to [0 \,..\, 1]$ invertible.

(f) $\mathbb{R}^{m\times n} \to \mathbb{R}^{n\times m} : A \mapsto A^{\mathrm{c}}$ (with A^{c} the (conjugate) transpose of the matrix A).

(g) $\mathbb{R} \to \mathbb{R}^2 : x \mapsto (x, \sin(x))$.

(h) $\Pi \to \Pi : p \mapsto gp$ for some $g \in \Pi$, with $gp : x \mapsto g(x)p(x)$.

2.9* *The linear image of a vector space is a vector space*: Let $f : X \to T$ be a map on some vector space X into some set T on which addition and multiplication by scalars is defined in such a way that

$$f(\alpha x + \beta y) = \alpha f(x) + \beta f(y), \qquad \alpha, \beta \in \mathbb{F},\ x, y \in X. \tag{2.12}$$

Prove that ran f *is a vector space (with respect to the addition and multiplication as restricted to* ran f*).* (See Problem 3.25 for an important application.)

Linear maps from $\mathbb{F}^n$ (aka column maps)

As a ready source of many examples, we now give a complete description of $L(\mathbb{F}^n, X)$. This description will make good use of the unorthodox notation $x\alpha := \alpha x$ for the product of the scalar α with the vector x introduced in (2.1).

(2.13) For any sequence $v_1, v_2, \ldots, v_n$ in the vector space X, the map

$$f : \mathbb{F}^n \to X : \mathbf{a} \mapsto v_1a_1 + v_2a_2 + \cdots + v_na_n$$

is linear.

Proof: The proof is a boring but necessary verification.

(a) additivity:

$$\begin{aligned}
f(\mathbf{a}+\mathbf{b}) &= v_1\,(\mathbf{a}+\mathbf{b})_1 + v_2\,(\mathbf{a}+\mathbf{b})_2 + \cdots + v_n\,(\mathbf{a}+\mathbf{b})_n \\
&\qquad \text{(definition of } f\text{)} \\
&= v_1\,(a_1+b_1) + v_2\,(a_2+b_2) + \cdots + v_n\,(a_n+b_n) \\
&\qquad \text{(addition of } n\text{-vectors)} \\
&= v_1a_1 + v_1b_1 \ + \ v_2a_2 + v_2b_2 \ + \ \cdots \ + \ v_na_n + v_nb_n \\
&\qquad \text{(multiplication by scalar distributes)} \\
&= v_1a_1 + v_2a_2 + \cdots + v_na_n \quad + \quad v_1b_1 + v_2b_2 + \cdots + v_nb_n \\
&\qquad \text{(vector addition commutes)} \\
&= f(\mathbf{a}) \ + \ f(\mathbf{b}) \\
&\qquad \text{(definition of } f\text{)}
\end{aligned}$$

(s) homogeneity:

$$\begin{aligned}
f(\beta\mathbf{a}) &= v_1\,(\beta\mathbf{a})_1 + v_2\,(\beta\mathbf{a})_2 + \cdots + v_n\,(\beta\mathbf{a})_n \\
&\qquad \text{(definition of } f\text{)} \\
&= v_1\beta a_1 + v_2\beta a_2 + \cdots + v_n\beta a_n \\
&\qquad \text{(multiplication of scalar with } n\text{-vectors)} \\
&= \beta(v_1a_1 + v_2a_2 + \cdots + v_na_n) \\
&\qquad \text{(multiplication by scalar distributes)} \\
&= \beta f(\mathbf{a}) \\
&\qquad \text{(definition of } f\text{)}
\end{aligned}$$

□

(2.14) Definition: The sum

$$v_1a_1 + v_2a_2 + \cdots + v_na_n$$

is called the **linear combination of the vectors** $v_1, v_2, \ldots, v_n$ **with weights** $a_1, \ldots, a_n$. I will use the suggestive abbreviation

$$[v_1, v_2, \ldots, v_n]\mathbf{a} := v_1a_1 + v_2a_2 + \cdots + v_na_n,$$

hence use

$$[v_1, v_2, \ldots, v_n]$$

for the map $V : \mathbb{F}^n \to X : \mathbf{a} \mapsto v_1a_1 + v_2a_2 + \cdots + v_na_n$. I call such a map a **column map**, and call v_j its j**th column**. Further, I denote its number of columns by

$$\#V.$$

The most important special case of this occurs when also X is a coordinate space, $X = \mathbb{F}^m$ say. In this case, each v_j is an m-vector, $\mathbf{v}_j$ say, and

$$\mathbf{v}_1 a_1 + \mathbf{v}_2 a_2 + \cdots + \mathbf{v}_n a_n = V\mathbf{a},$$

with V the $m \times n$-matrix with columns $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$. This explains why I chose to write the weights in the linear combination $v_1 a_1 + v_2 a_2 + \cdots + v_n a_n$ to the right of the vectors v_j rather than to the left. For, it suggests that working with the map $[v_1, v_2, \ldots, v_n]$ is rather like working with a matrix with columns $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$.

Note that `MATLAB` uses the notation $[\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n]$ for the matrix with columns $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n$, as do some textbooks. This stresses the fact that it is customary to think of the *matrix* $C \in \mathbb{F}^{m \times n}$ with columns $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n$ as the *linear map*

$$[\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n] : \mathbb{F}^n \to \mathbb{F}^m : \mathbf{x} \mapsto \mathbf{c}_1 x_1 + \mathbf{c}_2 x_2 + \cdots + \mathbf{c}_n x_n.$$

(2.15) Agreement: For any sequence $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ of m-vectors,

$$[\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]$$

denotes both the $m \times n$-matrix V with columns $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ and the linear map

$$V : \mathbb{F}^n \to \mathbb{F}^m : \mathbf{a} \mapsto [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]\mathbf{a} = \mathbf{v}_1 a_1 + \mathbf{v}_2 a_2 + \cdots + \mathbf{v}_n a_n.$$

Thus (see (2.17)Proposition below),

$$\mathbb{F}^{m \times n} = L(\mathbb{F}^n, \mathbb{F}^m).$$

Thus, a matrix $V \in \mathbb{F}^{m \times n}$ is associated with two rather different maps: (i) since it is an assignment with domain $\underline{m} \times \underline{n}$ and values in $\mathbb{F}$, we could think of it as a map on $\underline{m} \times \underline{n}$ to $\mathbb{F}$; (ii) since it is the n-list of its columns, we can think of it as the linear map from $\mathbb{F}^n$ to $\mathbb{F}^m$ that carries the n-vector $\mathbf{a}$ to the m-vector $V\mathbf{a} = \mathbf{v}_1 a_1 + \mathbf{v}_2 a_2 + \cdots + \mathbf{v}_n a_n$. From now on, we will stick to the second interpretation when we talk about the domain, the range, or the target, of a matrix. Thus, for $V \in \mathbb{F}^{m \times n}$, $\operatorname{dom} V = \mathbb{F}^n$ and $\operatorname{tar} V = \mathbb{F}^m$, and $\operatorname{ran} V \subset \mathbb{F}^m$. – If we want the first interpretation, we call $V \in \mathbb{F}^{m \times n}$ a (two-dimensional) **array** and write $V \in \mathbb{F}^{\underline{m} \times \underline{n}}$.

Next, we prove that there is nothing special about the linear maps of the form $[\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]$ from $\mathbb{F}^n$ into the vector space X since *every* linear map from $\mathbb{F}^n$ to X is necessarily of that form. The identity map

$$\mathrm{id}_n : \mathbb{F}^n \to \mathbb{F}^n : \mathbf{a} \mapsto \mathbf{a}$$

is of this form, i.e.,

$$\mathrm{id}_n = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n],$$

with $\mathbf{e}_j$ the j**th coordinate direction (vector)**, i.e.,

$$\mathbf{e}_j := (\underbrace{0, \ldots, 0}_{j-1 \text{ zeros}}, 1, 0, \ldots, 0)$$

the vector (with the appropriate number of entries) all of whose entries are 0, except for the jth, which is 1. Written out in painful detail, this says that

$$\forall \mathbf{a} \in \mathbb{F}^n, \quad \mathbf{a} = \mathbf{e}_1 a_1 + \mathbf{e}_2 a_2 + \cdots + \mathbf{e}_n a_n.$$

Further,

(2.16) Proposition: If $V = [v_1, v_2, \ldots, v_n] : \mathbb{F}^n \to X$ and $f \in L(X, Y)$, then $fV = [f(v_1), \ldots, f(v_n)]$.

Proof: If $\operatorname{dom} f = X$ and f is linear, then fV is linear and, for any $\mathbf{a} \in \mathbb{F}^n$,

$$\begin{aligned}(fV)\mathbf{a} &= f(V\mathbf{a}) = f(v_1 a_1 + v_2 a_2 + \cdots + v_n a_n)\\ &= f(v_1)a_1 + f(v_2)a_2 + \cdots + f(v_n)a_n \;=\; [f(v_1), \ldots, f(v_n)]\mathbf{a}.\end{aligned}$$

□

Consequently, for any $f \in L(\mathbb{F}^n, X)$,

$$f = f\ \mathrm{id}_n = f\,[\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n] = [f(\mathbf{e}_1), \ldots, f(\mathbf{e}_n)].$$

This proves:

(2.17) Proposition: The map f from $\mathbb{F}^n$ to the vector space X is linear if and only if

$$f = [f(\mathbf{e}_1), f(\mathbf{e}_2), \ldots, f(\mathbf{e}_n)].$$

Therefore,

$$L(\mathbb{F}^n, X) = \{[v_1, v_2, \ldots, v_n] : v_1, v_2, \ldots, v_n \in X\},$$

with the linear map $X^n \to L(\mathbb{F}^n, X) : (v_1, v_2, \ldots, v_n) \mapsto [v_1, v_2, \ldots, v_n]$ invertible (see Problem 2.21), hence $L(\mathbb{F}^n, X)$ and X^n are **isomorphic**. In symbols

$$L(\mathbb{F}^n, X) \simeq X^n.$$

As a simple example, recall from (2.8) the map

$$\mathbf{a}^{\mathrm{t}} : \mathbb{R}^n \to \mathbb{R} : \mathbf{x} \mapsto a_1x_1 + a_2x_2 + \cdots + a_nx_n = [a_1, \ldots, a_n]\mathbf{x},$$

and, in this case, $\mathbf{a}^{\mathrm{t}}\mathbf{e}_j = a_j$, all j. This confirms that $\mathbf{a}^{\mathrm{t}}$ is linear and shows that

(2.18) $$\mathbf{a}^{\mathrm{t}} = [a_1, \ldots, a_n] = [\mathbf{a}]^{\mathrm{t}}.$$

Notation: If V and W are both columns maps into the same vector space, then $[V, W]$ denotes the column map in which first all the columns of V are used and then all the columns of W. Also, I write

$$V \subset W$$

to mean that V is obtained by omitting (zero or more) columns from W; i.e., $V = W(\,:\,, \mathsf{c})$ for some sub-list c of $1{:}\#W$.

Finally, if W is a column map and M is a set, then I write

$$W \ \subset \ M$$

to mean that the columns of W are elements of M. For example:

(2.19) Proposition: A nonempty subset Z of the vector space Y is a linear subspace (of Y) if and only if, for all column maps W into Y, $W \subset Z \quad \Longrightarrow \quad \operatorname{ran} W \subset Z$.

In more traditional language, this proposition says that a nonempty subset Z of a vector space Y is a linear subspace of Y iff it is closed under formation of linear combinations. Indeed, since, for $x, y \in Y$ and $\alpha \in \mathbb{F}$, αy and $x + y$ are particular linear combinations of elements from Y, the "if" is clear. As to the "only if", it is clear by induction on $\#W$, since the case $\#W = 1$ is covered by Z being closed under multiplication by scalars, while, for $\#W > 1$, we can write $W = [U, V]$ with both $\#U, \#V < \#W$, hence, for any $W\mathbf{c} \in \operatorname{ran} W$, writing appropriately $\mathbf{c} = (\mathbf{a}, \mathbf{b})$, we have $W\mathbf{c} = U\mathbf{a} + V\mathbf{b} \in Z$ by induction hypothesis.

The important (2.16)Proposition is the reason we define the **product of matrices** the way we do, namely as

$$\forall i, j, \quad (AB)_{ij} := \sum_k A_{ik} B_{kj}.$$

For, if $A \in \mathbb{F}^{m \times n} = L(\mathbb{F}^n, \mathbb{F}^m)$ and $B = [\mathbf{b}_1, \ldots, \mathbf{b}_r] \in \mathbb{F}^{n \times r} = L(\mathbb{F}^r, \mathbb{F}^n)$, then $AB \in L(\mathbb{F}^r, \mathbb{F}^m) = \mathbb{F}^{m \times r}$, and

$$AB = A[\mathbf{b}_1, \ldots, \mathbf{b}_r] = [A\mathbf{b}_1, \ldots, A\mathbf{b}_r]. \tag{2.20}$$

Notice that the product AB of two maps A and B makes sense if and only if $\operatorname{dom} A \supset \operatorname{tar} B$. For matrices A and B, this means that the number of columns of A must equal the number of rows of B; we couldn't apply A to the columns of B otherwise.

In particular, *the 1-column matrix* $[A\mathbf{x}]$ *is the product of the matrix* A *with the 1-column matrix* $[\mathbf{x}]$, i.e.,

$$\forall (A, \mathbf{x}) \in \mathbb{F}^{m \times n} \times \mathbb{F}^n, \quad A[\mathbf{x}] = [A\mathbf{x}].$$

For this reason, most books on elementary linear algebra and most users of linear algebra *identify* the n-vector $\mathbf{x}$ with the $n \times 1$-matrix $[\mathbf{x}]$, hence write simply $\mathbf{x}$ for what I have denoted here by $[\mathbf{x}]$. I will feel free from now on to use the same identification. However, I will not be doctrinaire about it. In particular, I will continue to specify a particular n-vector $\mathbf{x}$ by writing down its entries in a list, like $\mathbf{x} = (x_1, x_2, \ldots)$, since that uses much less space than does the writing of

$$[\mathbf{x}] = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \end{bmatrix}.$$

It is consistent with the standard identification of the n-vector $\mathbf{x}$ with the $n \times 1$-matrix $[\mathbf{x}]$ to mean by $\mathbf{x}^{\mathrm{t}}$ the $1 \times n$-matrix $[\mathbf{x}]^{\mathrm{t}}$. Further, with $\mathbf{y}$ also an n-vector, one identifies the $(1,1)$-*matrix* $[\mathbf{x}]^{\mathrm{t}}[\mathbf{y}] = \mathbf{x}^{\mathrm{t}}\mathbf{y}$ with the *scalar*

$$\sum_j x_j y_j$$

which then also equals $\mathbf{y}^{\mathrm{t}}\mathbf{x}$. On the other hand, even when $\mathbf{x}$ and $\mathbf{y}$ are of different lengths, say $\mathbf{x} \in \mathbb{F}^n$ and $\mathbf{y} \in \mathbb{F}^m$ with $n \neq m$,

$$\mathbf{y}\mathbf{x}^{\mathrm{t}} \ := \ [\mathbf{y}][\mathbf{x}]^{\mathrm{t}} \ = \ (y_i x_j : (i,j) \in \underline{m} \times \underline{n})$$

is a well-defined $m \times n$-*matrix* (and identified with a scalar only if $m = n = 1$).

However, I will *not* use the terms ‘column vector’ or ‘row vector’, as they don’t make sense to me. Also, whenever I want to stress the fact that $\mathbf{x}$ or $\mathbf{x}^{\mathrm{t}}$ is meant to be a matrix, I will write $[\mathbf{x}]$ and $[\mathbf{x}]^{\mathrm{t}}$, respectively.

For example, what about the expression $\mathbf{x}\mathbf{y}^{\mathrm{t}}\mathbf{z}$ in case $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ are vectors? It makes sense only if $\mathbf{y}$ and $\mathbf{z}$ are vectors of the same length, say $\mathbf{y}, \mathbf{z} \in \mathbb{F}^n$. In that case, it is $[\mathbf{x}][\mathbf{y}]^{\mathrm{t}}[\mathbf{z}]$, and this we can compute in two ways: we can apply the matrix $\mathbf{x}\mathbf{y}^{\mathrm{t}}$ to the vector $\mathbf{z}$, or we can multiply the vector $\mathbf{x}$ with the scalar $\mathbf{y}^{\mathrm{t}}\mathbf{z}$. Either way, we obtain the vector $\mathbf{x}(\mathbf{y}^{\mathrm{t}}\mathbf{z}) = (\mathbf{y}^{\mathrm{t}}\mathbf{z})\mathbf{x}$, i.e., the $(\mathbf{y}^{\mathrm{t}}\mathbf{z})$-multiple of $\mathbf{x}$. However, while the product $\mathbf{x}(\mathbf{y}^{\mathrm{t}}\mathbf{z})$ of $\mathbf{x}$ with $(\mathbf{y}^{\mathrm{t}}\mathbf{z})$ makes sense both as a matrix product and as multiplication of the vector $\mathbf{x}$ by the scalar $\mathbf{y}^{\mathrm{t}}\mathbf{z}$, the product $(\mathbf{y}^{\mathrm{t}}\mathbf{z})\mathbf{x}$ *only* makes sense as a product of the scalar $\mathbf{y}^{\mathrm{t}}\mathbf{z}$ with the vector $\mathbf{x}$.

(2.21) Example: Here is an example, of help later. Consider a so-called **elementary row operation** on n-vectors, specifically the one that adds α times the kth entry to the ith entry. Is this a linear map? What is a formula for it?

We note that the kth entry of any n-vector $\mathbf{x}$ can be computed as ${\mathbf{e}_k}^{\mathrm{t}}\mathbf{x}$, while adding β to the ith entry of $\mathbf{x}$ is accomplished by adding $\mathbf{e}_i\beta$ to $\mathbf{x}$. Hence, adding α times the kth entry of $\mathbf{x}$ to its ith entry replaces $\mathbf{x}$ by $\mathbf{x} + \mathbf{e}_i(\alpha{\mathbf{e}_k}^{\mathrm{t}}\mathbf{x}) = \mathbf{x} + \alpha\mathbf{e}_i{\mathbf{e}_k}^{\mathrm{t}}\mathbf{x}$. This gives the handy formula

$$E_{\mathbf{e}_i,\mathbf{e}_k}(\alpha) := \mathrm{id}_n + \alpha\mathbf{e}_i{\mathbf{e}_k}^{\mathrm{t}} \tag{2.22}$$

for this elementary row operation (the use of 'row' here justified by the traditional view of an n-vector $\mathbf{x}$ as the $n \times 1$-matrix $[\mathbf{x}]$). Now, to check that $E_{\mathbf{e}_i,\mathbf{e}_k}(\alpha)$ is linear, we observe that it is the sum of two maps, and the first one, id_n, is certainly linear, while the second is the composition of the three maps,

$$\mathbf{e}_k{}^{\mathrm{t}} : \mathbb{F}^n \to \mathbb{F} \simeq \mathbb{F}^1 : \mathbf{z} \mapsto \mathbf{e}_k{}^{\mathrm{t}}\mathbf{z}, \quad [\mathbf{e}_i] : \mathbb{F}^1 \to \mathbb{F}^n : \beta \to \mathbf{e}_i\beta,$$

$$\alpha : \mathbb{F}^n \to \mathbb{F}^n : \mathbf{z} \mapsto \alpha\mathbf{z},$$

and each of these is linear (the last one because we assume $\mathbb{F}$ to be a *commutative* field).

Matrices of the form

$$E_{\mathbf{y},\mathbf{z}}(\alpha) := \mathrm{id} + \alpha\mathbf{y}\mathbf{z}^{\mathrm{t}} \tag{2.23}$$

are called **elementary**. They are very useful since, if invertible, their inverse has the same simple form; see (2.34)Proposition below. □

2.10 (a) Compute the number of additions/multiplications needed to apply $E_{\mathbf{e}_i,\mathbf{e}_k}(\alpha)$ as given by (2.22) to one n-vector $\mathbf{x}$ and compare with the numbers needed to apply an arbitrary matrix of order n to one n-vector. (b) How many nonzero entries does $E_{\mathbf{e}_i,\mathbf{e}_k}(\alpha)$, written out as a matrix, have? Is it worth the effort to write out $E_{\mathbf{e}_i,\mathbf{e}_k}(\alpha)$ as an $n \times n$-matrix?

2.11 Use the fact that the jth column of the matrix A is the image of $\mathbf{e}_j$ under the linear map A to construct the matrices that carry out the specified action.

(i) The matrix A of order 2 that rotates the plane clockwise 90 degrees;

(ii) The matrix B that reflects $\mathbb{R}^n$ across the hyperplane $\{\mathbf{x} \in \mathbb{R}^n : x_n = 0\}$;

(iii) The matrix C that keeps the hyperplane $\{\mathbf{x} \in \mathbb{R}^n : x_n = 0\}$ pointwise fixed, and maps $\mathbf{e}_n$ to $-\mathbf{e}_n$;

(iv) The matrix D of order 2 that keeps the y-axis fixed and maps $(1, 1)$ to $(2, 1)$.

(v) The matrix $A \in \mathbb{F}^{n\times n}$ that maps $\mathbf{e}_j$ to $\mathbf{e}_{\sigma(j)}$, $j = 1{:}n$, for some permutation σ of degree n. Any such A is called a **permutation matrix**.

2.12* Use the fact that the jth column of the matrix $A \in \mathbb{F}^{m\times n}$ is the image of $\mathbf{e}_j$ under A to derive the four matrices A^2, AB, BA, and B^2 for each of the given pair A and B: (i) $A = [\mathbf{e}_1, \mathbf{0}]$, $B = [\mathbf{0}, \mathbf{e}_1]$; (ii) $A = [\mathbf{e}_2, \mathbf{e}_1]$, $B = [\mathbf{e}_2, -\mathbf{e}_1]$; (iii) $A = [\mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_1]$, $B = A^2$.

2.13* Prove that $(\mathbf{e}_i\mathbf{e}_j{}^{\mathrm{t}})(\mathbf{e}_k\mathbf{e}_h{}^{\mathrm{t}}) = \delta_{jk}\mathbf{e}_i\mathbf{e}_h{}^{\mathrm{t}}$, with $\mathbf{e}_r \in \mathbb{F}^n$, $r = i, j, k, h$.

2.14 For each of the following pairs of matrices A, B, determine their products AB and BA if possible, or else state why that cannot be done.

(a) $A = \begin{bmatrix} 1 & -1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$, $B = \mathrm{id}_2$; (b) $A = \begin{bmatrix} 2 & 1 & 4 \\ 0 & 1 & 2 \end{bmatrix}$, $B = A^{\mathrm{t}}$; (c) $A = \begin{bmatrix} 2 & 1 & 4 \\ 0 & 1 & 2 \\ 0 & 0 & -1 \end{bmatrix}$, $B = \begin{bmatrix} -1 & -1 & 2 \\ 0 & 2 & -1 \\ 0 & 0 & 3 \end{bmatrix}$; (d) $A = \begin{bmatrix} 2+\mathrm{i} & 4-\mathrm{i} \\ 3-\mathrm{i} & 3+\mathrm{i} \end{bmatrix}$, $B = \begin{bmatrix} 2-\mathrm{i} & 3+\mathrm{i} & 3\mathrm{i} \\ 3-\mathrm{i} & 4+\mathrm{i} & 2 \end{bmatrix}$.

2.15 For any $A, B \in L(X)$, the products AB and BA are also linear maps on X, as are $A^2 := AA$ and $B^2 := BB$. Give an example of $A, B \in L(X)$ for which $(A + B)^2$ does not equal $A^2 + 2AB + B^2$. (Hint: Keep it as simple as possible, by choosing X to be $\mathbb{R}^2$, hence both A and B are 2-by-2 matrices.)

2.16 Give an example of matrices A and B for which both $AB = 0$ and $BA = 0$, while neither A nor B is a zero matrix.

2.17* Prove the formula

$$\begin{bmatrix} A & B \\ E & F \end{bmatrix} \begin{bmatrix} C & G \\ D & H \end{bmatrix} = \begin{bmatrix} AC + BD & AG + BH \\ EC + FH & EG + FH \end{bmatrix}$$

for the product of two **compatibly partitioned** matrices, i.e, when the products AC and FH are defined.

2.18* Let $A \in \mathbb{F}^{m \times n}$, $\mathbf{x} \in \mathbb{F}^n$, hence $A\mathbf{x}$ is well-defined. Show that

$$A\mathbf{x} = A(\mathtt{:}, \mathtt{f})x_{\mathtt{f}} + A(\mathtt{:}, \mathtt{g})x_{\mathtt{g}},$$

with $\mathtt{f}$ and $\mathtt{g}$ a **partitioning** of $\underline{n}$ in the sense that $(\mathtt{f}, \mathtt{g})$ is a permutation of degree n, i.e., both $\mathtt{f}$ and $\mathtt{g}$ are 1-1 and $\mathtt{g}$ contains all the elements of $\underline{n}$ not contained in $\mathtt{f}$, and recall from (1.17) that e.g., $x_{\mathtt{f}}$ is the list $\mathbf{x} \circ \mathtt{f} = (x_{f_1}, x_{f_2}, \ldots)$.

2.19* With the same setup as in Problem 2.18, and with $B \in \mathbb{F}^{n \times k}$, hence AB is defined, prove that

$$AB = A(\mathtt{:}, \mathtt{f})\, B(\mathtt{f}, \mathtt{:}) + A(\mathtt{:}, \mathtt{g})\, B(\mathtt{g}, \mathtt{:})$$

and relate this to Problem 2.17. For that, show that, with $\mathtt{a}$ a list into $\underline{m}$ and $\mathtt{b}$ a list into $\underline{k}$,

$$(AB)(\mathtt{a}, \mathtt{b}) = A(\mathtt{a}, \mathtt{:})\, B(\mathtt{:}, \mathtt{b}).$$

2.20 Prove that *both* $\mathbb{C} \to \mathbb{R} : z \mapsto \operatorname{Re} z$ *and* $\mathbb{C} \to \mathbb{R} : z \mapsto \operatorname{Im} z$ *are linear maps when we consider* $\mathbb{C}$ *as a vector space over the real scalar field.*

2.21* Recall from (2.17)Proposition that two vector spaces X and Y over the same scalar field are *isomorphic* if $L(X, Y)$ contains an invertible map, and this is indicated by writing $X \simeq Y$. Prove the claim made in the last display of (2.17)Proposition.

Linear maps from $\mathbb{F}^T$

We can, in the same way, establish that, for any *finite* set T, any linear map from the more general coordinate space $\mathbb{F}^T$ into a vector space X is necessarily of the form

$$\mathbb{F}^T \to X : \mathbf{a} \mapsto [v_t : t \in T]\mathbf{a} := \sum_{t \in T} v_t a_t$$

for some assignment $t \mapsto v_t$ with domain T and with values in X.

This is a convenient notation when there is no natural way to order the elements of T. This even makes good sense for an infinite T in which case the maps necessarily have the domain

$$(2.24) \qquad \mathbb{F}_0^T := \{\mathbf{a} \in \mathbb{F}^T : \#\{t \in T : a_t \neq 0\} < \infty\}.$$

A special case occurs with $T \subset X$ and the choice $v_t = t$, all $t \in T$, in which case the following notation is convenient:

$$(2.25) \qquad [T] : \mathbb{F}_0^T \to X : \mathbf{a} \mapsto \sum_{t \in T} t a_t.$$

The linear equation $A? = \mathbf{y}$, and $\operatorname{ran} A$ and $\operatorname{null} A$

We are ready to recognize and use the fact that the general system

$$
\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= y_1\\
a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= y_2\\
&\vdots\\
a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= y_m
\end{aligned}
\tag{2.26}
$$

of m linear equations in the n unknowns $x_1, \ldots, x_n$ is equivalent to the vector equation

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

provided

$$A := \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n}\\ a_{21} & a_{22} & \cdots & a_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ a_{m1} & a_{m2} & \cdots & a_{mn}\end{bmatrix},\ \mathbf{x} := (x_1, x_2, \ldots, x_n),\ \mathbf{y} := (y_1, y_2, \ldots, y_m).$$

Here, **equivalence** means that the entries $x_1, \ldots, x_n$ of the n-vector $\mathbf{x}$ solve the system of linear equations (2.26) if and only if $\mathbf{x}$ solves the vector equation $A? = \mathbf{y}$. This equivalence is not only a notational convenience. Switching from (2.26) to $A? = \mathbf{y}$ is the conceptual shift that started Linear Algebra. It shifts the focus, from the scalars $x_1, \ldots, x_n$, to the vector $\mathbf{x}$ formed by them, and to the map A given by the coefficients in (2.26), its range and nullspace (about to be defined), and this makes for simplicity, clarity, and generality.

To stress the generality, we now give a preliminary discussion of the equation

$$A? = y$$

in case A is a *linear* map, from the vector space X to the vector space Y say, with y some element of Y.

Existence of a solution for every $y \in Y$ is equivalent to having A be *onto*, i.e., to having $\operatorname{ran} A = Y$. The check whether A is onto is made easier to answer if we happen to know an *onto* column map $[v_1, \ldots, v_m] = V \in L(\mathbb{F}^m, Y)$. For, then we only have to check that the *finitely many* columns, $v_1, \ldots, v_m$, of V are in $\operatorname{ran} A$. Indeed, if some are not in $\operatorname{ran} A$, then, surely, A is not onto. However, if they all are in $\operatorname{ran} A$, then also $\operatorname{ran} V \subset \operatorname{ran} A$ by (2.7) and (2.19)Proposition, hence $Y = \operatorname{ran} V \subset \operatorname{ran} A \subset \operatorname{tar} A = Y$, i.e., $\operatorname{ran} A = Y$, i.e., A is onto.

(2.27) Proposition: If Y is the range of the column map V, then $A \in L(X, Y)$ is onto if and only if the finitely many columns of V are in $\operatorname{ran} A$.

Uniqueness of a solution for every $y \in Y$ is equivalent to having A be *1-1*, i.e., to have $Ax = Az$ imply that $x = z$. For a *linear* map $A : X \to Y$, we have $Ax = Az$ if and only if $A(x - z) = 0$. In other words, if $y = Ax$, then

$$A^{-1}\{y\} = x + \{z \in X : Az = 0\}. \tag{2.28}$$

In particular, A is 1-1 if and only if $\{z \in X : Az = 0\} = \{0\}$. Therefore, to check whether a *linear* map is 1-1, we only have to check whether it is 1-1 'at' one particular point, e.g., 'at' 0. For this reason, the set $A^{-1}\{0\} = \{z \in X : Az = 0\}$ of all elements of X mapped by A to 0 is singled out.

(2.29) Definition: The set

$$\operatorname{null} A := \{z \in \operatorname{dom} A : Az = 0\}$$

is called the **nullspace** or **kernel** of the linear map A.

(2.30) A linear map is 1-1 if and only if its nullspace is **trivial**, i.e., contains only the zero vector.
The nullspace of a linear map is a linear subspace (of its target).

Indeed, if A is a linear map and $Z := \operatorname{null} A$, then $A(Z + Z) = A(Z) + A(Z) = \{0\} + \{0\} = \{0\}$ and $A(\mathbb{F}Z) = \mathbb{F}A(Z) = \mathbb{F}\{0\} = \{0\}$.

Almost all linear subspaces you will meet will be of the form ran A or null A for some linear map A. These two ways of specifying a linear subspace are very different in character.

If we are told that our linear subspace Z of X is of the form null A, for a certain linear map A on X, then we know, offhand, exactly one element of Z for sure, namely the element 0 which lies in every linear subspace. On the other hand, it is easy to *test* whether a given $x \in X$ lies in $Z = \operatorname{null} A$: simply compute Ax and check whether it is the zero vector.

If we are told that our linear subspace Z of X is of the form ran A for some linear map A from some U into X, then we can 'write down' explicitly every element of ran A: they are all of the form Au for some $u \in \operatorname{dom} A$. On the other hand, it is much harder to *test* whether a given $x \in X$ lies in $Z = \operatorname{ran} A$: Now we have to check whether the equation $A? = x$ has a solution (in U).

As a simple example, the vector space $\Pi_{\le k}$ of all polynomials of degree $\le k$ is usually specified as the range of the column map

$$[()^0, ()^1, \ldots, ()^k] : \mathbb{R}^{k+1} \to \mathbb{R}^{\mathbb{R}},$$

with

$$()^j : \mathbb{R} \to \mathbb{R} : t \mapsto t^j$$

a convenient (though nonstandard!) notation for the **monomial of degree** j, i.e., as the collection of all real-valued functions that are of the form

$$t \mapsto a_0 + a_1 t + \cdots + a_k t^k$$

for some coefficient-vector $\mathbf{a}$. On the other hand, $\Pi_{\le k}$ can also be defined as $\operatorname{null} D^{k+1}$, i.e., as the collection of all real-valued functions that are $(k+1)$-times continuously differentiable and have their $(k+1)$st derivative identically zero.

(2.31) Remark: The nullspace $\operatorname{null} A$ of the linear map $A : X \to Y$ consists exactly of the solutions to the *homogeneous* equation

$$A? = 0.$$

The linear equation $A? = y$ is readily associated with a *homogeneous* linear equation, namely the equation

$$[A, y]? = 0,$$

with

$$[A, y] : X \times \mathbb{F} : (z, \alpha) \mapsto Az + y\alpha.$$

If $Ax = y$, then $(x, -1)$ is a nontrivial element of $\operatorname{null}[A, y]$. Conversely, if $(z, \alpha) \in \operatorname{null}[A, y]$ *and* $\alpha \ne 0$, then $z/(-\alpha)$ is a solution to $A? = y$. Hence, for the construction of solutions to linear equations, it is sufficient to know how to solve *homogeneous* linear equations, i.e., how to construct the nullspace of a linear map.

2.22 What can you conclude about the linear system $A? = y$ if all the elements of $\operatorname{null}[A, y]$ are of the form $[z, \alpha]$ with $\alpha = 0$?

2.23 For each of the following three systems of linear equations, determine A and $\mathbf{y}$ of the equivalent vector equation $A? = \mathbf{y}$.

$$(a) \begin{array}{rcrcr} 2x_1 & - & 3x_2 & = & 4 \\ 4x_1 & + & 2x_2 & = & -6 \end{array}; \quad (b) \begin{array}{rcrcr} 2u_1 & - & 3u_2 & = & 4 \\ 4u_1 & + & 2u_2 & = & -6 \end{array}; \quad (c) \begin{array}{rcrcrcr} & & & & -4c & = & 16 \\ 2a & + & 3b & & & = & 9 \end{array}.$$

2.24 For each of the following A and $\mathbf{y}$, write out a system of linear equations equivalent to the vector equations $A? = \mathbf{y}$.

(a) $A = \begin{bmatrix} 2 & 3 \\ 6 & 4 \\ \pi & -2 \end{bmatrix}$, $\mathbf{y} = (9, -\sqrt{3}, 1)$; (b) $A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{bmatrix}$, $\mathbf{y} = (10, 10)$;

(c) $A = [\,] \in \mathbb{R}^{0\times 3}$, $\mathbf{y} = () \in \mathbb{R}^0$.

2.25* Prove: (i) *for any* $A, B \in L(X)$, $\operatorname{null} A \cap \operatorname{null} B \subset \operatorname{null}(A + B)$. (ii) *for any* $A, B \in L(X)$ *with* $AB = BA$, $\operatorname{null} A + \operatorname{null} B \subset \operatorname{null}(AB)$.

Inverses

We have agreed to think of the *matrix* $A \in \mathbb{F}^{m\times n}$ as the *column map* $[A_{:1}, \ldots, A_{:n}]$, i.e., as the linear map

$$\mathbb{F}^n \to \mathbb{F}^m : \mathbf{a} \mapsto A\mathbf{a} := \sum_j A_{:j} a_j.$$

For this reason, it is also customary to refer to $\operatorname{ran} A$ of a matrix A as the **column space** of that matrix, while the range $\operatorname{ran} A^{\mathrm{t}}$ of its transpose is known as its **row space**. Further, we have found (see (2.20)) that the *matrix product* AB is also the *composition* $A \circ B$, i.e.,

$$(A \circ B)\mathbf{a} = A(B(\mathbf{a})) = (AB)\mathbf{a} = \sum_j (AB)_{:j} a_j.$$

In these terms, the identity map id_n on $\mathbb{F}^n$ corresponds to the **identity matrix** $[\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n]$, hence the name for the latter.

(2.32) Proposition: The inverse of a linear map is again a linear map.

Proof: Let $A \in L(X,Y)$ be invertible and $y, z \in Y$. By additivity of A, $A(A^{-1}y + A^{-1}z) = A(A^{-1}y) + A(A^{-1}z) = y + z$. Hence, applying A^{-1} to both sides, we get $A^{-1}y + A^{-1}z = A^{-1}(y+z)$, thus A^{-1} is additive. Also, from $A(\alpha A^{-1}y) = \alpha A(A^{-1}y) = \alpha y$, we conclude that $\alpha A^{-1}y = A^{-1}(\alpha y)$, hence A^{-1} is homogeneous. □

Thus, if $A \in \mathbb{F}^{n\times n}$ is invertible (as a linear map from $\mathbb{F}^n$ to $\mathbb{F}^n$), then also its inverse is a linear map from $\mathbb{F}^n$ to $\mathbb{F}^n$, hence a square matrix of order n. We call it the **inverse** matrix for A, and denote it by A^{-1}. Being the inverse for A, it is both a right and a left inverse for A, i.e., it satisfies

$$A^{-1}A = \mathrm{id}_n = AA^{-1}.$$

More generally, we would call $A \in \mathbb{F}^{m\times n}$ invertible if there were $B \in \mathbb{F}^{n\times m}$ so that

$$AB = \mathrm{id}_m \quad \text{and} \quad BA = \mathrm{id}_n.$$

However, we will soon prove (cf. (3.24)Corollary) that this can only happen when $m = n$.

This is related to the *pigeonhole principle for square matrices* (3.26)Theorem, i.e., to the fact that a linear map from $\mathbb{F}^n$ to $\mathbb{F}^n$ is 1-1 if and only if it is onto. This implies that if $A, B \in \mathbb{F}^{n\times n}$ and, e.g., $AB = \mathrm{id}_n$, hence

A is onto, then A must also be 1-1, hence invertible, and therefore its right inverse must be its inverse, therefore we must also have $BA = \mathrm{id}_n$. In short:

(2.33) Amazing Fact: If $A, B \in \mathbb{F}^{n\times n}$ and $AB = \mathrm{id}_n$, then also $BA = \mathrm{id}_n$.

To me, this continues to be one of the most remarkable results in basic Linear Algebra. Its proof uses nothing more than the identification of matrices with linear maps (between coordinate spaces) and the realization (see (3.13)Theorem) that if V is a 1-1 column map into the range of the column map W, then $\#V \le \#W$.

In preparation for the chapter on Elimination, and as an exercise in invertible matrices, we verify the following useful fact about elementary matrices which is also useful for the proof of its generalization, the Sherman-Morrison formula (see Problem 2.35).

(2.34) Proposition: For $\mathbf{y}, \mathbf{z} \in \mathbb{F}^n$ and $\alpha \in \mathbb{F}$, the elementary matrix

$$E_{\mathbf{y},\mathbf{z}}(\alpha) = \mathrm{id}_n + \alpha\mathbf{y}\mathbf{z}^{\mathrm{t}}$$

is invertible if and only if $1 + \alpha\mathbf{z}^{\mathrm{t}}\mathbf{y} \ne 0$, and, in that case

$$E_{\mathbf{y},\mathbf{z}}(\alpha)^{-1} = E_{\mathbf{y},\mathbf{z}}\Big(\frac{-\alpha}{1+\alpha\mathbf{z}^{\mathrm{t}}\mathbf{y}}\Big). \tag{2.35}$$

Proof: We compute $E_{\mathbf{y},\mathbf{z}}(\alpha)E_{\mathbf{y},\mathbf{z}}(\beta)$ for arbitrary α and β. Since

$$\alpha\mathbf{y}\mathbf{z}^{\mathrm{t}}\ \beta\mathbf{y}\mathbf{z}^{\mathrm{t}} = \alpha\beta\ (\mathbf{z}^{\mathrm{t}}\mathbf{y})\ \mathbf{y}\mathbf{z}^{\mathrm{t}},$$

we conclude that

$$E_{\mathbf{y},\mathbf{z}}(\alpha)E_{\mathbf{y},\mathbf{z}}(\beta) = (\mathrm{id}_n + \alpha\mathbf{y}\mathbf{z}^{\mathrm{t}})(\mathrm{id}_n + \beta\mathbf{y}\mathbf{z}^{\mathrm{t}}) = \mathrm{id}_n + (\alpha+\beta+\alpha\beta(\mathbf{z}^{\mathrm{t}}\mathbf{y}))\mathbf{y}\mathbf{z}^{\mathrm{t}}.$$

In particular, since the scalar factor $(\alpha+\beta+\alpha\beta(\mathbf{z}^{\mathrm{t}}\mathbf{y}))$ is symmetric in α and β, we conclude that

$$E_{\mathbf{y},\mathbf{z}}(\beta)E_{\mathbf{y},\mathbf{z}}(\alpha) = E_{\mathbf{y},\mathbf{z}}(\alpha)E_{\mathbf{y},\mathbf{z}}(\beta).$$

Further, if $1 + \alpha\mathbf{z}^{\mathrm{t}}\mathbf{y} \ne 0$, then the choice

$$\beta := \frac{-\alpha}{1+\alpha\mathbf{z}^{\mathrm{t}}\mathbf{y}}$$

will give $\alpha+\beta+\alpha\beta(\mathbf{z}^{\mathrm{t}}\mathbf{y}) = 0$, hence $E_{\mathbf{y},\mathbf{z}}(\beta)E_{\mathbf{y},\mathbf{z}}(\alpha) = E_{\mathbf{y},\mathbf{z}}(\alpha)E_{\mathbf{y},\mathbf{z}}(\beta) = \mathrm{id}_n$. This proves that $E_{\mathbf{y},\mathbf{z}}(\alpha)$ is invertible, with its inverse given by (2.35).

Conversely, assume that $1 + \alpha\mathbf{z}^{\mathrm{t}}\mathbf{y} = 0$. Then $\mathbf{y} \ne 0$, yet

$$E_{\mathbf{y},\mathbf{z}}(\alpha)\mathbf{y} = \mathbf{y} + \alpha(\mathbf{z}^{\mathrm{t}}\mathbf{y})\mathbf{y} = 0,$$

showing that $E_{\mathbf{y},\mathbf{z}}(\alpha)$ is not 1-1 in this case, hence not invertible. □

2.26 Prove: *If two matrices commute (i.e., $AB = BA$), then they are square matrices, of the same order.*

2.27 Give a noninvertible 2-by-2 matrix without any zero entries.

2.28 Prove that *the matrix $A := \begin{bmatrix} 1 & 2 \\ 4 & -1 \end{bmatrix}$ satisfies the equation $A^2 = 9\,\mathrm{id}_2$*. Use this to show that A is invertible, and to write down the matrix A^{-1}.

2.29 Prove: *The matrix $A := \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is invertible if and only if $ad \neq bc$, in which case $\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}/(ad - bc)$ is its inverse.*

2.30 Consider the map $f : \mathbb{C} \to \mathbb{R}^{2\times 2} : z = a + \mathrm{i}b \mapsto \begin{bmatrix} a & -b \\ b & a \end{bmatrix}$. Show that f is a 1-1 linear map when we think of $\mathbb{C}$ as a vector space over the real scalar field.

2.31 Let $A, B \in L(X)$. Show that $(AB)^2 = A^2B^2$ can hold without necessarily having $AB = BA$. Show also that $(AB)^2 = A^2B^2$ implies that $AB = BA$ in case both A and B are invertible.

2.32 Give an example of matrices A and B, for which both AB and BA are defined and for which $AB = \mathrm{id}$, but neither A nor B is invertible.

2.33 Prove: *If A and C are invertible matrices, and B has as many rows as does A and as many columns as does C, then also $[A, B; 0, C]$ is invertible and*

$$[A, B; 0, C]^{-1} = \begin{bmatrix} A & B \\ 0 & C \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} & -A^{-1}BC^{-1} \\ 0 & C^{-1} \end{bmatrix}.$$

2.34* A square matrix A is called **strictly diagonally dominant** if, for all i, $|A_{ii}| > \sum_{j\neq i} |A_{ij}|$. Prove: *a strictly diagonally dominant matrix is invertible.* (Hint: Prove the contrapositive: *if $\mathbf{0} \neq \mathbf{x} \in \mathrm{null}\, A$, then, for some i, $|A_{ii}| \leq \sum_{j\neq i} |A_{ij}|$.*)

2.35* Use (2.34)Proposition to prove the **Sherman-Morrison Formula**: *If $A \in \mathbb{F}^{n\times n}$ is invertible and $\mathbf{y}, \mathbf{z} \in \mathbb{F}^n$ are such that $\alpha := 1 + \mathbf{z}^{\mathrm{t}}A^{-1}\mathbf{y} \neq 0$, then $A + \mathbf{y}\mathbf{z}^{\mathrm{t}}$ is invertible, and*

$$(A + \mathbf{y}\mathbf{z}^{\mathrm{t}})^{-1} = A^{-1} - \alpha^{-1}A^{-1}\mathbf{y}\mathbf{z}^{\mathrm{t}}A^{-1}.$$

(Hint: $A + \mathbf{y}\mathbf{z}^{\mathrm{t}} = A(\,\mathrm{id} + (A^{-1}\mathbf{y})\mathbf{z}^{\mathrm{t}})$.)

2.36 Prove the **Woodbury** generalization of the Sherman-Morrison Formula: *If A and $\mathrm{id} + D^{\mathrm{t}}A^{-1}C$ are invertible, then so is $A + CD^{\mathrm{t}}$, and*

$$(A + CD^{\mathrm{t}})^{-1} = A^{-1} - A^{-1}C(\,\mathrm{id} + D^{\mathrm{t}}A^{-1}C)^{-1}D^{\mathrm{t}}A^{-1}.$$

2.37 T/F

(a) If $A \in L(X, Y)$, then the set of solutions of $A? = y$ is a linear subspace of X.

(b) Any column map having a 0 column fails to be 1-1.

(c) If the column map V is not 1-1, then one of its columns is 0.

(d) If Y_1 and Y_2 are linear subspaces of the vector space X, then so is $Y_1 \cup Y_2$.

(e) If Y is a subset of some vector space X, x, y, z are particular elements of X, and x and $2y - 3x$ are in Y, but $3y - 2x$ or y are not, then Y cannot be a linear subspace.

(f) If $A, B \in L(X, Y)$ are both invertible, then so is $A + B$.

(g) If $AB = 0$ for $A, B \in \mathbb{F}^{n\times n}$, then $B = 0$.

(h) If A and B are matrices with $AB = \mathrm{id}_m$ and $BA = \mathrm{id}_n$, then $B = A^{-1}$.

(i) If $A = \begin{bmatrix} B & C \\ 0 & 0 \end{bmatrix}$ with both A and B square matrices and 0 standing for zero matrices of the appropriate size, then $A^n = \begin{bmatrix} B^n & B^{n-1}C \\ 0 & 0 \end{bmatrix}$ for all n.

(j) If $A \in \mathbb{R}^{m\times n}$ and $A^{\mathrm{t}}A = 0$, then $A = 0$.

(k) If the matrix product AB is defined, then $(AB)^{\mathrm{t}} = A^{\mathrm{t}}B^{\mathrm{t}}$.

(l) If A is an invertible matrix, then so is A^{t}, and $(A^{\mathrm{t}})^{-1} = (A^{-1})^{\mathrm{t}}$.

(m) $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$ is an elementary matrix.

(n) If the scalar field $\mathbb{F}$ were not commutative, then the map $\mathbb{F}^n \to \mathbb{F}^n : \mathbf{x} \mapsto \alpha\mathbf{x}$, of multiplication by the scalar α, would not be linear.

3 The dimension of a vector space

Bases

The only vector spaces in which we can carry out calculations are the coordinate spaces $\mathbb{F}^n$. To calculate with other vector spaces, we have to relate them first to some coordinate space. This is true even when X is a proper subspace of $\mathbb{F}^n$, e.g., the nullspace of some matrix.

For example, we do not really compute with polynomials, we usually compute with the coefficients of the polynomial. Precisely (see (3.38)Proposition), one sets up the invertible linear map

$$\mathbb{F}^n \to \Pi_{<n} : \mathbf{a} \mapsto a_1 + a_2 t + a_3 t^2 + \cdots + a_n t^{n-1}$$

where I have, temporarily, followed the (ancient and sometimes confusing) custom of describing the *monomials* by the list of symbols $(\ , t, t^2, t^3, \ldots)$ rather than by the nonstandard symbols $()^j$, $j = 0, 1, 2, 3, \ldots$ introduced earlier. One adds polynomials by adding their coefficients, or evaluates polynomials from their coefficients, etc. . You may be so used to that, that you haven't even noticed until now that you do not work with the polynomials themselves, but only with their coefficients.

It is therefore a practically important goal to provide ways of **representing** the elements of a given vector space X by n-vectors. We do this by using linear maps from some $\mathbb{F}^n$ that have X as their range, i.e., we look for sequences $v_1, v_2, \ldots, v_n$ in X for which the linear map $[v_1, v_2, \ldots, v_n] : \mathbb{F}^n \to X$ is onto. If there is such a map for some n, then we call X **finitely generated**.

Among such onto maps $V \in L(\mathbb{F}^n, X)$, those that are also 1-1, hence invertible, are surely the most desirable ones since, for such V, there is, for any $x \in X$, exactly one $\mathbf{a} \in \mathbb{F}^n$ with $x = V\mathbf{a}$. Any *invertible* column map to X is, by definition, a **basis** for X.

Since $\mathrm{id}_n \in L(\mathbb{F}^n)$ is trivially invertible, it is a basis for $\mathbb{F}^n$. It is called the **natural basis for** $\mathbb{F}^n$.

Here is a small difficulty with this (and any other) definition of basis: What is the basis of the **trivial space**, i.e., the vector space that consists of the zero vector alone? It is a perfectly well-behaved vector space (though a bit limited – except as a challenge to textbook authors when it comes to discussing its basis).

We deal with it here by considering $V \in L(\mathbb{F}^n, X)$ even when $n = 0$. Since $\mathbb{F}^n$ consists of lists of n items (each item an element from $\mathbb{F}$), the peculiar space $\mathbb{F}^0$ must consist of lists of *no* items, i.e., of *empty* lists. There is only one empty list (of scalars), hence $\mathbb{F}^0$ has just one element, the empty list, $(\,)$, and this element is necessarily the neutral element (or, zero vector) for this space, i.e., $\mathbf{0} = ()$. Correspondingly, there is exactly one *linear* map from $\mathbb{F}^0$ into X, namely the map $\mathbb{F}^0 \to X : () \mapsto 0$. Since this is a linear map from $\mathbb{F}^0$, we call it the column map into X with *no* columns or the **empty column map**, and denote it by $[\,]$. Thus,

$$[\,] : \mathbb{F}^0 \to X : () = \mathbf{0} \mapsto 0. \tag{3.1}$$

Note that $[\,]$ *is 1-1*. Note also that the range of $[\,]$ consists of the trivial subspace, $\{0\}$. In particular, the column map $[\,]$ is *onto* $\{0\}$, hence is invertible, as map from $\mathbb{F}^0$ to $\{0\}$. It follows that $[\,]$ is a basis for $\{0\}$. Isn't Mathematics wonderful?! – As it turns out, the column map $[\,]$ will also be very helpful below.

Here are some standard terms related to bases of a vector space:

Definition: The range of $V := [v_1, v_2, \ldots, v_n]$ is called the **span of the sequence** $v_1, v_2, \ldots, v_n$:

$$\mathrm{span}(v_1, v_2, \ldots, v_n) := \mathrm{ran}\, V.$$

An element x of X is said to be **linearly dependent on** $v_1, v_2, \ldots, v_n$ in case $x \in \mathrm{ran}\, V$, i.e., in case x is a **linear combination of the** v_j. Otherwise x is said to be **linearly independent of** $v_1, v_2, \ldots, v_n$.

The sequence $v_1, v_2, \ldots, v_n$ is said to be **linearly independent** in case V is 1-1, i.e., in case $V\mathbf{a} = 0$ implies $\mathbf{a} = \mathbf{0}$ (i.e., the only way to write the zero vector as a linear combination of the v_j is to choose all the weights equal to 0).

The sequence $v_1, v_2, \ldots, v_n$ is said to be **spanning for** X in case V is onto, i.e., in case $\mathrm{span}(v_1, v_2, \ldots, v_n) = X$.

The sequence $v_1, v_2, \ldots, v_n$ is said to be a **basis for** X in case V is invertible, i.e., 1-1 and onto.

If V is invertible, then $V^{-1}x$ is an n-vector, called the **coordinate vector for** x **with respect to the basis** $v_1, v_2, \ldots, v_n$.

You may wonder why there are all these terms in use for the *sequence* $v_1, v_2, \ldots, v_n$, particularly when the corresponding terms for the *map* $V = [v_1, v_2, \ldots, v_n]$ are so much shorter and to the point. I don't know the answer. However, bear in mind that the terms commonly used are those for sequences. An even greater puzzle is the fact that many textbooks present bases as *sets* rather than *sequences*. At least, that is what they say. But, not surprisingly, whenever there is some action involving a basis, the basis is written $\{v_1, \ldots, v_n\}$, i.e., as a sequence in everything but in name. It is for you to ask such authors whether $\{3, 3\}$ is a basis for $\mathbb{R}^1 = \mathbb{R}$. They will say that it is not even though it is since, after all, $3 = 3$, hence $\{3, 3\} = \{3\}$.

A major use of the basis concept is the following which generalizes the way we earlier constructed arbitrary linear maps from $\mathbb{F}^n$.

(3.2) Proposition: Let $V = [v_1, \ldots, v_n]$ be a basis for the vector space X, and let Y be an arbitrary vector space. Any map $f : \{v_1, \ldots, v_n\} \to Y$ has exactly one extension to a linear map A from X to Y. In other words, we can choose the values of a linear map on the columns of a basis arbitrarily and, once chosen, this pins down the linear map everywhere.

Proof: The map $A := [f(v_1), \ldots, f(v_n)]V^{-1}$ is linear, from X to Y, and carries v_j to $f(v_j)$ since $V^{-1}v_j = \mathbf{e}_j$, all j. This shows existence. Further, if also $B \in L(X, Y)$ with $Bv_j = f(v_j)$, all j, then $BV = [f(v_1), \ldots, f(v_n)] = AV$, therefore $B = A$ (since V is invertible). □

3.1 Prove that the linear map $S := [\mathbf{e}_n, \mathbf{e}_{n-1}, \ldots, \mathbf{e}_1] \in L(\mathbb{F}^n)$ reverses the order of the entries of an n-vector, i.e., $S\mathbf{x} = (x_n, x_{n-1}, \ldots, x_1)$, hence is its own inverse.

3.2 Describe what the $n \times n$-matrix $A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}$ does to all the vectors $\mathbf{e}_j$, i.e., give a simple formula for $A\mathbf{e}_j$. Deduce from your formula that $\operatorname{ran} A^n = \{0\}$, hence that $A^n = 0$.

3.3 Prove: $A \in L(X)$ *commutes with every* $B \in L(X)$ *if and only if* $A = \alpha \operatorname{id}_X$, i.e., A is a scalar multiple of the identity.

3.4 Let $X \times Y$ be the product space of the vector spaces X and Y. The map $f : X \times Y \to \mathbb{F}$ is **bilinear** if it is linear in each slot, i.e., if the map $f(\cdot, y) : X \mapsto \mathbb{F} : x \mapsto f(x, y)$ is linear for every $y \in Y$, and the map $f(x, \cdot) : Y \to \mathbb{F} : y \mapsto f(x, y)$ is linear for every $x \in X$. (This use of $\cdot$ is known as **placeholder notation**.)

(i) Prove that, *for every* $A \in \mathbb{F}^{m\times n}$*, the map* $f_A : \mathbb{F}^m \times \mathbb{F}^n : (\mathbf{x}, \mathbf{y}) \mapsto \mathbf{y}^{\mathrm{t}} A\mathbf{x}$ *is bilinear.*

(ii) Prove that, *for every bilinear* $f : \mathbb{F}^m \times \mathbb{F}^n \to \mathbb{F}$*, there exists exactly one* $A \in \mathbb{F}^{m\times n}$ *with* $f_A = f$*.*

(iii) Prove that *the map* $A \mapsto f_A$ *is an invertible linear map on* $\mathbb{F}^{m\times n}$ *to the vector space* $BL(\mathbb{F}^m, \mathbb{F}^n)$ *of all bilinear maps on* $\mathbb{F}^m \times \mathbb{F}^n$ *under pointwise vector operations.*

3.5 `MATLAB`'s command `yy = interp1(x,y,xx,'spline')` returns the value(s) at `xx` of a certain function f that matches the data given by `x`, `y`, in the sense that $f($`x(i)`$) =$ `y(i)` for `i=1:n`, with `n` the length of both `x` and `y` (and assuming that the entries of `x` are pairwise distinct). (If you wanted to look at f on the interval $[a\,..\,b]$, you might choose `xx = linspace(a,b,N+1);` with `N` some suitably large number, and then `plot(xx,yy)`.)

(a) Generate some numerical evidence for the claim that (up to roundoff) the map `y` $\mapsto f$ provided by this command is linear.

(b) Assuming that the map is linear, deduce from the above description of the map that it must be 1-1, hence a basis for its range.

(c) Still assuming that the map `y` $\mapsto f$ provided by that command is indeed linear, hence a column map, provide a plot of each of its columns, as functions on the interval $[0\,..\,3]$, for the specific choice `0:3` for `x`.

(d) (quite open-ended) Determine as much as you can about the elements of the range of this column map.

(e) Is the map still linear if you replace `'spline'` by `'cubic'`?

Construction of a basis

Next, we consider the construction of a basis. This can be done either by *extending a 1-1 column map* V to a basis, or by *thinning an onto column map* W to a basis. For this, remember that, for two column maps V and W into some vector space X, we agreed to mean by $V \subset W$ that V can be obtained from W by thinning, i.e., by omitting zero or more columns from W, and W can be obtained from V by extending, i.e., by inserting zero or more columns.

Thinning an onto map to a basis is based on the following observation.

(3.3) Proposition: The column map $W = [w_1, w_2, \ldots, w_n]$ fails to be 1-1 if and only if, for some j, $w_j \in \operatorname{ran}[w_i : i \neq j]$, in which case $\operatorname{ran} W = \operatorname{ran}[w_i : i \neq j]$.

Proof: If W is not 1-1, then there is some nontrivial n-vector $\mathbf{a}$ in its nullspace, and then

$$(3.4) \qquad [w_i : i \neq j] a_{\underline{n}\backslash j} = \sum_{i\neq j} w_i a_i = -a_j w_j$$

for every $j \in \underline{n}$. Since $\mathbf{a} \neq \mathbf{0}$, there is some j for which $a_j \neq 0$, hence we can divide both sides of (3.4) by $-a_j$ to get that $w_j \in \operatorname{ran}[w_i : i \neq j]$. Conversely, if $w_j \in \operatorname{ran}[w_i : i \neq j]$, i.e., if $w_j = [w_i : i < j]\mathbf{b} + [w_i : i > j]\mathbf{c}$, then the

n-vector $\mathbf{a} := (\mathbf{b}, -1, \mathbf{c})$ is a nontrivial vector in $\operatorname{null} W$, i.e., W fails to be 1-1.

Finally, if $w_j = [w_i : i \neq j]\mathbf{b}$, then

$$\forall \mathbf{a} \in \mathbb{F}^n, \quad W\mathbf{a} = [w_i : i \neq j]a_{i \neq j} + a_j[w_i : i \neq j]\mathbf{b} \ \in \operatorname{ran}[w_i : i \neq j],$$

hence $\operatorname{ran} W \subseteq \operatorname{ran}[w_i : i \neq j] \subseteq \operatorname{ran} W$. □

In thinning out an onto column map, it turns out to be more convenient to focus on columns that are in the span of the columns to their left.

(3.5) Definition: We say that the jth column of the column map $V = [v_1, v_2, \ldots, v_n]$ is **free** in case $v_j \in \operatorname{ran}[v_i : i < j]$. Otherwise, we call the jth column **bound**.

For example, since $\operatorname{ran}[\,] = \{0\}$, the first column of a column map is free if and only if it is 0. The practical determination of the bound and free columns is taken up in the next chapter, in the discussion of elimination, the algorithm that gave rise to this terminology.

The proof of (3.3)Proposition also proves the following three propositions if we add the observation that every nonzero n-vector has a rightmost nonzero entry.

(3.6) Proposition: The kth column of the column map V is free if and only there exists $\mathbf{x} \in \operatorname{null} V$ with x_k its rightmost nonzero entry.

(3.7) Proposition: A column map fails to be 1-1 if and only if it has free columns.

(3.8) Proposition: If U is the map obtained from $V = [v_1, v_2, \ldots, v_n]$ by deleting a free column, then $\operatorname{ran} U = \operatorname{ran} V$.

The last two propositions have the following

(3.9) Corollary: Any column map can be thinned to a basis for its range.

Indeed, if U is the column map obtained from V by removing all free columns, then $\operatorname{ran} U = \operatorname{ran} V$ by (3.8)Proposition, while, by (3.7)Proposition, U is 1-1.

Now note that the classification of a column as free or bound depends entirely on the columns to its left. In fact, by (3.8)Proposition, it depends only on the *bound* columns to its left. Hence, removal of a free column will not alter the classification of the remaining columns. In particular, the order in which we remove free columns to thin an onto map V to a basis is immaterial. The resulting basis for $\operatorname{ran} V$ will consist of the bound columns of V (in their original order). This proves the following strengthening of (3.9)Corollary.

(3.10) The bound columns of a column map form a basis for its range.

(3.11) Corollary: Every 1-1 column map into a finitely generated vector space can be extended to a basis for that space.

Indeed, if V is a 1-1 column map into $X = \operatorname{ran} W$ for some column map W, then also $\operatorname{ran}[V, W] = X$, and since V is 1-1, all of its columns are bound in $[V, W]$, hence, by (3.10), for some $U \subset W$, $[V, U]$ is a basis for X.

A more careful argument along these lines gives the following.

(3.12) Steinitz Exchange: If V is a 1-1 column map into the range of a column map W, then there exists a column map $U \subset W$ with $\#U = \#W - \#V$ for which $\operatorname{ran}[V, U] = \operatorname{ran} W$.

Proof: By induction on $\#V$: If $\#V = 0$, then the conclusion follows from (3.9)Corollary. In the contrary case, let v be the last column of V, i.e.,

$V =: [V_1, v]$. Then, by induction, there is $U_1 \subset W$ with $\#U_1 = \#W - \#V_1$ so that $\operatorname{ran}[V_1, U_1] = \operatorname{ran} W$. Since $v \in \operatorname{ran} W = \operatorname{ran}[V_1, U_1]$, it follows from (3.3)Proposition that $[V_1, v, U_1]$ has a nontrivial nullspace, hence, by (3.7)Proposition, one of its columns must be free, yet $[V_1, v]$ is 1-1, hence a free column must occur in U_1 and, dropping one such, we obtain from U_1 the column map $U \subset U_1 \subset W$ with $\#U = \#U_1 - 1 = \#W - (\#V_1 + 1) = \#W - \#V$ and $\operatorname{ran}[V, U] = \operatorname{ran} W$ by (3.7)Proposition, thus advancing the induction. □

(3.13) Theorem: If V and W are column maps into the vector space X and V is 1-1 and W is onto, then $\#V \le \#W$.

Proof: By (3.12), there exists $U \subset W$ with $0 \le \#U = \#W - \#V$. □

3.6* Prove: *If V is* **maximally 1-1** *into X, meaning that $[V, w]$ fails to be 1-1 for every $w \in X$, then V is a basis for X.*

3.7* Prove: *If W is* **minimally onto** *X, meaning that no $V \subset W$ (other than W itself) is onto X, then W is a basis for X.*

Dimension

Since two bases of a vector space are both 1-1 and onto, (3.13)Theorem implies the following.

(3.14) Lemma: Any two bases for a vector space have the same number of columns.

This number of columns in any basis for X is denoted

$$\dim X$$

and is called the **dimension of** X.

Since id_n is a basis for $\mathbb{F}^n$ and has n columns, we conclude that the *n-dimensional coordinate space has, indeed, dimension n*. In effect, $\mathbb{F}^n$ is the prototypical vector space of dimension n. Any n-dimensional vector space X is connected to $\mathbb{F}^n$ by invertible linear maps, the bases for X.

Note that the trivial vector space, $\{0\}$, has dimension 0 since its (unique) basis has no columns.

(3.15) Example: The dimension of $\Pi_{\le k}(\mathbb{R}^d)$. The space $\Pi_{\le k}(\mathbb{R}^d)$ of **d-variate polynomials of degree** $\le k$ is, by definition, the range of the column map

$$V := [()^{\boldsymbol{\alpha}} : |\boldsymbol{\alpha}| \le k] : \mathbb{F}^{\{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|\le k\}} \to (\mathbb{R}^d \to \mathbb{R}) : \mathbf{a} \mapsto \sum_{|\boldsymbol{\alpha}|\le k} ()^{\boldsymbol{\alpha}} a_{\boldsymbol{\alpha}},$$

with

$$()^{\boldsymbol{\alpha}} : \mathbb{R}^d \to \mathbb{R} : \mathbf{t} \mapsto \mathbf{t}^{\boldsymbol{\alpha}} := t_1^{\alpha_1} \cdots t_d^{\alpha_d}$$

a nonstandard notation for the $\boldsymbol{\alpha}$-power function, with $\boldsymbol{\alpha} \in \mathbb{Z}_+^d$, i.e., $\boldsymbol{\alpha}$ any d-vector with nonnegative integer entries, and with $|\boldsymbol{\alpha}| := \sum_j \alpha_j$.

When $d = 1$, then V can be seen to be 1-1, hence a basis for $\Pi_{\le k}(\mathbb{R})$, by considering the 'data map'

$$Q : \Pi_{\le k} \to \mathbb{R}^{k+1} : p \mapsto (p(0), Dp(0), D^2p(0)/2, \ldots, D^kp(0)/k!),$$

for which we have $QV =$ id, hence V is 1-1.

An analogous argument, involving the 'data map'

$$p \mapsto (D^{\boldsymbol{\alpha}}p(0)/\boldsymbol{\alpha}! : \boldsymbol{\alpha} \in \mathbb{Z}_+^d, |\boldsymbol{\alpha}| \le k),$$

with $\boldsymbol{\alpha}! := \alpha_1! \cdots \alpha_d!$, shows that

$$\dim \Pi_{\le k}(\mathbb{R}^d) = \#\{\boldsymbol{\alpha} \in \mathbb{Z}_+^d : |\boldsymbol{\alpha}| \le k\},$$

and the latter number can be shown (see Problem 3.9) to equal $\binom{k+d}{d}$.

□

3.8 Prove that *the space $\Pi_{<3}(\mathbb{R}^2)$ of bivariate polynomials of total degree < 3 has dimension 6.*

3.9* Verify that $\#\{\boldsymbol{\alpha} \in \mathbb{Z}_+^d : |\boldsymbol{\alpha}| \le k\} = \binom{k+d}{d} = \binom{k+d}{k}$. (Hint: $\binom{s}{t}$ is the number of t-subsets of an s-set.)

3.10 Prove that *a vector space of dimension n has subspaces of dimension j for each $j = 0, 1, \ldots, n$.*

The dimension of $\mathbb{F}^T$

Recall from (2.2) that $\mathbb{F}^T$ is the set of all scalar-valued maps on the set T, with the set T, offhand, arbitrary.

The best known instance is n-dimensional coordinate space

$$\mathbb{F}^n := \mathbb{F}^{\underline{n}},$$

with $T = \underline{n} := \{1, 2, \ldots, n\}$. The vector space $\mathbb{F}^{m\times n}$ of all $(m \times n)$-matrices is another instance; here $T = \underline{m} \times \underline{n} := \{(i,j) : i \in \underline{m}, j \in \underline{n}\}$.

(3.16) Proposition: If T is a finite set, then $\dim \mathbb{F}^T = \#T$.

Proof: Since T is finite, $\#T =: n$ say, we can order its elements, i.e., there is an invertible map $s : \underline{n} \to T$ (in fact, there are $n! = 1 \cdot 2 \cdots n$ such). This induces the map

$$V : \mathbb{F}^n \to \mathbb{F}^T : f \mapsto f \circ s^{-1}$$

which is linear (since, in both spaces, the vector operations are pointwise), and is invertible since it has

$$\mathbb{F}^T \to \mathbb{F}^n : g \mapsto g \circ s$$

as its inverse. Hence, V is a basis for $\mathbb{F}^T$ (a **natural basis**). □

Note how we managed this without even exhibiting the columns of V. To be sure, the jth column V is the function $v_j : T \to \mathbb{F} : s_k \mapsto \delta_{kj}$ that maps s_j to 1 and maps any other $t \in T$ to 0.

(3.17) Corollary: $\dim \mathbb{F}^{m \times n} = mn$.

Proof: In this case, $\mathbb{F}^{m \times n} = \mathbb{F}^T$ with $T = \underline{m} \times \underline{n} := \{(i,j) : i \in \underline{m}; j \in \underline{n}\}$, hence $\#T = mn$. □

3.11* Prove that $[\mathbf{e}_i {\mathbf{e}_j}^{\mathrm{t}} : \mathbf{e}_i \in \mathbb{F}^m, i = 1{:}m; \mathbf{e}_j \in \mathbb{F}^n, j = 1{:}n]$ is a basis for $L(\mathbb{F}^n, \mathbb{F}^m)$.

3.12 Prove: *The dimension of the vector space of all upper triangular matrices of order n is $(n+1)n/2$.*

Some uses of the dimension concept

Here is a major use of the dimension concept as it relates to *vector spaces*.

(3.18) Proposition: If X and Y are vector spaces with $X \subset Y$ and $\dim Y < \infty$, then $\dim X \le \dim Y$, with equality iff $X = Y$.

Proof: Since there is *some* 1-1 column map into X (e.g., the unique linear map from $\mathbb{F}^0$ into X), while $\dim Y$ is an upper bound on the number of columns in any 1-1 column map into $X \subset Y$ (by (3.11)Corollary), there exists a maximally 1-1 column map V into X. By Problem 3.6, any such V is necessarily a basis for X, hence X is finitely generated. By (3.11)Corollary, we can extend V to a basis $[V, W]$ for Y. Hence, $\dim X \le \dim Y$ with equality iff $W = [\,]$, i.e., iff $X = Y$. □

(3.19) Corollary: If $\#T \not< \infty$, then $\mathbb{F}^T$ is not finite-dimensional.

Proof: For every finite $S \subset T$, $\mathbb{F}^T$ contains the linear subspace

$$\{f \in \mathbb{F}^T : f(t) = 0, \text{all } t \notin S\}$$

of dimension equal to $\dim \mathbb{F}^S = \#S$ by (3.16)Proposition. If $\#T \not< \infty$, then T contains finite subsets S of arbitrarily large size, hence $\mathbb{F}^T$ contains linear subspaces of arbitrarily large dimension, hence cannot itself be finite-dimensional, by (3.18)Proposition. □

Note the following important (nontrivial) part of (3.18)Proposition:

(3.20) Corollary: Any linear subspace of a finite-dimensional vector space is finite-dimensional.

(3.21) Proposition: Let X and Y be vector spaces over $\mathbb{F}$, and assume that X is finite-dimensional. Then $\dim X = \dim Y$ if and only if there exists an invertible $A \in L(X, Y)$, i.e., iff X and Y are isomorphic.

Proof: Let $n := \dim X$. Since $n < \infty$, there exists an invertible $V \in L(\mathbb{F}^n, X)$ (i.e., a basis for X). If now $A \in L(X, Y)$ is invertible, then AV is an invertible linear map from $\mathbb{F}^n$ to Y, hence $\dim Y = n = \dim X$. Conversely, if $\dim Y = \dim X$, then there exists an invertible $W \in L(\mathbb{F}^n, Y)$; but then WV^{-1} is an invertible linear map from X to Y. □

(3.22) Corollary: $\dim L(X, Y) = \dim X \cdot \dim Y$.

Proof: Assuming that $n := \dim X$ and $m := \dim Y$ are finite, we can represent every $A \in L(X, Y)$ as a matrix $\widehat{A} := W^{-1}AV \in \mathbb{F}^{m \times n}$, with V a basis for X and W a basis for Y. This sets up a map

$$R : L(X, Y) \to \mathbb{F}^{m \times n} : A \mapsto \widehat{A} = W^{-1}AV,$$

and this map is linear and invertible (indeed, its inverse is the map $\mathbb{F}^{m \times n} \to L(X, Y) : B \mapsto WBV^{-1}$). Consequently, by (3.21)Proposition, $L(X, Y)$ and $\mathbb{F}^{m \times n}$ have the same dimension while, by (3.17)Proposition, $\dim \mathbb{F}^{m \times n} = mn = \dim X \cdot \dim Y$. □

The dimension concept is usually applied to *linear maps* by way of the following formula.

(3.23) Dimension Formula: For any linear map A with finite-dimensional domain,

$$\dim \operatorname{dom} A = \dim \operatorname{ran} A + \dim \operatorname{null} A.$$

Proof: Since $\operatorname{dom} A$ is finite-dimensional, so is $\operatorname{null} A$ (by (3.20) Corollary), hence $\operatorname{null} A$ has a basis, $V \in L(\mathbb{F}^n, \operatorname{null} A)$ say. By (3.11)Corollary, we can extend this to a basis $[V, U]$ for $\operatorname{dom} A$. Let $r := \#U$. Then, $[V, U]$ is invertible and $\dim \operatorname{dom} A - \dim \operatorname{null} A = (n + r) - n = r$.

It remains to prove that $\dim \operatorname{ran} A = r$. For this, we prove that $AU : \mathbb{F}^r \to \operatorname{ran} A$ is invertible.

Since $A[V, U] = [AV, AU]$ maps onto $\operatorname{ran} A$ and $AV = 0$, already AU must map onto $\operatorname{ran} A$, i.e., AU is onto.

Moreover, AU is 1-1: For, if $AU\mathbf{a} = 0$, then $U\mathbf{a} \in \operatorname{null} A$, hence, since V maps onto $\operatorname{null} A$, there is some $\mathbf{b}$ so that $U\mathbf{a} = V\mathbf{b}$. This implies that $[V, U](\mathbf{b}, -\mathbf{a}) = \mathbf{0}$ and, since $[V, U]$ is 1-1, this shows that, in particular, $\mathbf{a} = \mathbf{0}$. □

3.13 Prove: *If the product AB of the two linear maps A and B is defined, then* $\dim \operatorname{ran}(AB) \le \min\{\dim \operatorname{ran} A, \dim \operatorname{ran} B\}$.

3.14 Prove: *If the product AB of the two linear maps A and B is defined, then* $\dim \operatorname{ran}(AB) = \dim \operatorname{ran} B - \dim(\operatorname{null} A \cap \operatorname{ran} B)$.

3.15 Give an example, of two square matrices A and B, that shows that $\dim \operatorname{ran}(AB)$ need not equal $\dim \operatorname{ran}(BA)$ when both AB and BA are defined.

(3.24) Corollary: Let $A \in L(X, Y)$.
(i) If $\dim X < \dim Y$, then A cannot be onto.
(ii) If $\dim X > \dim Y$, then A cannot be 1-1.
(iii) If $\dim X = \dim Y < \infty$, then A is onto if and only if A is 1-1. (This implies (2.33)!)

Proof: (i) $\dim \operatorname{ran} A \le \dim \operatorname{dom} A = \dim X < \dim Y = \dim \operatorname{tar} A$, hence $\operatorname{ran} A \ne \operatorname{tar} A$.

(ii) $\dim \operatorname{null} A = \dim \operatorname{dom} A - \dim \operatorname{ran} A = \dim X - \dim \operatorname{ran} A \ge \dim X - \dim Y > 0$, hence $\operatorname{null} A \ne \{0\}$.

(iii) If $\dim X = \dim Y$, then $\dim \operatorname{tar} A = \dim \operatorname{dom} A = \dim \operatorname{ran} A + \dim \operatorname{null} A$, hence A is onto (i.e., $\operatorname{tar} A = \operatorname{ran} A$) if and only if $\dim \operatorname{null} A = 0$, i.e., A is 1-1. □

For the special choice $X = \mathbb{F}^n$, $Y = \mathbb{F}^m$, (3.24)Corollary provides the following important matrix theorems.

(3.25) Theorem: Any matrix with more columns than rows has a nontrivial nullspace.

(3.26) Theorem (Pigeonhole Principle For Square Matrices): A square matrix is 1-1 if and only if it is onto.

For the next general result concerning the dimension concept, recall from (2.5)Proposition that both the sum

$$Y + Z := \{y + z : y \in Y, z \in Z\}$$

and the intersection $Y \cap Z$ of two linear subspaces is again a linear subspace.

(3.27) Proposition: If Y and Z are linear subspaces of the finite-dimensional vector space X, then

$$\dim(Y + Z) = \dim Y + \dim Z - \dim(Y \cap Z). \tag{3.28}$$

Proof: Consider the column map $A := [U, W]$ with U a basis for Y and W a basis for Z. Since $\dim \operatorname{dom} A = \#U + \#W = \dim Y + \dim Z$ and $\operatorname{ran} A = Y + Z$, the formula (3.28) follows from the (3.23)Dimension Formula, once we show that $\dim \operatorname{null} A = \dim(Y \cap Z)$.

For this, consider the map $Y \cap Z \to \operatorname{null} A : x \mapsto (U^{-1}x, -W^{-1}x)$. The map is into $\operatorname{null} A$, linear, and 1-1, and is onto since it is a left inverse for the linear map $\operatorname{null} A \to Y \cap Z : (\mathbf{a}, \mathbf{b}) \mapsto U\mathbf{a}(= -W\mathbf{b})$. Therefore, by (3.21)Proposition, $\dim(Y \cap Z) = \dim \operatorname{null} A$. □

Here are three corollaries of this basic proposition of use in the sequel.

(3.29) Corollary: If $[V, W]$ is 1-1, then $\operatorname{ran} V \cap \operatorname{ran} W$ is trivial.

Indeed, then $\dim(\operatorname{ran} V + \operatorname{ran} W) = \#V + \#W = \dim\operatorname{ran} V + \dim\operatorname{ran} W$.

(3.30) Corollary: If $\dim Y + \dim Z > \dim X$ for some linear subspaces Y and Z of the finite-dimensional vector space X, then $Y \cap Z$ is a *nontrivial* linear subspace, i.e., $Y \cap Z$ contains nonzero elements.

Indeed, $\dim(Y + Z) \le \dim X$ since $Y + Z$ is a linear subspace of X.

(3.31) Corollary: If Y and Z are linear subspaces of the finite-dimensional vector space X, and $Y \cap Z = \{0\}$, then

$$\dim Y + \dim Z = \dim(Y + Z) \le \dim X,$$

with equality if and only if $X = Y + Z$, in which case $\dim Z = \dim X - \dim Y =: \operatorname{codim} Y$ is called the **codimension** of Y (in X).

3.16 For each of the following linear maps, determine its range and its nullspace. Make as much use of the (3.23)Dimension Formula as possible. (You may, if need be, use the fact that, by (3.38)Proposition, $V_k := [()^0, ()^1, \ldots, ()^k]$ is a basis for $\Pi_{\le k}$.) (a) $D : \Pi_{\le k} \to \Pi_{<k} : p \mapsto Dp$, with Dp the first derivative of p. (b) $I : \Pi_{<k} \to \Pi_{\le k} : p \mapsto \int_0^{\cdot} p(s)\,\mathrm{d}s$, i.e., Ip is the primitive or antiderivative of p that vanishes at 0, i.e., $(Ip)(t) = \int_0^t p(s)\,\mathrm{d}s$. (c) $A : \Pi_{\le k} \to \Pi_{\le k} : p \mapsto Dp + p$.

3.17 Prove that $V := [()^0, ()^1, ()^2 - 1, 4()^3 - 3()^1, 8()^4 - 8()^2 + 1]$ is a basis for $\Pi_{<5}$.

3.18 Prove: *For any finite-dimensional linear subspace Y of the domain of a linear map A, $\dim A(Y) \le \dim Y$.*

3.19* Prove: *If V and W are 1-1 column maps into the vector space X, then $\operatorname{ran} V$ and $\operatorname{ran} W$ have a nontrivial intersection if and only if $[V, W]$ is not 1-1.*

3.20 Call $(Y_0, \ldots, Y_r)$ a **proper chain** in the vector space X if each Y_j is a subspace and $Y_0 \subsetneq Y_1 \subsetneq \cdots \subsetneq Y_r$. Prove that, *for any such proper chain, $r \le \dim X$, with equality if and only if $\dim Y_j = j$, $j = 0{:}(\dim X)$.*

3.21 Prove: *For any $A \in L(X, Y)$ and any linear subspace Z of X, $\dim A(Z) = \dim Z - \dim(Z \cap (\operatorname{null} A))$.*

3.22 The **defect** of a linear map is the dimension of its nullspace: $\operatorname{defect}(A) := \dim\operatorname{null} A$. (a) Prove that $\operatorname{defect}(B) \le \operatorname{defect}(AB) \le \operatorname{defect}(A) + \operatorname{defect}(B)$. (b) Prove: *If $\dim\operatorname{dom} B = \dim\operatorname{dom} A$, then also $\operatorname{defect}(A) \le \operatorname{defect}(AB)$.* (c) Give an example of linear maps A and B for which AB is defined and for which $\operatorname{defect}(A) > \operatorname{defect}(AB)$.

3.23 Let $A \in L(X, Y)$, $B \in L(X, Z)$, with Y finite-dimensional. Prove: *There exists $C \in L(Z, Y)$ with $A = CB$ if and only if $\operatorname{null} B \subset \operatorname{null} A$.*

3.24 Prove: *Assuming that the product ABC of three linear maps is defined,* $\dim\operatorname{ran}(AB) + \dim\operatorname{ran}(BC) \le \dim\operatorname{ran} B + \dim\operatorname{ran}(ABC)$.

3.25* Factor space: Let Y be a linear subspace of the vector space X and consider the collection

$$X/Y := \{x + Y : x \in X\}$$

of subsets of X, with

$$x + Y := \{x\} + Y = \{x + y : y \in Y\}.$$

(i) Prove that the map

$$f : X \to X/Y : x \mapsto x + Y$$

is linear with respect to the addition

$$M + N := \{m + n : m \in M, n \in N\}$$

and the multiplication by a scalar

$$\alpha M := \begin{cases} \{\alpha m : m \in M\}, & \text{if } \alpha \neq 0; \\ Y, & \text{if } \alpha = 0, \end{cases}$$

and has Y as its nullspace.

(ii) Prove that, *with these vector operations, X/Y is a linear space.* (X/Y is called a **factor space**.)

(iii) Prove that $\dim X/Y = \operatorname{codim} Y$ *in case X is finite-dimensional.*

Direct sums

A very useful coarsening of the basis concept concerns the sum of subspaces.

Let $Y_1, \ldots, Y_r$ be linear subspaces of the vector space X, let V_j be a column map onto Y_j, all j, and consider the column map

$$V := [V_1, \ldots, V_r].$$

To be sure, we could have also started with some arbitrary column map V into X, arbitrarily grouped its columns to obtain $V = [V_1, \ldots, V_r]$, and then defined $Y_j := \operatorname{ran} V_j$, all j.

Either way, any $\mathbf{a} \in \operatorname{dom} V$ is of the form $(\mathbf{a}_1, \ldots, \mathbf{a}_r)$ with $\mathbf{a}_j \in \operatorname{dom} V_j$, all j. Hence

$$\begin{aligned} \operatorname{ran} V &= \{V_1\mathbf{a}_1 + \cdots + V_r\mathbf{a}_r : \mathbf{a}_j \in \operatorname{dom} V_j, j \in \underline{r}\} \\ &= \{y_1 + \cdots + y_r : y_j \in Y_j, j \in \underline{r}\} =: Y_1 + \cdots + Y_r, \end{aligned}$$

the *sum* of the subspaces $Y_1, \ldots, Y_r$.

Think of this sum, as you may, as the range of the map

(3.32) $$A : Y_1 \times \cdots \times Y_r \to X : (y_1, \ldots, y_r) \mapsto y_1 + \cdots + y_r.$$

Having this map A onto says that every $x \in X$ can be written in the form $y_1 + \cdots + y_r$ with $y_j \in Y_j$, all j. In other words, X is the sum of the Y_j. In symbols,

$$X = Y_1 + \cdots + Y_r.$$

Having A also 1-1 says that there is *exactly one way* to write each $x \in X$ as such a sum. In this case, we write

$$X = Y_1 \dotplus \cdots \dotplus Y_r,$$

and say that X is the **direct sum** of the subspaces Y_j. Note the dot atop the plus sign, to indicate the special nature of this sum. Some books would use instead the encircled plus sign, $\oplus$, but we reserve that sign for an even more special direct sum in which the summands Y_j are 'orthogonal' to each other; see Chapter 6, on inner product spaces.

(3.33) Proposition: Let V_j be a basis for the linear subspace Y_j of the vector space X, $j \in \underline{r}$, and set $V := [V_1, \ldots, V_r]$. Then, the following are equivalent.

(i) $X = Y_1 \dot{+} \cdots \dot{+} Y_r$.

(ii) V is a basis for X.

(iii) $X = Y_1 + \cdots + Y_r$ and $\dim X \geq \dim Y_1 + \cdots + \dim Y_r$.

(iv) For each j, $Y_j \cap Y_{\backslash j} = \{0\}$, with $Y_{\backslash j} := Y_1 + \cdots + Y_{j-1} + Y_{j+1} + \cdots + Y_r$, and $\dim X \leq \dim Y_1 + \cdots + \dim Y_r$.

Proof: Since $\operatorname{dom} V = \operatorname{dom} V_1 \times \cdots \times \operatorname{dom} V_r$, and V_j is a basis for Y_j, all j, the linear map

$$C : \operatorname{dom} V \to Y_1 \times \cdots \times Y_r : \mathbf{a} = (\mathbf{a}_1, \ldots, \mathbf{a}_r) \mapsto (V_1\mathbf{a}_1, \ldots, V_r\mathbf{a}_r)$$

is invertible and $V = AC$, with A as given in (3.32). Hence, V is invertible if and only if A is invertible. This proves that (i) and (ii) are equivalent.

Also, (ii) implies (iii). As to (iii) implying (ii), the first assumption of (iii) says that V is onto X, and the second assumption says that $\dim \operatorname{dom} V = \#V \leq \dim X$, hence V is minimally onto and therefore a basis for X.

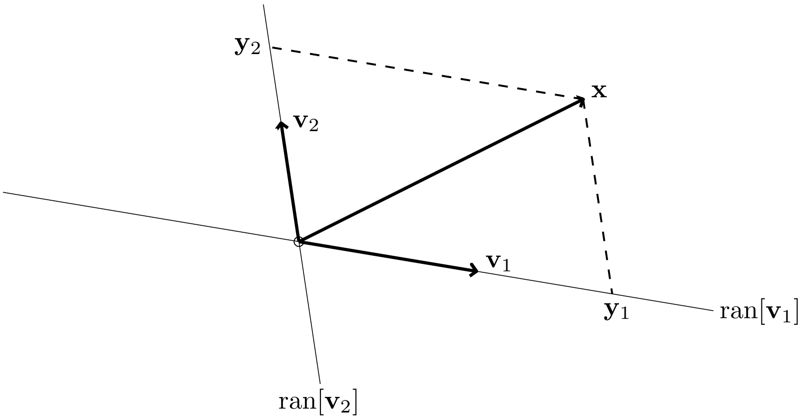
As to (ii) implying (iv), the first claim of (iv) is a special case of (3.29)Corollary, and the second claim is immediate.

Finally, as to (iv) implying (ii), assume that $0 = V\mathbf{a} = \sum_j V_j\mathbf{a}_j$. Then, for any j, $y := V_j\mathbf{a}_j = -\sum_{i \neq j} V_i\mathbf{a}_i \in Y_j \cap Y_{\backslash j}$, hence $y = 0$ by the first assumption and, since V_j is a basis for Y_j, hence 1-1, this implies that $\mathbf{a}_j = 0$. In other words, V is 1-1, while, by the second assumption, $\#V = \sum_j \dim Y_j \geq \dim X$, hence V is maximally 1-1, therefore a basis for X. □

(3.34) Corollary: If V is a basis for X, then, for any grouping $V =: [V_1, \ldots, V_r]$ of the columns of V, X is the direct sum of the linear subspaces $\operatorname{ran} V_j$, $j \in \underline{r}$.

One particular grouping is, of course, $V_j = [v_j]$, all j, in which case each $Y_j := \operatorname{ran} V_j$ is a one-dimensional linear subspace, i.e., a straight line through the origin, and we see $X = \operatorname{ran} V$ as the direct sum of these straight lines, each of which we are accustomed to think of as a **coordinate axis**.

This is illustrated in (3.35)Figure for the special case $\operatorname{ran} V = \mathbb{R}^2$, hence V has just two columns. We see each $\mathbf{x} \in \mathbb{R}^2$ written as the sum $\mathbf{x} = \mathbf{y}_1 + \mathbf{y}_2$, with $\mathbf{y}_j = a_j \mathbf{v}_j \in Y_j = \operatorname{ran}[\mathbf{v}_j]$ the **Y_j-component** of $\mathbf{x}$ (and, of course, $\mathbf{a} = (a_1, a_2)$ the coordinate vector of $\mathbf{x}$ with respect to the basis V).


$\mathbf{y}_2$
$\mathbf{x}$
$\mathbf{v}_2$
$\mathbf{v}_1$
$\mathbf{y}_1$
$\operatorname{ran}[\mathbf{v}_1]$
$\operatorname{ran}[\mathbf{v}_2]$

(3.35) Figure. A basis provides a coordinate system.

The direct sum construct is set up in just the same way, except that the Y_j may be planes or even higher-dimensional subspaces rather than just straight lines.

3.26 When X is the direct sum of Y and Z, then Z is said **to complement** Y or to be a **complement** of Y. With Y and Z linear subspaces of the finite-dimensional vector space X, prove the following assertions concerning complements.

(i) Y has a complement.

(ii) If both Z and Z_1 complement Y, then $\dim Z = \dim Z_1 = \operatorname{codim} Y$. In particular, $\operatorname{codim} Y = \dim X - \dim Y$.

(iii) $\operatorname{codim}(Y + Z) = \operatorname{codim} Y + \operatorname{codim} Z - \operatorname{codim}(Y \cap Z)$.

(iv) If Y has only one complement, then $Y = \{0\}$ or $Y = X$.

(v) If $\operatorname{codim} Y > \dim Z$, then $Y + Z \neq X$.

(vi) If $\dim Y > \operatorname{codim} Z$, then $Y \cap Z \neq \{0\}$.

3.27 Let $(d_1, \ldots, d_r)$ be a sequence of natural numbers, and let X be an n-dimensional vector space. There exists a direct sum decomposition

$$X = Y_1 \dotplus \cdots \dotplus Y_r$$

with $\dim Y_j = d_j$, all j, if and only if $\sum_j d_j = n$.

3.28* Prove: *If the vector space X is the direct sum of subspaces X_i, $i \in \underline{r}$, with each X_i the direct sum of subspaces X_{ij}, $j \in \underline{r_i}$, then X is the direct sum of X_{ij}, $j \in \underline{r_i}$, $i \in \underline{r}$.*

3.29 Let d be any scalar-valued map, defined on the collection of all linear subspaces of a finite-dimensional vector space X, that satisfies the following two conditions: (i) $Y \cap Z = \{0\} \implies d(Y + Z) = d(Y) + d(Z)$; (ii) $\dim Y = 1 \implies d(Y) = 1$.

Prove that $d(Y) = \dim Y$ *for every linear subspace* Y *of* X. Try to prove that, even without the assumption (ii), $d(Y) = \alpha \dim Y$ for some scalar α and all linear subspaces Y of X.

3.30 Prove that *the cartesian product* $Y_1 \times \cdots \times Y_r$ *of vector spaces, all over the same scalar field* $\mathbb{F}$*, becomes a vector space under pointwise or* **slotwise** *addition and multiplication by a scalar.*

This vector space is called the **product space** with **factor**s $Y_1, \ldots, Y_r$.

The only matrices whose invertibility can be ascertained at a glance

Here is an application of the pigeonhole principle for square matrices and the fact that, by (3.7)Proposition, a column map is not 1-1 if and only if it has a free column. The application concerns the invertibility of triangular matrices.

(3.36) Proposition: A square triangular matrix is invertible if and only if all its diagonal entries are nonzero.

Proof: Assume that the square matrix A is *upper* triangular, meaning that $i > j \Longrightarrow A_{ij} = 0$. E.g.,

$$A = \begin{bmatrix} * & * & \cdots & * & * & \cdots & * & * \\ & * & \cdots & * & * & \cdots & * & * \\ & & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & & & * & * & \cdots & * & * \\ & & & & A_{kk} & \cdots & * & * \\ & & & & & \cdots & * & * \\ & & & & & \ddots & \vdots & \vdots \\ & & & & & & * & * \\ & & & & & & & * \end{bmatrix},$$

with only the possibly nonzero entries indicated.

If all its diagonal elements are nonzero, then none of its columns can be free since, for each k, every element of $\operatorname{ran} A(:, 1{:}(k{-}1))$ has a zero kth entry while the kth entry of the kth column is nonzero, by assumption. Therefore, by (3.7)Proposition, A is 1-1, hence invertible by the pigeonhole principle for square matrices.

If, on the other hand, some diagonal element is zero, say $A_{kk} = 0$, then $\operatorname{ran} A(:, 1{:}k) \subset \operatorname{ran}[\mathbf{e}_1, \ldots, \mathbf{e}_{k-1}]$, a space of dimension $k-1$, hence $A(:, 1{:}k)$ must fail to be 1-1 by (3.24)Corollary(ii), therefore $A(:, 1{:}k)$ must have a free column, hence A is not invertible.

Now notice that, with

$$S := [\mathbf{e}_n, \mathbf{e}_{n-1}, \ldots, \mathbf{e}_1]$$

the linear map that reverses the order of the entries of an n-vector, hence is its own inverse, left and right multiplication by the invertible matrix S turns a lower triangular matrix A into an upper triangle matrix, whose diagonal entries are nonzero iff the diagonal entries of A are nonzero:

$$S\begin{bmatrix}3 & & \\ * & 2 & \\ * & * & 1\end{bmatrix}S = S\begin{bmatrix} & & 3\\ & 2 & *\\ 1 & * & *\end{bmatrix} = \begin{bmatrix}1 & * & *\\ & 2 & *\\ & & 3\end{bmatrix}.$$

Since SAS is invertible if and only if A is invertible, this proves that a *lower* triangular matrix is invertible if and only if all its diagonal entries are nonzero. □

Note that the same result holds concerning square matrices A that are triangular with respect to the **secondary diagonal**, $(A_{n1}, A_{n-1,2}, \ldots, A_{1n})$, as right or left multiplication by S turns these into upper triangular matrices and its secondary diagonal entries into diagonal entries.

Polynomial interpolation

If $V \in L(\mathbb{F}^n, X)$ and $Q \in L(X, \mathbb{F}^n)$, then QV is a linear map from $\mathbb{F}^n$ to $\mathbb{F}^n$, i.e., a square matrix, of order n. If QV is 1-1 or onto, then (3.26)Theorem tells us that QV is invertible. In particular, V is 1-1 and Q is onto, and so, for every $\mathbf{y} \in \mathbb{F}^n$, there exists exactly one $p \in \operatorname{ran} V$ for which $Qp = \mathbf{y}$. This is the essence of *interpolation*, to be pursued further in the discussion of the inverse of a basis, in Chapter 5. Here, we make use of the preceding (3.36)Proposition in a discussion of the most important example of interpolation.

(3.37) Example: Polynomial Interpolation Take $X = \mathbb{R}^{\mathbb{R}}$, $V = [()^0, ()^1, \ldots, ()^{k-1}]$, hence $\operatorname{ran} V$ equals $\Pi_{<k}$, the collection of all polynomials of degree $< k$. Further, take $Q : X \to \mathbb{R}^k : f \mapsto (f(\tau_1), \ldots, f(\tau_k))$ for some fixed sequence $(\tau_1, \ldots, \tau_k)$ of pairwise distinct points. Then the equation

$$QV? = Qf$$

asks for the (power) coefficients of a polynomial of degree $< k$ that agrees with the function f at the k distinct points $\tau_1, \ldots, \tau_k$.

We investigate whether QV is 1-1 or onto, hence invertible. For this, consider the matrix QW, with the columns of $W := [w_0, \ldots, w_{k-1}]$ the so-called **Newton polynomial**s

$$w_j(t) := (t - \tau_1) \cdots (t - \tau_j), \quad j = 0, 1, \ldots,$$

(with w_0, as the empty product, equal to $()^0$). Observe that $(QW)_{ij} = (Qw_{j-1})(\tau_i) = \prod_{0<h<j}(\tau_i - \tau_h) = 0$ if and only if $i < j$. Therefore, QW is square and lower triangular with nonzero diagonal entries, hence invertible by (3.36)Proposition, while w_{j-1} is a polynomial of degree $j-1<k$, hence $w_{j-1}=V\mathbf{c}_j$ for some k-vector $\mathbf{c}_j$. It follows that the invertible matrix QW equals

$$QW=[Qw_0,\ldots,Qw_{k-1}]=[QV\mathbf{c}_1,\ldots,QV\mathbf{c}_k]=(QV)[\mathbf{c}_1,\ldots,\mathbf{c}_k].$$

In particular, QV is onto, hence invertible, hence also V is 1-1, therefore invertible as a linear map from $\mathbb{R}^k$ to its range, $\Pi_{<k}$. We have proved:

(3.38) Proposition: For every $f:\mathbb{R}\to\mathbb{R}$ and every k distinct points $\tau_1,\ldots,\tau_k$ in $\mathbb{R}$, there is exactly one choice of coefficient vector $\mathbf{a}$ for which the polynomial $[()^0,\ldots,()^{k-1}]\mathbf{a}$ of degree $<k$ agrees with f at these τ_j.

In particular, (i) the column map $[()^0,\ldots,()^{k-1}]:\mathbb{R}^k\to\Pi_{<k}$ is invertible, and (ii) any polynomial of degree $<k$ with more than $k-1$ distinct zeros must be 0. (Do not confuse this simple result with the **Fundamental Theorem of Algebra** which asserts that every nonconstant polynomial with complex coefficients has a zero.)

3.31* Prove that $W:=[w_0,\ldots,w_{k-1}]$ is a basis for $\Pi_{<k}$ even if we drop the requirement that the τ_i are pairwise distinct.

3.32 Assume that $(\tau_1,\ldots,\tau_{2k+1})$ is nondecreasing. Prove that $W=[w_0,\ldots,w_k]$ *with* $w_j:t\mapsto(t-\tau_{j+1})\cdots(t-\tau_{j+k})$ *is a basis for* $\Pi_{\le k}$ *if and only if* $\tau_k<\tau_{k+1}$.

3.33 (a) Construct the unique element of ran$[()^0,()^2,()^4]$ that agrees with $()^1$ at the three points 0, 1, 2.

(b) Could (a) have been carried out if the pointset had been -1, 0, 1 (instead of 0, 1, 2)?

3.34 Let $\tau_1\ne\tau_2$. Prove that, *for an arbitrary* $\mathbf{a}\in\mathbb{R}^4$, *there exists exactly one cubic polynomial* p *for which*

$$(p(\tau_1),Dp(\tau_1),p(\tau_2),Dp(\tau_2))=\mathbf{a}.$$

(Hint: Try $W:=[()^0,(\cdot-\tau_1),(\cdot-\tau_1)^2,(\cdot-\tau_1)^2(\cdot-\tau_2)]$.)

3.35 T/F

(a) If one of the columns of a column map is 0, then the map cannot be 1-1.

(b) If the column map V into $\mathbb{R}^n$ is 1-1, then V has at most n columns.

(c) If the column map V into $\mathbb{R}^n$ is onto, then V has at most n columns.

(d) If a column map fails to be 1-1, then it has a zero column.

(e) If a vector space has only one basis, then it must be the trivial space.

(f) If a column of a matrix A is free, then it cannot be part of a basis for ran A.

4 Elimination, or: The determination of null A and ran A

Elimination and Backsubstitution

Elimination has as its goal an efficient description of the solution set for the *homogeneous* linear system $A? = \mathbf{0}$, i.e., of the nullspace of the matrix A. It identifies the free and bound columns of A, thereby (see (3.10)) also providing a basis for $\operatorname{ran} A$. Elimination is based on the following observation:

(4.1) Lemma: If B is obtained from A by subtracting some multiple of some row of A from some *other* row of A, then $\operatorname{null} B = \operatorname{null} A$.

Proof: Assume, more specifically, that B is obtained from A by subtracting α times row k from row i, for some $k \neq i$. Then, by (2.21)Example,

$$B = E_{\mathbf{e}_i,\mathbf{e}_k}(-\alpha)\, A,$$

with $E_{\mathbf{e}_i,\mathbf{e}_k}(-\alpha) = \operatorname{id}_m - \alpha \mathbf{e}_i {\mathbf{e}_k}^{\mathrm{t}}$. Consequently, $\operatorname{null} B \supset \operatorname{null} A$, and this holds even if $i = k$.

However, since $i \neq k$, we have ${\mathbf{e}_k}^{\mathrm{t}}\mathbf{e}_i = 0$, hence, for any α, $1+\alpha({\mathbf{e}_k}^{\mathrm{t}}\mathbf{e}_i) = 1 \neq 0$. Therefore, by (2.34), also

$$E_{\mathbf{e}_i,\mathbf{e}_k}(\alpha)\, B = A,$$

hence also $\operatorname{null} B \subset \operatorname{null} A$. □

One solves the homogeneous linear system $A? = \mathbf{0}$ by **elimination**. This is an *inductive* process, and it results in a classification of the unknowns as *free* or *bound*. A **bound** unknown has associated with it a **pivot row** or **pivot equation** which determines this unknown uniquely once all later unknowns are determined. Any unknown without a pivot equation is a **free** unknown; its value can be chosen arbitrarily. We will show (see (4.7)) that (not surprisingly) the jth *column* of A is bound (free) in the sense of (3.5)Definition exactly when the jth unknown is bound (free). The classification proceeds inductively, from the first to the last unknown or column, i.e., for $k = 1, 2, \ldots$, with the kth step as follows.

At the beginning of the kth **elimination step**, we have in hand a matrix B, called the **work-array**, which is **row-equivalent**[†] to our initial matrix A in the sense that $\operatorname{null} B = \operatorname{null} A$. Further, we have already classified each of the first $k-1$ unknowns as either bound or free, with each bound unknown associated with a particular row of B, its *pivot row*, and this row having a nonzero entry at the position of its associated bound unknown and zero entries for all previous unknowns. All other rows of B do not involve the unknowns already classified, i.e., they have nonzero entries only for unknowns not yet classified. Note that, with the choice $B := A$, this description also fits the situation at the beginning of the first step. We now classify the kth unknown and, correspondingly, change B, as follows:

bound case: We call the kth unknown **bound** (some would say **basic**) in case we can find some row $B_{h:}$ not yet used as pivot row for which $B_{hk} \neq 0$. We pick one such row and call it the **pivot row** for the kth unknown. Further, we use it to eliminate the kth unknown from all the rows $B_{i:}$ not yet used as pivot rows by the calculation

$$B_{i:} \leftarrow B_{i:} - \frac{B_{ik}}{B_{hk}} B_{h:} \; .$$

free case: In the contrary case, we call the kth unknown **free** (some would say **nonbasic**). No action is required in this case, since none of the rows not yet used as a pivot row involves the kth unknown.

By (4.1)Lemma, the changes (if any) made in B will not change $\operatorname{null} B$. This finishes the kth elimination step, with B now fitting the above description at the beginning of the $(k+1)$st step.

An informal description of this process involves two buckets, one containing the pivot equations found so far, the other the remaining equations. At the outset, all equations are in the second bucket. At the kth step, we look in the second bucket for an equation involving the kth unknown explicitly and, if there are such, we deposit one such in the first bucket as the pivot

[†] This terminology derives from the fact that two real matrices have the same nullspace iff they have the same row space; see Problem 6.16.

equation for the kth unknown but not before we have used it to eliminate the kth unknown from all the other equations still in that second bucket. This does not change the joint nullspace of all the equations, i.e., the set of all n-vectors satisfying all the equations in both buckets. The process ends if all equations, if any, in the second bucket are trivial, meaning that they involve none of the unknowns explicitly, hence the joint nullspace of all the equations in the two buckets equals the joint nullspace of the equations in the first bucket.

For future reference, here is a formal description of the entire algorithm. This description relies on an n-vector $\mathtt{p}$ to keep track of which row, if any, is used as pivot row for each of the unknowns. If row h is the pivot row for the kth unknown, then $p_k = h$ after the kth elimination step. Since $\mathtt{p}$ is initialized to have all its entries equal to 0, this means that, at any time, the rows h not yet used as pivot rows are exactly those for which h is not an entry of $\mathtt{p}$.

```
(4.2) Elimination Algorithm:
input: $A \in \mathbb{F}^{m \times n}$.
$B \leftarrow A$, $\mathtt{p} \leftarrow (0, \ldots, 0) \in \mathbb{Z}^n$.
for $k = 1{:}n$, do:
        for some $h \in \underline{m} \backslash \operatorname{ran} \mathtt{p}$ with $B_{hk} \neq 0$, do:
                $p_k \leftarrow h$
                for all $i \in \underline{m} \backslash \operatorname{ran} \mathtt{p}$, do:
                    $B_{i:} \leftarrow B_{i:} - \frac{B_{ik}}{B_{hk}} B_{h:}$
                enddo
        enddo
enddo
output: $B$, $\mathtt{p}$, and, possibly, free ← find(p==0), bound ← find(p>0).
```

Note that nothing is done at the kth step if there is no $h \notin \operatorname{ran} \mathtt{p}$ with $B_{hk} \neq 0$, i.e., if $B_{hk} = 0$ for all $h \notin \operatorname{ran} \mathtt{p}$. In particular, p_k will remain 0 in that case.

A numerical example: We start with

$$(4.3) \qquad A := \begin{bmatrix} 0 & 2 & 0 & 2 & 5 & 4 & 0 & 6 \\ 0 & 1 & 0 & 1 & 2 & 2 & 0 & 3 \\ 0 & 2 & 0 & 2 & 5 & 4 & -1 & 7 \\ 0 & 1 & 0 & 1 & 3 & 2 & -1 & 4 \end{bmatrix}, \qquad \mathtt{p} = (0, 0, 0, 0, 0, 0, 0, 0).$$

The first unknown is free. We take the second row as pivot row for the second unknown and eliminate it from the remaining rows, to get

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & \underline{\mathbf{1}} & 0 & 1 & 2 & 2 & 0 & 3 \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 & 1 \end{bmatrix}, \qquad \texttt{p} = (0,2,0,0,0,0,0,0).$$

Thus the third unknown is free as is the fourth, but the fifth is not, since there are nonzero entries in the fifth column of some row not yet used as pivot row, e.g., the first row. We choose the first row as pivot row for the fifth unknown and use it to eliminate this unknown from the remaining rows not yet used, i.e., from rows 3 and 4. This gives

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 & \underline{\mathbf{1}} & 0 & 0 & 0 \\ 0 & \mathbf{1} & 0 & 1 & 2 & 2 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}, \qquad \texttt{p} = (0,2,0,0,1,0,0,0).$$

The sixth unknown is free, but there are nonzero entries in the seventh column of the remaining rows not yet used, so the seventh unknown is bound, with, e.g., the fourth row as its pivot row. We use that row to eliminate the seventh unknown from the remaining row not yet used. This gives

$$(4.4) \qquad B = \begin{bmatrix} 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 \\ 0 & \mathbf{1} & 0 & 1 & 2 & 2 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \underline{\mathbf{-1}} & 1 \end{bmatrix}, \qquad \texttt{p} = (0,2,0,0,1,0,4,0).$$

With that, all rows not yet used are trivial, hence unusable as pivot rows. In particular, the eighth unknown is free, hence we have already in hand the final array.

Altogether, `bound` $= (2,5,7)$ (= `find(p>0)`) and `free` $= (1,3,4,6,8)$ (= `find(p==0)`). □

After the n steps of this elimination process (which started with $B = A$), we have in hand a matrix B with $\operatorname{null} B = \operatorname{null} A$ and with each unknown classified as bound or free. The two increasing sequences, `bound` and `free`, containing the indices of the bound and free unknowns respectively, will be much used in the sequel. Each bound unknown has associated with it a particular row of B, its pivot row. All rows of B not yet used as a pivot row (if any) are entirely zero.

Neat minds would reorder the rows of B, listing first the pivot rows in the order of their corresponding bound unknowns, followed by the rows not yet used as pivot rows and, in this way, obtain a **row echelon form** for A, i.e., a matrix in which the first nonzero entry (if any) in a row is to the right of the first nonzero entry in the preceding row. In any case, in determining $\mathtt{x} \in \operatorname{null} B$, we only have to pay attention to the pivot rows. This means that we can determine a particular element $\mathtt{x}$ of $\operatorname{null} B = \operatorname{null} A$ by *backsubstitution*, i.e., from its last entry to its first as follows:

For $k = n, n{-}1, \ldots, 1$, if the kth unknown is bound, i.e., $k \in \operatorname{ran} \mathtt{bound}$, determine x_k from its pivot equation, recognizable by the fact that its row in B is the only one that has its first or leftmost nonzero entry in the kth column, hence provides a formula for x_k in terms of $x_{k+1}, \ldots, x_n$; else, pick x_k arbitrarily (as then the kth unknown is free, i.e., $k \in \operatorname{ran} \mathtt{free}$).

A numerical example, continued: We try out backsubstitution on the earlier numerical example, using the final version of the work-array B as recorded in (4.4), with $n = 8$, and recalling that $\mathtt{free} = [1, 3, 4, 6, 8]$.

The last unknown is free, hence we leave its value open to choice, by using the symbol z_8 for its value, i.e., set $x_8 = z_8$.

The seventh unknown is bound, with the fourth row its pivot row, giving us $-x_7 + x_8 = 0$, or $x_7 = z_8$.

The sixth unknown is free, hence we set $x_6 = z_6$ for some value z_6 freely choosable.

The fifth unknown is bound, with the first row its pivot row, giving us $x_5 = 0$.

The fourth and third unknown are both free, hence we set $x_4 = z_4$ and $x_3 = z_3$ for some values z_3 and z_4 freely choosable.

The second unknown is bound, with the second row its pivot row, giving us $x_2 + x_4 + 2x_5 + 2x_6 + 3x_8 = 0$, or, substituting in the values of x_j, $j > 2$, already determined, $x_2 + z_4 + 2z_6 + 3z_8 = 0$, hence $x_2 = -z_4 - 2z_6 - 3z_8$.

The first unknown is free, hence we set $x_1 = z_1$ for some value z_1 freely choosable.

Thus the *general* solution of $B? = \mathbf{0}$, i.e., the general element of $\operatorname{null} B = \operatorname{null} A$ with A the matrix that started off the numerical example, is

$$\begin{aligned}
&(z_1, -z_4 - 2z_6 - 3z_8, z_3, z_4, 0, z_6, z_8, z_8) \\
&\quad = z_1\mathbf{e}_1 + z_3\mathbf{e}_3 + z_4(-\mathbf{e}_2 + \mathbf{e}_4) + z_6(-2\mathbf{e}_2 + \mathbf{e}_6) \\
&\qquad + z_8(-3\mathbf{e}_2 + \mathbf{e}_7 + \mathbf{e}_8)\,.
\end{aligned} \tag{4.5}$$

In other words, $\operatorname{null} A = \operatorname{ran} V$ with

$$V := [\mathbf{e}_1, \mathbf{e}_3, -\mathbf{e}_2 + \mathbf{e}_4, -2\mathbf{e}_2 + \mathbf{e}_6, -3\mathbf{e}_2 + \mathbf{e}_7 + \mathbf{e}_8]$$

1-1 since, e.g., $V(\mathtt{free}, :) = \mathrm{id}_5$, hence V is a basis for $\operatorname{null} A$. □

Here is a more formal description of backsubstitution, for future reference.

(4.6) Backsubstitution Algorithm:

```
input: $B \in \mathbb{F}^{m\times n}$ and p (both as output from (4.2)), $\mathbf{z} \in \mathbb{F}^n$.
x ← z
for $k = n{:}{-1}{:}1$, do:
        if $p_k \neq 0$, then $x_k \leftarrow -\left(\sum_{j>k} B_{p_k,j}x_j\right)/B_{p_k,k}$ endif
enddo
output: x, which is the unique solution of $A? = \mathbf{0}$ satisfying $x_i = z_i$ for all $i$ with $p_i = 0$.
```

Notice that, as in the example, the value of every free unknown is arbitrary and that, once these values are chosen, then the bound unknowns are uniquely determined by the requirement that we are seeking an element of $\operatorname{null} B = \operatorname{null} A$. In other words, the linear map

$$\operatorname{null} A \to \mathbb{F}^{\mathtt{free}} : \mathbf{x} \mapsto x_{\mathtt{free}}$$

is 1-1 and onto, hence invertible. Therefore, $\dim\operatorname{null} A = \#\mathtt{free}$, by (3.21).

By the (3.23)Dimension Formula, this implies that $\dim\operatorname{ran} A = \#\mathtt{bound}$, and this is no accident since, by (3.10), the bound columns of A are a basis for $\operatorname{ran} A$ while the following observation asserts that $A(:,\mathtt{bound})$ comprises the bound columns of A.

(4.7) Observation: The kth unknown of $A? = \mathbf{0}$ is free(bound) if and only if the kth column of A is free(bound).

Proof: It is sufficient to prove that the kth column of A is free if and only if the kth unknown of $A? = \mathbf{0}$ is free. For that, recall from (3.6)Proposition that the kth column of A is free if and only if there exists $\mathbf{x} \in \operatorname{null} A$ with x_k its rightmost nonzero entry.

Now observe that, for any k, if the kth entry, x_k, of an $\mathbf{x} \in \operatorname{null} B = \operatorname{null} A$ is nonzero, then either (a) the kth unknown is free; or else (b) the kth unknown is bound, but then necessarily $x_j \neq 0$ for some $j > k$. It follows that x_k can be the rightmost nonzero entry of such an $\mathbf{x}$ only if the kth unknown is free. Conversely, if the kth unknown is free, and $\mathbf{x}$ is the element of $\operatorname{null} B = \operatorname{null} A$ computed by setting $x_k = 1$ and setting all other free unknowns equal to 0, then x_k is necessarily the rightmost nonzero entry of $\mathbf{x}$ (since all free entries to the right of it were chosen to be zero, thus preventing any bound entry to the right of it from being nonzero). □

This simple observation gives a *characterization* of the sequence `free` entirely in terms of the nullspace of the matrix A we started with. This implies that *the classification into free and bound unknowns or columns is independent of the choice of pivot rows made during elimination*. More than that, since, for any 1-1 matrix M with m columns, $\operatorname{null}(MA) = \operatorname{null} A$, it implies that, for any such matrix MA, we get exactly the same sequences `free` and `bound` as we would get for A. This is the major reason for the uniqueness of a more disciplined echelon form, the ‘really reduced row echelon form’, to be discussed in the next section.

(4.8) Corollary: For every matrix A,

(i) the kth column of A is free if and only if it is a weighted sum of the columns strictly to the left of it, i.e., $A_{\mathbf{:}k} \in \operatorname{ran} A(\,:\,, 1{:}k{-}1)$;

(ii) $A(\,:\,, 1{:}k)$ is 1-1 if and only if all its columns are bound (in $A(\,:\,, 1{:}k)$, hence in A);

(iii) $\operatorname{null} A$ is nontrivial if and only if A has free columns.

4.1* Determine the bound and free columns for each of the following matrices A.

(a) $0 \in \mathbb{R}^{m\times n}$; (b) $[\mathbf{e}_1, \ldots, \mathbf{e}_n] \in \mathbb{R}^{n\times n}$; (c) $[\mathbf{e}_1, \mathbf{0}, \mathbf{e}_2, \mathbf{0}] \in \mathbb{R}^{6\times 4}$; (d) $\begin{bmatrix} 2 & 2 & 5 & 6 \\ 1 & 1 & -2 & 2 \end{bmatrix}$; (e) $\begin{bmatrix} 0 & 2 & 1 & 4 \\ 0 & 0 & 2 & 6 \\ 1 & 0 & -3 & 2 \end{bmatrix}$; (f) $[\mathbf{x}][\mathbf{x}]^{\mathrm{t}}$, with $\mathbf{x} = (1, 2, 3, 4)$.

4.2 Use (4.8)Corollary (iii) for a quick proof of (3.25)Theorem.

4.3* Prove: (1) *a column of a matrix is free regardless of where it appears in the matrix if and only if it is free as a first column, i.e., is a zero column*; (2) *free columns have no influence on whether a nonzero column is free or bound*; (3) *a column of a matrix is bound regardless of where it appears in the matrix if and only if it is bound as a last column.*

4.4 (4.8)Corollary assures you that $\mathbf{y} \in \operatorname{ran} A$ if and only if the last column of $[A, \mathbf{y}]$ is free. Use this fact to determine, for each of the following $\mathbf{y}$ and A, whether or not $\mathbf{y} \in \operatorname{ran} A$.

(a) $\mathbf{y} = (\pi, 1-\pi)$, $A = \begin{bmatrix} 1 & -2 \\ -1 & 2 \end{bmatrix}$; (b) $\mathbf{y} = \mathbf{e}_2$, $A = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 3 & -4 \\ 3 & 4 & -8 \end{bmatrix}$; (c) $\mathbf{y} = \mathbf{e}_2$, $A = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 3 & -4 \\ 3 & 4 & -7 \end{bmatrix}$.

4.5 Prove (4.1)Lemma directly, i.e., without using (2.34)Proposition. (Hint: Prove that $\operatorname{null} B \supset \operatorname{null} A$. Then prove that also A is obtainable from B by the same kind of step, hence also $\operatorname{null} A \supset \operatorname{null} B$.)

4.6* The previous homework uses the idea that the inverse of a map undoes what the map does. Use this idea for a proof of (2.34)Proposition.

4.7 Prove: *If M and A are matrices for which MA is defined and, furthermore, M is 1-1, then $MA? = 0$ has exactly the same free and bound unknowns as does $A? = 0$.*

4.8 Assuming the matrix A has exactly α bound columns and the matrix B has exactly β bound columns and both have the same number of rows, how many bound columns does the matrix $[A, B]$ have (a) at least? (b) at most? (c) How, if at all, would your answers to (a), (b) change if I told you that A has m rows?

4.9* Use (3.29)Corollary and elimination to determine for each of the matrices given whether $\operatorname{ran} A$ and $\operatorname{null} A$ have nontrivial intersection: (a) $A := \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$; (b) $A := \begin{bmatrix} -2 & -1 \\ 4 & 2 \end{bmatrix}$.

The really reduced row echelon form and other reduced forms

The construction of the *really reduced row echelon form* takes elimination four steps further, none of which changes the nullspace, since each of the first three steps amounts to changing the work-array B to EB, with E an evidently invertible matrix, hence $\operatorname{null} EB \subset \operatorname{null} B = \operatorname{null} E^{-1}EB \subset \operatorname{null} EB$, while the fourth step amounts to leaving off any row entirely zero:

(i) When the hth pivot row is found, and it is not the hth row, then it is exchanged with the current hth row to make it the hth row. (This keeps things neat; all the rows not yet used as pivot rows lie below all the rows already picked as pivot rows.) This doesn't change the nullspace of the work-array since the set of equations is unchanged. More formally, each such exchange changes the work-array B to EB with E the linear map that exchanges the ith entry with the jth which is invertible since it is its own inverse.

(ii) Each pivot row is divided by its **pivot element**, i.e., by its left-most nonzero entry. (This helps with the elimination of the corresponding unknown from other rows: if B_{hk} is the pivot element in question (i.e., $bound_h = k$, i.e., x_k is the hth bound unknown), then, after this normalization, one merely subtracts B_{ik} times $B_{h:}$ from $B_{i:}$ to eliminate the kth unknown from row i.) This doesn't change the nullspace of the work-array since multiplication of the new pivot row by the (former) pivot element reverses the action.

The work-array B at this point is said to have been obtained from A by **Gauss elimination with partial (row) pivoting**. Use of the next step completes what is known as **Gauss-Jordan elimination**.

(iii) One eliminates each bound unknown from *all* rows (other than its pivot row), i.e., also from pivot rows belonging to earlier bound unknowns, and not just from the rows not yet used as pivot rows. For real efficiency, though, this additional step should be carried out after elimination is completed; it starts with the elimination of the *last* bound unknown, proceeds to the second-last bound unknown, etc., and ends with the *second* bound unknown (the first bound unknown was eliminated from all other rows already).

The resulting matrix B is called the **reduced row echelon form for** A, and this is written:

$$B = \operatorname{rref}(A).$$

However, it turns out to be very neat to add the following final step:

(iv) Remove all rows that are entirely zero, thus getting the matrix

$$R := B(1{:}\#\texttt{bound}, :) =: \mathrm{rrref}(A)$$

which I call the *really reduced row echelon form* of A.

Here is a formal description (in which we talk about *the* rrref for A even though we prove its *uniqueness* only later; see (4.21)):

(4.9) Definition: We say that R is the **really reduced row echelon form for** $A \in \mathbb{F}^{m\times n}$ and write

$$R = \mathrm{rrref}(A),$$

in case $R \in \mathbb{F}^{r\times n}$ for some r and there is a strictly increasing r-sequence `bound` (provided by the `MATLAB` function **`rref`** along with $\mathrm{rref}(A)$) so that the following is true:

1. R is a **row echelon form for** A: This means that (i) $\mathrm{null}\, R = \mathrm{null}\, A$; and (ii) for each $k = bound_i$, $R_{i:}$ is the pivot row for the kth unknown, i.e., $R_{i:}$ is the unique row in R for which R_{ik} is the first (or, leftmost) nonzero entry.

2. R is **really reduced** or normalized, in the sense that $R(:, \texttt{bound})$ is the identity matrix, i.e., for each i, the pivot element $R_{i,bound_i}$ equals 1 and is the only nonzero entry in its column, and R has only these $r = \#\texttt{bound}$ rows.

A numerical example, continued: For the matrix A given in (4.3) of the earlier numerical example, the rref and the rrref look like this:

$$\begin{bmatrix} 0 & \mathbf{1} & 0 & 1 & 0 & 2 & 0 & 3 \\ 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & \mathbf{1} & 0 & 1 & 0 & 2 & 0 & 3 \\ 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & -1 \end{bmatrix}. \tag{4.10}$$

Recall (or observe directly) that, for this example, $\texttt{bound} = (2, 5, 7)$ and $\texttt{free} = (1, 3, 4, 6, 8)$. □

Finally, for most purposes, it is sufficient to have a `b`**-form** for A.

(4.11) Definition: We say that R is the **really reduced row echelon form for $A \in \mathbb{F}^{m\times n}$ with respect to the index sequence** `b` and write

$$R = \text{rrref}_{\texttt{b}}(A),$$

in case $R \in \mathbb{F}^{r\times n}$ for some r and satisfies the following two conditions:

(4.12)(i) $\operatorname{null} R = \operatorname{null} A$;

(4.12)(ii) $R(:,\texttt{b}) = \text{id}$.

We will also call such R more briefly a `b`**-form for** A. For example, rrref$A$ is a `bound`-form for A.

A matrix A may have a `b`-form for many different `b` and, as we shall see, only the two conditions (4.12)(i-ii) really matter when we are interested in a basis for $\operatorname{null} A$ or $\operatorname{ran} A$. Moreover, we have, in effect, a `b`-form for A in hand well before we get to rrref(A). For, there is no need to reorder the rows of the work-array; we drop every row entirely zero, then eliminate each bound unknown from all rows but its pivot row, being sure first to divide each pivot row by its pivot element, and then, with R the resulting array, have in hand a `b`-form for A, with `b` the permutation of `bound=find(p>0)` for which $R(:,\texttt{b}) = \text{id}$.

For the example worked out earlier, at the stage recorded in (4.4), we would drop the third row, divide the now third row by its pivot element, -1, and eliminate the fifth unknown from the second row, and note that, for the resulting matrix

$$\begin{bmatrix} 0&0&0&0&\mathbf{1}&0&0&0\\ 0&\mathbf{1}&0&1&2&2&0&3\\ 0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&-\mathbf{1}&1 \end{bmatrix} \rightarrow \begin{bmatrix} 0&0&0&0&\mathbf{1}&0&0&0\\ 0&\mathbf{1}&0&1&0&2&0&3\\ 0&0&0&0&0&0&\mathbf{1}&-1 \end{bmatrix} =: R,$$

the permutation `b:=(5,2,7)` of `bound = (2,5,7)` gives $R(:,\texttt{b}) = \text{id}$.

A numerical example, modified: We start again with

$$A := \begin{bmatrix} 0&2&0&2&5&4&0&6\\ 0&1&0&1&2&2&0&3\\ 0&2&0&2&5&4&-1&7\\ 0&1&0&1&3&2&-1&4 \end{bmatrix}, \qquad \texttt{p} = (0,0,0,0,0,0,0,0).$$

But, this time, we do elimination free-style, choosing the unknown to be eliminated and the corresponding pivot row capriciously. This produces the following, quite different, sequence of work-arrays B with the same nullspace as A:

Using row 3 to eliminate the 7th unknown from all the other rows (after dividing it by its 7th entry) gives

$$B = \begin{bmatrix} 0 & 2 & 0 & 2 & 5 & 4 & 0 & 6 \\ 0 & 1 & 0 & 1 & 2 & 2 & 0 & 3 \\ 0 & -2 & 0 & -2 & -5 & -4 & \underline{\mathbf{1}} & -7 \\ 0 & -1 & 0 & -1 & -2 & -2 & 0 & -3 \end{bmatrix}, \qquad \mathtt{p} = (0,0,0,0,0,0,3,0).$$

Using row 2 to eliminate the 8th unknown from all the other rows (after dividing it by its 8th entry) gives

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1/3 & 0 & 1/3 & 2/3 & 2/3 & 0 & \underline{\mathbf{1}} \\ 0 & 1/3 & 0 & 1/3 & -1/3 & 2/3 & \mathbf{1} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \qquad \mathtt{p} = (0,0,0,0,0,0,3,2).$$

Dropping the last row since it is entirely zero, then using the first row to eliminate the 5th unknown from all the other rows gives

$$(4.13)\quad B = \begin{bmatrix} 0 & 0 & 0 & 0 & \underline{\mathbf{1}} & 0 & 0 & 0 \\ 0 & 1/3 & 0 & 1/3 & 0 & 2/3 & 0 & \mathbf{1} \\ 0 & 1/3 & 0 & 1/3 & 0 & 2/3 & \mathbf{1} & 0 \end{bmatrix}, \qquad \mathtt{p} = (0,0,0,0,1,0,3,2).$$

This last version of B is a $\mathtt{b}$-form for A with $\mathtt{b} = (5,8,7)$, since $B(:,\mathtt{b}) = \mathrm{id}_3$ and we reached B from A by a sequence of invertible row operations and by dropping a zero row, hence $\operatorname{null} B = \operatorname{null} A$. □

Note that, in the 'capricious' elimination practiced in the preceding example, the set of unknowns ending up with a pivot row is, by the very capriciousness of my choices, quite different from the set of bound unknowns obtained with the elimination algorithm earlier. Therefore, also the set of unknowns ending up without a pivot row is quite different from the set of free unknowns obtained with the elimination algorithm earlier (although the two sets have a nontrivial intersection). Yet, as will be clear eventually, it is no accident that the number of unknowns with a pivot row is the same as was obtained by the elimination algorithm earlier, i.e., $\#\mathtt{b} = \#\mathtt{bound}$, since we already know by (3.10) that $A(:,\mathtt{bound})$ is a basis for $\operatorname{ran} A$ while, more generally, $A(:,\mathtt{b})$ is a basis $\operatorname{ran} A$, by (4.21)Proposition yet to be proved.

4.10 For each of the matrices A in Problem 4.1, determine its rrref.

4.11* Use elimination from right to left, i.e., starting at the last unknown, to get a $\mathtt{b}$-form for the matrices in Problem 4.1(d) and (e).

The basis for null A obtained from a b-form

We construct in this section, from a b-form R for A, a basis for null A.

In recognition of the special case $R = \text{rrref}(A)$, we use f for a sequence **complementary to** b in the sense that it contains all the indices in $\underline{n}$ that are not in b.

In MATLAB, one would obtain f from n and b by the commands `f = 1:n; f(b) = [];`

In the discussion, we use the following notation introduced in (1.17): If $\mathbf{x}$ is an n-vector and p is a list of length r with range in $\underline{n}$, then $x_{\mathtt{p}}$ is the r-vector

$$x_{\mathtt{p}} = (x_{p_i} : i \in \underline{r}).$$

Further, if p is 1-1 into $\underline{n}$, and q is complementary to p in the sense that it is also 1-1 into $\underline{n}$ and has in its range all the elements of $\underline{n}$ not in the range of p, then (see Problem 2.18), for any $B \in \mathbb{F}^{m\times n}$,

$$B\mathbf{x} = \sum_{j=1}^{n} B_{:j}x_j = \sum_{j\in\mathtt{p}} B_{:j}x_j + \sum_{j\in\mathtt{q}} B_{:j}x_j = B(\,:\,,\mathtt{p})x_{\mathtt{p}} + B(\,:\,,\mathtt{q})x_{\mathtt{q}}.$$

With this, by property (4.12)(i),

$$\mathbf{x} \in \text{null}\, A \quad\Longleftrightarrow\quad \mathbf{0} = R\mathbf{x} = R(\,:\,,\mathtt{b})x_{\mathtt{b}} \;+\; R(\,:\,,\mathtt{f})x_{\mathtt{f}}\,.$$

Since $R(\,:\,,\mathtt{b}) = \text{ id}$ by property (4.12)(ii), we conclude that

$$\mathbf{x} \in \text{null}\, A \quad\Longleftrightarrow\quad x_{\mathtt{b}} = -R(\,:\,,\mathtt{f})x_{\mathtt{f}}.$$

We can write this even more succinctly in matrix form as follows:

$$\text{null}\, A = \text{ran}\, C,$$

with C the $(n \times \#\mathtt{f})$-matrix whose 'f-rows' form an identity matrix, and the columns of whose 'b-rows' form the 'f-columns' of $-R$:

$$(4.14) \qquad C(\mathtt{f},\,:\,) = \text{ id}, \qquad C(\mathtt{b},\,:\,) = -R(\,:\,,\mathtt{f}).$$

Note that C is 1-1 since, by (4.14), $C\mathbf{a} = \mathbf{0}$ implies, in particular, that $\mathbf{a} = (C\mathbf{a})_{\mathtt{f}} = \mathbf{0}$. Hence, C is a basis for null A.

E.g., for the A of the earlier numerical example (4.3), we obtained

$$R = \text{rrref}(A) = \begin{bmatrix} 0 & \mathbf{1} & 0 & 1 & 0 & 2 & 0 & 3 \\ 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & -1 \end{bmatrix}$$

in (4.10), with $\texttt{b}= (2,5,7)$ and $\texttt{f}= (1,3,4,6,8)$, hence

$$C \;=\; \begin{bmatrix} \mathbf{1} & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & -2 & -3 \\ 0 & \mathbf{1} & 0 & 0 & 0 \\ 0 & 0 & \mathbf{1} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{1} & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & \mathbf{1} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & \mathbf{1} & 0 & 0 & 0 \\ 0 & 0 & \mathbf{1} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{1} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathbf{1} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ -0 & -0 & -1 & -2 & -3 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -0 & -0 & -0 & -0 & -0 \\ 0 & 0 & 0 & 0 & 0 \\ -0 & -0 & -0 & -0 & -(-1) \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

It is comforting to discover in this example in the columns of the matrix C the vectors that appear in the description (4.5) of the general element of $\operatorname{null} A$ worked out earlier.

Using instead the $\texttt{b}$-form for this A from (4.13),

$$R = \operatorname{rrref}_{\texttt{b}}(A) = \begin{bmatrix} 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 \\ 0 & 1/3 & 0 & 1/3 & 0 & 2/3 & 0 & \mathbf{1} \\ 0 & 1/3 & 0 & 1/3 & 0 & 2/3 & \mathbf{1} & 0 \end{bmatrix},$$

hence $\texttt{b}= (5,8,7)$ and $\texttt{f}= (1,2,3,4,6)$, gives the quite different matrix

$$C \;=\; \begin{bmatrix} \mathbf{1} & 0 & 0 & 0 & 0 \\ 0 & \mathbf{1} & 0 & 0 & 0 \\ 0 & 0 & \mathbf{1} & 0 & 0 \\ 0 & 0 & 0 & \mathbf{1} & 0 \\ -0 & -0 & -0 & -0 & -0 \\ 0 & 0 & 0 & 0 & \mathbf{1} \\ 0 & -1/3 & 0 & -1/3 & -2/3 \\ 0 & -1/3 & 0 & -1/3 & -2/3 \end{bmatrix}.$$

Finally, when $R = \operatorname{rrref}(A)$, hence $\texttt{f} = \texttt{free}$, then the resulting C is 'upper triangular' in the sense that then

$$(4.15) \qquad\qquad i > free_j \quad\Longrightarrow\quad C_{ij} = 0.$$

4.12 Determine a basis for the nullspace of $A := \begin{bmatrix} 2 & 3 \\ 2 & 3 \end{bmatrix}$ and use it to describe the solution set of the system $A? = (6,6)$. Draw a picture indicating both the solution set and $\operatorname{null} A$.

4.13 For each of the matrices A in Problem 4.1, give a basis for null A.

4.14* Use elimination to determine a necessary and sufficient condition on the scalars α and β for which the matrix $A := \begin{bmatrix} \alpha & 2 & \beta \\ 0 & 2 & 2 \\ 2 & 1 & 2 \end{bmatrix}$ is not 1-1.

The factorization $A = A(\,:\,,\mathtt{b})\mathrm{rrref}_{\mathtt{b}}(A)$

It turns out that any $\mathtt{b}$-form for A is also a factor of A in the following sense.

(4.16) Proposition: If A has a $\mathtt{b}$-form, then

$$A = A(\,:\,,\mathtt{b})\mathrm{rrref}_{\mathtt{b}}(A). \tag{4.17}$$

Proof: Recall from Problem 2.19 that, for $A \in \mathbb{F}^{m\times n}$ and $B \in \mathbb{F}^{n\times k}$ and $(\mathtt{p},\mathtt{q})$ a permutation of degree n,

$$AB \;=\; A(\,:\,,\mathtt{p})B(\mathtt{p},\,:\,) + A(\,:\,,\mathtt{q})B(\mathtt{q},\,:\,).$$

With this, observe that, with $R := \mathrm{rrref}_{\mathtt{b}}(A)$ and with C given by (4.14), hence $AC = 0$, we have

$$\begin{aligned} 0 = AC \;&=\; A(\,:\,,\mathtt{b})C(\mathtt{b},\,:\,) + A(\,:\,,\mathtt{f})C(\mathtt{f},\,:\,) \\ &=\; A(\,:\,,\mathtt{b})(-R(\,:\,,\mathtt{f})) + A(\,:\,,\mathtt{f}) \end{aligned}$$

showing that $A(\,:\,,\mathtt{b})R(\,:\,,\mathtt{f}) = A(\,:\,,\mathtt{f})$, while $A(\,:\,,\mathtt{b})R(\,:\,,\mathtt{b}) = A(\,:\,,\mathtt{b})$ since $R(\,:\,,\mathtt{b}) =$ id by (4.12)(ii). Since $\underline{n} = \mathrm{ran}\,\mathtt{b} \cup \mathrm{ran}\,\mathtt{f}$, this proves that $A(\,:\,,\mathtt{b})R(\,:\,,j) = A(\,:\,,j)$ for all $j \in \underline{n}$, hence proves (4.17). □

In particular, for $\mathtt{b} = \mathtt{bound}$,

$$A = A(\,:\,,\mathtt{bound})\,\mathrm{rrref}(A). \tag{4.18}$$

This says that we can view the task of constructing $R := \mathrm{rrref}(A)$ as finding the sequence $\mathtt{bound}$ indicating the bound columns of A, hence then know $R(\,:\,,\mathtt{bound}) =$ id, and then finding each of the remaining (free) columns of R as the solution of the linear system $A(\,:\,,\mathtt{bound})? = \mathbf{v}$, with $\mathbf{v}$ the corresponding (free) column of A.

Why should this linear system have solutions and, if it does, why should there be a *unique* solution? These questions are answered in the next section.

4.15 Verify that (4.17) holds for the two $\mathtt{b}$-forms worked out in Problem 4.11.

4.16 Prove: *If M is such that $MA = \mathrm{rrref}(A) =: R$, and* $\mathtt{bound}$ *is the increasing sequence of indices of bound columns of A, then M is a left inverse for $A(\,:\,,\mathtt{bound})$.*

The basis for $\operatorname{ran} A$ obtained from a $\mathtt{b}$-form

Let $R := \operatorname{rrref}_{\mathtt{b}}(A)$, hence R satisfies (4.12)(i–ii). Then we know from (4.17) that $A = A(:,\mathtt{b})R$. This factorization implies that $\operatorname{ran} A \subset \operatorname{ran} A(:,\mathtt{b})$, while certainly $\operatorname{ran} A(:,\mathtt{b}) \subset \operatorname{ran} A$. Hence

$$\operatorname{ran} A = \operatorname{ran} A(:,\mathtt{b}),$$

i.e., $A(:,\mathtt{b})$ *is onto* $\operatorname{ran} A$. Also, $A(:,\mathtt{b})$ *is 1-1*: For, if $A(:,\mathtt{b})\mathbf{a} = \mathbf{0}$, then the n-vector $\mathbf{x}$ with $x_{\mathtt{b}} = \mathbf{a}$ and with $x_{\mathtt{f}} = \mathbf{0}$ is in $\operatorname{null} A = \operatorname{null} R$, hence $\mathbf{0} = R\mathbf{x} = R(:,\mathtt{b})\mathbf{a} + R(:,\mathtt{f})\mathbf{0} = \mathbf{a}$ (since R is a $\mathtt{b}$-form, therefore $R(:,\mathtt{b}) = \mathrm{id}$). Consequently, $A(:,\mathtt{b})$ is a basis for $\operatorname{ran} A$.

Conversely, if, for some $\mathtt{b}$, $A(:,\mathtt{b})$ is a basis for $\operatorname{ran} A$, then

$$R := A(:,\mathtt{b})^{-1}A$$

is well-defined, and $\operatorname{null} R = \operatorname{null} A$ and $R(:,\mathtt{b}) = \mathrm{id}$, showing that R is the *unique* $\mathtt{b}$-form for A, i.e.,

$$\operatorname{rrref}_{\mathtt{b}}(A) = A(:,\mathtt{b})^{-1}A. \tag{4.19}$$

“Wait a moment!”, you now say,“Didn’t we learn the pigeon hole principle for linear maps and, in particular, for matrices, according to which an invertible matrix is necessarily square? So, what is meant by $A(:,\mathtt{b})^{-1}$ when $A(:,\mathtt{b})$ is not square?”

(4.20) Definition. If the matrix $V \in \mathbb{F}^{m\times n}$ is 1-1, hence a basis for its range, we denote by $V^{-1} \in L(\operatorname{ran} V, \mathbb{F}^n)$ the inverse of V as an element of $L(\mathbb{F}^n, \operatorname{ran} V)$.

To be sure, this V^{-1} is a matrix only when V is square, as then $\operatorname{ran} V = \mathbb{F}^n$, hence V^{-1} is the matrix of order n inverse to V. Yet, since we know this, we will not be confused by the notation to think, in the contrary case, that V^{-1} is an inverse *matrix.* Still, how one might construct V^{-1} in the contrary case needs to be discussed. This is the subject of the next chapter, which is devoted to the construction of the inverse of a basis, in general.

(4.21) Proposition: For any sequence $\mathtt{b}$, a matrix A has a $\mathtt{b}$-form if and only if its submatrix $A(:,\mathtt{b})$ is a 1-1 map onto $\operatorname{ran} A$, in which case the $\mathtt{b}$-form is the solution to the equation $A(:,\mathtt{b})? = A$, hence unique.

It follows that, for any matrix A, rrref(A) is uniquely determined by A since it is a `bound`-form for A, hence, by (4.21)Proposition, the unique `bound`-form for A while we know from (4.7) that the sequence `bound` depends only on null A.

Further, since rref(A) differs from rrref(A) only by those additional $\#A^{\mathrm{t}} - \#\mathtt{bound}$ zero rows, it follows that each A also has a *unique* rref.

4.17 For each of the matrices A in Problem 4.1, give a basis for ran A.

4.18 Let A be the $n \times n$ matrix $[\mathbf{0}, \mathbf{e}_1, \ldots, \mathbf{e}_{n-1}]$ (with $\mathbf{e}_j$ denoting the jth coordinate direction (vector), of the appropriate length). (a) What is its rref? (b) Which are its bound(free) columns? (c) Give a basis for null A and a basis for ran A.

4.19 Let A be the 6×3-matrix $[\mathbf{e}_3, \mathbf{e}_2, \mathbf{e}_1]$. (a) What is its rref? (b) Use (a) to prove that A is 1-1. (c) Construct a left inverse for A. (d) (off the wall:) Give a matrix P for which null P = ran A.

4.20 Let $B := A^{\mathrm{t}}$, with A the matrix in the previous problem. (a) What is its rref? (b) Use (a) to prove that B is onto. (c) Construct a right inverse for B.

4.21* Use the rref to prove that ran U = ran V, with

$$U := \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ -1 & 1 & 3 \end{bmatrix}, \qquad V := \begin{bmatrix} 1 & 2 \\ 2 & 4 \\ -4 & -5 \end{bmatrix}.$$

(Hints: Proving two sets to be equal usually involves showing that each is a subset of the other. In this case, applying elimination to $[V, U]$ as well as to $[U, V]$ should provide all the information you need.)

4.22* Show by an example that when A has a `b`-form, the matrix $A(:, \mathtt{b})$ need not be invertible (even though it is a 1-1 map onto ran A).

4.23 Show that *the matrix A has a `b`-form if and only if $A(:, \mathtt{b})$ is a maximally 1-1 submatrix of A, i.e., (i) $A(:, \mathtt{b})$ is 1-1 and (ii) for any $j \in 1{:}\#A$, $[A(:, \mathtt{b}), A_{:j}]$ is not 1-1.*

The rrref(A) and the solving of $A? = \mathbf{y}$

(4.8)Corollary(i) is exactly what we need when considering the linear system

$$(4.22) \qquad A? = \mathbf{y}$$

for given $A \in \mathbb{F}^{m \times n}$ and given $\mathbf{y} \in \mathbb{F}^m$. For, here we are hoping to write $\mathbf{y}$ as a linear combination of the columns of A, and (4.8) tells us that this is possible exactly when the last unknown in the *homogeneous* system

$$(4.23) \qquad [A, \mathbf{y}]? = \mathbf{0}$$

is free. Further, the factorization (4.18), applied to the **augmented** matrix $[A, \mathbf{y}]$, tells us how to write $\mathbf{y}$ as a linear combination of the columns of A in case that can be done. For, with $R = \mathrm{rrref}([A, \mathbf{y}])$, it tells us that

$$\mathbf{y} = [A, \mathbf{y}](:, \mathtt{bound}) R(:, n+1),$$

and this gives us $\mathbf{y}$ in terms of the columns of A precisely when $n + 1 \notin$ ran `bound`, i.e., when the $(n+1)$st unknown is free, hence $[A, \mathbf{y}](:, \mathtt{bound}) =$

$A(:,\texttt{bound})$.

(4.24) Proposition: For $A \in \mathbb{F}^{m\times n}$ and $\mathbf{y} \in \mathbb{F}^m$, the equation

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

has a solution if and only if the last column of $[A, \mathbf{y}]$ is free, in which case the last column of $\operatorname{rrref}([A, \mathbf{y}])$ provides the unique solution to

$$A(:,\texttt{bound})? = \mathbf{y}.$$

More generally, if $R = \operatorname{rrref}([A, B])$ for some arbitrary matrix $B \in \mathbb{F}^{m\times s}$ and all the unknowns corresponding to columns of B are free, then, by (4.18), applied to $[A, B]$ rather than A, we have

$$B = A(:,\texttt{bound})R(:, n + (1{:}s)).$$

4.24 Prove that $\operatorname{rrref}(\operatorname{id}_n) = \operatorname{id}_n$.

A numerical example, continued: Recall our earlier example in which we used elimination to convert a given matrix to its rrref, as follows:

$$\begin{bmatrix} 0 & 2 & 0 & 2 & 5 & 4 & 0 & 6 \\ 0 & 1 & 0 & 1 & 2 & 2 & 0 & 3 \\ 0 & 2 & 0 & 2 & 5 & 4 & -1 & 7 \\ 0 & 1 & 0 & 1 & 3 & 2 & -1 & 4 \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 \\ 0 & \mathbf{1} & 0 & 1 & 2 & 2 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\mathbf{1} & 1 \end{bmatrix}$$
$$\rightarrow \begin{bmatrix} 0 & \mathbf{1} & 0 & 1 & 0 & 2 & 0 & 3 \\ 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & -1 \end{bmatrix},$$

hence $\texttt{bound} = (2, 5, 7)$, $\texttt{free} = (1, 3, 4, 6, 8)$. Now, the elimination algorithm is entirely unaware of how we got the initial matrix. In particular, we are free to interpret in various ways the array on the left as being of the form $[A, B]$. As soon as we specify the number of columns, in A or B, we know A and B exactly.

First, choose B to be a one-column matrix. Then, since the last unknown is free, we conclude that

$$(6, 3, 7, 4) = A(:,\texttt{bound})R_{:8} = \begin{bmatrix} 2 & 5 & 0 \\ 1 & 2 & 0 \\ 2 & 5 & -1 \\ 1 & 3 & -1 \end{bmatrix} (3, 0, -1).$$

If we choose B to be a three-column matrix instead, then the linear system $A? = B$ is unsolvable since now one of the columns of B (the second one) corresponds to a bound unknown. What about the other two columns of this B? The first one corresponds to a free unknown, hence is a weighted sum of the columns to the left of it, hence is in $\operatorname{ran} A$. But the last one fails to be in $\operatorname{ran} A$ since its unknown is free only because of the presence of the seventh column, and this seventh column is *not* a weighted sum of the columns to the left of it, hence neither is the eighth column. Indeed, the corresponding column of R has its last entry nonzero, showing that the *bound*$_3$-column of A is needed to write the last column of A as a weighted sum of columns to the left of it. □

4.25 Use elimination to show that $\begin{bmatrix} 2 & -1 & 0 \\ 1 & 2 & 1 \\ 0 & 2 & -1 \end{bmatrix}$ is 1-1 and onto.

4.26 Use elimination to settle the following assertions, concerning the linear system $A? = \mathbf{y}$, with the (square) matrix A and the right side $\mathbf{y}$ given by

$$[A, \mathbf{y}] := \begin{bmatrix} 1 & -2 & 3 & 1 \\ 2 & k & 6 & 6 \\ -1 & 3 & k-3 & 0 \end{bmatrix}.$$

(a) If $k = 0$, then the system has an infinite number of solutions. (b) For another specific value of k, which you must find, the system has no solutions. (c) For all other values of k, the system has a unique solution.

(To be sure, there probably is some preliminary work to do, after which it is straightforward to answer all three questions.)

Constructing the inverse of a matrix by elimination

If $A \in \mathbb{F}^{n\times n}$ is invertible, then the first n columns of $[A, \mathrm{id}_n]$ are necessarily bound and the remaining n columns are necessarily free. Therefore, if $R := \mathrm{rrref}([A, \mathrm{id}_n])$, then $R = [\,\mathrm{id}_n, ?]$ and, with (4.18), necessarily $[A, \mathrm{id}_n] = AR = [A\,\mathrm{id}_n, A?]$, hence $? = A^{-1}$, i.e., $R = [\,\mathrm{id}_n, A^{-1}]$.

Practical note: Although `MATLAB` provides the function `inv(A)` to generate the inverse of `A`, there is usually no reason to compute the inverse of a matrix, nor would you solve the linear system $A? = \mathbf{y}$ in practice by computing $\mathrm{rref}([A, \mathbf{y}])$ or by computing `inv(A)*y`. Rather, in `MATLAB` you would compute the solution of `A`$? =$`y` as `A\y`. For this, `MATLAB` also uses elimination, but in a more sophisticated form, to keep rounding error effects as small as possible. In effect, the choice of pivot rows is more elaborate than we discussed above.

4.27 Prove: *If the product AB of two square matrices A and B is invertible, then so are both matrices.*

4.28 Here are three questions that can be settled *without doing any arithmetic*. Please do so.

(i) Can both of the following equalities be right?

$$\begin{bmatrix} -5 & 2 \\ 3 & -1 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 3 & 5 \end{bmatrix} = \mathrm{id}_2 = \begin{bmatrix} 1 & 2 \\ 3 & 5 \end{bmatrix}\begin{bmatrix} -4 & 2 \\ 3 & 5 \end{bmatrix}$$

(ii) How does one find the coordinates of $\mathbf{e}_1 \in \mathbb{R}^2$ with respect to the vector sequence $(1,3), (2,5)$ (i.e., numbers α, β for which $\mathbf{e}_1 = (1,3)\alpha + (2,5)\beta$), given that

$$AV := \begin{bmatrix} -5 & 2 \\ 3 & -1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 5 \end{bmatrix} = \mathrm{id}_2 ?$$

(iii) How does one conclude at a glance that the following equation must be wrong?

$$\begin{bmatrix} -5 & 2 \\ 3 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 1 \\ 3 & 5 & 0 \end{bmatrix} = \mathrm{id}_3 ?$$

4.29 For each of the following matrices A, use elimination (to the extent necessary) to (a) determine whether it is invertible and, if it is, to (b) construct the inverse.

(a) $\begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \end{bmatrix}$; (b) $\begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 3 & 4 \end{bmatrix}$; (c) $\begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 5 \end{bmatrix}$; (d) $\begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 4 \end{bmatrix}$; (e) $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 8 \end{bmatrix}$;

(f) $[\mathbf{e}_1 - \mathbf{e}_3, \mathbf{e}_2, \mathbf{e}_3 + \mathbf{e}_4, \mathbf{e}_4] \in \mathbb{R}^{4\times 4}$.

4.30 Prove that $A \in \mathbb{F}^n$ *is invertible iff* $\mathrm{rrref}(A) = \mathrm{id}_n$.

4.31* One way to solve *Laplace's equation*, $\Delta f := D_1^2 f + D_2^2 f = y$ on some domain G in $\mathbb{R}^2$ with $f = g$ on the boundary of G numerically is to choose a regular grid $T = \{(ih, jh) : i \in I, j \in J\}$ of points, with I and J chosen so that (ih, jh) is either strictly inside G or else is next to one such, and then to try to compute $u \in \mathbb{R}^T$ so that

$$(u(\mathbf{t} + (h,0)) + u(\mathbf{t} - (h,0)) + u(\mathbf{t} + (0,h)) + u(\mathbf{t} - (0,h)))/4 - u(\mathbf{t}) = y(\mathbf{t})$$

for all $\mathbf{t} \in T_G$, with T_G the set of points of T strictly inside G, while, for the other points in T, $u(\mathbf{t})$ is determined from the given boundary values g in a linear manner, e.g., by interpolation, i.e., formally, $u(\mathbf{t}) = L_{\mathbf{t}} g$ for all $\mathbf{t} \in T \backslash T_G$ for some scalar valued linear maps $L_{\mathbf{t}}$.

Prove that *the resulting linear system $Au = y$ for the 'vector' $u = (u(\mathbf{t}) : \mathbf{t} \in T)$ (containing one equation for each $\mathbf{t} \in T$) has exactly one solution.* (Hint: if $u(\mathbf{t}) = \max u(T)$ for some $\mathbf{t}$ inside G for a solution u of the corresponding homogeneous system $A? = 0$, then, $u(\mathbf{t})$ being the average of its four neighbors, those neighbors must have the same value.)

4.32 Let $L \in \mathbb{R}^{n\times n}$ be the lower triangular matrix with all diagonal entries equal to 1 and all the strictly lower triangular entries equal to -1, and let $n > 1$. Prove that $(L^{-1})_{n1} = 2^{n-2}$.

4.33* (a) Verify that elimination of the first unknown from the second row of the matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ produces the row-equivalent matrix $B := \begin{bmatrix} a & b \\ 0 & d - c(1/a)b \end{bmatrix}$ under the assumption that $a \neq 0$.

(b) Generalize this formula to the case of a block matrix, i.e., when a, b, c, d are matrices of compatible sizes (with a and d square) and, correspondingly, $1/a$ is interpreted as the matrix a^{-1}, assuming, of course, that a is invertible.

(c) The *matrix* $d - ca^{-1}b$ in (b) is called the **Schur complement** of the submatrix a in A and is denoted A/a (not to be confused with the result of `MATLAB`'s right division operator). Prove that, *if A is also invertible, then the Schur complement A/a of the invertible submatrix a of A is invertible, and its inverse equals the corresponding block of A^{-1}.*

(d) Prove that *the Schur complement is unique*, i.e., if $\begin{bmatrix} X & 0 \\ Y & \mathrm{id} \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} a' & b' \\ 0 & d' \end{bmatrix}$ then $d' = d - ca^{-1}b$.

Elimination in vector spaces

In the discussion of the fact (3.10) that the bound columns of $W \in L(\mathbb{F}^m, X)$ form a basis for $\operatorname{ran} W$, we left unanswered the unspoken question of just how one would tell which columns of W are bound.

The answer is immediate in case $X \subset \mathbb{F}^r$ for some r, for then W is just an $r \times m$-matrix, and elimination does the trick since it is designed to determine the bound columns of a matrix. It works just as well when X is, more generally, a subset of $\mathbb{F}^T$ for some set T, as long as T is finite, since we can then apply elimination to the 'matrix'

$$\delta_T W := (w_j(t) : (t, j) \in T \times \underline{m}) \tag{4.25}$$

whose rows are indexed by the (finitely many) elements of T.

Elimination even works when T is not finite, since looking for a pivot row in the matrix (4.25) with *infinitely* many rows is only a *practical* difficulty. If τ_i is the row 'index' of the pivot row for the ith bound column of W, $i = 1{:}r$, then we know that W has the same nullspace as the (finite-rowed) matrix $(w_j(\tau_i) : i = 1{:}r, j = 1{:}m)$. This proves, for arbitrary T, the following important

(4.26) Proposition: For any $W \in L(\mathbb{F}^m, \mathbb{F}^T)$, there exists a sequence $(\tau_1, \ldots, \tau_r)$ in T, with r equal to the number of bound columns in W, so that $\operatorname{null} W$ is equal to the nullspace of the matrix $(w_j(\tau_i) : i = 1{:}r, j = 1{:}m)$.

In particular, W is 1-1 if and only if the matrix $(w_j(\tau_i) : i, j = 1{:}m)$ is invertible for some sequence $(\tau_1, \ldots, \tau_m)$ in T.

If T is not finite, then we may not be able to determine in finite time whether or not a given column is bound since we may have to look at infinitely many rows not yet used as pivot rows. The only efficient way around this is to have W given to us in the form

$$W = UA,$$

with U some 1-1 column map, hence A a matrix. Under these circumstances, the kth column of W is free if and only if the kth column of A is free, and the latter we can determine by elimination applied to A.

Indeed, if U is 1-1, then both W and A have the same nullspace, hence, by (3.6)Proposition, the kth column of W is bound if and only if the kth column of A is bound.

As an example, consider $W = [w_1, w_2, w_3, w_4]$, with $w_j : \mathbb{R} \to \mathbb{R} : t \mapsto \sin(t - j)$, $j = 1, 2, 3, 4$. Hence, by the addition formula,

$$W = UA, \quad \text{with } U := [\sin, \cos], \quad \text{and}$$

$$A := \begin{bmatrix} \cos(-1) & \cos(-2) & \cos(-3) & \cos(-4) \\ \sin(-1) & \sin(-2) & \sin(-3) & \sin(-4) \end{bmatrix},$$

and we see at once that U is 1-1 (e.g. from the fact that $QU = \mathrm{id}_2$, with $Q : f \mapsto (f(\pi/2), f(0)))$. We also see at once that the first two columns of A are bound (e.g., since $\cos(1)\cos(2) < 0$ while $\sin(1)\sin(2) > 0$), hence the remaining columns of A must be free (since there are no rows left to bind them). Consequently, the first two columns of W are bound, while the last two columns are free.

Note that, necessarily, U is a basis for $\operatorname{ran} W$ since $W = UA$ implies that $\operatorname{ran} W \subset \operatorname{ran} U$, hence having two columns of W bound implies that $2 \le \dim\operatorname{ran} W \le \dim\operatorname{ran} U \le \#U = 2$, and so U is 1-1 onto $\operatorname{ran} W$.

In general, it may be hard to find such a handy factorization $W = UA$ for given $W \in L(\mathbb{F}^m, X)$. In that case, we may have to *discretize* our problem by finding somehow some $Q \in L(X, \mathbb{F}^n)$ that is 1-1 on $\operatorname{ran} W$. With such a 'data map' Q in hand, we know that $\operatorname{null} W$ equals the nullspace of the *matrix* QW. In particular, the kth column of W is bound if and only if the kth column of the *matrix* QW is bound, and elimination applied to QW will ferret out all those columns.

The need for suitable 'data maps' here in the general case is one of many reasons why we now turn to the study of this second way of connecting our vector space X to some coordinate space, namely via linear maps from X to $\mathbb{F}^n$.

4.34 For each of the following column maps $V = [v_1, \ldots, v_r]$ into the vector space $\Pi_{<5}$ of all real polynomials of degree < 5, determine whether or not it is 1-1 and/or onto.

(a) $[()^3 - ()^1 + 1, ()^2 + 2()^1 + 1, ()^1 - 1]$; (b) $[()^4 - ()^1, ()^3 + 2, ()^2 + ()^1 - 1, ()^1 + 1]$; (c) $[1 + ()^4, ()^4 + ()^3, ()^3 + ()^2, ()^2 + ()^1, ()^1 + 1]$.

4.35 For each of the specific column maps $V = [f_j : j = 0{:}r]$ given below (with f_j certain real-valued functions on the real line), determine which columns are bound and which are free. Use this information to determine (i) a basis for $\operatorname{ran} V$; and (ii) the smallest n so that $f_n \in \operatorname{ran}[f_0, f_1, \ldots, f_{n-1}]$.

(a) $r = 6$, and $f_j : t \mapsto (t - j)^2$, all j.

(b) $r = 4$ and $f_j : t \mapsto \sin(t - j)$, all j.

(c) $r = 4$ and $f_j : t \mapsto \exp(t - j)$, all j. (If you know enough about the exponential function, then you need not carry out any calculation on this problem.)

4.36 T/F

(a) $\begin{bmatrix} 1 & 0 & 1 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ is in row echelon form.

(b) If all unknowns in the linear system $A? = \mathbf{0}$ are free, then $A = 0$;

(c) If all unknowns in the linear system $A? = \mathbf{0}$ are bound, then A is invertible.

(d) If some unknowns in the linear system $A? = \mathbf{0}$ are free, then A cannot be invertible.

(e) The inverse of an upper triangular matrix is lower triangular.

(f) A linear system of n equations in $n + 1$ unknowns always has solutions.

(g) Any square matrix in row echelon form is upper triangular.

(h) If A and B are square matrices of the same order, then $AB? = \mathbf{0}$ has the same number of bound unknowns as does $BA? = \mathbf{0}$.

(i) If A and B are square matrices of the same order, and AB is invertible, then also BA is invertible.

(j) If $\operatorname{null} A = \operatorname{null} B$, then $A? = \mathbf{0}$ and $B? = \mathbf{0}$ have the same free and bound unknowns.

5 The inverse of a basis, and interpolation

Linear maps into $\mathbb{F}^n$ (aka row maps)

There are two ways to connect a given vector space X with the coordinate space $\mathbb{F}^n$ in a linear way, namely by a linear map from $\mathbb{F}^n$ to X, and by a linear map to $\mathbb{F}^n$ from X. By now, you are thoroughly familiar with the first kind, the column maps. It is time to learn something about the other kind.

A very important example of such a map is the inverse of a basis $V : \mathbb{F}^n \to X$ for the vector space X. This inverse is also known as the **coordinate map** for that basis because it provides, for each $x \in X$, its **coordinates with respect to the basis**, i.e., the n-vector $\mathbf{a} := V^{-1}x$ for which $x = V\mathbf{a}$. In effect, every *invertible* linear map from X to $\mathbb{F}^n$ is a coordinate map, namely the coordinate map for its inverse. However, (nearly) every linear map from X to $\mathbb{F}^n$, invertible or not, is of interest, as a means of extracting numerical information from the elements of X. For, we can, offhand, only compute with numbers, hence can 'compute' with elements of an abstract vector space only in terms of numerical data about them.

Any linear map from the vector space X to $\mathbb{F}^n$ is necessarily of the form

$$f : X \to \mathbb{F}^n : x \mapsto (f_i(x) : i = 1{:}n),$$

with each $f_i := \mathbf{e}_i{}^{\mathrm{t}} \circ f$ the composition of two linear maps, hence linear, therefore a **linear functional** on X, i.e., a scalar-valued linear map on X.

5.1 For each of the following maps, determine whether or not it is a linear functional. (a) $\Pi_{\le k} \to \mathbb{R} : p \mapsto \deg p$; (b) $\mathbb{R}^3 \to \mathbb{R} : \mathbf{x} \mapsto 3x_1 - 2x_3$; (c) $C([a\,..\,b]) \to \mathbb{R} : f \mapsto \max_{a\le t\le b} f(t)$; (d) $C([a..b]) \to \mathbb{R} : f \mapsto \int_a^b f(s)w(s)\,\mathrm{d}s$, with $w \in C([a..b])$; (e) $C^{(2)}(\mathbb{R}) \to \mathbb{R} : f \mapsto a(t)D^2f(t) + b(t)Df(t) + c(t)f(t)$, for some functions a, b, c defined on $[a\,..\,b]$ and some $t \in [a\,..\,b]$. (f) $C^{(2)}(\mathbb{R}) \to C(\mathbb{R}) : f \mapsto aD^2f + bDf + cf$, for some $a, b, c \in C(\mathbb{R})$.

Here are some standard examples of linear functionals. Assume that X is a space of functions, hence X is a linear subspace of $\mathbb{F}^T$ for some set T.

Then, for each $t \in T$,

$$\delta_t : X \to \mathbb{F} : x \mapsto x(t)$$

is a linear functional on X, the linear functional of evaluation at t. For any n-sequence $(s_1, \ldots, s_n)$ in T,

$$X \to \mathbb{F}^n : f \mapsto (f(s_1), \ldots, f(s_n))$$

is a standard linear map from X to $\mathbb{F}^n$.

If, more concretely, X is a linear subspace of $C^{(n-1)}[a\,..\,b]$ and $s \in [a\,..\,b]$, then

$$X \to \mathbb{F}^n : f \mapsto (f(s), Df(s), \ldots, D^{n-1}f(s))$$

is another standard linear map from such X to $\mathbb{F}^n$.

Finally, if $X = \mathbb{F}^m$, then any linear map from X to $\mathbb{F}^n$ is necessarily a matrix. But it is convenient to write this matrix in the form A^{t} for some $A \in \mathbb{F}^{m\times n}$, as such A^{t} acts on $X = \mathbb{F}^m$ via the rule

$$X \mapsto \mathbb{F}^n : \mathbf{x} \mapsto A^{\mathrm{t}}\mathbf{x} = (A(\,:\,,j)^{\mathrm{t}}\mathbf{x} : j = 1{:}n).$$

Because of this last example, we will call all linear maps from a vector space to a coordinate space **row maps**, and use the notation

$$(5.1) \qquad \Lambda^{\mathrm{t}} : X \to \mathbb{F}^n : x \mapsto (\lambda_i x : i = 1{:}n) =: [\lambda_1, \ldots, \lambda_n]^{\mathrm{t}} x,$$

calling the linear functional λ_i the ith **row** of this map. We will also call such maps **data maps** since they extract numerical information from the elements of X. There is no hope of doing any practical work with the vector space X unless we have a ready supply of such data maps on X. For, by and large, we can only compute with numbers.

(5.2) Proposition: If $\Lambda^{\mathrm{t}} = [\lambda_1, \lambda_2, \ldots, \lambda_n]^{\mathrm{t}} : X \to \mathbb{F}^n$ and $B \in L(U, X)$, then $\Lambda^{\mathrm{t}} B = [\lambda_1 B, \ldots, \lambda_n B]^{\mathrm{t}}$.

This raises the question what, in this setting, the map $[\lambda_1, \lambda_2, \ldots, \lambda_n]$ might be. Well, it is a column map whose columns are linear functionals on X, hence it is a column map into the space $L(X, \mathbb{F})$ of all linear functionals on X. This space is the **dual** of X, in symbols

$$X' := L(X, \mathbb{F}),$$

to be further discussed in Chapter 9.

A formula for the coordinate map

Let $V \in L(\mathbb{F}^n, X)$ be a basis for the vector space X. How do we find the coordinates

$$(5.3) \qquad \mathbf{a} = V^{-1}x$$

for given $x \in X$?

Offhand, we solve the (linear) equation $V? = x$ for $\mathbf{a}$. Since V is a basis, we know that this equation has exactly one solution. But that is not the same thing as having a concrete formula for $\mathbf{a}$ in terms of x.

If $X = \mathbb{F}^n$, then V^{-1} is a matrix; in this case, (5.3) is an explicit formula. However, even if $X \subset \mathbb{F}^n$ but $X \neq \mathbb{F}^n$, then (5.3) is merely a formal expression.

(5.4) Example: If V is a basis for some linear subspace X of $\mathbb{F}^n$, then we can obtain a formula for V^{-1} via elimination as follows.

Act as if V were invertible, i.e., apply elimination to $[V, \mathrm{id}_n]$. Let $r := \#V$. Since V is 1-1, the first r columns in $[V, \mathrm{id}_n]$ are bound, hence we are able to produce, via elimination, a row-equivalent matrix R for which $R(\mathtt{q}, 1{:}r) = \mathrm{id}_r$, for some r-sequence $\mathtt{q}$. Since we obtain R from $[V, \mathrm{id}_n]$ by (invertible) row operations, we know that $R = M[V, \mathrm{id}_n] = [MV, M]$ for some invertible matrix M. In particular,

$$\mathrm{id}_r = R(\mathtt{q}, 1{:}r) = (MV)(\mathtt{q}, :) = M(\mathtt{q}, :)V,$$

showing $M(\mathtt{q}, :) = R(\mathtt{q}, r + (1{:}n))$ to be a left inverse for V, hence equal to V^{-1} when restricted to $\operatorname{ran} V$.

Suppose, in particular, that we carry elimination all the way through, to obtain $R = \mathrm{rref}([V, \mathrm{id}_n]) =: M[V, \mathrm{id}_n] = [MV, M]$ for some invertible matrix $M \in \mathbb{F}^{n \times n}$ (as must happen since $[V, \mathrm{id}_n]$ is onto). Then, $R(:, \mathtt{bound}) = \mathrm{id}_n$ with $\mathtt{bound}$ indicating the bound columns of $[V, \mathrm{id}_n]$, therefore, $\mathtt{q} := \mathit{bound}_{1:r} = 1{:}r$, hence $M(\mathtt{q}, :)V = \mathrm{id}_r$, i.e., $M(\mathtt{q}, :)$ is a left inverse for V, and $M(\mathtt{q}, \mathtt{b}) = 0$, with $r + \mathtt{b}$ the bound columns of $[V, \mathrm{id}_n]$ *other than* the columns of V. Hence, with $r + \mathtt{f}$ the free columns of $[V, \mathrm{id}_n]$ and for this choice of M, we get

$$V^{-1}\mathtt{x} = M(\mathtt{q}, :)\mathtt{x} = M(\mathtt{q}, \mathtt{f})x_{\mathtt{f}}, \qquad \mathtt{x} \in X = \operatorname{ran} V.$$

In effect, we have replaced here the equation $V? = \mathtt{x}$ by the *equivalent* equation

$$V(\mathtt{f}, :)? = x_{\mathtt{f}}$$

whose coefficient matrix is invertible. In particular, $\#\mathtt{f} = \#V$; see Problem 5.3. □

5.2 For each of the following bases V of the linear subspace $\operatorname{ran} V$ of $\mathbb{F}^n$, give a matrix U for which Ux gives the coordinates of $x \in \operatorname{ran} V$ with respect to the basis V. How would you check your answer?

$$\text{(a) } V = \begin{bmatrix} 2 \\ 2 \end{bmatrix}; \text{ (b) } V = [\mathbf{e}_2, \mathbf{e}_1, \mathbf{e}_3] \in \mathbb{R}^{3\times 3}; \text{ (c) } V = \begin{bmatrix} 1 & 2 \\ 2 & 4 \\ 0 & 6 \end{bmatrix}; \text{ (d) } V = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ -1 & 1 \\ 2 & -2 \end{bmatrix}.$$

5.3* Prove the claim at the end of (5.4)Example.

The reduction in (5.4)Example, of the abstract linear equation $V? = x$ to a uniquely solvable square linear system, also works in the general setting.

To obtain a concrete expression, we **discretize** the abstract equation $V? = x$ by considering instead the *numerical* equation

$$\Lambda^{\mathrm{t}} V? = \Lambda^{\mathrm{t}} x$$

for some suitable data map $\Lambda^{\mathrm{t}} \in L(Y, \mathbb{F}^n)$ defined on some vector space $Y \supset X$. Here, 'suitable' means that the *matrix* $\Lambda^{\mathrm{t}} V$ is invertible, for then the unique solution of this equation must be the sought-for coordinate vector for $x \in X$ with respect to the basis V, i.e.,

$$\mathbf{a} = V^{-1}x = (\Lambda^{\mathrm{t}} V)^{-1} \Lambda^{\mathrm{t}} x.$$

In (5.4)Example, we simply chose the linear map $\mathbf{y} \mapsto y_{\mathtt{f}}$ as our Λ^{t}, i.e., $\Lambda^{\mathrm{t}} = \mathrm{id}_n(\mathtt{f}, :) = [\mathbf{e}_j : j \in \mathtt{f}]^{\mathrm{t}}$, with $\mathtt{f}$ chosen, in effect, to ensure that $\Lambda^{\mathrm{t}} V = V(\mathtt{f}, :)$ is invertible. We indeed obtained there V^{-1} as

$$\mathbf{x} \mapsto M(\mathtt{q}, \mathtt{f}) x_{\mathtt{f}} = V(\mathtt{f}, :)^{-1} x_{\mathtt{f}} = (\Lambda^{\mathrm{t}} V)^{-1} \Lambda^{\mathrm{t}} \mathbf{x}.$$

How would one find a 'suitable' data map in general? That depends on the particular circumstances. For example, if $V \in L(\mathbb{F}^n, Y)$ and $\Lambda^{\mathrm{t}} \in L(Y, \mathbb{F}^n)$, and we somehow know that Λ^{t} maps $X := \operatorname{ran} V = V(\mathbb{F}^n)$ *onto* $\mathbb{F}^n$, then we know that $\Lambda^{\mathrm{t}} V$ maps $\mathbb{F}^n$ onto $\mathbb{F}^n$, hence, being a square matrix, $\Lambda^{\mathrm{t}} V$ must be invertible. Conversely, if $\Lambda^{\mathrm{t}} V$ is invertible, then V must be 1-1, hence provides a basis for its range, and Λ^{t} must map $\operatorname{ran} V$ onto $\mathbb{F}^n$.

(5.5) Proposition: If the linear map $V : \mathbb{F}^n \to X \subset Y$ is onto, and $\Lambda^{\mathrm{t}} \in L(Y, \mathbb{F}^n)$ is such that their (square) **Gramian matrix**, $\Lambda^{\mathrm{t}} V$, is 1-1 or onto, hence invertible, then V is a basis for X, and its inverse is

$$V^{-1} : X \to \mathbb{F}^n : x \mapsto (\Lambda^{\mathrm{t}} V)^{-1} \Lambda^{\mathrm{t}} x.$$

(5.6) Proposition: The inverse $\mathrm{M}^{\mathrm{t}} =: [\mu_1, \ldots, \mu_n]^{\mathrm{t}}$ of any basis V of the n-dimensional vector space X provides the basis $\mathrm{M} := [\mu_1, \ldots, \mu_n]$ for the dual space X', called the basis **dual to** V.

Indeed, let $\mathrm{M}^{\mathrm{t}} := V^{-1}$ and consider, for $\mathbf{a} \in \mathbb{F}^n$ and $x \in X$, $(\mathrm{M}\mathbf{a})x = \sum_i a_i \mu_i x = \mathbf{a}^{\mathrm{t}}\mathrm{M}^{\mathrm{t}}x$, hence $(\mathrm{M}\mathbf{a})V = \mathbf{a}^{\mathrm{t}}(\mathrm{M}^{\mathrm{t}}V) = \mathbf{a}^{\mathrm{t}}$, showing that the linear map $\mathrm{M} : \mathbb{F}^n \to X' = L(X, \mathbb{F}) : \mathbf{a} \mapsto [\mu_1, \ldots, \mu_n]\mathbf{a}$ is 1-1. Since, by (3.22), its target, $X' = L(X, \mathbb{F})$, has dimension n, it follows, by (3.24)Corollary, that M is also onto X', hence a basis for X'.

5.4* Prove: *For every nonzero x in the finite-dimensional vector space X, there exists $\lambda \in X'$ with $\lambda x \neq 0$.*

Change of basis

To be sure, under the assumptions of (5.5)Proposition, we also know that Λ^{t} maps X onto $\mathbb{F}^n$, hence, since both X and $\mathbb{F}^n$ are of the same finite dimension, the restriction of Λ^{t} to X must be invertible as a linear map to $\mathbb{F}^n$. Consequently, there must be an invertible $W \in L(\mathbb{F}^n, X)$, i.e., a basis W for X, with $\Lambda^{\mathrm{t}}W = \mathrm{id}_n$.

Hence, the right side in our numerical equation $\Lambda^{\mathrm{t}}V? = \Lambda^{\mathrm{t}}x$ is the coordinate vector for $x \in X$ with respect to this basis W for X. In other words, our great scheme for computing the coordinates of $x \in X$ with respect to the basis V for X requires us to know the coordinates of x with respect to some basis for X. In other words, the entire calculation is just a *change of basis*, with $\Lambda^{\mathrm{t}}V = W^{-1}V$ the socalled **transition matrix** that carries the V-coordinates of x to the W-coordinates of x.

However, this in no way diminishes its importance. For, we really have no choice in this matter. We cannot compute without having numbers to start with. Also, we often have ready access to the coordinate vector $\Lambda^{\mathrm{t}}x$ without having in hand the corresponding basis W. At the same time, we may much prefer to know the coordinates of x with respect to a different basis.

For example, we know from (3.38)Proposition that, with $(\tau_1, \ldots, \tau_k)$ any sequence of pairwise distinct real numbers, the linear map $\Lambda^{\mathrm{t}} : p \mapsto (p(\tau_1), \ldots, p(\tau_k))$ is 1-1 on the k-dimensional space $\Pi_{<k}$, hence provides the coordinates of $p \in \Pi_{<k}$ with respect to a certain basis W of $\Pi_{<k}$, namely the socalled **Lagrange basis** whose columns can be verified to be the so-called **Lagrange fundamental polynomials**

$$(5.7) \qquad \ell_j : t \mapsto \prod_{h \neq j} \frac{t - \tau_h}{\tau_j - \tau_h}, \qquad j = 1{:}k.$$

However, you can imagine that it is a challenge to differentiate or integrate a polynomial written in terms of this basis. Much better for that to have the coordinates of the polynomial with respect to the power basis $V = [()^0, \ldots, ()^{k-1}]$.

5.5 What are the coordinates of $p \in \Pi_{\leq k}$ with respect to the Lagrange basis for $\Pi_{\leq k}$ for the points $\tau_1, \ldots, \tau_k$?

5.6 Find the value at 0 of the quadratic polynomial p, for which $p(-1) = p(1) = 3$ and $Dp(1) = 6$.

5.7 Find a formula for $p(0)$ in terms of $p(-1)$, $p(1)$ and $Dp(1)$, assuming that p is a quadratic polynomial.

5.8 Find the coordinates for the polynomial $q(t) = 3 - 4t + 2t^2$ with respect to the basis $W := [()^0, ()^0 + ()^1, ()^0 + ()^1 + ()^2]$ of the space of quadratic polynomials. (Hint: you are given the coordinates for q wrto $V := [()^0, ()^1, ()^2] = W(W^{-1}V)$ and can easily determine $(W^{-1}V)^{-1} = V^{-1}W$.)

5.9* Let $v_1, \ldots, v_n$ be a sequence of $(n-1)$-times continuously differentiable functions, all defined on the interval $[a \,.\,.\, b]$. For $x \in [a \,.\,.\, b]$, the matrix

$$W(v_1, \ldots, v_n; x) := (D^{i-1}v_j(x) : i, j = 1{:}n)$$

is called the **Wronski matrix at** x for the sequence $(v_j : j = 1{:}n)$.

Prove that $V := [v_1, \ldots, v_n]$ *is 1-1 in case, for some* $x \in [a \,.\,.\, b]$, $W(v_1, \ldots, v_n; x)$ *is invertible.* (Hint: Consider the Gram matrix $\Lambda^{\mathrm{t}}V$ with $\Lambda^{\mathrm{t}}f := (f(x), f'(x), \ldots, D^{n-1}f(x))$.)

Interpolation and linear projectors

As the discussion of polynomial interpolation on pages 64ff already intimates, our formula in (5.5) for the inverse of a basis $V \in L(\mathbb{F}^n, X)$ can be much more than that. It is useful for *interpolation* in the following way. Assuming that $\Lambda^{\mathrm{t}}V$ is invertible, it follows that, for any $y \in Y$, $x = V(\Lambda^{\mathrm{t}}V)^{-1}\Lambda^{\mathrm{t}}y$ is the unique element in X that **agrees with** y **at** Λ^{t} in the sense that

$$\Lambda^{\mathrm{t}}x = \Lambda^{\mathrm{t}}y.$$

To recall the specifics of polynomial interpolation from pages 64ff, if $X = \Pi_{<k}$ and $\Lambda^{\mathrm{t}} : g \mapsto (g(\tau_i) : i = 1{:}k)$, with $\tau_1 < \cdots < \tau_k$, then, by (3.38)Proposition, for arbitrary $g : \mathbb{R} \to \mathbb{R}$, there is exactly one polynomial p of degree $< k$ for which $p(\tau_i) = g(\tau_i)$, $i = 1{:}k$.

One can readily imagine other examples.

Example: In **Hermite interpolation**, one specifies not only values but also derivatives. For example, in **two-point** Hermite interpolation from $\Pi_{<k}$, one picks two points, $t \neq u$, and two nonnegative integers r and s with $r + 1 \ + s + 1 = k$, and defines

$$\Lambda^{\mathrm{t}} : g \mapsto (g(t), Dg(t), \ldots, D^r g(t), g(u), Dg(u), \ldots, D^s g(u)).$$

Now the requirement that $\Lambda^{\mathrm{t}}p = \Lambda^{\mathrm{t}}g$ amounts to looking for $p \in \Pi_{<k}$ that agrees with g in the sense that p and g have the same derivative values of order $0, 1, \ldots, r$ at t and the same derivative values of order $0, 1, \ldots, s$ at u. □

Example: Recall from Calculus the bivariate **Taylor series**

$$\begin{aligned}g(s,t) = g(0) + D_s g(0)\, s + D_t g(0)\, t + \\ + \left({D_s}^2 g(0) s^2 + D_s D_t g(0) st + D_t D_s g(0) ts + {D_t}^2 g(0) t^2\right)/2 + h.o.t.\end{aligned}$$

In particular, for any smooth function g,

$$\begin{aligned}p : (s,t) \mapsto g(0) + D_s g(0)\, s + D_t g(0)\, t + \\ + \left({D_s}^2 g(0) s^2 + 2 D_s D_t g(0) st + {D_t}^2 g(0) t^2\right)/2\end{aligned}$$

is the unique quadratic polynomial that matean hes the information about g given by the data map

$$\Lambda^{\mathrm{t}} : g \mapsto (g(0), D_s g(0), D_t g(0), {D_s}^2 g(0), D_s D_t g(0), {D_t}^2 g(0)).$$

□

Example: When dealing with **Fourier series**, one uses the data map

$$\Lambda^{\mathrm{t}} : g \mapsto (\int_0^{2\pi} g(t)\, \mathrm{cis}(jt)\, \mathrm{d}t : j = 0{:}N),$$

with cis standing for ‘cosine or sine’. One looks for a **trigonometric polynomial**

$$p = [\mathrm{cis}(j\cdot) : j = 0{:}N]\mathbf{a}$$

that satisfies $\Lambda^{\mathrm{t}} p = \Lambda^{\mathrm{t}} g$, and finds it in the **truncated Fourier series** for g. □

Directly from (5.5)Proposition, we obtain (under the assumptions of that proposition) the following pretty formula

$$x = Py := V(\Lambda^{\mathrm{t}} V)^{-1} \Lambda^{\mathrm{t}} y \tag{5.8}$$

for the interpolant $x \in X$ to given $y \in Y$ with respect to the data map Λ^{t}. The linear map $P := V(\Lambda^{\mathrm{t}} V)^{-1} \Lambda^{\mathrm{t}}$ so defined on Y is very special:

(5.9) Proposition: Let the linear map $V : \mathbb{F}^n \to Y$ be onto $X \subset Y$, and let $\Lambda^{\mathrm{t}} \in L(Y, \mathbb{F}^n)$ be such that their Gramian matrix, $\Lambda^{\mathrm{t}} V$, is invertible. Then V is 1-1 and Λ^{t} is onto, and $P := V(\Lambda^{\mathrm{t}} V)^{-1} \Lambda^{\mathrm{t}}$ is a linear map on Y with the following properties:

(i) P is the identity on $X = \mathrm{ran}\, V$.

(ii) $\mathrm{ran}\, P = \mathrm{ran}\, V = X$.

(iii) P is a **projector** or **idempotent**, i.e., $PP = P$, hence $P(\,\mathrm{id} - P) = 0$.

(iv) $\mathrm{null}\, P = \mathrm{null}\, \Lambda^{\mathrm{t}} = \mathrm{ran}(\,\mathrm{id} - P)$.

(v) Y is the direct sum of $\mathrm{ran}\, P$ and $\mathrm{null}\, P$, i.e., $Y = \mathrm{ran}\, P \dotplus \mathrm{null}\, P$.

Proof: (i) $PV = V(\Lambda^{\rm t}V)^{-1}\Lambda^{\rm t}\, V = V\,{\rm id} = V$, hence $P(V\mathbf{a}) = V\mathbf{a}$ for all $\mathbf{a} \in \mathbb{F}^n$.

(ii) Since $P = VA$ for some A, we have that $\operatorname{ran} P \subset \operatorname{ran} V$, while $PV = V$ implies that $\operatorname{ran} P \supset \operatorname{ran} V$.

(iii) By (i) and (ii), P is the identity on its range, hence, in particular, $PP = P$, or, equivalently, $P(\,{\rm id} - P) = 0$.

(iv) The fact that $P = A\Lambda^{\rm t}$ for some A implies that $\operatorname{null} P \supset \operatorname{null} \Lambda^{\rm t}$, while also

$$\Lambda^{\rm t}P = \Lambda^{\rm t}\, V(\Lambda^{\rm t}V)^{-1}\Lambda^{\rm t} = {\rm id}_n\Lambda^{\rm t} = \Lambda^{\rm t},$$

hence also $\operatorname{null} P \subset \operatorname{null} \Lambda^{\rm t}$. As to $\operatorname{null} P = \operatorname{ran}(\,{\rm id} - P)$, note that $x \in \operatorname{null} P$ implies that $x = x - Px = (\,{\rm id} - P)x \in \operatorname{ran}(\,{\rm id} - P)$, while, conversely, $\operatorname{null} P \supset \operatorname{ran}(\,{\rm id} - P)$ since, by (iii), $P(\,{\rm id} - P) = 0$.

(v) For any $y \in Y$, $y = Py + (\,{\rm id} - P)y \in \operatorname{ran} P + \operatorname{null} P$, by (iv), hence $Y = \operatorname{ran} P + \operatorname{null} P$. If also $y = x + z$ for some $x \in \operatorname{ran} P$ and some $z \in \operatorname{null} P$, then, by (i) and (iv), $Py = P(x + z) = Px + Pz = x$, therefore also $z = y - x = y - Py = (\,{\rm id} - P)y$, showing such a decomposition to be unique. □

5.10 Prove: *If $P \in L(X)$ is an invertible linear projector, then $P = {\rm id}$.*

5.11* Prove: *If $P \in L(X)$, then the following are equivalent: (i) P is the identity on $\operatorname{ran} P$; (ii) $P^2 = P$; (iii) $P(\,{\rm id} - P) = 0$; (iv) $\operatorname{null} P = \operatorname{ran}(\,{\rm id} - P)$, and if any of these hold, then $X = \operatorname{ran} P \dot{+} \operatorname{null} P$.*

5.12* Prove: *If $P, Q \in L(X)$ are linear projectors and $\operatorname{ran} P \subset \operatorname{ran} Q$ and $\operatorname{null} P \subset \operatorname{null} Q$, then $P = Q$.*

5.13* Let $P \in L(X)$. (i) Prove that *P is a projector if and only if $R := {\rm id} - 2P$ is* **involutory** *or* **self-inverse** *(meaning that $RR = {\rm id}$)*. (This would be wrong if our field of scalars had **characteristic 2**, i.e., if $1 + 1 = 0$.) (ii) For the linear projector P of (5.10)Example, work out the corresponding map R, and add to (5.11)Figure the point Ry.

5.14 Consider the linear map Q given on $X = \mathbb{R}^{\mathbb{R}}$ by $Qf(t) = (f(t) + f(-t))/2$. Prove that *Q is a linear projector*. Also, give a succinct description of its range and its nullspace. (Hint: consider the map $F : X \to X$ defined by $(Ff)(t) = -f(t)$.)

(5.10) Example: We specialize the general situation of (5.9)Proposition to the case $V : \mathbb{R}^1 \to X \subset \mathbb{R}^2$, so we can draw a figure like (5.11)Figure.

Take $Y = \mathbb{R}^2$, and let $\mathbf{v} \in \mathbb{R}^2\backslash\{\mathbf{0}\}$, hence $X := \operatorname{ran} V$ with $V := [\mathbf{v}]$ is 1-dimensional. The general linear map $\Lambda^{\rm t} : \mathbb{R}^2 \to \mathbb{R}^1$ is of the form $[\mathbf{w}]^{\rm t}$ for some $\mathbf{w} \in \mathbb{R}^2$, and the requirement that $\Lambda^{\rm t}V$ be invertible reduces to the requirement that $[\mathbf{w}]^{\rm t}[\mathbf{v}] = \mathbf{w}^{\rm t}\mathbf{v} \neq 0$.

With $V = [\mathbf{v}]$ and $\Lambda^{\rm t} = [\mathbf{w}]^{\rm t}$ so chosen, the linear projector P becomes

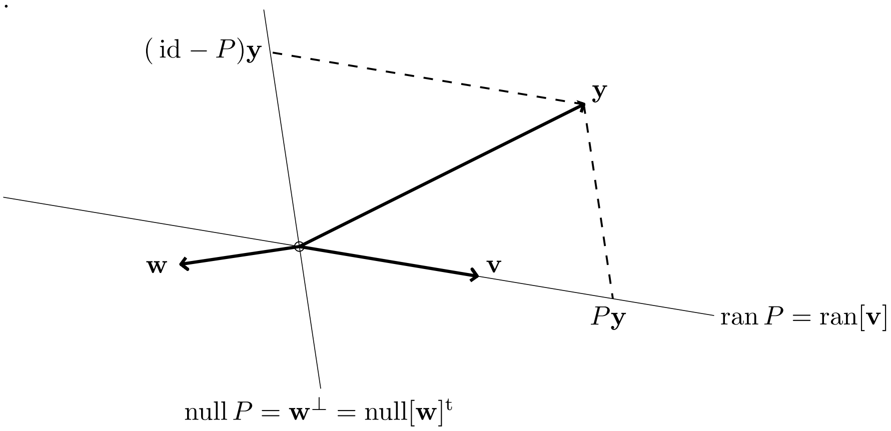
$$P := \frac{\mathbf{v}\mathbf{w}^{\rm t}}{\mathbf{w}^{\rm t}\mathbf{v}} : \mathbf{y} \mapsto \mathbf{v}\frac{\mathbf{w}^{\rm t}\mathbf{y}}{\mathbf{w}^{\rm t}\mathbf{v}}.$$

We verify directly that

$$PP = \frac{\mathbf{v}\mathbf{w}^{\rm t}}{\mathbf{w}^{\rm t}\mathbf{v}}\frac{\mathbf{v}\mathbf{w}^{\rm t}}{\mathbf{w}^{\rm t}\mathbf{v}} = \frac{\mathbf{v}\,\mathbf{w}^{\rm t}\mathbf{v}\,\mathbf{w}^{\rm t}}{(\mathbf{w}^{\rm t}\mathbf{v})\,(\mathbf{w}^{\rm t}\mathbf{v})} = \frac{\mathbf{v}\mathbf{w}^{\rm t}}{\mathbf{w}^{\rm t}\mathbf{v}} = P,$$

i.e., that P is a linear projector. Its range equals $\operatorname{ran}[\mathbf{v}]$, i.e., the straight line through the origin in the direction of $\mathbf{v}$. Its nullspace equals $\operatorname{null}[\mathbf{w}]^{\mathrm{t}}$ and this is necessarily also 1-dimensional, by (3.23)Dimension Formula, hence is the straight line through the origin perpendicular to $\mathbf{w}$. The two lines have only the origin in common since $\mathbf{y} \in \operatorname{ran} P \cap \operatorname{null} P$ implies that $\mathbf{y} = \mathbf{v}\alpha$ for some scalar α, therefore $0 = \mathbf{w}^{\mathrm{t}}\mathbf{y} = \mathbf{w}^{\mathrm{t}}\mathbf{v}\alpha$ and this implies that $\alpha = 0$ since $\mathbf{w}^{\mathrm{t}}\mathbf{v} \neq 0$ by assumption.

.


$(\mathrm{id} - P)\mathbf{y}$
$\mathbf{y}$
$\mathbf{w}$
$\mathbf{v}$
$P\mathbf{y}$
$\operatorname{ran} P = \operatorname{ran}[\mathbf{v}]$
$\operatorname{null} P = \mathbf{w}^{\perp} = \operatorname{null}[\mathbf{w}]^{\mathrm{t}}$

(5.11) Figure. The direct sum decomposition provided by the linear projector $P : \mathbb{R}^2 \to \mathbb{R}^2 : \mathbf{y} \mapsto \mathbf{v}(\mathbf{w}^{\mathrm{t}}\mathbf{y}/\mathbf{w}^{\mathrm{t}}\mathbf{w})$. Compare this to (3.35)Figure.

We can locate the two summands in the split

$$\mathbf{y} = P\mathbf{y} + (\mathrm{id} - P)\mathbf{y}$$

graphically (see (5.11)Figure): To find $P\mathbf{y}$, draw the line through $\mathbf{y}$ parallel to $\operatorname{null} P$; its unique intersection with $\operatorname{ran} P = \operatorname{ran}[\mathbf{v}]$ is $P\mathbf{y}$. The process of locating $(\mathrm{id} - P)\mathbf{y}$ is the same, with the roles of $\operatorname{ran} P$ and $\operatorname{null} P$ reversed: Now draw the line through $\mathbf{y}$ parallel to $\operatorname{ran} P$; its unique intersection with $\operatorname{null} P$ is the element $(\mathrm{id} - P)\mathbf{y}$.

This shows graphically that, for each $\mathbf{y}$, $P\mathbf{y}$ is the unique element of $\operatorname{ran} P$ for which $\mathbf{w}^{\mathrm{t}}P\mathbf{y} = \mathbf{w}^{\mathrm{t}}\mathbf{y}$, i.e., the unique point in the intersection of $\operatorname{ran} P$ and $\mathbf{y} + \operatorname{null}[\mathbf{w}]^{\mathrm{t}}$. □

It is useful to note that, for any linear projector P, also $(\mathrm{id} - P)$ is a linear projector (since $(\mathrm{id} - P)(\mathrm{id} - P) = \mathrm{id} - P - P + PP = \mathrm{id} - P$), and that any direct sum decomposition $Y = X \dotplus Z$ of a finite-dimensional Y necessarily has $X = \operatorname{ran} P$ and $Z = \operatorname{null} P$ for some linear projector P. The following is a more general such claim, of use later.

(5.12) Proposition: Let $X_1, \ldots, X_r$ be linear subspaces of the finite-dimensional vector space Y. Then the following are equivalent.

(i) Y is the direct sum of the X_j, i.e., $Y = X_1 \dot{+} \cdots \dot{+} X_r$.

(ii) There exist $P_j \in L(Y)$ with $\operatorname{ran} P_j = X_j$ so that

$$\text{(5.13)} \qquad \mathrm{id}_Y = P_1 + \cdots + P_r$$

and

$$\text{(5.14)} \qquad P_j P_k = \begin{cases} P_j = P_k & \text{if } j = k; \\ 0 & \text{otherwise.} \end{cases}$$

In particular, each P_j is a linear projector.

Also, the conditions in (ii) uniquely determine the P_j.

Proof: Let V_j be a basis for X_j, all j. By (3.33)Proposition, (i) is equivalent to having $V := [V_1, \ldots, V_r]$ be a basis for Y.

'(i) $\Longrightarrow$ (ii)': By assumption, V is a basis for Y. Let $V^{-1} =: \Lambda^{\mathrm{t}} =: [\Lambda_1, \ldots, \Lambda_r]^{\mathrm{t}}$ be its inverse, grouped correspondingly. Then

$$\mathrm{id}_{\dim Y} = \Lambda^{\mathrm{t}} V = [\Lambda_1, \ldots, \Lambda_r]^{\mathrm{t}} [V_1, \ldots, V_r] = ({\Lambda_i}^{\mathrm{t}} V_j : i, j = 1{:}r),$$

i.e.,

$${\Lambda_i}^{\mathrm{t}} V_j = \begin{cases} \mathrm{id} & \text{if } i = j; \\ 0 & \text{otherwise.} \end{cases}$$

Hence, the linear maps

$$P_j := V_j {\Lambda_j}^{\mathrm{t}}, \quad j = 1{:}r,$$

satisfy (5.14), and $\operatorname{ran} P_j = X_j$, for all j. But also

$$\mathrm{id}_Y = V \Lambda^{\mathrm{t}} = [V_1, \ldots, V_r][\Lambda_1, \ldots, \Lambda_r]^{\mathrm{t}} = \sum_j V_j {\Lambda_j}^{\mathrm{t}},$$

showing (5.13).

'(ii) $\Longrightarrow$ (i)': By assumption, $\operatorname{ran} P_j = \operatorname{ran} V_j$, all j. Therefore, by assumption (5.14),

$$\text{(5.15)} \qquad P_j V_i = \begin{cases} V_j & \text{if } j = i; \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, $0 = V\mathbf{a} = \sum_i V_i \mathbf{a}_i$ implies, for any particular j, that $0 = P_j \mathbf{0} = P_j V \mathbf{a} = \sum_i P_j V_i \mathbf{a}_i = P_j V_j \mathbf{a}_j = V_j \mathbf{a}_j$, hence $\mathbf{a}_j = \mathbf{0}$ (since V_j is 1-1). It follows that V is 1-1. On the other hand, the assumption (5.13) implies that V is onto. Hence, V is a basis for Y.

Finally, to prove the uniqueness of the P_j satisfying (ii), notice that (5.15) pins down P_j on all the columns of V. Since (ii) implies that V is a basis for Y, this therefore determines P_j uniquely (by (3.2)Proposition). □

Returning to the issue of interpolation, this gives the following

(5.16) Corollary: If $V \in L(\mathbb{F}^n, Y)$ is 1-1, and $\Lambda^{\rm t} \in L(Y, \mathbb{F}^n)$ is such that $\operatorname{ran} V \cap \operatorname{null} \Lambda^{\rm t} = \{0\}$, then $P := V(\Lambda^{\rm t} V)^{-1}\Lambda^{\rm t}$ is well-defined; it is the *unique* linear projector P with

$$\operatorname{ran} P = \operatorname{ran} V, \qquad \operatorname{null} P = \operatorname{null} \Lambda^{\rm t}. \tag{5.17}$$

In particular, then $\Lambda^{\rm t}$ is onto, and

$$Y = \operatorname{ran} V \dotplus \operatorname{null} \Lambda^{\rm t}. \tag{5.18}$$

For an arbitrary abstract vector space, it may be very hard to come up with suitable concrete data maps. For that reason, we now consider a particular kind of vector space for which it is very easy to provide suitable data maps, namely the inner product spaces.

6 Inner product spaces

Definition and examples

Inner product spaces are vector spaces with an additional operation, the *inner product*. Here is the definition.

(6.1) Definition: An **inner product space** is a vector space Y (over the field $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$) and an **inner product**, meaning a map

$$\langle\ ,\ \rangle : Y \times Y \to \mathbb{F} : (x, y) \mapsto \langle x, y\rangle$$

that is

(a) **positive definite**, i.e., $\|x\|^2 := \langle x, x\rangle \ge 0$, with equality iff $x = 0$;

(b) **linear in its first argument**, i.e., $\langle\cdot, y\rangle \in L(Y, \mathbb{F})$;

(c) **hermitian**, or **skew-symmetric**, i.e., $\langle y, x\rangle = \overline{\langle x, y\rangle}$.

You already know an inner product space, namely n-dimensional Euclidean space, i.e., the space $\mathbb{F}^n$ of n-vectors with the inner product

$$\langle \mathbf{x}, \mathbf{y}\rangle := \overline{\mathbf{y}}^{\mathrm{t}}\mathbf{x} = \sum_j x_j\overline{y_j} =: \mathbf{y}^{\mathrm{c}}\mathbf{x},$$

though you may know it under the name **scalar product** or **dot product**. In particular, (b) and (c) are evident in this case. As to (a), observe that, for any complex number $z = u + \mathrm{i}v$,

$$\overline{z}z = (u - \mathrm{i}v)(u + \mathrm{i}v) = u^2 + v^2 = |z|^2 \ge 0,$$

with equality if and only if $u = 0 = v$, i.e., $z = 0$. Hence, for any $\mathbf{x} \in \mathbb{F}^n$,

$$\langle \mathbf{x}, \mathbf{x} \rangle = \overline{\mathbf{x}}^{\mathrm{t}}\mathbf{x} = |x_1|^2 + \cdots + |x_n|^2 \geq 0,$$

with equality iff all the x_j are zero, i.e., $\mathbf{x} = \mathbf{0}$.

Of course, if the scalar field is $\mathbb{R}$, we can forget about taking complex conjugates since then $\overline{\mathbf{x}} = \mathbf{x}$. But if $\mathbb{F} = \mathbb{C}$, then it is essential that we define $\langle \mathbf{x}, \mathbf{y} \rangle$ as $\mathbf{y}^{\mathrm{c}}\mathbf{x} = \overline{\mathbf{y}}^{\mathrm{t}}\mathbf{x}$ rather than as $\mathbf{y}^{\mathrm{t}}\mathbf{x}$ since we would not get positive definiteness otherwise. Indeed, if z is a complex number, then there is no reason to think that z^2 is nonnegative, and the following calculation

$$(1, \mathrm{i})^{\mathrm{t}}(1, \mathrm{i}) = 1^2 + (\mathrm{i})^2 = 1 - 1 = 0$$

shows that, for a complex $\mathbf{x}$, $\mathbf{x}^{\mathrm{t}}\mathbf{x}$ can be zero without $\mathbf{x}$ being zero.

So, why not simply stick with $\mathbb{F} = \mathbb{R}$? Work on eigenvalues requires consideration of *complex* scalars (since it relies on zeros of polynomials, and a polynomial may have complex zeros even if all its coefficients are real). For this reason, we have taken the trouble all along to take into account the possibility that $\mathbb{F}$ might be $\mathbb{C}$. It is a minor nuisance at this point, but will save time later.

Another example of an inner product space of great practical interest is the space $Y = \overset{\circ}{C}$ of all continuous 2π-periodic functions, with the inner product

$$\langle f, g \rangle := \int_0^{2\pi} f(t)\overline{g(t)}\,\mathrm{d}t.$$

Of course, we can also think of the space $C([a \,..\, b])$ as an inner product space, with respect to the inner product

$$\langle f, g \rangle := \int_a^b f(t)\overline{g(t)}\,\mathrm{d}t.$$

Often, it is even useful to consider on $C([a \,..\, b])$ the more general inner product

$$\langle f, g \rangle := \int_a^b f(t)\overline{g(t)}w(t)\,\mathrm{d}t$$

with w some positive function on $[a. .b]$, and there are analogous inner product spaces consisting of functions of several variables.

In order to stress the fact that a general inner product space Y behaves just like $\mathbb{F}^n$ with the standard inner product, I will use the notation

$$\forall y \in Y, \quad y^{\mathrm{c}} : Y \to \mathbb{F} : x \mapsto \langle x, y \rangle,$$

for the linear functional provided, according to (6.1)(b), by the inner product, hence will feel free to write $y^{\mathrm{c}}x$ rather than $\langle x, y \rangle$ for the inner product of x with y. Correspondingly, you can read the rest of this chapter as if we were just talking about the familiar space of n-vectors with the dot product, yet be certain that, when the time comes, you will have in hand very useful facts about an arbitrary inner product space, for example the space $\overset{\circ}{C}$.

The conjugate transpose

Here is the promised ready supply of data maps available for an inner product space.

Any column map $W = [w_1, \ldots, w_n] \in L(\mathbb{F}^n, Y)$ into an inner product space Y provides the corresponding data map

$$W^{\mathrm{c}} : Y \mapsto \mathbb{F}^n : x \mapsto ({w_j}^{\mathrm{c}} x : j = 1{:}n), \tag{6.2}$$

called the **conjugate transpose** or **Hermitian** of W.

The terminology comes from the special case $Y = \mathbb{F}^m$. In that case, $W \in \mathbb{F}^{m\times n}$, and then W^{c} is, indeed, just the conjugate transpose of the matrix W since then $w_j = W_{:j} =: \mathbf{w}_j$, hence

$${\mathbf{w}_j}^{\mathrm{c}}\mathbf{x} = {W_{:j}}^{\mathrm{c}}\mathbf{x} = \sum_k \overline{W_{kj}} x_k = \sum_k (W^{\mathrm{c}})_{jk} x_k = (W^{\mathrm{c}}\mathbf{x})_j.$$

With that, if $W \in L(\mathbb{F}^n, Y)$ and $A \in \mathbb{F}^{n\times m}$, then, as

$$WA = [\sum_k \mathbf{w}_k A_{kj} : j = 1{:}n],$$

it follows that, for $j = 1{:}n$,

$$((WA)^{\mathrm{c}}\mathbf{x})_j = (\sum_k \mathbf{w}_k A_{kj})^{\mathrm{c}}\mathbf{x} = \sum_k (A^{\mathrm{c}})_{jk} {\mathbf{w}_k}^{\mathrm{c}}\mathbf{x} = (A^{\mathrm{c}}(W^{\mathrm{c}}\mathbf{x}))_j.$$

This proves

(6.3) Observation: If $W \in L(\mathbb{F}^n, Y)$ and $A \in \mathbb{F}^{n\times m}$, then $WA \in L(\mathbb{F}^m, Y)$ and $(WA)^{\mathrm{c}} = A^{\mathrm{c}} W^{\mathrm{c}}$.

More than that, the conjugate transpose of a column map is a special case of the conjugate transpose A^{c} of a linear map A from an inner product space X to an inner product space Y defined as follows.

(6.4) Definition: Let X and Y be inner product spaces (over the same field $\mathbb{F}$) with inner products $\langle,\rangle_X$ and $\langle,\rangle_Y$, respectively. The **conjugate transpose** of $A \in L(X, Y)$ is the unique map $A^{\mathrm{c}} : Y \to X$ (necessarily linear) for which

$$\forall (x, y) \in X\times Y, \quad \langle x, A^{\mathrm{c}} y\rangle_X = \langle Ax, y\rangle_Y. \tag{6.5}$$

Indeed, if also $\langle x, z\rangle_X = \langle Ax, y\rangle_Y$ for all $x \in X$, then $\langle x, z - A^{\mathrm{c}}y\rangle_X = 0$ for all $x \in X$, including $x = z - A^{\mathrm{c}}y$, hence, by the definiteness of the inner product, $z - A^{\mathrm{c}}y = 0$, showing that $A^{\mathrm{c}}y$ is uniquely determined by (6.5). Since, for arbitrary $x \in X$, $y, z \in Y$ and $\alpha \in \mathbb{F}$, $\langle Ax, y + \alpha z\rangle_Y = \langle Ax, y\rangle_Y + \overline{\alpha}\langle Ax, z\rangle_Y = \langle x, A^{\mathrm{c}}y\rangle_X + \overline{\alpha}\langle x, A^{\mathrm{c}}z\rangle_X = \langle x, A^{\mathrm{c}}y + \alpha A^{\mathrm{c}}z\rangle_X$, therefore, by the uniqueness,

$$A^{\mathrm{c}}(y + \alpha z) = A^{\mathrm{c}}y + \alpha A^{\mathrm{c}}z,$$

i.e., A^{c} is a linear map. Also, the conjugate transpose of an n-column map into Y is, indeed, the conjugate transpose in the sense of (6.4) (with $X = \mathbb{F}^n$), and

$$(BA)^{\mathrm{c}} = A^{\mathrm{c}}B^{\mathrm{c}} \tag{6.6}$$

in case BA makes sense, hence, in particular,

$$A^{-\mathrm{c}} := (A^{-1})^{\mathrm{c}} = (A^{\mathrm{c}})^{-1}. \tag{6.7}$$

The only fly in the ointment is the fact that, for some $A \in L(X, Y)$, there may not be any map $A^{\mathrm{c}} : Y \to X$ satisfying (6.5) unless X is 'complete', a condition that is beyond the scope of this book. However, if both X and Y are finite-dimensional inner-product spaces, then, with V and W bases for X and Y, respectively, we can write any $A \in L(X, Y)$ as $A = W\widehat{A}V^{-1}$ (using the *matrix* $\widehat{A} := W^{-1}AV$), hence, with (6.6), have available the formula

$$A^{\mathrm{c}} = (W\widehat{A}V^{-1})^{\mathrm{c}} = V^{-\mathrm{c}}\widehat{A}^{\mathrm{c}}W^{\mathrm{c}}$$

for the conjugate transpose of A, – another nice illustration of the power of the basis concept.

With that, we are ready for the essential fact about the conjugate transpose needed now.

(6.8) Lemma: If the range of the 1-1 column map V is contained in the range of some column map W, then $W^{\mathrm{c}}V$ is 1-1, hence W^{c} is 1-1 on $\operatorname{ran} V$.

Proof: Assume that $W^{\mathrm{c}}V\mathbf{a} = 0$ and let $b := V\mathbf{a}$. Then $b \in \operatorname{ran} V \subset \operatorname{ran} W$, hence we must have $b = W\mathbf{c}$ for some vector $\mathbf{c}$. Therefore, using (6.3),

$$0 = \mathbf{c}^{\mathrm{c}}\mathbf{0} = \mathbf{c}^{\mathrm{c}}\, W^{\mathrm{c}}V\mathbf{a} = (W\mathbf{c})^{\mathrm{c}}V\mathbf{a} = b^{\mathrm{c}}b.$$

By the definiteness of the inner product, this implies that $b = 0$, i.e., $V\mathbf{a} = 0$, therefore that $\mathbf{a} = \mathbf{0}$, since V is assumed to be 1-1. □

By taking now, in particular, $W = V$, it follows that, for any basis V of the linear subspace X of the inner product space Y, the linear map $(V^cV)^{-1}V^c$ is well-defined, hence provides a formula for V^{-1}.

In MATLAB, the conjugate transpose of a matrix `A` is obtained as `A'`, hence the corresponding formula is `inv(V'*V)*V'`. It is, in effect, used there to carry out the operation `V\` for a matrix `V` that is merely 1-1.

6.1 Prove (6.6) and (6.7).

6.2 Prove that, *for any inner product space X, and any $x \in X$, $[x]^c = x^c$ and that, for any $A \in L(X,Y)$, with Y an inner product space for which A^c exists, $[Ax]^c = x^cA^c$.*

Orthogonal projectors and closest points

Continuing with V a basis for the linear subspace X of the inner product space Y, we recognize that, with the choice $\Lambda^t = V^c$, we have in hand a special case of the situation described in (5.16)Corollary. We conclude that the linear projector

$$P_V := V(V^cV)^{-1}V^c$$

is well-defined. Moreover, by (5.9), $\operatorname{null} P_V = \operatorname{null} V^c = \{y \in Y : V^cy = 0\}$. Since $x \in \operatorname{ran} P_V = \operatorname{ran} V$ is necessarily of the form $x = V\mathbf{a}$, it follows that, for any $x \in \operatorname{ran} P_V$ and any $y \in \operatorname{null} P_V$,

$$x^cy = (V\mathbf{a})^cy = \mathbf{a}^c(V^cy) = 0.$$

In other words, *$\operatorname{ran} P_V$ and $\operatorname{null} P_V = \operatorname{ran}(\operatorname{id} - P_V)$ are perpendicular or orthogonal to each other*, in the sense of the following definition.

(6.9) Definition: We say that the elements u, v of the inner product space Y are **orthogonal** or **perpendicular** to each other, and write this

$$u \perp v,$$

in case $\langle u, v\rangle = 0$.

More generally, for any $F, G \subset Y$, we write $F \perp G$ to mean that, $\forall (f,g) \in F\times G$, $f \perp g$.

The **orthogonal complement**

$$F^\perp := \{y \in Y : y \perp F\}$$

of F is the largest set G perpendicular to F.

Note that $u \perp v$ iff $v \perp u$ since $\langle v, u\rangle = \overline{\langle u, v\rangle}$.

Because of the orthogonality

$$\operatorname{null} P_V = \operatorname{ran}(\operatorname{id} - P_V) \;\perp\; \operatorname{ran} P_V$$

just proved, P_V is called the **orthogonal** projector onto $\operatorname{ran} V$. Correspondingly, we write

$$Y = \operatorname{ran} P_V \oplus \operatorname{null} P_V \tag{6.10}$$

to stress the fact that, in this case, the summands in this direct sum are orthogonal to each other. Since they sum to Y, it follows (see Problem 6.12 below) that each is the orthogonal complement of the other.

This orthogonality, as we show in a moment, has the wonderful consequence that, for any $y \in Y$, $P_V y$ is the unique element of $\operatorname{ran} P_V = \operatorname{ran} V$ that is closest to y in the sense of the **(Euclidean) norm**

$$\|\cdot\| : Y \to \mathbb{R} : y \mapsto \sqrt{y^{\mathrm{c}} y}. \tag{6.11}$$

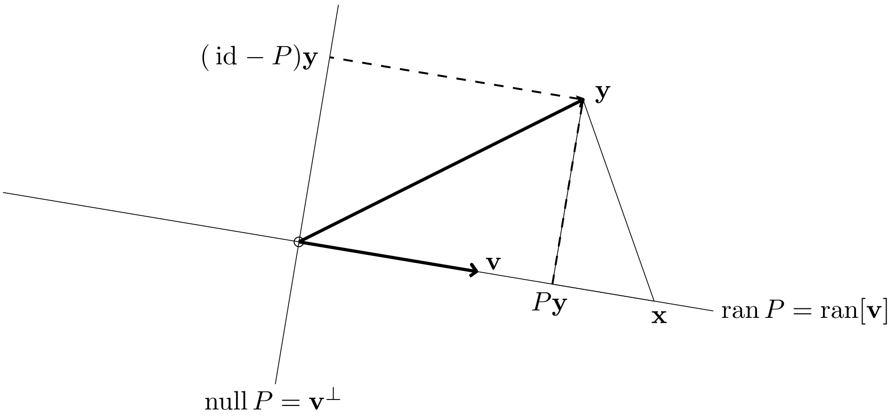
Thus, for every $y \in Y$, our formula for the coordinate vector $\mathbf{a} = (V^{\mathrm{c}} V)^{-1} V^{\mathrm{c}} y$ of $y \in \operatorname{ran} V$ with respect to V gives the coordinates of the point in $\operatorname{ran} V$ closest to y. If $y \in \operatorname{ran} V$, then this is, of course, y itself.

(6.12) Example: We continue with (5.10)Example. In that example, the choice $\Lambda^{\mathrm{t}} = V^{\mathrm{c}}$ amounts to choosing $\mathbf{w} = \mathbf{v}$. Now P becomes $P = \mathbf{v}\mathbf{v}^{\mathrm{c}}/\mathbf{v}^{\mathrm{c}}\mathbf{v}$, and, correspondingly,

$$P\mathbf{y} = \mathbf{v}\frac{\mathbf{v}^{\mathrm{c}}\mathbf{y}}{\mathbf{v}^{\mathrm{c}}\mathbf{v}},$$

which we recognize as the standard formula for the orthogonal projection of the vector $\mathbf{y}$ onto the line spanned by the vector $\mathbf{v}$.

Correspondingly, (5.11)Figure changes to the following.


$(\mathrm{id} - P)\mathbf{y}$
$\mathbf{y}$
$\mathbf{v}$
$P\mathbf{y}$
$\mathbf{x}$
$\operatorname{ran} P = \operatorname{ran}[\mathbf{v}]$
$\operatorname{null} P = \mathbf{v}^{\perp}$

(6.13) Figure. If $\mathbf{y} - P\mathbf{y}$ is perpendicular to $\operatorname{ran} P$, then $P\mathbf{y}$ is the closest point to $\mathbf{y}$ from $\operatorname{ran} P$ since then, for any $\mathbf{x} \in \operatorname{ran} P$, $\|\mathbf{y} - \mathbf{x}\|^2 = \|\mathbf{y} - P\mathbf{y}\|^2 + \|\mathbf{x} - P\mathbf{y}\|^2$.

□

The *proof* that, for any $y \in Y$, $P_V y$ is the unique element of $\operatorname{ran} V$ closest to y in the sense of the norm (6.11) is based on nothing more than the following little calculation.

$$\|u+v\|^2 = (u+v)^{\mathrm{c}}(u+v) = \|u\|^2 + v^{\mathrm{c}}u + u^{\mathrm{c}}v + \|v\|^2. \tag{6.14}$$

Since $v^{\mathrm{c}}u = \overline{u^{\mathrm{c}}v}$, this proves

(6.15) Pythagoras: $u \perp v \quad \Longrightarrow \quad \|u+v\|^2 = \|u\|^2 + \|v\|^2.$

Since, for any $x \in X$, $y - x = (y - P_V y) + (P_V y - x)$, while $(y - P_V y) \in \operatorname{null} P_V \perp \operatorname{ran} P_V = X \ni (P_V y - x)$ we conclude that

$$\|y-x\|^2 = \|y - P_V y\|^2 + \|P_V y - x\|^2. \tag{6.16}$$

Here, the first term on the right is *independent of* x. This shows that $\|y - x\|$ is uniquely minimized over $x \in X$ by the choice $x = P_V y$, as we claimed.

Here is the formal statement.

(6.17) Theorem: For any basis V for the linear subspace X of the inner product space Y, the linear map

$$P_V = V(V^{\mathrm{c}}V)^{-1}V^{\mathrm{c}}$$

equals P_X, the **orthogonal projector onto** X, in the sense that, for all $y \in Y$, $P_V y \in X$ and $y - P_V y \perp X$.

Therefore, Y is the **orthogonal direct sum**

$$Y = \operatorname{ran} V \oplus \operatorname{null} V^{\mathrm{c}} = \operatorname{ran} P_V \oplus \operatorname{null} P_V = X \oplus \operatorname{ran}(\operatorname{id} - P_V),$$

and

$$\forall (y,x) \in Y \times X, \quad \|y - x\| \ge \|y - P_V y\|,$$

with equality if and only if $x = P_V y$.

Incidentally, by choosing $x = 0$ in (6.16) – legitimate since $\operatorname{ran} V$ is a linear subspace – we find the following very useful fact.

(6.18) Proposition: For any 1-1 column map V into Y and any $y \in Y$,

$$\|y\| \ge \|P_V y\|,$$

with equality if and only if $y = P_V y$, i.e., if and only if $y \in \operatorname{ran} V$.

This says that P_V *strictly reduces norms*, except for those elements that it doesn't change at all.

6.3 Construct the orthogonal projection of the vector $(1, 1, 1)$ onto the line $L = \operatorname{ran}[1; -1; 1]$.

6.4 Construct the orthogonal projection of the vector $\mathbf{x} := (1, 1, 1)$ onto the straight line $\mathbf{y} + \operatorname{ran}[\mathbf{v}]$, with $\mathbf{y} = (2, 0, 1)$ and $\mathbf{v} = (1, -1, 1)$, i.e., the point $\mathbf{y} + \alpha\mathbf{v}$ that minimizes $\|\mathbf{x} - (\mathbf{y} + \alpha\mathbf{v})\|$ over all $\alpha \in \mathbb{R}$.

6.5 Compute the distance between the two straight lines $\mathbf{y} + \operatorname{ran}[\mathbf{v}]$ and $\mathbf{z} + \operatorname{ran}[\mathbf{w}]$, with $\mathbf{y} = (2, 0, 1)$, $\mathbf{v} = (1, 1, 1)$, $\mathbf{z} = (-1, 1, -1)$ and $\mathbf{w} = (0, 1, 1)$. (Hint: you want to minimize $\|\mathbf{y} + \alpha\mathbf{v} - (\mathbf{z} + \beta\mathbf{w})\|$ over α, β.)

6.6 With $\mathbf{v}_1 = (1, 2, 2)$, $\mathbf{v}_2 = (-2, 2, -1)$, (a) construct the matrix that provides the orthogonal projection onto the subspace $\operatorname{ran}[\mathbf{v}_1, \mathbf{v}_2]$ of $\mathbb{R}^3$; (b) compute the orthogonal projection of the vector $\mathbf{y} = (1, 1, 1)$ onto $\operatorname{ran}[\mathbf{v}_1, \mathbf{v}_2]$.

6.7* Taking for granted that the space $Y := C([-1\,..\,1])$ of real-valued continuous functions on the interval $[-1..1]$ is an inner product space with respect to the inner product

$$\langle f, g\rangle := \int_{-1}^{1} f(t)g(t)\,\mathrm{d}t,$$

do the following: (a) Construct (a formula for) the orthogonal projector onto $X := \Pi_{<2}$, using the power basis, $V = [()^0, ()^1]$ for X. (b) Use your formula to compute the orthogonal projection of $()^2$ onto $\Pi_{<2}$.

6.8 (a) Prove: *If* $\mathbb{F} = \mathbb{R}$, *then* $u \perp v$ *if and only if* $\|u + v\|^2 = \|u\|^2 + \|v\|^2$. (b) What goes wrong with your argument when $\mathbb{F} = \mathbb{C}$?

6.9 For each of the following maps $f : \mathbb{F}^n \times \mathbb{F}^n \to \mathbb{F}$, determine whether or not it is an inner product.

(a) $\mathbb{F} = \mathbb{R}$, $n = 3$, and $f(\mathbf{x}, \mathbf{y}) = x_1y_1 + x_3y_3$; (b) $\mathbb{F} = \mathbb{R}$, $n = 3$, and $f(\mathbf{x}, \mathbf{y}) = x_1y_1 - x_2y_2 + x_3y_3$; (c) $\mathbb{F} = \mathbb{R}$, $n = 2$, and $f(\mathbf{x}, \mathbf{y}) = x_1^2 + y_1^2 + x_2y_2$; (d) $\mathbb{F} = \mathbb{C}$, $n = 3$, and $f(\mathbf{x}, \mathbf{y}) = x_1y_1 + x_2y_2 + x_3y_3$; (e) $\mathbb{F} = \mathbb{R}$, $n = 3$, and $f(\mathbf{x}, \mathbf{y}) = x_1y_2 + x_2y_3 + x_3y_1$;

6.10 Prove that, *for any invertible* $A \in \mathbb{F}^{n\times n}$, $\langle\cdot,\cdot\rangle : \mathbb{F}^n \times \mathbb{F}^n \to \mathbb{F} : (\mathbf{x}, \mathbf{y}) \mapsto (A\mathbf{y})^{\mathrm{c}}A\mathbf{x} = \mathbf{y}^{\mathrm{c}}(A^{\mathrm{c}}A)\mathbf{x}$ *is an inner product on* $\mathbb{F}^n$.

6.11 Prove that, *for any subset* F *of the inner product space* Y, *the orthogonal complement* $F^\perp$ *is a linear subspace.* (Hint: $F^\perp = \cap_{f\in F} \operatorname{null} f^{\mathrm{c}}$.)

6.12* Prove that, *whenever* $Y = X \oplus Z$, *then* $X^\perp = Z$ *and* $Z^\perp = X$.

6.13 Prove that, *if* X *is a linear subspace of an inner product space* Y, *and* P *is a linear projector on* X *with* $\operatorname{ran} P \subset X$ *and* $\operatorname{ran}(\operatorname{id} - P) \subset X^\perp$, *then* $P = P_X$, *the orthogonal projector onto* X.

6.14 Prove that, *for any linear subspace* X *of a finite-dimensional inner product space* Y, $(\operatorname{id} - P_X) = P_{X^\perp}$.

6.15* Prove that, *for any finite-dimensional linear subspace* X *of an inner product space* Y, $(X^\perp)^\perp = X$.

6.16* Use Problem 6.15 to prove that *two real matrices are row-equivalent (i.e., have the same nullspace) if and only if they have the same row space.*

6.17* An **isometry** or **rigid motion** in an inner product space X is any map $f : X \to X$ that preserves distances, i.e., for which $\|f(x) - f(y)\| = \|x - y\|$ for all $x, y \in X$. Prove that *any rigid motion on a real inner product space* X *that maps the origin to itself is necessarily a linear map.* (Hint: you might prove first that, for any $x \neq y$ and any $\alpha \in \mathbb{R}$, the point $(1 - \alpha)x + \alpha y$ is the unique point in X whose distance from x is $|\alpha|\,\|y - x\|$ and from y is $|1 - \alpha|\,\|y - x\|$.)

Least-squares

We continue our discussion of the orthoprojector $P_V = V(V^{\mathrm{c}}V)^{-1}V^{\mathrm{c}}$ on the inner product space Y onto the subspace X with basis V. Note that, for $y \in Y$, $P_V y = V\mathbf{a}$ with $\mathbf{a}$ the unique solution to the linear equation

$$V^{\mathrm{c}}V? = V^{\mathrm{c}}y.$$

This equation is also referred to as the **normal equation** since it requires that $V^{\mathrm{c}}(y - V\mathbf{a}) = 0$, i.e., that the residual, $y - V\mathbf{a}$, be perpendicular or *normal* to every column of V, hence to all of $\operatorname{ran} V$ (see (6.13)Figure). In effect, given that the equation $V? = y$ doesn't have a solution for $y \in Y\backslash X$, our particular $V\mathbf{a} = P_V y$ gives us the closest thing to a solution.

In particular, if $\mathbf{y} \in Y = \mathbb{R}^n$ and $V \in \mathbb{R}^{n\times r}$ is 1-1, then $P_V\mathbf{y}$ minimizes $\|\mathbf{y} - V\mathbf{a}\|$ over all $\mathbf{a} \in \mathbb{R}^r$. For that reason, the coefficient vector $\mathbf{a} := V^{-1}P_V\mathbf{y}$ is called the **least-squares solution** to the (usually inconsistent or overdetermined) linear system $V? = \mathbf{y}$.

In MATLAB, the vector $P_V\mathbf{y}$ is computed as V*(V\y), in line with the fact mentioned earlier that the action of the matrix $(V^{\mathrm{c}}V)^{-1}V^{\mathrm{c}}$ is provided by the operator V\, i.e., (up to roundoff and for any vector y) the three vectors

```
a1 = V\y, a2 = inv(V'*V)*V'*y, a3 = (V'*V)\(V'*y)
```

are all the same. However, the first way is preferable since it avoids actually forming the matrix V'*V (or its inverse) and, therefore, is less prone to roundoff effects.

A practically very important special case of this occurs when $X = \operatorname{ran} V$ consists of functions on some domain T and, for some finite subset S of T,

$$\delta_S : X \to \mathbb{R}^S : f \mapsto (f(s) : s \in S)$$

is 1-1. A good example of this would be $T = [a\,..\,b]$, $\#S \ge k$, and $X = \Pi_{<k}$. Then

$$(6.19) \qquad \langle f, g\rangle_S := \sum_{s\in S} f(s)g(s) = (\delta_S f)^{\mathrm{t}}(\delta_S g)$$

is an inner product on X since it is evidently linear in the first argument and also hermitian and nonnegative, and is definite since $\langle f, f\rangle_S = 0$ implies $\delta_S f = 0$, hence $f = 0$ since δ_S is assumed to be 1-1. Then, for arbitrary $g \in \mathbb{R}^S$, and with $V =: [v_1, \ldots, v_r]$, we can compute

$$V^{\mathrm{c}}g := (\langle g, v_j\rangle_S : j = 1{:}r) = (\delta_S V)^{\mathrm{t}}g,$$

hence can construct

$$P_{V,S}\,g := V(V^{\mathrm{c}}V)^{-1}V^{\mathrm{c}}g$$

as the unique element $V\mathbf{a}$ of $\operatorname{ran} V$ closest to g in the sense that the sum of squares $\sum_{s\in S}|g(s)-(V\mathbf{a})(s)|^2$ is as small as possible. For this reason, $P_{V,S}g$ is also called the **discrete least-squares approximation** from $\operatorname{ran} V$ to g, or, more explicitly, to the data $((s,g(s)) : s \in S)$. If $\#V = \#S$, then $P_{V,S}g$ is the unique interpolant to these data from $\operatorname{ran} V$.

In any calculation of such a discrete least-squares approximation, we would, of course, have to list the elements of S in some fashion, say as the entries s_j of the sequence $(s_1,\ldots,s_n)$. Then we can think of δ_S as the data map into $\mathbb{R}^n$ given by $f \mapsto (f(s_j) : j = 1{:}n)$. Correspondingly, $\delta_S V$ becomes an $n \times r$-matrix, and this matrix is 1-1, by the assumption that δ_S is 1-1 on $X = \operatorname{ran} V$. Further, the coefficient vector $\mathbf{a} := (V^{\mathrm{c}}V)^{-1}V^{\mathrm{c}}g$ for $P_{V,S}g$ is the least-squares solution to the linear equation

$$\delta_S V? = g$$

which seeks a coefficient vector $\mathbf{a}$ so that $V\mathbf{a}$ interpolates to the data $((s_j, g(s_j)) : j = 1{:}n)$. Such an interpolant exists if and only if the matrix $\delta_S V$ is invertible. Otherwise, one has to be content with a least-squares solution, i.e., a discrete least-squares approximation to these data, from $\operatorname{ran} V$.

6.18* Compute the discrete least squares approximation by straight lines (i.e., from $\Pi_{<2}$) to the data (j, j^2), $j = 1{:}10$ using (a) the basis $[()^0, ()^1]$; (b) the basis $[()^0, ()^1 - 5.5()^0]$. (c) Why might one prefer (b) to (a)?

Orthonormal column maps

The formula

$$P_V = V(V^{\mathrm{c}}V)^{-1}V^{\mathrm{c}}$$

for the orthogonal projector onto the range of the 1-1 column map V becomes particularly simple in case

$$V^{\mathrm{c}}V = \mathrm{id}; \tag{6.20}$$

it then reduces to

$$P_V = VV^{\mathrm{c}}.$$

We call V **orthonormal** (or, **o.n.**, for short) in this case since, written out entry by entry, (6.20) reads

$$\langle v_j, v_k\rangle = \left\{\begin{matrix} 1 & \text{if } j = k; \\ 0 & \text{otherwise,}\end{matrix}\right\} =: \delta_{jk}.$$

In other words, each column of V is *normalized*, meaning that it has norm 1, and different columns are orthogonal to each other. Such bases are special in that they provide their own inverse, i.e.,

$$\forall x \in \operatorname{ran} V, \quad x = V(V^{\mathrm{c}}x).$$

The term 'orthonormal' can be confusing, given that earlier we mentioned the normal equation, $V^{\rm c}V? = V^{\rm c}y$, socalled because it expresses the condition that the residual, $y - V\mathbf{a}$, be orthogonal or 'normal' to the columns of V. In fact, *norma* is the Latin name for a mason's tool for checking that a wall is at right angles to the ground. In the same way, the *normal* to a surface at a point is a vector at right angles to the surface at that point. Nevertheless, to **normalize** the vector y does not mean to change it into a vector that is perpendicular to some subspace or set. Rather, it means to divide it by its norm, thereby obtaining the **normalized vector** or **direction**[†] $y/\|y\|$ that points in the same direction as y but has norm 1. To be sure, this can only be done for $y \neq 0$ and then $\|\, y/\|y\|\, \| = 1$ because the Euclidean norm is **absolutely homogeneous**, meaning that

$$\forall(\alpha, y) \in \mathbb{F}\times Y, \quad \|\alpha y\| = |\alpha|\|y\|. \tag{6.21}$$

We now show that every finite-dimensional linear subspace of an inner-product space Y has o.n. bases.

(6.22) Proposition: For every 1-1 $V \in L(\mathbb{F}^n, Y)$, there exists an o.n. $Q \in L(\mathbb{F}^n, Y)$ so that, for all j, $\operatorname{ran}[q_1, q_2, \ldots, q_j] = \operatorname{ran}[v_1, v_2, \ldots, v_j]$, hence $V = QR$ with R (invertible and) upper triangular, a **QR factorization** for V.

Proof: For $j = 1{:}n$, define $u_j := v_j - P_{V_{<j}} v_j$, with $V_{<j} := V_{j-1} := [v_1, \ldots, v_{j-1}]$. By (6.17)Theorem, $u_j \perp \operatorname{ran} V_{<j}$, all j, hence $u_j \perp u_k$ for $j \neq k$. Also, each u_j is nonzero (since $u_j = V_j(\mathbf{a}, 1)$ for some $\mathbf{a} \in \mathbb{F}^{j-1}$, and V_j is 1-1), hence $q_j := u_j/\|u_j\|$ is well-defined and, still, $q_j \perp q_k$ for $j \neq k$.

It follows that $Q := [q_1, \ldots, q_n]$ is o.n., hence, in particular, 1-1. Finally, since $q_j = u_j/\|u_j\| \in \operatorname{ran} V_j$, it follows that, for each j, the 1-1 map $[q_1, \ldots, q_j]$ has its range in the j-dimensional space $\operatorname{ran} V_j$, hence must be a basis for it. □

Since $Q_{<j} = [q_1, \ldots, q_{j-1}]$ is an o.n. basis for $\operatorname{ran} V_{<j}$, it is of help in constructing q_j since it gives

(6.23)
$$u_j = v_j - P_{V_{<j}} v_j, \quad \text{with } P_{V_{<j}} v_j = P_{Q_{<j}} v_j = \sum_{k<j} q_k \langle v_j, q_k\rangle = \sum_{k<j} u_k \frac{\langle v_j, u_k\rangle}{\langle u_k, u_k\rangle}.$$

For this reason, it is customary to construct the u_j or the q_j's one by one, from the first to the last, using (6.23). This process is called **Gram-Schmidt**

[†] In this book, we avoid using the term *unit vector* for what we have just agreed to call a *direction* since *unit vector* means to some only one of the coordinate directions $\mathbf{e}_j$.

orthogonalization. To be sure, as (6.23) shows, there is no real need (other than neatness) to compute the q_j from the u_j and, by skipping the calculation of q_j, one avoids taking square-roots.

Since any 1-1 column map into a finite-dimensional vector space can be extended to a basis for that vector space, we have also proved the following.

(6.24) Corollary: Every o.n. column map Q into a finite-dimensional inner product space can be extended to an o.n. basis for that space.

Given any 1-1 matrix V, the MATLAB command [q,r] = qr(V,0) provides an o.n. basis, q, for ran V, along with the upper triangular matrix r for which q*r equals V. The (simpler) statement [Q,R]=qr(V) provides a *unitary*, i.e., a *square* o.n., matrix Q and an upper triangular matrix R so that Q*R equals V. If V is itself square, then q equals Q. In the contrary case, Q equals [q,U] for some o.n. basis U of the orthogonal complement of ranV. Finally, the simplest possible statement, p = qr(V), gives the most complicated result, namely a matrix p of the same size as V that contains r in its upper triangular part and complete information about the various Householder matrices used in its strictly lower triangular part.

While, for each $j = 1{:}\#\mathtt{V}$, $\mathrm{ran}\mathtt{V}(\,:\,,1{:}j) = \mathrm{ran}\mathtt{Q}(\,:\,,1{:}j)$, the construction of q or Q does not involve the Gram-Schmidt algorithm, as that algorithm is not reliable numerically when applied to an arbitrary 1-1 matrix V. Rather, the matrix V is factored column by column with the aid of certain elementary matrices, the so-called **Householder reflections** $\mathrm{id} - 2\mathbf{w}\mathbf{w}^{\mathrm{c}}/\mathbf{w}^{\mathrm{c}}\mathbf{w}$.

As already observed, it is customary to call a square o.n. matrix **unitary**. It is also customary to call a *real* unitary matrix **orthogonal**. However, the columns of such an 'orthogonal matrix' are not just orthogonal to each other, they are also normalized. Thus it would be better to call such a matrix 'orthonormal', freeing the term 'orthogonal matrix' to denote one whose columns are merely orthogonal to each other. But such naming conventions are hard to change. I will simply not use the term 'orthogonal matrix', but use 'real unitary matrix' instead.

An o.n. column map Q has many special properties, all of which derive from the defining property, $Q^{\mathrm{c}}Q = \mathrm{id}$, by the observation that therefore, for any $\mathbf{a}, \mathbf{b} \in \mathbb{F}^n$,

$$(6.25) \qquad \langle Q\mathbf{a}, Q\mathbf{b}\rangle \;=\; \langle Q^{\mathrm{c}}Q\mathbf{a}, \mathbf{b}\rangle = \langle \mathbf{a}, \mathbf{b}\rangle.$$

This says that Q is **inner-product preserving**. In particular, any o.n. $Q \in L(\mathbb{F}^n, X)$ is an **isometry** in the sense that

$$(6.26) \qquad \forall \mathbf{a}, \mathbf{b} \in \mathbb{F}^n, \quad \|Q\mathbf{a} - Q\mathbf{b}\| = \|\mathbf{a} - \mathbf{b}\|.$$

More than that, any o.n. $Q \in L(\mathbb{F}^n, X)$ is **angle-preserving** since a standard definition of the **angle** φ between two real nonzero n-vectors $\mathbf{x}$ and

$\mathbf{y}$ is the following implicit one:

$$\cos(\varphi) := \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \|\mathbf{y}\|}.$$

To be sure, this definition makes sense only if we can be sure that the right-hand side lies in the interval $[-1\,..\,1]$. But this is a consequence of the

Cauchy-Bunyakovski-Schwarz or CBS Inequality: For any u, v in the inner product space Y,

$$|\langle u, v \rangle| = |v^{\mathrm{c}} u| \le \|u\| \, \|v\|, \tag{6.27}$$

with equality if and only if $[u, v]$ is not 1-1.

Be sure to remember not only the *in*equality, but also exactly when it is an *equal*ity.

Proof: If $v = 0$, then there is equality in (6.27) and $[u, v]$ is not 1-1. Otherwise, $v \ne 0$ and, in that case, by (6.18)Proposition, the orthogonal projection $P_{[v]} u = v(v^{\mathrm{c}} u)/\|v\|^2$ onto ran$[v]$ of an arbitrary $u \in Y$ has norm smaller than $\|u\|$ unless $u = P_{[v]} u$. In other words, $|v^{\mathrm{c}} u|/\|v\| = \|v(v^{\mathrm{c}} u)/\|v\|^2\| \le \|u\|$, showing that (6.27) holds in this case, with equality if and only if $u \in \operatorname{ran}[v]$. □

6.19 Prove (6.21).

6.20 Prove that $V = \begin{bmatrix} 1 & 1 & 1 \\ -1 & 1 & -1 \\ 0 & 1 & 2 \end{bmatrix}$ is a basis for $\mathbb{R}^3$ and compute the coordinates of $\mathbf{x} := (1, 1, 1)$ with respect to V.

6.21 Verify that $V = \begin{bmatrix} 1 & -1 & 1 \\ 1 & -1 & -1 \\ 1 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$ is an orthogonal basis for its range, and extend it to an orthogonal basis for $\mathbb{R}^4$.

6.22 (a) Use the calculations in Problem 6.18 to construct an orthogonal basis for $\Pi_{<3}$ from the power basis $V = [()^0, ()^1, ()^2]$ with respect to the (discrete) inner product in Problem 6.18.

(b) Use (a) to compute the discrete least-squares approximation from $\Pi_{<3}$ to the data (j, j^3), $j = 1{:}10$.

6.23 Use the result of Problem 6.7 to construct an o.n. basis for $\Pi_{<3}$ wrto the inner product $\langle f, g \rangle := \int_{-1}^{1} f(t) g(t) \, \mathrm{d}t$.

6.24 What is the angle between $(1, 2, 2)$ and $(3, -1, -2)$?

6.25 Consider the **Vandermonde** matrix

$$A := [\delta_{z_0}, \ldots, \delta_{z_k}]^{\mathrm{c}} [()^0, \ldots, ()^k] = (z_i^j : i, j = 0{:}k)$$

for some sequence $z_0, \ldots, z_k$ of complex numbers.

Prove that A is a scalar multiple of a unitary matrix if and only if, for some real α,

$$(z_0, \ldots, z_k) = (\exp(2\pi\mathrm{i}(\alpha + j/(k+1))) : j = 0{:}k).$$

$\operatorname{ran} A$ and $\operatorname{null} A^{\mathrm{c}}$ form an orthogonal direct sum for $\operatorname{tar} A$

The two basic linear subspaces associated with $A \in L(X, Y)$ are its range, $\operatorname{ran} A$, and its kernel or nullspace, $\operatorname{null} A$. However, when X and Y are finite-dimensional inner product spaces, it is also possible and very useful to consider the range of A and the nullspace of the (conjugate) transpose A^{c} of A together. For, then, by the definiteness of the inner product, $A^{\mathrm{c}} y = 0$ iff $\langle x, A^{\mathrm{c}} y \rangle = 0$ for all $x \in X$, while, by (6.5), $\langle x, A^{\mathrm{c}} y \rangle = \langle Ax, y \rangle$, hence

$$\operatorname{null} A^{\mathrm{c}} = \{y \in Y : y \perp \operatorname{ran} A\}.$$

Recalling the notation

$$M^{\perp} := \{y \in Y : y \perp M\}$$

for the *orthogonal complement* of the subset M of Y, we get the following.

(6.28) Proposition: For any $A \in L(X, Y)$, $(\operatorname{ran} A)^{\perp} = \operatorname{null} A^{\mathrm{c}}$.

(6.29) Corollary: For any $A \in L(X, Y)$, Y is the *orthogonal* direct sum $Y = \operatorname{ran} A \oplus \operatorname{null} A^{\mathrm{c}}$. Hence

$$\dim \operatorname{tar} A = \dim \operatorname{ran} A + \dim \operatorname{null} A^{\mathrm{c}}.$$

Proof: Let V be any basis for $\operatorname{ran} A$. By (6.17)Theorem,

$$Y = \operatorname{ran} V \oplus \operatorname{null} V^{\mathrm{c}},$$

while, by choice of V, $\operatorname{ran} V = \operatorname{ran} A$, and so, by (6.28), $\operatorname{null} V^{\mathrm{c}} = (\operatorname{ran} V)^{\perp} = (\operatorname{ran} A)^{\perp} = \operatorname{null} A^{\mathrm{c}}$. □

In particular, A is onto if and only if A^{c} is 1-1. Further, since $(A^{\mathrm{c}})^{\mathrm{c}} = A$, we also have the following complementary statement.

(6.30) Corollary: For any $A \in L(X,Y)$, X is the *orthogonal* direct sum $X = \operatorname{ran} A^{\mathrm{c}} \oplus \operatorname{null} A$. Hence,

$$\dim\operatorname{dom} A = \dim\operatorname{ran} A^{\mathrm{c}} + \dim\operatorname{null} A.$$

In particular, A^{c} is onto if and only if A is 1-1. Also, on comparing (6.30) with the (3.23)Dimension Formula, we see that $\dim\operatorname{ran} A = \dim\operatorname{ran} A^{\mathrm{c}}$.

The fact (see (6.29)Corollary) that $\operatorname{tar} A = \operatorname{ran} A \oplus \operatorname{null} A^{\mathrm{c}}$ is often used as a characterization of the elements $y \in \operatorname{tar} A$ for which the equation $A? = y$ has a solution. For, it says that $y \in \operatorname{ran} A$ if and only if $y \perp \operatorname{null} A^{\mathrm{c}}$. Of course, since $\operatorname{null} A^{\mathrm{c}}$ consists exactly of those vectors that are orthogonal to all the columns of A, this is just a special case of the fact (see Problem 6.15) that the orthogonal complement of the orthogonal complement of a linear subspace is that linear subspace itself.

The inner product space $\mathbb{F}^{m\times n}$ and the trace of a matrix

At the outset of this book, we introduced the space $\mathbb{F}^{m\times n}$ as a special case of the space $\mathbb{F}^T$ of all scalar-valued functions on some set T, namely with

$$T = \underline{m} \times \underline{n}.$$

This set being finite, there is a natural inner product on $\mathbb{F}^{m\times n}$, namely

$$\langle A, B\rangle := \sum_{i,j} \overline{B_{ij}} A_{ij}.$$

This inner product can also be written in the form

$$\langle A, B\rangle = \sum_{i,j} (B^{\mathrm{c}})_{ji} A_{ij} = \sum_j (B^{\mathrm{c}} A)_{jj} = \operatorname{trace}(B^{\mathrm{c}} A).$$

Here, the **trace** of a square matrix C is, by definition, the sum of its diagonal entries,

$$\operatorname{trace} C := \sum_j C_{jj}.$$

Note that

$$\operatorname{trace}(AB^{\mathrm{c}}) = \sum_{i,j} A_{ij}\overline{B_{ij}} = \sum_{j,i} \overline{B_{ij}} A_{ij} = \operatorname{trace}(B^{\mathrm{c}} A). \tag{6.31}$$

The norm in this inner product space is called the **Frobenius norm**,

$$\|A\|_F^2 := \operatorname{trace} A^{\mathrm{c}} A = \sum_{i,j} |A_{ij}|^2. \tag{6.32}$$

The Frobenius norm is **compatible** with the Euclidean norm $\|\ \|_2$ on $\mathbb{F}^n$ and $\mathbb{F}^m$ in the sense that (according to Problem 6.30)

$$\|A\mathbf{x}\|_2 \le \|A\|_F \|\mathbf{x}\|_2, \quad \mathbf{x} \in \mathbb{F}^n. \tag{6.33}$$

Not surprisingly, the map $\mathbb{F}^{m\times n} \to \mathbb{F}^{n\times m} : A \mapsto A^{\mathrm{t}}$ is unitary, i.e., inner-product preserving:

$$\langle A^{\mathrm{t}}, B^{\mathrm{t}} \rangle = \sum_{i,j} \overline{(B^{\mathrm{t}})_{ij}} (A^{\mathrm{t}})_{ij} = \sum_{i,j} \overline{B_{ji}} A_{ji} = \langle A, B \rangle. \tag{6.34}$$

6.26 Prove that $\operatorname{trace} A = \langle A, \operatorname{id} \rangle$.

6.27 (6.31) shows that $\operatorname{trace}(AB) = \operatorname{trace}(BA)$. Give an example to show that $\operatorname{trace}(AB)$ does not in general equal $(\operatorname{trace} A)(\operatorname{trace} B)$.

6.28 Verify that $\operatorname{trace} : \mathbb{F}^{n\times n} \to \mathbb{F} : A \mapsto \operatorname{trace} A$ is a linear map that satisfies $\operatorname{trace}(AB) = \operatorname{trace}(BA)$.

6.29 Prove: *If $f : \mathbb{F}^{n\times n} \to \mathbb{F}$ is a linear map that satisfies $f(AB) = f(BA)$ for all $A, B \in \mathbb{F}^{n\times n}$ and is normalized so that $f(\operatorname{id}_n) = n$, then $f = \operatorname{trace}$.* (Hint: Problem 2.13.)

6.30* Prove (6.33).

6.31 T/F

(a) $(\mathbf{x}, \mathbf{y}) \mapsto \mathbf{y}^{\mathrm{c}} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \mathbf{x}$ is an inner product on $\mathbb{R}^2$;

(b) $\|x + y\|^2 \le \|x\|^2 + \|y\|^2$.

(c) For any $A \in \mathbb{F}^{m\times n}$, $\operatorname{trace} A^{\mathrm{c}} A \ge 0$ with eqality iff $A = 0$.

7 Norms, map norms, and the condition of a basis

Assume that V is a basis for the nontrivial linear subspace X of the inner product space Y. The coordinate vector $\mathbf{a}$ for $x \in X$ is the unique solution of the equation

$$V? = x.$$

We may not be able to compute the solution exactly. Even if we know the entries of the solution exactly, as common fractions, say, we may not be able to use them exactly if we use some floating-point arithmetic, as is common. It is for this reason that one is interested in gauging the effect of an erroneous coordinate vector $\widehat{\mathbf{a}}$ on the accuracy of $V\widehat{\mathbf{a}}$ as a representation for $x = V\mathbf{a}$.

How to judge the error by the residual

Since, presumably, we do not know $\mathbf{a}$, we cannot compute the **error**

$$\varepsilon := \mathbf{a} - \widehat{\mathbf{a}};$$

we can only compute the **residual**

$$r := x - V\widehat{\mathbf{a}}.$$

Nevertheless, can we judge the error by the residual? Does a 'small' **relative residual**

$$\|r\|/\|x\|$$

imply a 'small' **relative error**

$$\|\varepsilon\|/\|\mathbf{a}\| \ ?$$

By definition, the **condition** (or, **condition number**) $\kappa(V)$ of the basis V is the greatest factor by which the relative error, $\|\boldsymbol{\varepsilon}\|/\|\mathbf{a}\|$, can exceed the relative residual, $\|r\|/\|x\| = \|V\boldsymbol{\varepsilon}\|/\|V\mathbf{a}\|$; i.e.,

$$(7.1) \qquad \kappa(V) := \sup_{\mathbf{a},\boldsymbol{\varepsilon}} \frac{\|\boldsymbol{\varepsilon}\|/\|\mathbf{a}\|}{\|V\boldsymbol{\varepsilon}\|/\|V\mathbf{a}\|}$$

(here and below, the supremum (or infimum) is only taken over well-defined expressions, i.e., in this case, both $\mathbf{a}$ and $V\boldsymbol{\varepsilon}$ are restricted to be nonzero). However, by interchanging here the roles of $\mathbf{a}$ and $\boldsymbol{\varepsilon}$ and then taking reciprocals, this also says that

$$1/\kappa(V) = \inf_{\boldsymbol{\varepsilon},\mathbf{a}} \frac{\|\boldsymbol{\varepsilon}\|/\|\mathbf{a}\|}{\|V\boldsymbol{\varepsilon}\|/\|V\mathbf{a}\|}.$$

Hence, altogether,

$$(7.2) \qquad \frac{1}{\kappa(V)}\frac{\|r\|}{\|x\|} \le \frac{\|\boldsymbol{\varepsilon}\|}{\|\mathbf{a}\|} \le \kappa(V)\frac{\|r\|}{\|x\|}.$$

In other words, *the larger the condition number, the less information about the size of the relative error is provided by the size of the relative residual.*

For a better feel for the condition number, note that we can also write the formula (7.1) for $\kappa(V)$ in the following fashion:

$$\kappa(V) = \sup_{\boldsymbol{\varepsilon}} \frac{\|\boldsymbol{\varepsilon}\|}{\|V\boldsymbol{\varepsilon}\|} \sup_{\mathbf{a}} \frac{\|V\mathbf{a}\|}{\|\mathbf{a}\|}.$$

Also,

$$\|V\mathbf{a}\|/\|\mathbf{a}\| = \|V(\mathbf{a}/\|\mathbf{a}\|)\|,$$

with $\mathbf{a}/\|\mathbf{a}\|$ *normalized*, i.e., of norm 1. Hence, altogether,

$$(7.3) \qquad \kappa(V) = \frac{\sup\{\|V\mathbf{a}\| : \|\mathbf{a}\| = 1\}}{\inf\{\|V\mathbf{a}\| : \|\mathbf{a}\| = 1\}}.$$

This says that we can visualize the condition number $\kappa(V)$ in the following way, as illustrated in (7.5)Figure. Consider the image

$$(7.4) \qquad \{V\mathbf{a} : \|\mathbf{a}\| = 1\}$$

under V of the **unit sphere**

$$\{\mathbf{a} \in \mathbb{F}^n : \|\mathbf{a}\| = 1\}$$

in $\mathbb{F}^n$. It will be some kind of ellipsoid, symmetric with respect to the origin. In particular, by (17.6)Theorem, there will be a point $\mathbf{a}_{\max}$ with $\|\mathbf{a}_{\max}\| = 1$

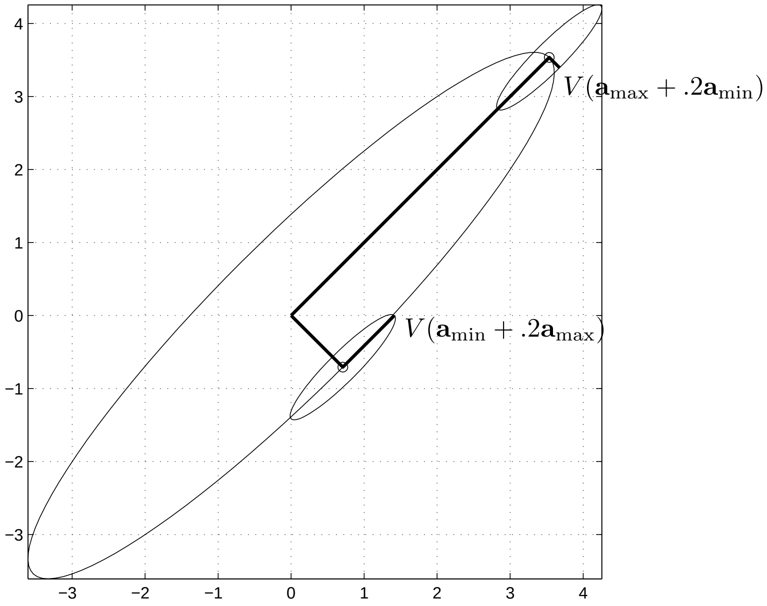
for which $V\mathbf{a}_{\max}$ will be as far from the origin as possible. There will also be a point $\mathbf{a}_{\min}$ with $\|\mathbf{a}_{\min}\| = 1$ for which $V\mathbf{a}_{\min}$ will be as close to the origin as possible. In other words,

$$\kappa(V) = \|V\mathbf{a}_{\max}\|/\|V\mathbf{a}_{\min}\|,$$

saying that the condition number gives the ratio of the largest to the smallest diameter of the ellipsoid (7.4). The larger the condition number, the skinnier is the ellipsoid.

In particular, if $\mathbf{a} = \mathbf{a}_{\max}$ while $\boldsymbol{\varepsilon} = \mathbf{a}_{\min}$, then the relative error is 1 while the relative residual is $\|V\mathbf{a}_{\min}\|/\|V\mathbf{a}_{\max}\|$, and this is tiny to the extent that the ellipsoid is 'skinny'.

On the other hand, if $\mathbf{a} = \mathbf{a}_{\min}$ while $\boldsymbol{\varepsilon} = \mathbf{a}_{\max}$, then the relative error is still 1, but now the relative residual is $\|V\mathbf{a}_{\max}\|/\|V\mathbf{a}_{\min}\|$, and this is large to the extent that the ellipsoid is 'skinny'.



(7.5) Figure. Extreme effects of a 20% relative error on the relative residual, for $V = \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix}$.

The worst-conditioned column maps V are those that fail to be 1-1 since, for them, $V\mathbf{a}_{\min} = 0$, hence $\kappa(V) = \infty$.

On the other extreme, it follows directly from (7.3) that $\kappa(V) \ge 1$, and this lower bound is reached by any o.n. basis V since any o.n. basis is an isometry, by (6.26), i.e., $\|V\mathbf{a}\| = \|\mathbf{a}\|$ for all $\mathbf{a} \in \mathbb{F}^n$. Thus o.n. bases are best-conditioned, and rightfully prized for that. It was for this reason that we took the trouble to prove that every finite-dimensional linear subspace of an inner product space has o.n. bases, and even discussed just how to construct such bases.

The map norm

As we now explain, the numbers $\|V\mathbf{a}_{\max}\| = \max\{\|V\mathbf{a}\| : \|\mathbf{a}\| = 1\}$ and $1/\|V\mathbf{a}_{\min}\| = 1/\min\{\|V\mathbf{a}\| : \|\mathbf{a}\| = 1\}$ both are examples of a map norm according to the following

(7.6) Definition: The **map norm**, $\|A\|$, of $A \in L(X,Y)$ is the smallest nonnegative number c for which

$$\forall x \in X, \quad \|Ax\| \le c\|x\|.$$

If X is trivial, then $\|A\| = 0$ for the sole $A \in L(X,Y)$. Otherwise

$$\|A\| = \sup_{x \ne 0} \|Ax\|/\|x\| = \sup\{\|Ax\| : \|x\| = 1\}. \tag{7.7}$$

Here, the last equality follows from the absolute homogeneity of the norm and the homogeneity of A which combine to permit the conclusions that

$$\|Ax\|/\|x\| = \|A(x/\|x\|)\| \quad \text{and} \quad \|(x/\|x\|)\| = 1.$$

In this book, we are only interested in *finite-dimensional* X and, for such X,

$$\|A\| = \max_{x \ne 0} \|Ax\|/\|x\| = \max\{\|Ax\| : \|x\| = 1\}. \tag{7.8}$$

The reason for this is beyond the scope of this book, but is now stated for the record:

(7.9) Fact: If X is a finite-dimensional normed vector space and $A \in L(X,Y)$ for some normed vector space Y, then

$$F : x \mapsto \|Ax\|$$

is continuous and the **unit sphere**

$$\{x \in X : \|x\| = 1\}$$

is compact, hence F achieves its maximum value, $\|A\|$, on that sphere.

See page 276 for more details. For the same reason, F also achieves its minimum value on the unit sphere, and this justifies the existence of $\mathbf{a}_{\max}$ and $\mathbf{a}_{\min}$ in the preceding section.

We conclude that *determination* of the map norm is a two-part process, as formalized in the following.

(7.10) Calculation of $\|A\|$: The number c equals the norm $\|A\|$ if and only if

(i) for all x, $\|Ax\| \le c\|x\|$; and

(ii) for some $x \ne 0$, $\|Ax\| \ge c\|x\|$.

The first says that $\|A\| \le c$, while second says that $\|A\| \ge c$, hence, together they say that $\|A\| = c$. See, e.g., the answer to Problem 7.13 for an illustration.

(7.11) Example: We compute $\|A\|$ in case $A \in \mathbb{F}^{m\times n}$ is of the simple form

$$A = [\mathbf{v}][\mathbf{w}]^{\mathrm{c}} = \mathbf{v}\mathbf{w}^{\mathrm{c}}$$

for some $\mathbf{v} \in \mathbb{F}^m$ and some $\mathbf{w} \in \mathbb{F}^n$. Since $A\mathbf{x} = (\mathbf{v}\mathbf{w}^{\mathrm{c}})\mathbf{x} = \mathbf{v}(\mathbf{w}^{\mathrm{c}}\mathbf{x})$, we have

$$\|(\mathbf{v}\mathbf{w}^{\mathrm{c}})\mathbf{x}\| = \|\mathbf{v}\||\mathbf{w}^{\mathrm{c}}\mathbf{x}| \le \|\mathbf{v}\|\|\mathbf{w}\|\|\mathbf{x}\|,$$

the equality by the absolute homogeneity of the norm, and the inequality by (6.27)Cauchy's Inequality. This shows that $\|\mathbf{v}\mathbf{w}^{\mathrm{c}}\| \le \|\mathbf{v}\|\|\mathbf{w}\|$. On the other hand, for the specific choice $\mathbf{x} = \mathbf{w}$, we get

$$(\mathbf{v}\mathbf{w}^{\mathrm{c}})\mathbf{w} = \mathbf{v}(\mathbf{w}^{\mathrm{c}}\mathbf{w}) = \mathbf{v}\|\mathbf{w}\|^2,$$

hence $\|\mathbf{v}\mathbf{w}^{\mathrm{c}}\|\|\mathbf{w}\| \ge \|\mathbf{v}\mathbf{w}^{\mathrm{c}}\mathbf{w}\| = \|\mathbf{v}\|\|\mathbf{w}\|^2$, therefore, assuming that $\mathbf{w} \ne \mathbf{0}$, $\|\mathbf{v}\mathbf{w}^{\mathrm{c}}\| \ge \|\mathbf{v}\|\|\mathbf{w}\|$, while this inequality holds trivially if $\mathbf{w} = \mathbf{0}$. So, altogether, we have that

$$\|\mathbf{v}\mathbf{w}^{\mathrm{c}}\| = \|\mathbf{v}\|\|\mathbf{w}\|.$$

Note that this, incidentally, proves that, for any $\mathbf{v} \in \mathbb{F}^n$,

$$\|[\mathbf{v}]\| = \|\mathbf{v}\| = \|[\mathbf{v}]^{\mathrm{c}}\|. \tag{7.12}$$

□

As another example, note that, if also $B \in L(Y, Z)$ for some inner product space Z, then BA is defined and

$$\|(BA)x\| = \|B(Ax)\| \le \|B\|\, \|Ax\| \le \|B\|\, \|A\| \|x\|.$$

Therefore,

$$\|BA\| \le \|B\|\, \|A\|. \tag{7.13}$$

We are ready to discuss the condition (7.3) of a basis V in terms of map norms.

Directly from (7.8), $\max\{\|V\mathbf{a}\| : \|\mathbf{a}\| = 1\} = \|V\|$.

(7.14) Proposition: If $A \in L(X, Y)$ is invertible and $X \ne \{0\}$ is finite-dimensional, then

$$\|A^{-1}\| = 1/\min\{\|Ax\| : \|x\| = 1\}.$$

Proof: Since A is invertible, $y \in Y$ is nonzero if and only if $y = Ax$ for some nonzero $x \in X$. Hence,

$$\|A^{-1}\| = \max_{y \ne 0} \frac{\|A^{-1}y\|}{\|y\|} = \max_{x \ne 0} \frac{\|A^{-1}Ax\|}{\|Ax\|} = 1/\min_{x \ne 0} \frac{\|Ax\|}{\|x\|},$$

and this equals $1/\min\{\|Ax\| : \|x\| = 1\}$ by the absolute homogeneity of the norm and the homogeneity of A. □

In particular, $1/\|A^{-1}\|$ is the largest number c for which

$$\forall x \in X, \quad c\|x\| \le \|Ax\|.$$

We conclude that

$$\kappa(V) = \|V\| \|V^{-1}\|. \tag{7.15}$$

7.1 Complement (7.14)Proposition by discussing the situation when $X = \{0\}$.

7.2 Prove that $\kappa(V) \ge 1$ for any basis V with at least one column.

7.3 Determine $\kappa([\,])$.

7.4* Discuss a relationship, if any, between the condition of a product and the condition of its factors.

Vector norms and their associated map norms

`MATLAB` provides the map norm of the matrix `A` by the statement `norm(A)` (or by the statement `norm(A,2)`, indicating that there are other map norms available).

The `norm` command gives the Euclidean norm when its argument is a 'vector'. Specifically, `norm(v)` and `norm(v,2)` both give $\|\mathbf{v}\| = \sqrt{\mathbf{v}^c\mathbf{v}}$. However, since in (present-day) `MATLAB`, everything is a matrix, there is room here for confusion since experimentation shows that `MATLAB` defines a 'vector' to be any 1-column matrix and any 1-row matrix. Fortunately, there is no problem with this, since, by (7.12), the norm of the *vector* $\mathbf{v}$ equals the norm of the *matrices* $[\mathbf{v}]$ and $\mathbf{v}^c$.

The best explicit expression available for $\|A\|$ for an arbitrary $A \in \mathbb{F}^{m\times n}$ is the following:

$$\|A\| = \sigma_1(A) = \sqrt{\rho(A^cA)}. \tag{7.16}$$

The first equality is (8.16), with $\sigma_1(A)$, by definition, the largest 'singular value' of A, and, by Problem 10.7, $\sigma_1(A)^2$ is the (absolutely) largest 'eigenvalue' of A^cA, hence the 'spectral radius' $\rho(A^cA)$ of A^cA, i.e., the smallest possible radius of a disk centered at the origin that contains all the 'eigenvalues' of A^cA. In general, one can only compute approximations to this number.

7.5 Prove that the Frobenius norm $\|A\|_F$ of $A \in \mathbb{F}^{m\times n}$ (see (6.32)) is an upper bound for its map norm $\|A\|$ associated with the Euclidean norm. (Hint: Problem 6.30.)

For this reason (and others), other vector norms are in common use, among them the **max-norm**

$$\forall \mathbf{x} \in \mathbb{F}^n, \quad \|\mathbf{x}\|_\infty := \max_j |x_j|,$$

for which the associated map norm is easily computable. It is

$$\|A\|_\infty := \max_{\mathbf{x}\neq 0} \|A\mathbf{x}\|_\infty/\|\mathbf{x}\|_\infty = \max_i \sum_j |A_{ij}| = \max_i \|A_{i:}\|_1, \tag{7.17}$$

with

$$\|\mathbf{v}\|_1 := \sum_j |v_j| \tag{7.18}$$

yet another vector norm, the socalled **1-norm**. The map norm associated with the 1-norm is also easily computable. It is (see Problem 7.7)

$$\begin{aligned}\|A\|_1 := \max_{\mathbf{x}\neq 0} \|A\mathbf{x}\|_1/\|\mathbf{x}\|_1 &= \max_j \sum_i |A_{ij}| \\ &= \max_j \|A_{:j}\|_1 = \|A^t\|_\infty = \|A^c\|_\infty.\end{aligned} \tag{7.19}$$

In this connection, the Euclidean norm is also known as the **2-norm**, since

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}^{\mathrm{c}}\mathbf{x}} = \sqrt{\sum_j |x_j|^2} =: \|\mathbf{x}\|_2.$$

Therefore, when it is important, one writes the corresponding map-norm with a subscript 2, too. For example, compare (7.19) with

$$\|A\| = \|A\|_2 = \|A^{\mathrm{c}}\|_2 = \|A^{\mathrm{t}}\|_2. \tag{7.20}$$

For the proof of these identities, recall from (6.27) that

$$\|\mathbf{x}\|_2 = \max_{\mathbf{y}\neq 0} |\langle \mathbf{x}, \mathbf{y}\rangle| / \|\mathbf{y}\|_2, \qquad \mathbf{x} \in \mathbb{F}^n. \tag{7.21}$$

Hence,
(7.22)

$$\begin{aligned}\|A\|_2 = \max_{\mathbf{x}\neq 0} \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2} &= \max_{\mathbf{x}\neq 0}\max_{\mathbf{y}\neq 0} \frac{|\langle A\mathbf{x}, \mathbf{y}\rangle|}{\|\mathbf{x}\|_2\|\mathbf{y}\|_2} \\ &= \max_{\mathbf{y}\neq 0}\max_{\mathbf{x}\neq 0} \frac{|\langle \mathbf{x}, A^{\mathrm{c}}\mathbf{y}\rangle|}{\|\mathbf{x}\|_2\|\mathbf{y}\|_2} = \max_{\mathbf{y}\neq 0} \frac{\|A^{\mathrm{c}}\mathbf{y}\|_2}{\|\mathbf{y}\|_2} = \|A^{\mathrm{c}}\|_2.\end{aligned}$$

The equality $\|A^{\mathrm{t}}\| = \|A^{\mathrm{c}}\|$ holds in any of the map-norms discussed since they all depend only on the absolute values of the entries of the matrix A.

The `MATLAB` statement `norm(A,inf)` provides the norm $\|\mathtt{A}\|_\infty$ in case `A` is a 'matrix', i.e., not a 'vector'. If `A` happens to equal $[\mathbf{v}]$ or $[\mathbf{v}]^{\mathrm{t}}$ for some vector $\mathbf{v}$, then `norm(A,inf)` returns the max-norm of that vector, i.e., the number $\|\mathbf{v}\|_\infty$. By (7.17), this is ok if `A` $= [\mathbf{v}]$, but gives, in general, the wrong result if `A` $= \mathbf{v}^{\mathrm{t}}$. This is an additional reason for sticking with the rule of using only $(n, 1)$-matrices for representing n-vectors in `MATLAB`.

The 1-norm, $\|\mathtt{A}\|_1$, is supplied by the statement `norm(A,1)`.

All three (vector-)norms mentioned so far are, indeed, norms in the sense of the following definition.

(7.23) Definition: The map $\|\ \| : X \to \mathbb{R} : x \mapsto \|x\|$ is a **vector norm** provided it is

(i) **positive definite**, i.e., $\forall x \in X$, $\|x\| \geq 0$ with equality if and only if $x = 0$;

(ii) **absolutely homogeneous**, i.e., $\forall(\alpha, x) \in \mathbb{F} \times X$, $\|\alpha x\| = |\alpha|\|x\|$;

(iii) **subadditive**, i.e., $\forall x, y \in X$, $\|x + y\| \leq \|x\| + \|y\|$.

This last inequality is called the **triangle inequality**, and the vector space X supplied with a vector norm is called a **normed vector space**.

The absolute value is a vector norm for the vector space $\mathbb{F} = \mathbb{F}^1$. From this, it is immediate that both the max-norm and the 1-norm are vector norms for $\mathbb{F}^n$. As to the norm $x \mapsto \sqrt{x^c x}$ on an inner product space and, in particular, the Euclidean or 2-norm on $\mathbb{F}^n$, only the triangle inequality might still be in doubt, but it is an immediate consequence of (6.27)Cauchy's Inequality, which gives that

$$\langle x, y\rangle + \langle y, x\rangle = 2\,\mathrm{Re}\langle x, y\rangle \le 2|\langle x, y\rangle| \le 2\|x\|\|y\|,$$

and therefore:

$$\|x + y\|^2 = \|x\|^2 + \langle x, y\rangle + \langle y, x\rangle + \|y\|^2 \le (\|x\| + \|y\|)^2.$$

Also, for X finite-dimensional, and both X and Y normed vector spaces, with norms $\|\,\|_X$ and $\|\,\|_Y$ respectively, the vector space $L(X, Y)$ is a normed vector space with respect to the corresponding map norm

$$\text{(7.24)} \qquad \|A\| := \|A\|_{X,Y} := \max_{x\ne 0} \frac{\|Ax\|_Y}{\|x\|_X}.$$

All statements about the map norm $\|A\| = \|A\|_2$ made in the preceding section hold for any of the map norms $\|A\|_{X,Y}$ since their proofs there use only the fact that $x \mapsto \sqrt{x^c x}$ is a norm according to (7.23)Definition. In particular, we will feel free to consider

$$\kappa(A)_p := \|A\|_p \|A^{-1}\|_p, \quad p = 1, 2, \infty, \quad A \in \mathbb{F}^{n\times n}.$$

Why all these different norms? Each norm associates with a vector just one number, and, as with bases, any particular situation may best be handled by a particular norm.

For example, in considering the condition of the power basis $V := [()^{j-1} : j = 1{:}k]$ for $\Pi_{<k}$, we might be more interested in measuring the size of the residual $p - V\widehat{\mathbf{a}}$ in terms of the max-norm

$$\|f\|_{[c..d]} := \max\{|f(t)| : c \le t \le d\}$$

over the interval $[c\,..\,d]$ of interest, rather than in the averaging way supplied by the corresponding 2-norm

$$\left(\int_c^d |f(t)|^2 \,\mathrm{d}t\right)^{1/2}.$$

In any case, any two norms on a finite-dimensional vector space are equivalent in the following sense.

(7.25) Proposition: For any two norms, $\|\,\|'$ and $\|\,\|''$, on a finite-dimensional vector space X, there exists a positive constant c so that

$$\forall x \in X, \quad \|x\|'' \le c\|x\|'.$$

This is just the statement that the map norm

$$\|\operatorname{id}_X\| := \max_{x\ne 0} \|x\|''/\|x\|'$$

is finite.

For example, for any $\mathbf{x} \in \mathbb{F}^n$,

(7.26)
$$\|\mathbf{x}\|_1 \le \sqrt{n}\|\mathbf{x}\|_2, \quad \text{and} \quad \|\mathbf{x}\|_2 \le \sqrt{n}\|\mathbf{x}\|_\infty, \quad \text{while} \quad \|\mathbf{x}\|_1 \ge \|\mathbf{x}\|_2 \ge \|\mathbf{x}\|_\infty.$$

Finally, given that it is very easy to compute the max-norm $\|A\|_\infty$ of $A \in \mathbb{F}^{m\times n}$ and much harder to compute the 2-norm $\|A\| = \|A\|_2$, why does one bother at all with the 2-norm? One very important reason is the availability of a large variety of *isometries*, i.e., matrices A with

$$\forall \mathbf{x}, \mathbf{y}, \quad \|A\mathbf{x} - A\mathbf{y}\| = \|\mathbf{x} - \mathbf{y}\|.$$

Each of these provides an o.n. basis for its range, and, by (6.22)Proposition, each finite-dimensional linear subspace of an inner product space has o.n. bases.

In contrast, the only $A \in \mathbb{F}^{n\times n}$ that are isometries in the max-norm, i.e., satisfy

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{F}^n, \quad \|A\mathbf{x} - A\mathbf{y}\|_\infty = \|\mathbf{x} - \mathbf{y}\|_\infty,$$

are of the form

$$\operatorname{diag}(\varepsilon_1, \ldots, \varepsilon_n)P,$$

with P a permutation matrix and each ε_j a scalar of absolute value 1.

For this reason, we continue to rely on the 2-norm. In fact, any norm without a subscript or other adornment is meant to be the 2-norm (or, more generally, the norm in the relevant inner product space).

7.6 Prove that *a linear map A on the normed vector space X is an isometry iff it is* **norm-preserving**, *i.e., $\|Ax\| = \|x\|$ for all $x \in X$.*

7.7* Prove that, for any $A \in \mathbb{F}^{m\times n}$, (1) $\max_{\mathbf{x}\ne\mathbf{0}} \|A\mathbf{x}\|_\infty/\|\mathbf{x}\|_\infty = \max_i \|A_{i\,:}\|_1 =: \|A(i_{\max}, :)\|_1$, hence $\|A\mathbf{x}\|_\infty = \|A\|_\infty\|\mathbf{x}\|_\infty$ for $\mathbf{x} := \overline{(\operatorname{signum}(A(i_{\max}, j))} : j = 1{:}n)$, and (2) $\max_{\mathbf{x}\ne\mathbf{0}} \|A\mathbf{x}\|_1/\|\mathbf{x}\|_1 = \max_j \|A_{:\,j}\|_1 =: \|A(:, j_{\max})\|_1$, with $\|A\mathbf{x}\|_1 = \|A\|_1\|\mathbf{x}\|_1$ for $\mathbf{x} = \mathbf{e}_{j_{\max}}$.

7.8 Prove that, for any $\alpha \in \mathbb{F}$, the linear map $M_\alpha : X \to X : x \mapsto \alpha x$ on the normed vector space $X \neq \{0\}$ has map norm $|\alpha|$.

7.9 Prove that, for any diagonal matrix $D \in \mathbb{F}^{m\times n}$ and for $p = 1, 2, \infty$, $\|D\|_p = \max_j |D_{jj}|$.

7.10 Consider the vector space $C(K)$, of all continuous real-valued functions on some compact subset K of $\mathbb{R}^n$. Show that $C(K)$ is a normed vector space with respect to the norm $\|f\| = \|f\|_K := \max_{t\in K} |f(t)|$.

Any linear map close enough to an invertible linear map is invertible

As a simple but important use of map norms, we now investigate their use in proving the invertibility of a linear map by showing that it is close enough to a linear map known to be invertible.

Assume that $A \in L(X, Y)$ is invertible, with X and Y finite-dimensional normed vector spaces. Then, by (3.21)Proposition, $\dim X = \dim Y$, hence the invertibility of $B \in L(X, Y)$ is established once B is shown to be 1-1. For this, observe that

$$(7.27) \qquad \|Bx\| \geq \|Ax\| - \|Bx - Ax\| \geq (1/\|A^{-1}\| - \|B - A\|)\|x\|,$$

using the fact that

$$(7.28) \qquad \min_x \frac{\|Ax\|}{\|x\|} = 1/\|A^{-1}\|$$

(see, e.g., (7.14)Proposition). Hence, if $1/\|A^{-1}\| > \|B - A\|$, then $Bx = 0$ implies, with (7.27), that $x = 0$, i.e., B is 1-1, hence invertible and, in that case (using (7.28) for B instead of A), (7.27) implies that $1/\|B^{-1}\| \geq 1/\|A^{-1}\| - \|B - A\|$. This proves the following useful

(7.29) Proposition. *If $A \in L(X, Y)$ is invertible, with X, hence Y, finite-dimensional normed vector spaces, then any $B \in L(X, Y)$ with $\|B - A\| < \|A^{-1}\|^{-1}$ is invertible, and*

$$\|B^{-1}\| \leq \|A^{-1}\|/(1 - \|B - A\|\,\|A^{-1}\|).$$

7.11* Let $(0\,..\,1] \to L(X, Y) : t \mapsto B_t$ be a given map into the vector space $L(X, Y)$ normed with the map norm derived from the norms of the finite-dimensional normed vector spaces X and Y. Prove: *If $A := \lim_{t\to 0} B_t$ exists (i.e., for some $A \in L(X, Y)$, $\lim_{t\to 0} \|A - B_t\| = 0$) and is invertible, then, for all sufficiently small t, B_t is invertible, and $\lim_{t\to 0}(B_t)^{-1} = A^{-1}$.* (Hint: $B^{-1} - A^{-1} = A^{-1}(A - B)B^{-1}$)

Bounding the interpolation error: Lebesgue's inequality

Recall from (5.9)Proposition the setup of interpolation: We are given the column map V into the vector space Y and a corresponding row map Λ^{t}

on Y so that their Gramian, $\Lambda^{\rm t}V$, is invertible, therefore V is 1-1 and $\Lambda^{\rm t}$ is onto, and $P := V(\Lambda^{\rm t}V)^{-1}\Lambda^{\rm t}$ is the linear projector with $\operatorname{ran} P = \operatorname{ran} V$ and $\operatorname{null} P = \operatorname{null} \Lambda^{\rm t}$, and so, for each $y \in Y$, Py is the unique element of $\operatorname{ran} V$ that interpolates y in the sense that $\Lambda^{\rm t}(Py) = \Lambda^{\rm t}y$.

Now assume that Y is a normed vector space, with vector norm $\|\cdot\|$. We derive a useful bound on the norm of the **interpolation error**, $y - Py$, in terms of the map norm $\|P\| = \max_{y \in Y\backslash 0} \|Py\|/\|y\|$ of P. Since

$$y - Py = (\,\mathrm{id} - P)y = (\,\mathrm{id} - P)(y - x)$$

for all $x \in \operatorname{ran} P = \operatorname{ran} V$, we obtain

$$\|y - Py\| \le \|\,\mathrm{id} - P\|\|y - x\|$$

for all $x \in \operatorname{ran} P$, and this proves **Lebesgue's Inequality**

$$\|y - Py\| \le \|\,\mathrm{id} - P\| \operatorname{dist}(y, \operatorname{ran} P), \tag{7.30}$$

with

$$\operatorname{dist}(y, X) := \inf_{x \in X} \|y - x\|$$

the **distance** of y from the set X. Since the map norm is a norm, it satisfies the triangle inequality, hence

$$\|\,\mathrm{id} - P\| \le 1 + \|P\|.$$

This leads to the conclusion that such an interpolation scheme P is **near-optimal** in the sense that its interpolation error is close to the smallest possible error of approximation from $\operatorname{ran} P$ to the extent that $\|P\|$ is close to 1.

As a very simple example, consider polynomial interpolation to data at the endpoints of an interval $[a\,.\,.\,b]$, i.e.,

$$(Py)(t) = \frac{y(a)(b - t) + y(b)(t - a)}{b - a}, \quad t \in [a\,.\,.\,b],$$

with $Y = C([a\,.\,.\,b])$ with the max-norm $\|y\| = \|y\|_\infty = \max(|y([a\,.\,.\,b])|)$. Then

$$\|Py\|_\infty = \max(|y(a)|, |y(b)|) \le \|y\|_\infty,$$

therefore $\|P\| = 1$. It follows that, in this case, the interpolation error is at most twice the error achievable by any approximation to y by a straight line.

7.12 Prove that, *for a linear projector P, $\|P\| < 1$ implies that $P = 0$.*

7.13* Prove that the linear projector $P_{\boldsymbol{\tau}}$ of polynomial interpolation at the pairwise distinct points $\tau_1, \ldots, \tau_k$ in the interval $[a\,.\,.\,b]$ has, as a linear map on the normed vector space $Y = C([a\,.\,.\,b])$ with norm $\|\cdot\|_\infty$, the map norm $\|P_{\boldsymbol{\tau}}\| = \|L_{\boldsymbol{\tau}}\|_\infty$, with $L_{\boldsymbol{\tau}} := \sum_i |\ell_i|$ where the ℓ_i are the columns of the corresponding Lagrange basis for $\Pi_{<k}$ (see (5.7)). $L_{\boldsymbol{\tau}}$ is known as the **Lebesgue function** of this interpolation process.

8 Factorization and rank

The need for factoring linear maps

In order to compute with a linear map $A \in L(X, Y)$, we have to factor it through a coordinate space. This means that we have to write it as

$$A = V\Lambda^{\mathrm{t}}, \qquad \text{with } \ V \in L(\mathbb{F}^r, Y), \ \text{hence } \Lambda^{\mathrm{t}} \in L(X, \mathbb{F}^r)\ .$$

The following picture might be helpful:

$$\begin{tikzcd}
X \arrow[rr, "A"] \arrow[dr, "\Lambda^{\mathrm{t}}"'] & & Y \\
& \mathbb{F}^r \arrow[ur, "V"'] &
\end{tikzcd}$$

For example, recall how you apply the linear map D of differentiation to a polynomial $p \in \Pi_{\le k}$: First you get the polynomial coefficients of that polynomial, and then you write down Dp in terms of those coefficients.

To test my claim, carry out the following thought experiment: You know that there is exactly one polynomial p of degree $\le k$ that matches given ordinates at given $k+1$ distinct abscissae, i.e., that satisfies

$$p(\tau_i) = y_i, \qquad i = 0{:}k$$

for given data $(\tau_i, y_i), i = 0{:}k$. Now, try, e.g., to compute the first derivative of the polynomial p of degree ≤ 3 that satisfies $p(j) = (-1)^j$, $j = 1, 2, 3, 4$. Can you do it without factoring the linear map $D : \Pi_{<4} \to \Pi_{<4}$ through some coordinate space?

As another *example*, recall how we dealt with **coordinate maps**, i.e., the inverse of a basis. We saw that, even though a basis $V : \mathbb{F}^n \to \mathbb{F}^m$ for some linear subspace X of $\mathbb{F}^m$ is a concrete matrix, its inverse, V^{-1} is, offhand, just a formal expression. For actual work, we made use of any *matrix* $\Lambda^{\mathrm{t}} : \mathbb{F}^m \to \mathbb{F}^n$ that is 1-1 on X, thereby obtaining the *factorization*

$$V^{-1} = (\Lambda^{\mathrm{t}} V)^{-1} \Lambda^{\mathrm{t}}$$

in which $\Lambda^{\mathrm{t}} V$ is a square matrix, hence $(\Lambda^{\mathrm{t}} V)^{-1}$ is also a matrix.

The smaller one can make $\#V$ in a factorization $A = V\Lambda^{\mathrm{t}}$ of $A \in L(X, Y)$, the cheaper is the calculation of A.

(8.1) Definition: The smallest r for which $A \in L(X, Y)$ can be factored as $A = V\Lambda^{\mathrm{t}}$ with $V \in L(\mathbb{F}^r, Y)$ (hence $\Lambda^{\mathrm{t}} \in L(X, \mathbb{F}^r)$) is called the **rank** of A. This is written

$$r = \operatorname{rank} A.$$

Any factorization $A = V\Lambda^{\mathrm{t}}$ with $\#V = \operatorname{rank} A$ is called **minimal**.

As an *example*,

$$A := \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} [\,1 \quad 2 \quad 3 \quad 4 \quad 5\,],$$

hence this A has rank 1 (since we can write it as $A = V\Lambda^{\mathrm{t}}$ with $\operatorname{dom} V = \mathbb{F}^1$, but we couldn't do it with $\operatorname{dom} V = \mathbb{F}^0$). To calculate $A\mathbf{x}$, we merely need to calculate the number $\alpha := (1, 2, 3, 4, 5)^{\mathrm{t}}\mathbf{x}$, and then obtain $A\mathbf{x}$ as the particular scalar multiple $\mathbf{y}\alpha$ of the vector $\mathbf{y} := (1, 1, 1, 1)$. That is much cheaper than computing the matrix product of the 4×5-matrix A with the 1-column matrix $[\mathbf{x}]$.

As the example illustrates, any matrix

$$A := [\mathbf{v}][\mathbf{w}]^{\mathrm{t}} = \mathbf{v}\mathbf{w}^{\mathrm{t}}$$

with $\mathbf{v} \in \mathbb{F}^m$ and $\mathbf{w} \in \mathbb{F}^n$ has rank 1 unless it is trivial, i.e., unless either $\mathbf{v}$ or $\mathbf{w}$ is the zero vector. This explains why an *elementary* matrix $\mathrm{id} + \mathbf{v}\mathbf{w}^{\mathrm{t}}$ is also called a **rank-one perturbation of the identity**.

The only linear map of rank 0 is the zero map. If A is not the zero map, then its range contains some nonzero vector, hence so must the range of any V for which $A = V\Lambda^{\mathrm{t}}$ with $\operatorname{dom} V = \mathbb{F}^r$, therefore such r must be > 0.

As another *example*, for any vector space X,

$$\dim X = \operatorname{rank} \mathrm{id}_X.$$

Indeed, if $n = \dim X$, then, for any basis $V \in L(\mathbb{F}^n, X)$ for X, $\mathrm{id}_X = VV^{-1}$, therefore $\operatorname{rank} \mathrm{id}_X \le n$, while, for any factorization $\mathrm{id}_X = V\Lambda^{\mathrm{t}}$ for some $V \in L(\mathbb{F}^r, X)$, V must necessarily be onto, hence $\dim X \le r$, by (3.9)Corollary, and therefore $\dim X \le \operatorname{rank} \mathrm{id}_X$. In fact, it is possible to make the rank concept the primary one and *define* $\dim X$ as the rank of id_X.

When A is an $m \times n$-matrix, then, trivially, $A = A\,\mathrm{id}_n = \mathrm{id}_m A$, hence $\operatorname{rank} A \le \min\{m, n\}$.

At times, particularly when A is a matrix, it is convenient to write the factorization $A = V\Lambda^{\mathrm{t}}$ more explicitly as

$$A =: [v_1, v_2, \ldots, v_r][\lambda_1, \lambda_2, \ldots, \lambda_r]^{\mathrm{t}} = \sum_{j=1}^{r} [v_j]\lambda_j. \tag{8.2}$$

Since each of the maps

$$v_j\lambda_j := [v_j]\lambda_j = [v_j] \circ \lambda_j : x \mapsto (\lambda_j x)v_j$$

has rank ≤ 1, this shows that *the rank of A gives the smallest number of terms necessary to write A as a sum of rank-one maps.*

(8.3) Proposition: $A = V\Lambda^{\mathrm{t}}$ is minimal if and only if V is a basis for $\operatorname{ran} A$. In particular,

$$\operatorname{rank} A = \dim \operatorname{ran} A.$$

Proof: Let $A = V\Lambda^{\mathrm{t}}$. Then $\operatorname{ran} A \subset \operatorname{ran} V$, hence

$$\dim \operatorname{ran} A \le \dim \operatorname{ran} V \le \#V,$$

with equality in the first $\le$ iff $\operatorname{ran} A = \operatorname{ran} V$ (by (3.18)Proposition), and in the second $\le$ iff V is 1-1. Thus, $\dim \operatorname{ran} A \le \#V$, with equality iff V is a basis for $\operatorname{ran} A$. □

On can prove in a similar way that $A = V\Lambda^{\mathrm{t}}$ *is minimal if and only if* Λ^{t} *is onto and* $\operatorname{null} \Lambda^{\mathrm{t}} = \operatorname{null} A$; see Problem 8.5.

(8.4) Corollary: The factorization $A = A(:, \texttt{bound})\operatorname{rrref}(A)$ provided by elimination (see (4.18)) is minimal.

(8.5) Corollary: If $A = V\Lambda^{\rm t}$ is minimal and A is invertible, then also V and $\Lambda^{\rm t}$ are invertible.

Proof: By (8.3)Proposition, $V \in L(\mathbb{F}^r, Y)$ is a basis for $\operatorname{ran} A$, while $\operatorname{ran} A = Y$ since A is invertible. Hence, V is invertible. Therefore, also $\Lambda^{\rm t} = V^{-1}A$ is invertible. □

But note that the matrix $[\,1\,] = [\,1 \quad 0\,]\begin{bmatrix}1\\0\end{bmatrix}$ is invertible, even though neither of its two factors is.

8.1 Determine a minimal factorization for the matrix

$$A := \begin{bmatrix} 1 & 2 & 0 & 3 & 4\\ 2 & 4 & 0 & 6 & 8\\ 1 & 1 & 0 & 1 & 1\\ 8 & 7 & 0 & 6 & 5\end{bmatrix}.$$

8.2 With A the matrix of the previous problem, give a basis for $\operatorname{ran} A$ and a basis for $\operatorname{ran} A^{\rm t}$.

8.3 Give an example of a pair of matrices, A and B, of order 4, each of rank 2, yet $\operatorname{ran} A \cap \operatorname{ran} B = \{0\}$.

8.4 Prove: *For any two linear maps A and B for which AB is defined,* $\operatorname{rank}(AB) \le \min\{\operatorname{rank} A, \operatorname{rank} B\}$. (Hint: If $A = V_A\Lambda_A{}^{\rm t}$ and $B = V_B\Lambda_B{}^{\rm t}$, then $AB = V_A(\Lambda_A{}^{\rm t}V_B\Lambda_B{}^{\rm t})$ $= (V_A\Lambda_A{}^{\rm t}V_B)\Lambda_B{}^{\rm t}$. Totally different hint: Use the (3.23)Dimension Formula together with the fact that $\operatorname{rank} C = \dim \operatorname{ran} C$.)

8.5* Prove that the following three statements are equivalent: (a) $A = V\Lambda^{\rm t}$ is minimal; (b) V is 1-1 and $\operatorname{ran} V = \operatorname{ran} A$; (c) $\Lambda^{\rm t}$ is onto and $\operatorname{null}\Lambda^{\rm t} = \operatorname{null} A$.

8.6 Prove: *If $A = V\Lambda^{\rm t}$ is a minimal factorization and A is a projector (i.e., $A^2 = A$), then* $\Lambda^{\rm t}V = \operatorname{id}$. (Hint: Problem 1.28.)

The trace of a linear map

Each $A \in L(X)$ can be factored in possibly many different ways as

$$A = V\Lambda^{\rm t} = [v_1, \ldots, v_n][\lambda_1, \ldots, \lambda_n]^{\rm t}$$

for some n (necessarily $\ge \operatorname{rank} A$). It may therefore be surprising that, nevertheless, *the number*

$$\sum_j \lambda_j v_j \;=\; \operatorname{trace}(\Lambda^{\rm t}V)$$

only depends on A. For the proof, let W be a basis for X. Then

$$\widehat{A} := W^{-1}AW = W^{-1}V\Lambda^{\rm t}W,$$

while

$$\Lambda^{\rm t}WW^{-1}V = \Lambda^{\rm t}V.$$

Hence, by (6.31),

$$\operatorname{trace}(\widehat{A}) = \operatorname{trace}(W^{-1}V\Lambda^{\rm t}W) = \operatorname{trace}(\Lambda^{\rm t}WW^{-1}V) = \operatorname{trace}(\Lambda^{\rm t}V).$$

By holding our factorization $A = V\Lambda^{\rm t}$ fixed, this implies that $\operatorname{trace}(\widehat{A})$ does not depend on the particular basis W for X we happen to use here, hence only depends on the linear map A. With that, holding now the linear map A and this basis W, hence the matrix $\widehat{A}$, fixed, we see that also $\operatorname{trace}(\Lambda^{\rm t}V)$ does not depend on the particular factorization $A = V\Lambda^{\rm t}$ we picked, but only depends on A. This number is called the **trace of** A, written

$$\operatorname{trace}(A).$$

The problems provide the basic properties of the trace of a linear map.

8.7 $\operatorname{trace}(\operatorname{id}_X) = \dim X$.

8.8 If $P \in L(X)$ is a projector (i.e., $P^2 = P$), then $\operatorname{trace}(P) = \dim \operatorname{ran} P$.

8.9 $A \mapsto \operatorname{trace}(A)$ is the unique scalar-valued linear map on $L(X)$ for which $\operatorname{trace}([x]\lambda) = \lambda x$ for all $x \in X$ and $\lambda \in X'$.

8.10 If $A \in L(X, Y)$ and $B \in L(Y, X)$, then (both AB and BA are defined and) $\operatorname{trace}(AB) = \operatorname{trace}(BA)$.

8.11 Prove that, for column maps V, W into X, and row maps $\Lambda^{\rm t}$, $\mathrm{M}^{\rm t}$ from X, $V\Lambda^{\rm t} = W\mathrm{M}^{\rm t}$ implies that $\operatorname{trace}(\Lambda^{\rm t}V) = \operatorname{trace}(\mathrm{M}^{\rm t}W)$ even if X is not finite-dimensional.

The rank of a matrix and of its (conjugate) transpose

In this section, let A' denote either the transpose or the conjugate transpose of the matrix A. Then, either way, $A = VW'$ iff $A' = WV'$. This trivial observation implies all kinds of things about the relationship between a matrix and its (conjugate) transpose.

As a starter, it says that $A = VW'$ is minimal if and only if $A' = WV'$ is minimal. Therefore:

(8.6) Proposition: $\operatorname{rank} A = \operatorname{rank} A^{\rm c} = \operatorname{rank} A^{\rm t}$.

(8.7) Corollary: If A is a matrix, then $\dim \operatorname{ran} A = \dim \operatorname{ran} A^{\rm c} = \dim \operatorname{ran} A^{\rm t}$.

(8.8) Corollary: For any matrix A, A' is 1-1 (onto) if and only if A is onto (1-1).

Proof: If $A \in \mathbb{F}^{m\times n}$, then A is onto iff $\operatorname{rank} A = m$ iff $\operatorname{rank} A' = m$ iff the natural factorization $A' = A' \operatorname{id}_m$ is minimal, i.e., iff A' is 1-1.

The other equivalence follows from this since $(A')' = A$. □

For a different proof of these results, see the comments that follow (6.29)Corollary and (6.30)Corollary.

Elimination as factorization

The description (4.2) of elimination does not rely on any particular ordering of the rows of the given $(m\times n)$-matrix A. At any stage, it only distinguishes between pivot rows and those rows not yet used as pivot rows. We may therefore imagine that we initially place the rows of A into the work-array B in exactly the order in which they are going to be used as pivot rows, followed, in any order whatsoever, by those rows (if any) that are never going to be used as pivot rows.

In terms of the n-vector `p` provided by the (4.2)Elimination Algorithm, this means that we start with $B = A($`q`$, :)$, with `q` obtained from `p` by

```
q = p(find(p>0)); 1:m; ans(q) = []; q = [q, ans];
```

Indeed, to recall, `p(j)` is positive if and only if the `j`th unknown is bound, in which case row `p(j)` is the pivot row for that unknown. Thus the assignment `q = p(find(p>0))` initializes `q` so that `A(q,\all)` contains the pivot rows in order of their use. With that, `1:m; ans(q) = [];` leaves, in `ans`, the indices of all rows not used as pivot rows.

Note that `q` is a permutation of order m. Hence $B = QA$, with Q the corresponding permutation matrix, meaning the matrix $Q =$ `I(q,\all)` obtained from the identity matrix `I` := `eye(m)` by the very same reordering.

We prefer to write this as $A = PB$, with P the inverse of Q, hence obtainable from `q` by

```
P = eye(m); P(q,\all) = P;
```

With that done, we have, at the beginning of the algorithm,

$$B = P^{-1}A$$

for some permutation matrix P, and all the work in the algorithm consists of repeatedly subtracting some multiple α of some row h of B from some *later* row, i.e., some row i with $i > h$. In terms of matrices, this means the repeated replacement

$$B \leftarrow E_{\mathbf{e}_i,\mathbf{e}_h}(-\alpha)B$$

with $i > h$. Since, by (2.34), $E_{\mathbf{e}_i,\mathbf{e}_h}(-\alpha)^{-1} = E_{\mathbf{e}_i,\mathbf{e}_h}(\alpha)$, this implies that

$$A = PLU,$$

with L the product of all those elementary matrices $E_{\mathbf{e}_i,\mathbf{e}_h}(\alpha)$ (in the appropriate order), and U the final state of the work-array B. Specifically, U is in row-echelon form (as defined in (4.9)); in particular, U is upper triangular.

Each $E_{\mathbf{e}_i,\mathbf{e}_h}(\alpha)$ is **unit lower triangular**, i.e., of the form $\mathrm{id} + N$ with N **strictly lower triangular**, i.e.,

$$N_{rs} \neq 0 \quad \Longrightarrow \quad r > s.$$

For, because of the initial ordering of the rows in B, only $E_{\mathbf{e}_i,\mathbf{e}_h}(\alpha)$ with $i > h$ appear. This implies that L, as the product of unit lower triangular matrices, is itself unit lower triangular.

If we apply the elimination algorithm to the matrix $[A, C]$, with $A \in \mathbb{F}^{m\times m}$ invertible, then the first m columns are bound, hence the remaining columns are free. In particular, both P and L in the resulting factorization depend only on A and not at all on C.

In particular, in solving $A? = \mathbf{y}$, there is no need to subject all of $[A, \mathbf{y}]$ to the elimination algorithm. If elimination just applied to A gives the factorization

$$A = PLU \tag{8.9}$$

for an invertible A, then we can find the unique solution $\mathbf{x}$ to the equation $A? = \mathbf{y}$ by the two-step process:

$$\begin{aligned} \mathbf{c} &\leftarrow L^{-1}P^{-1}\mathbf{y} \\ \mathbf{x} &\leftarrow U^{-1}\mathbf{c} \end{aligned}$$

and these two steps are easily carried out. The first step amounts to subjecting the rows of the matrix $[\mathbf{y}]$ to all the row operations (including reordering) used during elimination applied to A. The second step is handled by the Backsubstitution Algorithm (4.6), with input $B = [U, \mathbf{c}]$, $\mathtt{p} = (1, 2, \ldots, m, 0)$, and $\mathbf{z} = (0, \ldots, 0, -1)$.

Once it is understood that the purpose of elimination for solving $A? = \mathbf{y}$ is the factorization of A into a product of "easily" invertible factors, then it is possible to seek factorizations that might serve the same goal in a better way. The best-known alternative is the QR factorization, in which one obtains

$$A = QR,$$

with R upper triangular and Q o.n., i.e., $Q^{\mathrm{c}}Q = \mathrm{id}$. Such a factorization is obtained by doing elimination a column at a time, usually with the aid of **Householder matrices**. These are elementary matrices of the form

$$H_{\mathbf{w}} := E_{\mathbf{w},\mathbf{w}}(-2/\mathbf{w}^{\mathrm{c}}\mathbf{w}) = \mathrm{id} - \frac{2}{\mathbf{w}^{\mathrm{c}}\mathbf{w}}\mathbf{w}\mathbf{w}^{\mathrm{c}},$$

and are easily seen to be **self-inverse** or **involutory** (i.e., $H_{\mathbf{w}}H_{\mathbf{w}} = \mathrm{id}$), **hermitian** (i.e., $H_{\mathbf{w}}{}^{\mathrm{c}} = H_{\mathbf{w}}$), hence **unitary** (i.e., $H_{\mathbf{w}}{}^{\mathrm{c}}H_{\mathbf{w}} = \mathrm{id} = H_{\mathbf{w}}H_{\mathbf{w}}{}^{\mathrm{c}}$).

While the computational cost of constructing the QR factorization is roughly double that needed for the PLU factorization, the QR factorization has the advantage of being more impervious to the effects of rounding errors. Precisely, the relative rounding error effects in the derivation of the triangular linear system $Q^{-1}A? = Q^{-1}\mathbf{y}$ from the original linear system $A? = \mathbf{y}$ are, in general, much smaller because Q, hence Q^{-1}, is an isometry, than they are in the derivation of the linear system $PL^{-1}A? = PL^{-1}\mathbf{y}$, and, while the condition of the upper triangular matrix $Q^{-1}A$ is that of A, the condition of $PL^{-1}A$ is likely larger than that.

8.12 Prove: *If $L_1D_1U_1 = A = L_2D_2U_2$, with L_i unit lower triangular, D_i invertible diagonal, and U_i unit upper triangular matrices, then $L_1 = L_2$, $D_1 = D_2$, and $U_1 = U_2$.*

SVD

Let $A = VW^{\mathrm{c}}$ be a minimal factorization for the $m \times n$-matrix A of rank r. Then $A^{\mathrm{c}} = WV^{\mathrm{c}}$ is a minimal factorization for A^{c}. By (8.3), this implies that V is a basis for $\operatorname{ran} A$ and W is a basis for $\operatorname{ran} A^{\mathrm{c}}$.

Can we choose both these bases to be o.n.?

Well, if both V and W are o.n., then, for any $\mathbf{x}$, $\|A\mathbf{x}\| = \|VW^{\mathrm{c}}\mathbf{x}\| = \|W^{\mathrm{c}}\mathbf{x}\|$, while, for $\mathbf{x} \in \operatorname{ran} A^{\mathrm{c}}$, $\mathbf{x} = WW^{\mathrm{c}}\mathbf{x}$, hence $\|\mathbf{x}\| = \|W^{\mathrm{c}}\mathbf{x}\|$. Therefore, altogether, in such a case, A is an isometry on $\operatorname{ran} A^{\mathrm{c}}$, a very special situation.

Nevertheless and, perhaps, surprisingly, there is an o.n. basis W for $\operatorname{ran} A^{\mathrm{c}}$ for which the columns of AW are *orthogonal*, i.e., $AW = V\Sigma$ with V o.n. and Σ diagonal, hence $A = V\Sigma W^{\mathrm{c}}$ with also V o.n.

(8.10) Theorem: For every $A \in \mathbb{F}^{m\times n}$, there exist o.n. bases V and W for $\operatorname{ran} A$ and $\operatorname{ran} A^{\mathrm{c}}$, respectively, and a diagonal matrix Σ with positive diagonal entries, so that

(8.11) $$A = V\Sigma W^{\mathrm{c}}.$$

Proof: For efficiency, the proof given here uses results, concerning the 'eigenstructure' of hermitian positive definite matrices, that are established only later in this book. This may help to motivate the study to come of such 'eigenstructure' of matrices.

For motivation of the proof, assume for the moment that $A = V\Sigma W^{\mathrm{c}}$ is a factorization of the kind we claim to exist. Then, with $\Sigma =: \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$, it follows that

$$A^{\mathrm{c}}A = W\Sigma^{\mathrm{c}}V^{\mathrm{c}}\; V\Sigma W^{\mathrm{c}} = W\Sigma^{\mathrm{c}}\Sigma W^{\mathrm{c}},$$

hence

(8.12) $$A^{\mathrm{c}}AW = W\mathrm{T}, \quad \text{with} \quad \mathrm{T} := \operatorname{diag}(\tau_1, \ldots, \tau_r)$$

and W o.n., and the $\tau_j = \overline{\sigma_j}\sigma_j = |\sigma_j|^2$ all positive.

Just such an o.n. $W \in \mathbb{F}^{n\times r}$ and positive scalars τ_j do exist by (12.2) Corollary and (14.2)Proposition, since the matrix $A^{\rm c}A$ is **hermitian** (i.e., $(A^{\rm c}A)^{\rm c} = A^{\rm c}A$) and **positive semidefinite** (i.e., $\langle A^{\rm c}A\mathbf{x}, \mathbf{x}\rangle \ge 0$ for all $\mathbf{x}$) and has rank r.

With W and the τ_j so chosen, it follows that W is an o.n. basis for $\operatorname{ran} A^{\rm c}$, since (8.12) implies that $\operatorname{ran} W \subset \operatorname{ran} A^{\rm c}$, and W is a 1-1 column map of order $r = \dim \operatorname{ran} A^{\rm c}$. Further, $U := AW$ satisfies $U^{\rm c}U = W^{\rm c}A^{\rm c}AW = W^{\rm c}W\mathrm{T} = \mathrm{T}$, hence

$$V := AW\Sigma^{-1}, \quad \text{with } \ \Sigma := \mathrm{T}^{1/2} := \operatorname{diag}(\sqrt{\tau_j} : j = 1{:}r),$$

is o.n., and so $V\Sigma W^{\rm c} = A$, because $WW^{\rm c} = P := P_{\operatorname{ran} A^{\rm c}}$, hence $\operatorname{ran}(\,\mathrm{id} - P) = \operatorname{null} P = \operatorname{ran} A^{{\rm c}\perp} = \operatorname{null} A$, and so $AWW^{\rm c} = AP = A(P + (\,\mathrm{id} - P)) = A$. □

It is customary to order the numbers

$$\sigma_j := \sqrt{\tau_j}, \quad j = 1{:}r.$$

Specifically, one assumes that

$$\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r.$$

These numbers σ_j are called the (nonzero) **singular value**s of A, and with this ordering, the factorization

$$A = \sum_{j=1}^{\operatorname{rank} A} \mathbf{v}_j \sigma_j {\mathbf{w}_j}^{\rm c}$$

is called a (**reduced**) **singular value decomposition** or **svd** for A.

Offhand, a svd is *not* unique. E.g., *any* o.n. basis V for $\mathbb{F}^n$ provides the svd $V\,\mathrm{id}_n V^{\rm c}$ for id_n.

Some prefer to have a factorization $A = \tilde V\tilde\Sigma\tilde W^{\rm c}$ in which both $\tilde V$ and $\tilde W$ are o.n. bases for all of $\mathbb{F}^m$ and $\mathbb{F}^n$, respectively (rather than just for $\operatorname{ran} A$ and $\operatorname{ran} A^{\rm c}$, respectively). This can always be achieved by extending V and W from (8.11) in any manner whatsoever to o.n. bases $\tilde V := [V, V_1]$ and $\tilde W := [W, W_1]$ and, correspondingly, extending Σ to

$$\tilde\Sigma := \operatorname{diag}(\Sigma, 0) = \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \ \in \ \mathbb{F}^{m\times n}$$

by the adjunction of blocks of 0 of appropriate size. With this, we have

$$(8.13) \qquad A = \tilde V\tilde\Sigma\tilde W^{\rm c} \ = \sum_{j=1}^{\min\{m,n\}} \mathbf{v}_j \sigma_j {\mathbf{w}_j}^{\rm c},$$

and the diagonal entries

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0 = \sigma_{r+1} = \cdots = \sigma_{\min\{m,n\}}$$

of $\tilde{\Sigma}$ are altogether referred to as the **singular value**s of A. Note that this sequence is still ordered. We will refer to (8.13) as a **Singular Value Decomposition** or **SVD**.

The `MATLAB` command `svd(A)` returns the SVD rather than the svd of `A` when issued in the form `[V,S,W] = svd(A)`. Specifically, `A` = `V*S*W'`, with `V` and `W` both unitary, of order m and n, respectively, if `A` is an $m \times n$-matrix. By itself, `svd(A)` returns, in a one-column matrix, the (ordered) sequence of singular values of `A`.

The Pseudo-inverse

Here is a first of many uses to which the svd has been put. It concerns the solution of the equation

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

in case A is not invertible (for whatever reason). In a previous chapter (see page 107), we looked in this case for a solution of the 'projected' problem

$$A? = P_{\operatorname{ran} A}\,\mathbf{y} =: \widehat{\mathbf{y}} \tag{8.14}$$

for the simple reason that any solution $\mathbf{x}$ of this equation makes the **residual** $\|A\mathbf{x} - \mathbf{y}\|_2$ as small as it can be made by any x. For this reason, any solution of (8.14) is called a **least-squares solution** for $A? = \mathbf{y}$.

If now A is 1-1, then (8.14) has exactly one solution. The question is what to do in the contrary case. One proposal is to get the **best least-squares solution**, i.e., the least-squares solution of minimal norm. The svd for A makes it easy to find this particular solution.

If $A = V\Sigma W^{\mathrm{c}}$ is a svd for A, then V is an o.n. basis for $\operatorname{ran} A$, hence

$$\widehat{\mathbf{y}} = P_{\operatorname{ran} A}\,\mathbf{y} = VV^{\mathrm{c}}\mathbf{y}.$$

Therefore, (8.14) is equivalent to the equation

$$V\Sigma W^{\mathrm{c}}? = VV^{\mathrm{c}}\mathbf{y}.$$

Since V is o.n., hence 1-1, and Σ is invertible, this equation is, in turn, equivalent to

$$W^{\mathrm{c}}? = \Sigma^{-1}V^{\mathrm{c}}\mathbf{y},$$

hence, since also W is o.n., hence 1-1, to

$$WW^{\mathrm{c}}? = W\Sigma^{-1}V^{\mathrm{c}}\mathbf{y}. \tag{8.15}$$

Since W is o.n., $WW^{\mathrm{c}} = P_W$ is an o.n. projector, hence, by (6.18)Proposition, strictly reduces norms unless it is applied to something in its range. Since the right-hand side of (8.15) is in $\operatorname{ran} W$, it follows that the solution of smallest norm of (8.15), i.e., the best least-squares solution of $A? = \mathbf{y}$, is that right-hand side, i.e., the vector

$$\widehat{\mathbf{x}} := A^{+}\mathbf{y},$$

with the matrix

$$A^{+} := W\Sigma^{-1}V^{\mathrm{c}}$$

the **Moore-Penrose pseudo-inverse** of A.

Note that

$$A^{+}A = W\Sigma^{-1}V^{\mathrm{c}}V\Sigma W^{\mathrm{c}} = WW^{\mathrm{c}},$$

hence A^{+} is a left inverse for A in case W is square, i.e., in case $\operatorname{rank} A = \#A$. Similarly,

$$AA^{+} = V\Sigma W^{\mathrm{c}}W\Sigma^{-1}V^{\mathrm{c}} = VV^{\mathrm{c}},$$

hence A^{+} is a right inverse for A in case V is square, i.e., in case $\operatorname{rank} A = \#A^{\mathrm{c}}$. In any case,

$$A^{+}A = P_{\operatorname{ran} A^{\mathrm{c}}}, \qquad AA^{+} = P_{\operatorname{ran} A},$$

therefore, in particular,

$$AA^{+}A = A.$$

2-norm and 2-condition of a matrix

Recall from (6.26) that o.n. matrices are 2-norm-preserving, i.e.,

$$\|\mathbf{x}\|_2 = \|U\mathbf{x}\|_2, \qquad \mathbf{x} \in \mathbb{F}^n,\ \text{o.n. } U \in \mathbb{F}^{m\times n}.$$

This implies that

$$\|TB\|_2 = \|B\|_2 = \|BU^{\mathrm{c}}\|_2, \qquad \text{o.n. } T \in \mathbb{F}^{r\times m},\ B \in \mathbb{F}^{m\times n},\ \text{o.n. } U \in \mathbb{F}^{r\times n}.$$

Indeed,

$$\|TB\|_2 = \max_{\mathbf{x}\neq 0} \frac{\|TB\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \max_{\mathbf{x}\neq 0} \frac{\|B\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \|B\|_2.$$

By (7.22), this implies that also

$$\|BU^{\mathrm{c}}\|_2 = \|UB^{\mathrm{c}}\|_2 = \|B^{\mathrm{c}}\|_2 = \|B\|_2.$$

It follows that, with $A = V\Sigma W^{\mathrm{c}} \in \mathbb{F}^{m\times n}$ a svd for A,

$$\|A\|_2 = \|\Sigma\|_2 = \sigma_1, \tag{8.16}$$

the last equality because of the fact that $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$ with $\sigma_1 \ge \sigma_2 \ge \cdots \ge 0$.

Assume that, in addition, A is invertible, therefore $r = \operatorname{rank} A = n = m$, making also V and W square, hence A^{+} is both a left and a right inverse for A, therefore necessarily $A^{-1} = A^{+} = V\Sigma^{-1}W^{\mathrm{c}}$. It follows that $\|A^{-1}\|_2 = 1/\sigma_n$. Hence, the 2-condition of $A \in \mathbb{F}^{n\times n}$ is

$$\kappa_2(A) = \|A\|_2\|A^{-1}\|_2 = \sigma_1/\sigma_n,$$

and this is how this condition number is frequently *defined*.

The effective rank of a noisy matrix

The problem to be addressed here is the following. If we construct a matrix in the computer, we have to deal with the fact that the entries of the constructed matrix are not quite exact; rounding errors during the calculations may have added some noise. This is even true for a matrix merely entered into the computer, in case some of its entries cannot be represented exactly by the floating point arithmetic used (as is the case, e.g., for the number .1 or the number 1/3 in any of the standard binary-based floatingpoint arithmetics).

This makes it impossible to use, e.g., the rref algorithm to determine the rank of the underlying matrix. However, if one has some notion of the size of the noise involved, then one can use the svd to determine a sharp *lower* bound on the rank of the underlying matrix, because of the following.

(8.17) Proposition: If $A = V\Sigma W^{\text{c}}$ is a svd for A and $\operatorname{rank}(A) > k$, then $\min\{\|A - B\|_2 : \operatorname{rank}(B) \le k\} = \sigma_{k+1} = \|A - A_k\|_2$, with

$$A_k := \sum_{j=1}^{k} \mathbf{v}_j \sigma_j {\mathbf{w}_j}^{\text{c}}.$$

Proof: If $B \in \mathbb{F}^{m\times n}$ with $\operatorname{rank}(B) \le k$, then $\dim \operatorname{null}(B) > n - (k+1) = \dim \mathbb{F}^n - \dim \operatorname{ran} W_{k+1}$, with

$$W_{k+1} := [\mathbf{w}_1, \ldots, \mathbf{w}_{k+1}].$$

Therefore, by (3.30)Corollary, the intersection $\operatorname{null}(B) \cap \operatorname{ran} W_{k+1}$ contains a vector $\mathbf{z}$ of norm 1. Then $B\mathbf{z} = \mathbf{0}$, and $W^{\text{c}}\mathbf{z} = {W_{k+1}}^{\text{c}}\mathbf{z}$, and $\|{W_{k+1}}^{\text{c}}\mathbf{z}\|_2 = \|\mathbf{z}\|_2 = 1$. Therefore, $A\mathbf{z} = V\Sigma W^{\text{c}}\mathbf{z} = V_{k+1}\Sigma_{k+1}{W_{k+1}}^{\text{c}}\mathbf{z}$, hence

$$\begin{aligned}\|A - B\|_2 &\ge \|A\mathbf{z} - B\mathbf{z}\|_2 = \|A\mathbf{z}\|_2 = \|\Sigma_{k+1}{W_{k+1}}^{\text{c}}\mathbf{z}\|_2 \\ &\ge \sigma_{k+1}\|{W_{k+1}}^{\text{c}}\mathbf{z}\|_2 = \sigma_{k+1}.\end{aligned}$$

On the other hand, for the specific choice $B = A_k$, we get $\|A - A_k\|_2 = \sigma_{k+1}$ by (8.16), since $A - A_k = \sum_{j>k} \mathbf{v}_j\sigma_j{\mathbf{w}_j}^{\text{c}}$ is a svd for it, hence its largest singular value is σ_{k+1}. □

In particular, if we have in hand a svd

$$A + E = V \operatorname{diag}(\hat\sigma_1, \ldots, \hat\sigma_{\hat r})W^{\text{c}}$$

for the *perturbed* matrix $A + E$, then we know, by (8.17)Proposition, that any matrix of rank $\le k$ differs from $A + E$ by at least $\hat\sigma_{k+1}$; hence A cannot be of rank k unless $\hat\sigma_{k+1} \le \|E\|_2$. Since $\hat\sigma_1 \ge \hat\sigma_2 \ge \cdots$, there is a largest k

for which $\hat\sigma_k > \|E\|_2$ and that k is the smallest j for which $\hat\sigma_{j+1} \le \|E\|_2$. So, if we know (or believe) that $\|E\|_2 \le \varepsilon$, then the best we can say about the rank of A is that it must be at least

$$r_\varepsilon := \max\{j : \hat\sigma_j > \varepsilon\}.$$

For example, the matrix

$$A = \begin{bmatrix} 2/3 & 1 & 1/3 \\ 4/3 & 2 & 2/3 \\ 1 & 1 & 1 \end{bmatrix}$$

is readily transformed by elimination into the matrix

$$B = \begin{bmatrix} 0 & 1/3 & -1/3 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix},$$

hence has rank 2. However, on entering A into a computer correct to four decimal places after the decimal point, we get (more or less) the matrix

$$A_c = \begin{bmatrix} .6667 & 1 & .3333 \\ 1.3333 & 2 & .6667 \\ 1 & 1 & 1 \end{bmatrix},$$

and for it, `MATLAB` correctly returns id_3 as its `rref`. However, the singular values of A_c, as returned by `svd`, are

```
(3.2340..., 0.5645..., 0.000054...)
```

indicating that there is a rank-2 matrix B with $\|A_c - B\|_2 < .000055$. Since entries of A_c are only accurate to within 0.00005, the safe conclusion is that A has rank ≥ 2; it happens to have rank 2 in this particular example.

The polar decomposition

The svd can also be very helpful in establishing results of a more theoretical flavor, as the following discussion is intended to illustrate.

This discussion concerns a useful extension to square matrices of the polar form (see page 273)

$$z = |z| \exp(\mathrm{i}\varphi)$$

of a complex number z, i.e., a factorization of z into a nonnegative number $|z| = \sqrt{z\overline{z}}$ (its modulus or absolute value) and a number whose absolute value is equal to 1, a socalled **unimodular** number.

There is, for any $A \in \mathbb{C}^{n\times n}$, a corresponding decomposition

$$A = \sqrt{AA^{\mathrm{c}}}\,E, \tag{8.18}$$

called a **polar decomposition**, with $\sqrt{AA^c}$ 'nonnegative' in the sense that it is hermitian and positive semidefinite, and E 'unimodular' in the sense that it is unitary, hence norm-preserving, i.e., an isometry.

A polar decomposition is almost immediate, given that we already have a SVD $A = \tilde{V}\tilde{\Sigma}\tilde{W}^c$ for A (see (8.13)) in hand. Indeed, from that,

$$A = \tilde{V}\tilde{\Sigma}\tilde{V}^c \, \tilde{V}\tilde{W}^c,$$

with $P := \tilde{V}\tilde{\Sigma}\tilde{V}^c$ evidently hermitian, and also positive semidefinite since

$$\langle P\mathbf{x}, \mathbf{x}\rangle = \mathbf{x}^c\tilde{V}\tilde{\Sigma}\tilde{V}^c\mathbf{x} = \sum_j \tilde{\sigma}_j |(\tilde{V}^c\mathbf{x})_j|^2$$

is nonnegative for all $\mathbf{x}$, given that $\tilde{\sigma}_j \ge 0$ for all j; and

$$P^2 = \tilde{V}\tilde{\Sigma}\tilde{V}^c\tilde{V}\tilde{\Sigma}\tilde{V}^c = \tilde{V}\tilde{\Sigma}\tilde{\Sigma}^c\tilde{V}^c = \tilde{V}\tilde{\Sigma}\tilde{W}^c\tilde{W}\tilde{\Sigma}^c\tilde{V}^c = AA^c;$$

and, finally, $E := \tilde{V}\tilde{W}^c$ unitary as the product of unitary maps.

Equivalence and similarity

The SVD provides a particularly useful example of *equivalence.* The linear maps A and $\widehat{A}$ are called **equivalent** if there are *invertible* linear maps V and W so that

$$A \;=\; V\widehat{A}W^{-1}.$$

Since both V and W are invertible, such equivalent linear maps share all essential properties, such as their rank, being 1-1, or onto, or invertible.

Equivalence is particularly useful when the domains of V and W are coordinate spaces, i.e., when V and W are *bases*, and, correspondingly, $\widehat{A}$ is a matrix, as in the following diagram:

$$\begin{tikzcd}
X \arrow[r, "A"] & Y \\
\mathbb{F}^n \arrow[u, "W"] \arrow[r, "\widehat{A}"'] & \mathbb{F}^m \arrow[u, "V"']
\end{tikzcd}$$

In this situation, $\widehat{A} = V^{-1}AW$ is called a **matrix representation for** A.

For example, we noted earlier that the matrix

$$\widehat{D}_k := \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 2 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & k \end{bmatrix}$$

is the standard matrix representation used in Calculus for the linear map $D : \Pi_{\le k} \to \Pi_{<k}$ of differentiation of polynomials of degree $\le k$.

In practice, one looks, for given $A \in L(X, Y)$, for matrix representations $\widehat{A}$ that are as simple as possible. If that means a matrix with as many zero entries as possible and, moreover, all the nonzero entries the same, say equal to 1, then a simplest such matrix representation is of the form

$$\widehat{A} = \operatorname{diag}(\operatorname{id}_{\operatorname{rank} A}, 0) = \begin{bmatrix} \operatorname{id}_{\operatorname{rank} A} & 0 \\ 0 & 0 \end{bmatrix},$$

with 0 indicating zero matrices of the appropriate size to make $\widehat{A}$ of size $\dim \operatorname{tar} A \times \dim \operatorname{dom} A$.

The situation becomes much more interesting and challenging when $\operatorname{dom} A = \operatorname{tar} A$ and, correspondingly, we insist that also $V = W$. Linear maps A and $\widehat{A}$ for which there exists an invertible linear map V with

$$A = V\widehat{A}V^{-1}$$

are called *similar*. Such similarity will drive much of the rest of this book.

8.13 For the given linear maps $A, B, C : \mathbb{F}^{2\times 3}$, find their matrix representation with respect to the basis $V = [\mathbf{e}_1 + \mathbf{e}_2, \mathbf{e}_2 + \mathbf{e}_3, \mathbf{e}_3 + \mathbf{e}_1]$ for $\mathbb{F}^3$ and $W := \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}$ for $\mathbb{F}^2$: (a) $A\mathbf{x} = (5x_1 + 2x_2 + 7x_3, x_1 + 2x_2 - x_3)$; (b) $B\mathbf{x} = (x_1 + x_2 + x_3, x_2 - x_3)$; (c) $C\mathbf{x} = (-x_1 - x_2 - x_3, x_3)$.

8.14 What is the matrix representation of the linear map $\mathbb{C} \to \mathbb{C} : x \mapsto zx$ with respect to the basis $[1, \mathrm{i}]$ for $\mathbb{C}$ (as a vector space with $\mathbb{F} = \mathbb{R}$) and with $z =: a + \mathrm{i}b$ a given complex number?

8.15 T/F

(a) If A, B, M are matrices such that $\operatorname{rank} AM = \operatorname{rank} B$, then M is invertible.

(b) If M is invertible and $AM = B$, then $\operatorname{rank} A = \operatorname{rank} B$.

(c) If M is invertible and $MA = B$, then $\operatorname{rank} A = \operatorname{rank} B$.

Peter: Homework is too sparse and too simple for the complexity of this chapter

9 Duality

This short chapter can be skipped without loss of continuity. Much of it can serve as a review of what has been covered so far. It owes much to the intriguing book [GL].

Complementary mathematical concepts

Duality concerns mathematical concepts that come in pairs, that *complement* one another. Examples of interest in this book include:

- $\subset$ and $\supset$;
- A *subset* S of T and its **complement**, $\backslash S := T \backslash S$;
- $\cap$ and $\cup$;
- $\forall$ and $\exists$;
- *1-1* and *onto*;
- *right* and *left* inverse;
- *bound* and *free*;
- *nullspace* and *range* of a linear map;
- an *invertible map* and its *inverse*;
- *column map* and *row map*;
- *synthesis* and *analysis*;
- a *basis* and its *inverse*;
- *columns* and *rows* of a matrix;
- a *matrix* and its (conjugate) *transpose*;
- a linear *subspace* and one of its *complements*;
- dim and codim;
- the *vector space* X and its **dual**, $X' := L(X, \mathbb{F})$;
- the linear map $A \in L(X, Y)$ and its **dual**, $A' : Y' \to X' : \lambda \mapsto \lambda A$;
- a *norm* on the vector space X and the *dual norm* on X'.

Each such pair expresses a kind of *symmetry.* Such symmetry provides, with each result, also its 'dual', i.e., the result obtained by replacing one or more concepts appropriately by its complement. This leads to efficiency, both in the proving and in the remembering of results.

A classical example is that of *points* and *lines* in a geometry, and results concerning lines through points. E.g., *through every two distinct points there goes exactly one line*; its 'dual' statement is: *any two nonparallel lines have exactly one point in common.*

Another classical example is **DeMorgan's Law**, according to which any statement concerning the union, intersection and containment of subsets is true if and only if its 'dual' statement is true, i.e., the statement obtained by replacing each set by its complement and replacing $(\subset, \supset, \cap, \cup)$ by $(\supset, \subset, \cup, \cap)$, respectively. For example, the two 'distributive' laws

$$(R \cap S) \cup T = (R \cup T) \cap (S \cup T), \quad (R \cup S) \cap T = (R \cap T) \cup (S \cap T)$$

are 'dual' to each other. Again, having verified that *the intersection of a collection of sets is the largest set contained in all of them*, we have, by 'duality', also verified that *the union of a collection of sets is the smallest set containing all of them.*

Here are some specific examples concerning the material covered in this book so far.

Let V, W be column maps. *If $V \subset W$ and W is 1-1, then so is V.* Its 'dual': *If $V \supset W$ and W is onto, then so is V.* This makes maximally 1-1 maps and minimally onto maps particularly interesting as, by now, you know very well: *A column map is maximally 1-1 if and only if it is minimally onto if and only if it is a basis.*

Let $A \in \mathbb{F}^{m\times n}$. Then, *$A$ is 1-1(onto) if and only if A^{t} is onto(1-1).* In terms of the rows and columns of the matrix A and in more traditional terms, this says that the columns form a linearly independent (spanning) sequence if and only if the rows form a spanning (linearly independent) sequence. This is a special case of the result that $\operatorname{null} A = (\operatorname{ran} A^{\mathrm{c}})^{\perp}$, hence that $\dim \operatorname{null} A = \operatorname{codim} \operatorname{ran} A^{\mathrm{c}}$. By going from A to A^{c}, and from a subspace to its orthogonal complement, we obtain from these the 'dual' result that $\operatorname{ran} A = (\operatorname{null} A^{\mathrm{c}})^{\perp}$, hence that $\dim \operatorname{ran} A = \operatorname{codim} \operatorname{null} A^{\mathrm{c}}$.

Recall from (4.18) the factorization $A = A(:, \texttt{bound})\operatorname{rrref}(A)$. It supplies the corresponding factorization $A^{\mathrm{t}} = A^{\mathrm{t}}(:, \texttt{rbound})\operatorname{rrref}(A^{\mathrm{t}})$ with $\texttt{rbound}$ the index sequence of bound columns of A^{t}, i.e. of bound *rows* of A. By combining these two factorizations, we get the more symmetric factorization

$$A = (\operatorname{rrref}(A^{\mathrm{t}}))^{\mathrm{t}} A(\texttt{rbound}, \texttt{bound})\operatorname{rrref}(A),$$

which is called the **car**-factorization by some.

9.1 Prove that, for any $A \in L(X, Y)$, $\operatorname{codim} \operatorname{null} A = \dim \operatorname{ran} A$.

9.2 In the list of pairs of complementary concepts, given at the beginning of this chapter, many of the pairs have been ordered so as to have the first term in each pair naturally correspond to the first term in any related pair.

For example, a right (left) inverse is necessarily 1-1 (onto).

Discover as many such correspondences as you can.

The dual of a vector space

The **dual of the vector space** X is, by definition, the vector space

$$X' := L(X, \mathbb{F})$$

of all linear maps from X into the underlying scalar field. Each such map is called a **linear functional** on X. (The term 'functional' is used to indicate a map, on a vector space, whose target is the underlying scalar field. Some books use the term 'form' instead.)

We have made much use of linear functionals, namely as the rows $\lambda_1, \ldots, \lambda_n$ of specific row maps (or data maps)

$$\Lambda^{\mathrm{t}} = [\lambda_1, \ldots, \lambda_n]^{\mathrm{t}} \in L(X, \mathbb{F}^n)$$

from the vector space X to n-dimensional coordinate space.

Example: If $X = \mathbb{F}^n$, then

$$X' = L(\mathbb{F}^n, \mathbb{F}) = \mathbb{F}^{1\times n} \sim \mathbb{F}^n,$$

and it has become standard to identify $(\mathbb{F}^n)'$ with $\mathbb{F}^n$ via

$$\mathbb{F}^n \to (\mathbb{F}^n)' : \mathbf{a} \mapsto \mathbf{a}^{\mathrm{t}}.$$

While this identification is often quite convenient, be aware that, strictly speaking, $\mathbb{F}^n$ and its dual are quite different objects. □

Here is a quick discussion of X' for an arbitrary finite-dimensional vector space, X. X being finite-dimensional, it has a basis, $V \in L(\mathbb{F}^n, X)$ say. Let

$$V^{-1} =: \Lambda^{\mathrm{t}} =: [\lambda_1, \ldots, \lambda_n]^{\mathrm{t}}$$

be its inverse. Each of its rows λ_i is a linear functional on X, hence

$$\Lambda := [\lambda_1, \ldots, \lambda_n]$$

is a column map into X'.

Λ is 1-1: Indeed, if $\Lambda\mathbf{a} = 0$, then $\sum_i a_i\lambda_i$ is the zero functional, hence, in particular, $\sum_i a_i\lambda_i\mathbf{v}_j = 0$ for all columns $\mathbf{v}_j$ of V. This implies that $\mathbf{0} = (\sum_i a_i\lambda_i\mathbf{v}_j : j = 1{:}n) = \mathbf{a}^{\mathrm{t}}(\Lambda^{\mathrm{t}}V) = \mathbf{a}^{\mathrm{t}}\,\mathrm{id}_n = \mathbf{a}^{\mathrm{t}}$, hence $\mathbf{a} = \mathbf{0}$.

It follows that $\dim \operatorname{ran} \Lambda = \dim \operatorname{dom} \Lambda = n$, hence we will know that Λ is also onto as soon as we know that the dimension of its target is $\le n$, i.e.,

$$\dim X' \le n.$$

For the proof of this inequality, observe that, for each $\lambda \in X'$, the composition λV is a linear map from $\mathbb{F}^n$ to $\mathbb{F}$, hence a 1-by-n matrix. Moreover, the resulting map

$$X' \to \mathbb{F}^{1\times n} \sim \mathbb{F}^n : \lambda \to \lambda V$$

is linear. It is also 1-1 since $\lambda V = 0$ implies that $\lambda = 0$ since V is invertible. Hence, indeed, $\dim X' \le n$.

(9.1) Proposition: For each basis V of the n-dimensional vector space X, the rows of its inverse, $V^{-1} =: \Lambda^{\rm t} =: [\lambda_1, \ldots, \lambda_n]^{\rm t}$, provide the columns for the basis $\Lambda = [\lambda_1, \ldots, \lambda_n]$ for X', which earlier we called the basis *dual to* V. In particular, $\dim X' = \dim X$.

The two maps Λ and V are said to be **bi-orthonormal** to signify that

$$\lambda_i v_j = \delta_{ij}, \qquad i, j = 1{:}n.$$

Here is the ‘dual’ claim.

(9.2) Proposition: Let X be an n-dimensional linear subspace of the vector space Y. Then, for each $\Lambda^{\rm t} \in L(Y, \mathbb{F}^n)$ that is 1-1 on X, there exists exactly one basis, V, for X that is bi-orthonormal to Λ.

For every $\lambda \in Y'$, there exists exactly one $\mathbf{a} \in \mathbb{F}^n$ so that

$$\lambda = \Lambda\mathbf{a} \quad \text{on } X. \tag{9.3}$$

In particular, each $\lambda \in X'$ has a unique such **representation** $\Lambda\mathbf{a}$ in $\operatorname{ran} \Lambda$.

Proof: Since $\dim X = \dim \operatorname{tar} \Lambda^{\rm t}$ and the restriction of

$$\Lambda^{\rm t} =: [\lambda_1, \ldots, \lambda_n]^{\rm t}$$

to X is 1-1, it must be invertible, i.e., there exists exactly one basis V for X with $\Lambda^{\rm t} V = \mathrm{id}_n$. This implies that Λ is bi-orthonormal to V but does not imply that Λ is a dual basis for V since we only assumed that Λ maps into

Y', not into X'. Still, it follows by (9.1)Proposition that $R_X\Lambda$ is a dual basis for V, with R_X the linear map

$$R_X : Y' \to X' : \lambda \mapsto \lambda|_X.$$

In particular, every $\lambda \in X'$ has exactly one representation in the form $R_X\Lambda\mathbf{a}$, and this equals $\Lambda\mathbf{a}$ on X, and this holds for $R_X\lambda \in X'$ for any $\lambda \in Y'$, and since $R_X\lambda = \lambda$ on X, this proves (9.3). □

If X is not finite-dimensional, it may be harder to provide a complete description of its dual. In fact, in that case, one calls X' the **algebraic dual** and, for even some very common vector spaces, like $C([a\,..\,b])$, there is no constructive description of its algebraic dual. If X is a normed vector space, one focuses attention instead on its **topological dual**. The topological dual consists of all *continuous* linear functionals, i.e., of all linear functionals λ whose norm $\|\lambda\| := \sup_{x\in X} |\lambda x|/\|x\|$ is finite, and this goes beyond the level of this book. Suffice it to say that a normed vector space is finite-dimensional if and only if its algebraic dual coincides with its topological dual. See (7.9)Fact for the ”only if”.

The very definition of $0 \in L(X, \mathbb{F})$ ensures that $\lambda \in X'$ is 0 if and only if $\lambda x = 0$ for all $x \in X$. What about its dual statement: $x \in X$ *is* 0 *if and only if* $\lambda x = 0$ *for all* $\lambda \in X'$? For an arbitrary vector space, this turns out to require the Axiom of Choice. However, if X is a linear subspace of $\mathbb{F}^T$ for some set T, then, in particular,

$$\delta_t : X \to \mathbb{F} : x \mapsto x(t)$$

is a linear functional on X, hence the vanishing at x of all linear functionals in X' implies that, in particular, $x(t) = 0$ for all $t \in T$, hence $x = 0$.

(9.4) Fact: For any x in the vector space X, $x = 0$ if and only if $\lambda x = 0$ for all $\lambda \in X'$.

Proof: Only the ‘only if’ needs proof. If X is finite-dimensional, then, by (9.1), the condition $\lambda x = 0$ for all $\lambda \in X'$ is equivalent, for any particular basis V for X with dual basis Λ for X', to having $\mathbf{b}^{\mathrm{t}}\Lambda^{\mathrm{t}}V\mathbf{a} = 0$ for all $\mathbf{b} \in \mathbb{F}^n$ and for $x =: V\mathbf{a}$. Since $\Lambda^{\mathrm{t}}V = \mathrm{id}_n$, it follows that $\mathbf{a} = \Lambda^{\mathrm{t}}V\mathbf{a}$ must be zero, hence $x = 0$.

If X is not finite-dimensional, then, for any nonzero $x \in X$, the linear subspace $\mathrm{ran}[x]$ has on it the linear functional μ that carries x to 1. By (9.5)Fact (whose proof uses the Axiom of Choice), there exists an extension λ of μ to all of Y and, by construction, $\lambda x = 1 \neq 0$. In other words, if $x \neq 0$ then $\lambda x \neq 0$ for some $\lambda \in X'$. □

Finally, one often needs (as we did just now in the preceding proof) the following

(9.5) Fact: Every linear functional on some linear subspace of a vector space can be extended to a linear functional on the whole vector space.

Proof: If X is a linear subspace of the finite-dimensional vector space Y, then there is a basis $[V, W]$ for Y with V a basis for X. If now $\lambda \in X'$, then there is a unique $\mu \in Y'$ with $\mu[V, W] = [\lambda V, 0]$, and it extends λ to all of Y.

If Y is not finite-dimensional, then the Axiom of Choice is needed in the proof. □

9.3 Use the proof of (9.2) to show that the linear map R_X introduced there is onto. Why does this not supply a proof of (9.5)Fact in case the linear subspace in question is finite-dimensional?

The dual of an inner product space

We introduced inner-product spaces as spaces with a ready supply of linear functionals. Specifically, the very definition of an inner product $\langle , \rangle$ on the vector space Y requires that, for each $y \in Y$, $y^{\mathrm{c}} := \langle \cdot, y \rangle$ be a linear functional on Y. This sets up a map

$$^{\mathrm{c}} : Y \to Y' : y \mapsto y^{\mathrm{c}}$$

from the inner product space to its dual. This map is additive. It is also homogeneous in case $\mathbb{F} = \mathbb{R}$. If $\mathbb{F} = \mathbb{C}$, then the map is **skew-homogeneous**, meaning that

$$(\alpha y)^{\mathrm{c}} = \overline{\alpha} y^{\mathrm{c}}, \qquad \alpha \in \mathbb{F},\ y \in Y.$$

Either way, this map is 1-1 if and only if its nullspace is trivial. But, since $y^{\mathrm{c}} = 0$ implies, in particular, that $y^{\mathrm{c}} y = 0$, the positive definiteness required of the inner product guarantees that then $y = 0$, hence the map $y \mapsto y^{\mathrm{c}}$ is 1-1.

If now $n := \dim Y < \infty$, then, by (9.1)Proposition, $\dim Y' = \dim Y = n$, hence, by the (3.23)Dimension Formula, $y \mapsto y^{\mathrm{c}}$ must also be onto. This proves

(9.6) Proposition: If Y is a finite-dimensional inner product space, then every $\lambda \in Y'$ can be written in exactly one way as $\lambda = y^{\mathrm{c}}$ for some $y \in Y$.

We say in this case that y^{c} **represent**s λ.

If Y is not finite-dimensional, then the conclusion of this proposition still holds, provided we consider only the topological dual of Y and provided Y is 'complete', the very concept we declared beyond the scope of this book when, earlier, we discussed the Hermitian (a.k.a. conjugate transpose) of a linear map between two inner product spaces.

The dual of a linear map

Any $A \in L(X, Y)$ induces in a natural way the linear map

$$A' : Y' \to X' : \lambda \mapsto \lambda A.$$

This map is called the **dual** to A.

If also $B \in L(Y, Z)$, then $BA \in L(X, Z)$ and, for every $\lambda \in Z'$, $\lambda(BA) = (\lambda B)A = A'(B'(\lambda))$, hence

$$(BA)' = A'B', \quad A \in L(X, Y),\ B \in L(Y, Z). \tag{9.7}$$

If both X and Y are coordinate spaces, hence A is a matrix, then, with the identification of a coordinate space with its dual via the map $\mathbb{F}^n \to (\mathbb{F}^n)' : \mathbf{a} \mapsto \mathbf{a}^{\mathrm{t}}$, the dual of A coincides with its transpose, i.e.,

$$A' = A^{\mathrm{t}}, \qquad A \in \mathbb{F}^{m \times n} = L(\mathbb{F}^n, \mathbb{F}^m).$$

If $Y = \mathbb{F}^m$, hence A is a row map, $A = \Lambda^{\mathrm{t}} = [\lambda_1, \ldots, \lambda_m]^{\mathrm{t}}$ say, then, with the identification of $(\mathbb{F}^m)'$ with $\mathbb{F}^m$, $(\Lambda^{\mathrm{t}})'$ becomes the column map

$$(\Lambda^{\mathrm{t}})' = [\lambda_1, \ldots, \lambda_m] = \Lambda.$$

In this way, we now recognize a row map on X as the **pre-dual** of a column map into X'.

If $X = \mathbb{F}^n$, hence A is a column map, $A = V = [v_1, \ldots, v_n]$ say, then, with the identification of $(\mathbb{F}^n)'$ with $\mathbb{F}^n$, V' becomes a row map on Y', namely the row map that associates $\lambda \in Y'$ with the n-vector $(\lambda v_j : j = 1{:}n)$. Its rows are the linear functionals

$$Y' \to \mathbb{F} : \lambda \mapsto \lambda v_j$$

on Y' 'induced' by the columns of V. Each of these rows is therefore a linear functional on Y', i.e., an element of $(Y')'$, the **bidual** of Y.

(9.8) Proposition: Let $A \in L(X, Y)$.

(a) If A is onto, then A' is 1-1.

(b) If A is of finite rank and 1-1, then A' is onto.

Proof: If A is onto, then $\lambda A = 0$ implies that $\lambda = 0$, hence A' is 1-1. If A is of finite rank, let $V \in L(\mathbb{F}^n, Y)$ be a basis for $\operatorname{ran} A$, and let $\mathrm{M}^{\mathrm{t}} : \operatorname{ran} A \to \mathbb{F}^n$ be its inverse. By (9.5)Fact, each of the rows of M^{t} has an extension to an element of Y'. In other words, there exists $\Lambda \in L(\mathbb{F}^n, Y')$ for which Λ^{t} agrees on $\operatorname{ran} A$ with $\mathrm{M}^{\mathrm{t}} = V^{-1}$. Also, A being 1-1 implies the existence of a column map W into X with $V = AW$, hence W is necessarily a basis for X, and $\mathrm{id} = \Lambda^{\mathrm{t}} V = \Lambda^{\mathrm{t}} AW$. By (9.1)Proposition, it follows that $(\Lambda^{\mathrm{t}} A)^{\mathrm{t}} = A^{\mathrm{t}} \Lambda$ is a basis for X', hence A' is onto. □

10 The powers of a linear map and its spectrum

If $\operatorname{tar} A = \operatorname{dom} A$, then we can form the powers

$$A^k := \underbrace{AA \cdots A}_{k \text{ factors}}$$

of A. Here are some examples that show the importance of understanding the powers of a linear map.

Examples

Fixed-point iteration: A standard method for solving a large linear system $A? = \mathbf{y}$ (with $A \in \mathbb{F}^{n\times n}$) is to split the matrix A suitably as

$$A = M - N$$

with M 'easily invertible', and to generate the sequence $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots$ of approximate solutions by the **iteration**

$$\mathbf{x}_k := M^{-1}(N\mathbf{x}_{k-1} + \mathbf{y}), \quad k = 1, 2, \ldots\ . \tag{10.1}$$

Assuming this iteration to converge, with $\mathbf{x} := \lim_{k\to\infty} \mathbf{x}_k$ its limit, it follows that

$$\mathbf{x} = M^{-1}(N\mathbf{x} + \mathbf{y}), \tag{10.2}$$

hence that $M\mathbf{x} = N\mathbf{x} + \mathbf{y}$, therefore finally that $A\mathbf{x} = (M - N)\mathbf{x} = \mathbf{y}$, i.e., the limit solves our original problem $A? = \mathbf{y}$.

Let $\varepsilon_k := \mathbf{x} - \mathbf{x}_k$ be the **error** in our kth approximate solution. Then on subtracting the iteration equation (10.1) from the exact equation (10.2), we find that

$$\varepsilon_k = \mathbf{x} - \mathbf{x}_k = M^{-1}(N\mathbf{x} + \mathbf{y} \ - (N\mathbf{x}_{k-1} + \mathbf{y})) = M^{-1}N\varepsilon_{k-1}.$$

Therefore, by induction,

$$\varepsilon_k = B^k\varepsilon_0, \quad \text{with } B := M^{-1}N$$

the **iteration map**. Since we presumably don't know the solution $\mathbf{x}$, we have no way of choosing the **initial guess** $\mathbf{x}_0$ in any special way. For convergence, we must therefore demand that

$$\lim_{k\to\infty} B^k\mathbf{z} = 0 \quad \text{for all } \mathbf{z} \in \mathbb{F}^n.$$

It turns out that this will happen if and only if all eigenvalues of B are less than 1 in absolute value.

random walk: Consider a random walk on a graph G. The specifics of such a random walk are given by a **stochastic** matrix M of order n, with n the number of vertices in the graph. This means that all the entries of M are nonnegative, and all the entries in each row add up to 1, i.e.,

$$M \geq 0, \qquad M\mathbf{e} = \mathbf{e},$$

with $\mathbf{e}$ the vector with all entries equal to 1,

$$\mathbf{e} := (1, 1, 1, \ldots, 1).$$

The entries of M are interpreted as probabilities: M_{ij} gives the probability that, on finding ourselves at vertex i, we would proceed to vertex j. Thus, the probability that, after two steps, we would have gone from vertex i to vertex j is the sum of the probabilities that we would have gone from i to some k in the first step and thence to j in the second step, i.e., the number

$$\sum_k M_{ik}M_{kj} = (M^2)_{ij}.$$

More generally, the probability that we have gone after m steps from vertex i to vertex j is the number $(M^m)_{ij}$, i.e., the (i, j)-entry of the mth power of the matrix M.

A study of the powers of such a stochastic matrix reveals that, for large m, all the rows of M^m look more and more alike. Precisely, for each row i,

$$\lim_{m\to\infty} (M^m)_{i:} = \mathbf{x}_\infty$$

for a certain (i-independent) vector $\mathbf{x}_\infty$ with nonnegative entries that sum to one; this is part of the so-called Perron-Frobenius Theory. In terms of the random walk, this means that, for large m, the probability that we will be at vertex j after m steps is more or less independent of the vertex we started off from. One can find this limiting probability distribution $\mathbf{x}_\infty$ as a properly scaled eigenvector of the transpose M^{t} of M belonging to the eigenvalue 1.

As the simple example $M = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ shows, the last paragraph isn't quite correct. Look for the discussion of the Perron-Frobenius theorem later in this book (see pages 207ff).

polynomials in a map: Once we know the powers A^k of A, we can also construct polynomials in A, in the following way. If p is the polynomial

$$p : t \mapsto c_0 + c_1 t + c_2 t^2 + \cdots + c_k t^k,$$

then we define the linear map $p(A)$ to be what we get when we substitute A for t:

$$p(A) := c_0 \,\mathrm{id} + c_1 A + c_2 A^2 + \cdots + c_k A^k.$$

We can even consider power series. The most important example is the **matrix exponential**:

$$\exp(A) := \mathrm{id} + A + A^2/2 + A^3/6 + \cdots + A^k/k! + \cdots. \tag{10.3}$$

The matrix exponential is used in solving the first-order system

$$D\mathbf{y}(t) = A\mathbf{y}(t) \text{ for } t > 0, \quad \mathbf{y}(0) = \mathbf{b} \tag{10.4}$$

of constant-coefficient ordinary differential equations. Here A is a square matrix, of order n say, and $\mathbf{y}(t)$ is an n-vector that depends on t. Further,

$$D\mathbf{y}(t) := \lim_{h\to 0}(\mathbf{y}(t+h) - \mathbf{y}(t))/h$$

is the first derivative at t of the vector-valued function $\mathbf{y}$. One verifies that the particular function

$$\mathbf{y}(t) := \exp(tA)\mathbf{b}, \quad t \ge 0,$$

solves the differential equation (10.4). Practical application does require efficient ways for evaluating the power series

$$\exp((tA)) := \mathrm{id} + tA + (tA)^2/2 + (tA)^3/6 + \cdots + (tA)^k/k! + \cdots,$$

hence for computing the powers of tA.

Eigenvalues and eigenvectors

The calculation of $A^k x$ is simplest if A maps x to a scalar multiple of itself, i.e., if

$$Ax = \mu x = x\mu$$

for some scalar μ. For, in that case, $A^2x = A(Ax) = A(x\mu) = Ax\mu = x\,\mu^2$ and, more generally,

$$(10.5) \qquad Ax = x\mu \quad \Longrightarrow \quad A^k x = x\,\mu^k, \quad k = 0, 1, 2, \ldots .$$

If $x = 0$, this will be so for any scalar μ. If $x \neq 0$, then this will be true for at most one scalar μ. That scalar is called an *eigenvalue for* A with associated *eigenvector* x.

(10.6) Definition: Let $A \in L(X)$. Any scalar μ for which there is a *nontrivial* vector $x \in X$ so that $Ax = x\mu$ is called an **eigenvalue** of A, with (μ, x) the corresponding **eigenpair**. The collection of all eigenvalues of A is called the **spectrum** of A and is denoted spct(A). Thus

$$\operatorname{spct}(A) = \{\mu \in \mathbb{F} : \ A - \mu \operatorname{id} \ \text{ is not invertible}\}.$$

All the elements of null$(A - \mu \operatorname{id})\backslash 0$ are called the **eigenvectors** of A *associated with* μ. The number

$$\rho(A) := \max|\operatorname{spct}(A)| = \max\{|\mu| : \mu \in \operatorname{spct}(A)\}$$

is called the **spectral radius of** A.

Since $\mu \in \operatorname{spct}(A)$ exactly when $(A - \mu \operatorname{id})$ is not invertible, this puts a premium on knowing whether or not a given linear map is invertible. We pointed out in Chapter 3 that the only matrices for which we could tell this at a glance are the triangular matrices. To recall, by (3.36)Proposition, a triangular matrix is invertible if and only if none of its diagonal entries is zero. Since $(A - \mu \operatorname{id})$ is triangular for any μ in case A is triangular, this gives the important

(10.7) Proposition: For any triangular matrix of order n, $\operatorname{spct}(A) = \{A_{jj} : j = 1{:}n\}$.

In the best of circumstances, there is an entire basis $V = [v_1, v_2, \ldots, v_n]$ for $X = \operatorname{dom} A$ consisting of eigenvectors for A. In this case, it is very easy to compute $A^k x$ for any $x \in X$. For, in this situation, $Av_j = v_j\mu_j$, $j = 1{:}n$, hence

$$AV = [Av_1, \ldots, Av_n] = [v_1\mu_1, \ldots, v_n\mu_n] = V\mathrm{M},$$

with M the *diagonal* matrix

$$\mathrm{M} := \operatorname{diag}(\mu_1, \ldots, \mu_n).$$

Therefore, for any k,

$$A^k V = V\mathrm{M}^k = V \operatorname{diag}(\mu_1^k, \ldots, \mu_n^k).$$

Also, since V is a basis for X, any $x \in X$ can be written (uniquely) as $x = V\mathbf{a}$ for some n-vector $\mathbf{a}$ and thus

$$A^k x = A^k V\mathbf{a} = V\mathrm{M}^k\mathbf{a} = v_1\mu_1^k a_1 + v_2\mu_2^k a_2 + \cdots + v_n\mu_n^k a_n$$

for any k. For example, for such a matrix and for any t,

$$\exp(tA) = V\exp(t\mathrm{M})V^{-1} = V \operatorname{diag}(\ldots, \exp(t\mu_j), \ldots)V^{-1}.$$

To be sure, if A is not 1-1, then at least one of the μ_j must be zero, but this doesn't change the fact that M is a diagonal matrix.

(10.8) Example: The matrix $A := \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ maps the 2-vector $\mathbf{x} := (1, 1)$ to $3\mathbf{x}$ and the 2-vector $\mathbf{y} := (1, -1)$ to itself. Hence, $A[\mathbf{x}, \mathbf{y}] = [3\mathbf{x}, \mathbf{y}] = [\mathbf{x}, \mathbf{y}] \operatorname{diag}(3, 1)$ or

$$A = V \operatorname{diag}(3, 1)V^{-1}, \quad \text{with } V := [\mathbf{x}, \mathbf{y}] = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Elimination gives

$$\begin{aligned}[V,\ \mathrm{id}] \ &= \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & -1 & 0 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} \mathbf{1} & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \end{bmatrix} \rightarrow \\ &\rightarrow \begin{bmatrix} 1 & 0 & 1/2 & 1/2 \\ 0 & \mathbf{-2} & -1 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & 1/2 & 1/2 \\ 0 & 1 & 1/2 & -1/2 \end{bmatrix},\end{aligned}$$

hence

$$V^{-1} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}/2.$$

It follows that, for any k,

$$A^k = V \operatorname{diag}(3^k, 1)V^{-1} = \begin{bmatrix} 3^k & 1 \\ 3^k & -1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}/2 = \begin{bmatrix} 3^k + 1 & 3^k - 1 \\ 3^k - 1 & 3^k + 1 \end{bmatrix}/2.$$

In particular,

$$A^{-1} = \begin{bmatrix} 1/3 + 1 & 1/3 - 1 \\ 1/3 - 1 & 1/3 + 1 \end{bmatrix}/2 = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}/3.$$

Also,

$$\exp(tA) = V \operatorname{diag}(\mathrm{e}^{3t}, \mathrm{e}^t)V^{-1} = \begin{bmatrix} \mathrm{e}^{3t} + \mathrm{e}^t & \mathrm{e}^{3t} - \mathrm{e}^t \\ \mathrm{e}^{3t} - \mathrm{e}^t & \mathrm{e}^{3t} + \mathrm{e}^t \end{bmatrix}.$$

□

10.1* Let $A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$. (i) Find a basis V and a diagonal matrix M so that $A = V\mathrm{M}V^{-1}$. (ii) Determine the matrix $\exp(A)$.

10.2 Let $A = \begin{bmatrix} 4 & 1 & -1 \\ 2 & 5 & -2 \\ 1 & 1 & 2 \end{bmatrix}$.

Use elimination to determine *all* eigenvectors for this A belonging to the eigenvalue 3, and all eigenvectors belonging to the eigenvalue 5. (It is sufficient to give a basis for $\mathrm{null}(A - 3\,\mathrm{id})$ and for $\mathrm{null}(A - 5\,\mathrm{id})$.)

10.3 If A is a triangular matrix, then one of its eigenvectors can be determined without any calculation. Which one?

10.4

(a) Prove that the matrix $A = \begin{bmatrix} 4 & 1 & -1 \\ 2 & 5 & -2 \\ 1 & 1 & 2 \end{bmatrix}$ maps the vector space $Y := \mathrm{ran}\,V$ with $V := \begin{bmatrix} 0 & 2 \\ 3 & 1 \\ 1 & 1 \end{bmatrix}$ into itself, hence the **restriction** of A to Y, i.e.,

$$A_{|Y} := B : Y \to Y : y \mapsto Ay$$

is a well-defined linear map. (You will have to verify that $\mathrm{ran}\,AV \subseteq \mathrm{ran}\,V$; looking at $\mathrm{rref}([V\ AV])$ should help.)

(b) Determine the matrix representation of B with respect to the basis V for $\mathrm{dom}\,B = Y$, i.e., compute the matrix $V^{-1}BV$. (Hint: (5.4)Example tells you how to read off this matrix from the calculations in (a).)

(c) Determine the spectrum of the linear map $B = A_{|Y}$ defined in (a). (Your answer in (b) could be helpful here since similar maps have the same spectrum.)

10.5 Prove that 0 is the only eigenvalue of the matrix $A = \begin{bmatrix} 0 & 1 & 2 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{bmatrix}$ and that, up to scalar multiples, $\mathbf{e}_1$ is the only eigenvector for A.

10.6 Let $\mu \in \mathrm{spct}(A)$ (hence $Ax = \mu x$ for some $x \neq 0$). Prove:

(i) For any scalar α, $\alpha\mu \in \mathrm{spct}(\alpha A)$.

(ii) For any scalar α, $\mu + \alpha \in \mathrm{spct}(A + \alpha\,\mathrm{id})$.

(iii) For any natural number k, $\mu^k \in \mathrm{spct}(A^k)$.

(iv) If A is invertible, then $\mu \neq 0$ and $\mu^{-1} \in \mathrm{spct}(A^{-1})$.

(v) If A is a matrix, then $\mu \in \mathrm{spct}(A^{\mathrm{t}})$ and $\overline{\mu} \in \mathrm{spct}(A^{\mathrm{c}})$.

10.7* Prove that, for any $A \in \mathbb{F}^{m\times n}$, the spectral radius of $A^{\mathrm{c}}A$ equals $\sigma_1(A)^2$, the square of the largest singular value of A.

Diagonalizability

(10.9) Definition: A linear map $A \in L(X)$ is called **diagonalizable** if it has an **eigenbasis**, i.e., if there is a basis for its domain X consisting entirely of eigenvectors for A.

(10.10) Lemma: If V_μ is a basis for $\operatorname{null}(A - \mu\,\mathrm{id})$, then $[V_\mu : \mu \in \operatorname{spct}(A)]$ is 1-1.

Proof: Note that, for any $\mu \in \operatorname{spct}(A)$ and any ν,

$$(A - \nu\,\mathrm{id})V_\mu = (\mu - \nu)V_\mu,$$

and, in particular, $(A - \mu\,\mathrm{id})V_\mu = 0$. Hence, if $\sum_\mu V_\mu \mathbf{a}_\mu = 0$, then, for each $\mu \in \operatorname{spct}(A)$, after applying to both sides of this equation the product of all $(A - \nu\,\mathrm{id})$ with $\nu \in \operatorname{spct}(A)\backslash\mu$, and using the fact that these factors commute with one another (see (10.21)Lemma), we are left with the equation $(\prod_{\nu\neq\mu}(\mu - \nu))V_\mu \mathbf{a}_\mu = \mathbf{0}$, and this implies that $\mathbf{a}_\mu = \mathbf{0}$ since V_μ is 1-1 by assumption. In short, $[V_\mu : \mu \in \operatorname{spct}(A)]\mathbf{a} = \mathbf{0}$ implies $\mathbf{a} = \mathbf{0}$. □

(10.11) Corollary: $\#\operatorname{spct}(A) \le \dim\operatorname{dom} A$, with equality only if A is diagonalizable.

(10.12) Proposition: A linear map $A \in L(X)$ is diagonalizable if and only if

$$\dim X = \sum_{\mu\in\operatorname{spct}(A)} \dim\operatorname{null}(A - \mu\,\mathrm{id}). \tag{10.13}$$

Proof: By (10.10)Lemma, (10.13) implies that $\operatorname{dom} A$ has a basis consisting of eigenvectors for A.

Conversely, if V is a basis for $X = \operatorname{dom} A$ consisting entirely of eigenvectors for A, then $AV = V\mathrm{M}$ for some diagonal matrix

$$\mathrm{M} =: \operatorname{diag}(\mu_1, \ldots, \mu_n),$$

hence, for any scalar μ, $(A - \mu\,\mathrm{id}) = V(\mathrm{M} - \mu\,\mathrm{id})V^{-1}$. In particular, $\operatorname{null}(A - \mu\,\mathrm{id}) = \operatorname{ran}[v_j : \mu = \mu_j]$, hence $\sum_{\mu\in\operatorname{spct}(A)} \dim\operatorname{null}(A - \mu\,\mathrm{id}) = \sum_{\mu\in\operatorname{spct}(A)} \#\{j : \mu_j = \mu\} = n = \#V = \dim X$. □

(10.12)Proposition readily identifies a circumstance under which A is *not* diagonalizable, namely when $\operatorname{null}(A-\mu\,\mathrm{id})\cap\operatorname{ran}(A-\mu\,\mathrm{id})\neq\{0\}$ for some μ. For, with V_ν a basis for $\operatorname{null}(A-\nu\,\mathrm{id})$ for any $\nu\in\operatorname{spct}(A)$, we compute $AV_\nu=\nu V_\nu$, hence $(A-\mu\,\mathrm{id})V_\nu=(\nu-\mu)V_\nu$ and therefore, for any $\nu\neq\mu$, $V_\nu=(A-\mu\,\mathrm{id})V_\nu/(\nu-\mu)\subset\operatorname{ran}(A-\mu\,\mathrm{id})$. This places all the columns of the 1-1 map $V_{\backslash\mu}:=[V_\nu:\nu\neq\mu]$ in $\operatorname{ran}(A-\mu\,\mathrm{id})$ while, by (10.10)Lemma, $\operatorname{ran}V_\mu\cap\operatorname{ran}V_{\backslash\mu}$ is trivial. Hence, if $\operatorname{ran}V_\mu=\operatorname{null}(A-\mu\,\mathrm{id})$ has nontrivial intersection with $\operatorname{ran}(A-\mu\,\mathrm{id})$, then $\operatorname{ran}V_{\backslash\mu}$ cannot be all of $\operatorname{ran}(A-\mu\,\mathrm{id})$, and therefore

$$\begin{aligned}\sum_{\nu\neq\mu}\dim\operatorname{null}(A-\nu\,\mathrm{id}) &= \#V_{\backslash\mu}\\ &< \dim\operatorname{ran}(A-\mu\,\mathrm{id})=\dim X-\dim\operatorname{null}(A-\mu\,\mathrm{id}),\end{aligned}$$

hence, by (10.12)Proposition, such A is not diagonalizable.

This has motivated the following

(10.14) Definition: The scalar μ is a **defective eigenvalue** of A if

$$\operatorname{null}(A-\mu\,\mathrm{id})\cap\operatorname{ran}(A-\mu\,\mathrm{id})\neq\{0\}.$$

Any such μ certainly is an eigenvalue (since, in particular, $\operatorname{null}(A-\mu\,\mathrm{id})\neq\{0\}$), but I don't care for such *negative labeling*; if it were up to me, I would call such μ an **interesting eigenvalue**, since the existence of such eigenvalues makes for a richer theory. Note that, by (3.27)Proposition, μ is a defective eigenvalue for A iff, for some, hence for every, bases V and W for $\operatorname{ran}(A-\mu\,\mathrm{id})$ and $\operatorname{null}(A-\mu\,\mathrm{id})$ respectively, $[V,W]$ is not 1-1.

(10.15) Corollary: If A has a defective eigenvalue, then A is not diagonalizable.

10.8 Prove: if $A\in L(X)$ is diagonalizable and $\#\operatorname{spct}(A)=1$, then $A=\mu\,\mathrm{id}_X$ for some $\mu\in\mathbb{F}$.

10.9 What is a simplest matrix A with $\operatorname{spct}(A)=\{1,2,3\}$?

10.10 For each of the following matrices $A\in\mathbb{F}^{2\times2}$, determine whether or not 0 is a defective eigenvalue (give a reason for your answer). For a mechanical approach, see Problem 4.9. (a) $A=0$. (b) $A=\begin{bmatrix}1&2\\2&4\end{bmatrix}$. (c) $A=\begin{bmatrix}-2&-1\\4&2\end{bmatrix}$. (d) $A=\mathrm{id}_2$.

10.11* Prove that, *for every linear map A on the finite-dimensional vector space X, if A is diagonalizable, then so is $p(A)$ for every polynomial p.*

10.12 Prove that *any linear projector P on a finite-dimensional vector space X is diagonalizable*. (Hint: Show that, for any basis U for $\operatorname{ran} P$ and any basis W for $\operatorname{null} P$, $V := [U, W]$ is a basis for X, and that all the columns of V are eigenvectors for P. All of this should follow from the fact that $P^2 = P$.)

10.13 Prove that any linear involutory map R on a finite-dimensional vector space X is diagonalizable. (Hint: Problem 5.13.)

Are all square matrices diagonalizable?

By (10.15)Corollary, this will be so only if all square matrices have only nondefective eigenvalues.

(10.16) Example: The simplest example of a matrix with a defective eigenvalue is provided by the matrix

$$N := \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} = [\mathbf{0}, \mathbf{e}_1].$$

By (10.7)Proposition, $\operatorname{spct}(N) = \{0\}$. Yet $\operatorname{null} N = \operatorname{ran}[\mathbf{e}_1] = \operatorname{ran} N$, hence the only eigenvalue of N is defective, and N fails to be diagonalizable, by (10.15)Corollary.

Of course, for this simple matrix, one can see directly that it cannot be diagonalizable, since, if it were, then some basis V for $\mathbb{R}^2$ would consist entirely of eigenvectors for the sole eigenvalue, 0, for N, hence, for this basis, $NV = 0$, therefore $N = 0$, contrary to fact. □

We will see shortly that, on a finite-dimensional vector space over the complex scalars, almost all linear maps are diagonalizable, and all linear maps are almost diagonalizable.

Does every square matrix have an eigenvalue?

Since an eigenvalue for A is any *scalar* μ for which $\operatorname{null}(A - \mu \operatorname{id})$ is not trivial, the answer necessarily depends on what we mean by a scalar.

If we only allow *real* scalars, i.e., if $\mathbb{F} = \mathbb{R}$, then not every matrix has eigenvalues. The simplest example is a rotation of the plane, e.g., the matrix

$$A := \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} = [\mathbf{e}_2, -\mathbf{e}_1].$$

This linear map rotates every $\mathbf{x} \in \mathbb{R}^2$ 90 degrees counter-clockwise, hence the only vector $\mathbf{x}$ mapped by it to a scalar multiple of itself is the zero vector. In other words, this linear map has no eigenvectors, hence no eigenvalues.

The situation is different when we also allow *complex* scalars, i.e., when $\mathbb{F} = \mathbb{C}$, and this is the reason why we considered complex scalars all along in this book. Now every (square) matrix has eigenvalues, as follows from the following simple argument.

(10.17) Theorem: Any linear map A on some nontrivial finite-dimensional vector space X over the *complex* scalar field $\mathbb{F} = \mathbb{C}$ has eigenvalues.

Proof: Let $n := \dim X$, pick any $x \in X\backslash 0$ and consider the column map

$$K := [x, Ax, A^2x, \dots, A^nx].$$

Since $\#K > \dim \operatorname{tar} K$, K cannot be 1-1. This implies that some column of K is free. Let A^dx be the first free column, i.e., the first column that is in the range of the columns preceding it. Then $\operatorname{null} K$ contains exactly one vector of the form

$$\mathbf{a} = (a_0, a_1, \dots, a_{d-1}, 1, 0, \dots, 0),$$

and this is the vector we choose. Then, writing the equation $K\mathbf{a} = 0$ out in full, we get

$$a_0x + a_1Ax + \cdots + a_{d-1}A^{d-1}x + A^dx \ = \ 0. \tag{10.18}$$

Now here comes the trick: Consider the *polynomial*

$$p : t \mapsto a_0 + a_1t + \cdots + a_{d-1}t^{d-1} + t^d. \tag{10.19}$$

Then, substituting for t our map A, we get the linear map

$$p(A) := a_0 \operatorname{id} + a_1A + \cdots + a_{d-1}A^{d-1} + A^d.$$

With this, (10.18) can be written, very concisely,

$$p(A)x = 0.$$

This is not just notational convenience. Since $a_d = 1$, p isn't the zero polynomial, and since $x \neq 0$, d must be greater than 0, i.e., p cannot be just a constant polynomial. Thus, by the *Fundamental Theorem of Algebra*, p has zeros. More precisely,

$$p(t) = (t - z_1)(t - z_2)\cdots(t - z_d)$$

for certain (possibly *complex*) scalars $z_1, \dots, z_d$. This implies (see (10.21) Lemma below) that

$$p(A) = (A - z_1 \operatorname{id})(A - z_2 \operatorname{id})\cdots(A - z_d \operatorname{id}).$$

Now, $p(A)$ is not 1-1 since it maps the nonzero vector x to zero. Therefore, *not all the maps* $(A - z_j \operatorname{id})$, $j = 1{:}d$, *can be 1-1.* In other words, for some j, $(A - z_j \operatorname{id})$ fails to be 1-1, i.e., has a nontrivial nullspace, and that makes z_j an eigenvalue for A. □

(10.20) Example: Let's try this out on our earlier example, the rotation matrix

$$A := [\mathbf{e}_2, -\mathbf{e}_1].$$

Choosing $x = \mathbf{e}_1$, we have

$$[\mathbf{e}_1, A\mathbf{e}_1, A^2\mathbf{e}_1] = [\mathbf{e}_1, \mathbf{e}_2, -\mathbf{e}_1],$$

hence the first free column is $A^2\mathbf{e}_1 = -\mathbf{e}_1$, or

$$\mathbf{e}_1 + A^2\mathbf{e}_1 = 0.$$

Thus the polynomial of interest is

$$p : t \mapsto 1 + t^2 = (t - \mathrm{i})(t + \mathrm{i}),$$

with

$$\mathrm{i} := \sqrt{-1}$$

the *imaginary unit* (see pages 272ff on complex numbers). In fact, we conclude, with $\mathbf{y} := (A + \mathrm{i}\,\mathrm{id})\mathbf{e}_1$, that $(A - \mathrm{i}\,\mathrm{id})\mathbf{y} = p(A)\mathbf{e}_1 = \mathbf{0}$, while $\mathbf{y} = A\mathbf{e}_1 + \mathrm{i}\mathbf{e}_1 = \mathbf{e}_2 + \mathrm{i}\mathbf{e}_1 \neq 0$, showing that $(\mathrm{i}, \mathbf{e}_2 + \mathrm{i}\mathbf{e}_1)$ is an eigenpair for this A.

Polynomials in a linear map, Krylov subspaces, and the minimal polynomials

The proofs of (10.10)Lemma and of (10.17)Theorem use in an essential way the following fact.

(10.21) Lemma: If r is the product of the polynomials p and q, i.e., $r(t) = p(t)q(t)$ for all t, then, for any linear map $A \in L(X)$,

$$r(A) = p(A)q(A) = q(A)p(A).$$

Proof: If you wanted to check that $r(t) = p(t)q(t)$ for the polynomials r, p, q, you would multiply p and q term by term, collect like terms and then compare coefficients with those of r. For example, if $p(t) = t^2 + t + 1$ and $q(t) = t - 1$, then

$$\begin{aligned} p(t)q(t) = (t^2 + t + 1)(t - 1) &= t^2(t - 1) + t(t - 1) + (t - 1) \\ &= t^3 - t^2 + t^2 - t + t - 1 = t^3 - 1, \end{aligned}$$

i.e., the product of these two polynomials is the polynomial r given by $r(t) = t^3 - 1$. The only facts you use are: (i) free reordering of terms (commutativity of addition), and (ii) things like $tt = t^2$, i.e., the fact that

$$t^i t^j = t^{i+j}.$$

Both of these facts hold if we replace t by A. □

Here is a further use of this lemma. We now prove that the polynomial p constructed in the proof of (10.17) has the property that every one of its roots is an eigenvalue for A. This is due to the fact that we constructed it in the form (10.19) with d the *smallest* integer for which $A^d x \in \operatorname{ran}[x, Ax, \ldots, A^{d-1}x]$. Thus, with μ any zero of p, we can write

(10.22) $$p(t) = (t - \mu)q(t)$$

for some polynomial q necessarily of the form

$$q(t) = b_0 + b_1 t + \cdots + b_{d-2}t^{d-2} + t^{d-1}.$$

The crucial point here is that q is of degree $< d$. This implies that $q(A)x \neq 0$ since, otherwise, $(b_0, b_1, \ldots, 1)$ would be a nontrivial vector in $\operatorname{null}[x, Ax, \ldots, A^{d-1}x]$ and this would contradict the choice of d as the index for which $A^d x$ is the *first* free column in $[x, Ax, A^2x, \ldots]$. Since

$$0 = p(A)x = (A - \mu \,\mathrm{id})q(A)x,$$

it follows that μ is an eigenvalue for A with associated eigenvector $q(A)x$.

This is exactly how we got an eigenvector for the eigenvalue i in (10.20) Example.

(10.23) Example: As another example, consider again the matrix $A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ from (10.8)Example. We choose $x = \mathbf{e}_1$ and consider

$$[x, Ax, \ldots, A^n x] = [\mathbf{e}_1, A\mathbf{e}_1, A(A\mathbf{e}_1)] = \begin{bmatrix} 1 & 2 & 5 \\ 0 & 1 & 4 \end{bmatrix}.$$

Since $[\mathbf{e}_1, A\mathbf{e}_1, A^2\mathbf{e}_1]$ is in row echelon form, we conclude that the first two columns are bound. Elimination gives the rref

$$\begin{bmatrix} 1 & 0 & -3 \\ 0 & 1 & 4 \end{bmatrix},$$

hence $(3, -4, 1) \in \operatorname{null}[\mathbf{e}_1, A\mathbf{e}_1, A^2\mathbf{e}_1]$. Correspondingly, $p(A)\mathbf{e}_1 = \mathbf{0}$, with

$$p(t) = 3 - 4t + t^2 = (t - 3)(t - 1).$$

Consequently, $\mu = 3$ is an eigenvalue for A, with corresponding eigenvector

$$(A - \ \mathrm{id})\mathbf{e}_1 = (1, 1);$$

also, $\mu = 1$ is an eigenvalue for A, with corresponding eigenvector

$$(A - 3\,\mathrm{id})\mathbf{e}_1 = (-1, 1).$$

Note that the resulting basis $\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$ for $\mathbb{F}^2$ consisting of eigenvectors for A differs in some detail from the one we found in (10.8)Example. After all, if v is an eigenvector, then so is αv for any nonzero scalar α. □

10.14* The **Fibonacci sequence** $f := (f_0, f_1, f_2, \ldots)$ is defined by its *two-term recurrence*:

$$f_{k+1} = f_k + f_{k-1}, \quad k = 1, 2, \ldots; \quad (f_0, f_1) := (0, 1).$$

Thus, $f_{k:k+1} = Af_{k-1:k}$ with

$$A := \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}.$$

Therefore, by induction on k, $f_{k:k+1} = A^k \mathbf{e}_2$.

(a) Use elimination as in (10.23)Example to determine a diagonal matrix M and basis V for $\mathbb{F}^2$ for which $AV = V$M.

(b) Use the eigenstructure of A found in (a) to compute f_{50}.

Here is some standard language concerning the items in our discussion so far. One calls $(x, Ax, A^2x, \ldots)$ the **Krylov sequence for** A **at** x, and calls

$$K_{A,x} := \operatorname{ran}[x, Ax, A^2x, \ldots]$$

the **Krylov subspace for** A **at** x. $K_{A,x}$ is A**-invariant**, meaning that $A(K_{A,x}) \subset K_{A,x}$. This implies that the restriction

$$A_x := A|_{K_{A,x}}$$

of A to $K_{A,x}$ is in $L(K_{A,x})$. Any eigenvalue of A_x is an eigenvalue of A. However, if $\dim K_{A,x}$ is much smaller than $\dim X$, then we would expect it to be much easier to find eigenvalues for A_x than it is to find eigenvalues for A.

To find out more about the structure of $K_{A,x}$ and of A_x, we now consider polynomials p for which $p(A)x = 0$. Any such nontrivial polynomial is called an **annihilating polynomial for** A **at** x. We may assume without loss of generality that this polynomial is **monic**, i.e., its highest nonzero coefficient is 1, since we can always achieve this by dividing the polynomial by its highest nonzero coefficient without changing the fact that it is an annihilating polynomial for A at x. When $K_{A,x} = X$, then x is called a **cyclic vector for** A, and A is called **non-derogatory** in case it has a cyclic vector. Such annihilating polynomials simplify our dealings with $K_{A,x}$ because of the following.

(10.24) Lemma: If the annihilating polynomial, p, for A at x has degree k, then $K_{A,x} = \operatorname{ran}[x, Ax, \ldots, A^{k-1}x]$.

Proof: If $y \in K_{A,x}$, then y is a weighted sum of vectors A^jx, hence can be written as $y = h(A)x$ for some polynomial h. By the Euclidean algorithm (see page 281), there exist polynomials q and r, with $\deg r < \deg p$ so that $h = qp + r$, therefore, by (10.21)Lemma and since $p(A)x = 0$,

$$y = h(A)x = q(A)p(A)x + r(A)x = r(A)x \in \operatorname{ran}[x, Ax, \ldots, A^{\deg r}x]$$

while $\operatorname{ran}[x, Ax, \ldots, A^{\deg r}x] \subset \operatorname{ran}[x, Ax, \ldots, A^{k-1}x]$ since $\deg r < \deg p = k$. □

This encourages us to choose k as small as possible or, equivalently, to choose for p the monic annihilating polynomial for A at x of smallest degree. By the Lemma just proved, that degree cannot be smaller than d, with $A^d x$ the first or leftmost free column in $[x, Ax, \ldots]$, yet as we saw in the proof of (10.17)Theorem, there is a unique monic annihilating polynomial of that degree d. Further, for that choice of d, all the columns of $[x, Ax, \ldots, A^{d-1}x]$ are bound, hence we conclude from (10.24)Lemma that $[x, Ax, \ldots, A^{d-1}x]$ is a basis for the Krylov subspace $K_{A,x}$.

Here, for the record, is a formal account of what we have proved so far.

(10.25) Proposition: For every $A \in L(X)$ with $\dim X < \infty$ and every $x \in X\backslash 0$, there is a unique monic polynomial p of smallest degree for which $p(A)x = 0$. This polynomial is called the **minimal polynomial for** A **at** x and is denoted

$$p_{A,x}.$$

It can be constructed in the form

$$p_{A,x}(t) = a_0 + a_1 t + \cdots + a_{d-1} t^{d-1} + t^d,$$

with $A^d x$ the first or leftmost free column of $[x, Ax, A^2x, \ldots]$, hence $(a_0, \ldots, a_{d-1}, 1) \in \text{null}[x, Ax, \ldots, A^d x]$.

For this choice of d, $[x, Ax, \ldots, A^{d-1}x]$ is a basis for the Krylov subspace $K_{A,x} = \text{ran}[x, Ax, A^2x, \ldots]$, hence

$$K_{A,x} = \{q(A)x : q \in \Pi_{<d}\}.$$

Assuming that X is a vector space over $\mathbb{F} = \mathbb{C}$, every zero μ of $p_{A,x}$ is an eigenvalue of A, with associated eigenvector $q(A)x$, where $p_{A,x}(t) =: (t - \mu)q(t)$. (See pages 280ff on Horner's method for the standard way to compute q from $p_{A,x}$ and μ.)

For *example*, consider the *permutation matrix* $P = [\mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_1]$ and take $\mathbf{x} = \mathbf{e}_1$. Then

$$[x, Px, P^2x, P^3x] = [\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_1].$$

Hence, P^3x is the first free column here. The element in the nullspace corresponding to it is the vector $(-1, 0, 0, 1)$. Hence, the minimal polynomial for P at $x = \mathbf{e}_1$ is of degree 3; it is the polynomial $p(t) = t^3 - 1$. It has the zero $\mu = 1$, which therefore is an eigenvalue of P. A corresponding eigenvector is obtainable in the form $q(P)\mathbf{e}_1$ with $q(t) := p(t)/(t-1) = t^2 + t + 1$, hence the eigenvector is $\mathbf{e}_3 + \mathbf{e}_2 + \mathbf{e}_1$.

Here are two additional useful properties of $p_{A,x}$.

(10.26) Proposition: Every annihilating polynomial p for A at x is of the form $p = qp_{A,x}$ for some polynomial q, i.e.,

$$\mathcal{I}_{A,x} := \{p \in \Pi : p(A)x = 0\} \;=\; \Pi p_{A,x} := \{qp_{A,x} : q \in \Pi\}.$$

Proof: Let p be a polynomial. By the Euclidean algorithm (see page 281), there exist polynomials q and r so that $p = qp_{A,x} + r$ with $\deg r < \deg p_{A,x}$. Further, if $p \in \mathcal{I}_{A,x}$, then also $r = p - qp_{A,x} \in \mathcal{I}_{A,x}$, hence r must be the zero polynomial since $p_{A,x}$ is of minimal (positive) degree in $\mathcal{I}_{A,x}$. □

In particular, $p_{A,x}$ divides the **minimal polynomial of** A, meaning the monic polynomial p of smallest degree that annihilates A, i.e., satisfies $p(A) = 0$. This polynomial is customarily denoted by

$$p_A.$$

(10.27) Proposition: For any $A \in L(X)$ and any $x \in X\backslash 0$, any eigenvalue μ of A with eigenvector y in $K_{A,x}$ is a zero of $p_{A,x}$.

Also, $p_{A,x} = p_{A_x,x}$ is the minimal polynomial for $A_x := A|_{K_{A,x}}$.

Proof: Let $d := \deg p_{A,x}$ and assume that $Ay = \mu y$ for some scalar μ and some nonzero $y \in K_{A,x} = \{q(A)x : q \in \Pi_{<d}\}$, the equality by (10.25)Proposition. Then $y = q(A)x$ for some nontrivial polynomial q of degree $< d$, while $0 = (A - \mu\,\mathrm{id})y = (A - \mu\,\mathrm{id})q(A)x$, hence $r := (\cdot - \mu)q$ is a nontrivial annihilating polynomial for A at x, of degree $\le d$, hence necessarily a (nontrivial) scalar multiple of $p_{A,x}$ by the minimality of $p_{A,x}$, therefore $p_{A,x}(\mu) = 0$.

Since $A_x = A$ on $K_{A,x}$, $[x, Ax, A^2x, \ldots] = [x, A_x x, A_x^2 x, \ldots]$, hence $p_{A_x,x} = p_{A,x}$. Further, by (10.21)Lemma,

$$p_{A,x}(A)q(A)x = q(A)p_{A,x}(A)x = q(A)0 = 0, \qquad q \in \Pi,$$

hence $K_{A,x} = \{q(A)x : q \in \Pi\} \subset \operatorname{null} p_{A,x}(A)$, therefore $p_{A,x} = 0$ on $K_{A,x}$. Any other nontrivial polynomial p for which $p(A)$ is zero on $K_{A,x}$ must necessarily have $p(A)x = 0$, hence be divisible by $p_{A,x}$, by (10.26)Proposition, therefore of degree $\ge \deg p_{A,x} = \dim K_{A,x}$, showing $p_{A,x}$ to be the monic annihilating polynomial for A_x of *minimal degree.* □

10.15 Use Elimination as in (10.23) to determine all the eigenvalues and, for each eigenvalue, a corresponding eigenvector, for each of the following matrices: (i) $\begin{bmatrix} 7 & -4 \\ 5 & -2 \end{bmatrix}$; (ii) $[0, \mathbf{e}_1, \mathbf{e}_2] \in \mathbb{R}^{3\times 3}$ (try $\mathbf{x} = \mathbf{e}_3$); (iii) $\begin{bmatrix} -1 & 1 & -3 \\ 20 & 5 & 10 \\ 2 & -2 & 6 \end{bmatrix}$.

10.16* Let (μ, x) be an eigenpair for $A \in L(X)$. Prove: *(i) For any polynomial p, $(p(\mu), x)$ is an eigenpair for $p(A)$; (ii) If $B \in L(X)$ with $AB = BA$ and $Bx \neq 0$, then (μ, Bx) is an eigenpair for A.*

10.17*

(a) Prove: *If p is any nontrivial polynomial and A is any square matrix for which $p(A) = 0$, then* $\operatorname{spct}(A) \subseteq \{\mu \in \mathbb{C} : p(\mu) = 0\}$. (Hint: Problem 10.16(i).)

(b) What can you conclude about $\operatorname{spct}(A)$ in case you know that A is *idempotent*, i.e., a linear projector, i.e., $A^2 = A$?

(c) What can you conclude about $\operatorname{spct}(A)$ in case you know that A is **nilpotent**, i.e., $A^q = 0$ for some integer q?

(d) What can you conclude about $\operatorname{spct}(A)$ in case you know that A is **involutory**, i.e., $A^{-1} = A$?

(e) What is the spectrum of the linear map $D : \Pi_{\le k} \to \Pi_{\le k}$ of differentiation, as a map on polynomials of degree $\le k$?

10.18* Prove: *Let $A \in L(X)$ and $x \in X\backslash 0$. If the corresponding minimal polynomial $p_{A,x}$ for A at x has degree $d = \dim X$, then* $\operatorname{spct}(A) \subset \{z \in \mathbb{F} : p_{A,x}(z) = 0\}$*, with equality in case $\mathbb{F} = \mathbb{C}$.*

10.19 Assume that $\mathbb{F} = \mathbb{C}$ and use the minimal polynomial at $\mathbf{e}_1$ to determine the spectrum of the following matrices: (i) $[\mathbf{e}_2, 0]$; (ii) $[\mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_1]$; (iii) $[\mathbf{e}_2, \mathbf{e}_2]$; (iv) $[\mathbf{e}_2, \mathbf{e}_1, 2\mathbf{e}_3]$.

10.20* The **companion matrix** for the monic polynomial $p : t \mapsto a_1 + a_2 t + \cdots + a_n t^{n-1} + t^n$ is, by definition, the matrix $A_p := [\mathbf{e}_2, \ldots, \mathbf{e}_n, -\mathbf{a}] \in \mathbb{F}^{n\times n}$. (a) Prove that p is the minimal polynomial for A at $\mathbf{e}_1$. (b) Use (a) and `MATLAB`'s `eig` command to find all the zeros of the polynomial $p : t \mapsto 1 + t + t^2 + \cdots + t^9$. Check your answer.

10.21* Let $x \in X\backslash 0$ and $A \in L(X)$, and $A_x := A\big|_{K_{A,x}}$. Prove that $A_x \in L(K_{A,x})$ and that, if $\mathbb{F} = \mathbb{C}$, then $\operatorname{spct}(A_x) = \{z \in \mathbb{C} : p_{A,x}(z) = 0\}$.

10.22* Prove that $A \in L(X)$ is non-derogatory if and only if, for some $x \in X$, $\deg p_{A,x} = \dim X$, in which case x is a cyclic vector for A.

10.23 Let $A \in L(X)$ be non-derogatory, $\dim X = n$. Prove:

(i) *$C(A) := \{B \in L(X) : AB = BA\} = \Pi(A) := \{p(A) : p \in \Pi\}$.*

(ii) *The map $f : p \mapsto p(A)$ is linear and carries $\Pi_{<n}$ 1-1 onto $C(A)$.*

10.24

(i) Show that any $A \in \mathbb{F}^{2\times 2}\backslash\{0\}$ must be non-derogatory or else be a scalar multiple of the identity. (Hint: find a nontrivial vector that is not an eigenvector.)

(ii) Show that any linear projector on the linear space X is derogatory unless $\dim X \le 2$.

10.25 Let A be a matrix of order n, let $\mathbf{x} \in \mathbb{F}^n\backslash 0$, and let P be the orthogonal projector of $\mathbb{F}^n$ onto the space $Y := \operatorname{ran}[\mathbf{x}, A\mathbf{x}, \ldots, A^{r-1}\mathbf{x}]$, the Krylov subspace of order r for A generated by $\mathbf{x}$. Assume that Y is r-dimensional, and let $PA^r\mathbf{x} =: \sum_{j<r} a_j A^j \mathbf{x}$. (i) Prove that $K := [\mathbf{x}, PA\mathbf{x}, (PA)^2\mathbf{x}, \ldots, (PA)^r\mathbf{x}] = [\mathbf{x}, A\mathbf{x}, \ldots, A^{r-1}\mathbf{x}, PA^r\mathbf{x}]$. (ii) Prove that $q(t) := t^r - \sum_{j<r} a_j t^j$ is the minimal polynomial at $\mathbf{x}$ for the linear map $PA : Y \to Y : \mathbf{y} \mapsto PA\mathbf{y}$. (iii) Conclude that q is the unique monic polynomial of degree r for which $\|q(A)\mathbf{x}\|_2$ is as small as possible.

10.26 Prove: *(i) for any $A, B \in L(X)$,* $\operatorname{null} A \cap \operatorname{null} B \subset \operatorname{null}(A + B)$*; (ii) for any $A, B \in L(X)$ with $AB = BA$,* $\operatorname{null} A + \operatorname{null} B \subset \operatorname{null}(AB)$.

10.27 Prove: *If g is the greatest common divisor of the nontrivial polynomials $p_1, \ldots,$*

p_r and m is their least common multiple, then, for any $A \in L(X)$, $\operatorname{null} g(A) = \cap_j \operatorname{null} p_j(A)$ and $\operatorname{null} m(A) = \sum_j \operatorname{null} p_j(A)$. (Hint: Problem 17.6.)

It is enough to understand the eigenstructure of matrices

So far, we know how to find *some* eigenvalues and corresponding eigenvectors for a given $A \in L(X)$, making use of minimal polynomials at some chosen $x \in X\backslash 0$ found by elimination. But can we be sure to find all the eigenvalues that way? By (10.11)Corollary, we know that we have found them all if we have found $n := \dim X$ of them. But if we find fewer than that, then we can't be sure.

The standard approach to finding the entire spectrum of A is by searching for linear maps B that have the same spectrum as A but carry that spectrum more openly, like triangular matrices (see (10.7)Proposition). This search makes essential use of the notion of similarity.

(10.28) Definition: We say that $A \in L(X)$ and $B \in L(Y)$ are **similar to each other** and write

$$A \sim B$$

in case there is an invertible $V \in L(Y, X)$ so that

$$A = VBV^{-1}.$$

In particular, *a linear map is diagonalizable if and only if it is similar to a diagonal matrix.*

In trying to decide whether or not a given linear map A is diagonalizable, it is sufficient to decide this question for any convenient linear map B similar to A. For, if such a B is diagonalizable, i.e., similar to a diagonal matrix, then A is similar to that very same diagonal matrix. This follows from the fact that similarity is an equivalence relation:

(10.29) Proposition: Similarity is an **equivalence relation**. Specifically,

(i) $A \sim A$ (**reflexive**);

(ii) $A \sim B$ implies $B \sim A$ (**symmetric**);

(iii) $A \sim B$ and $B \sim C$ implies $A \sim C$ (**transitive**).

Proof: $A \sim A$, since $A = \operatorname{id} A \operatorname{id}$. Also, if $A = VBV^{-1}$ for some invertible V, then also $W := V^{-1}$ is invertible, and $B = V^{-1}AV = WAW^{-1}$. Finally, if $A = VBV^{-1}$ and $B = WCW^{-1}$, then $U := VW$ is also invertible, and $A = V(WCW^{-1})V^{-1} = UCU^{-1}$. □

Now, any linear map $A \in L(X)$ on a *finite-dimensional* vector space X is similar (in many ways if X is not trivial) to a *matrix*. Indeed, for any basis V for X, $\widehat{A} := V^{-1}AV$ is a matrix similar to A. The map $\widehat{A}$ so defined is indeed a matrix since both its domain and its target is a coordinate space (the same one, in fact; hence $\widehat{A}$ is a *square* matrix). We conclude that, in looking for ways to decide whether or not a linear map is diagonalizable, it is sufficient to know how to do this for square *matrices*.

Every complex (square) matrix is similar to an upper triangular matrix

While having in hand a diagonal matrix similar to a given $A \in L(X)$ is very nice indeed, for most purposes it is sufficient to have in hand an *upper triangular* matrix similar to A. There are several reasons for this.

One reason is that, as soon as we have an upper triangular matrix similar to A, then (see (10.33)Corollary) we can easily manufacture from this a matrix similar to A and with off-diagonal elements as small as we please (except that, in general, we can't make them all zero).

A more fundamental reason is that, once we have an upper triangular matrix similar to A, then we know the entire spectrum of A since, by (10.7)Proposition, the spectrum of a triangular matrix is the set of its diagonal entries. Here are the various facts.

(10.30) Proposition: If A and $\widehat{A}$ are similar, then $\operatorname{spct}(A) = \operatorname{spct}(\widehat{A})$.

Proof: If $\widehat{A} = V^{-1}AV$ for some invertible V, then, for any scalar μ, $\widehat{A} - \mu\,\mathrm{id} = V^{-1}(A - \mu\,\mathrm{id})V$. In particular, $\widehat{A} - \mu\,\mathrm{id}$ is not invertible (i.e., $\mu \in \operatorname{spct}(\widehat{A})$) if and only if $A - \mu\,\mathrm{id}$ is not invertible (i.e., $\mu \in \operatorname{spct}(A)$). □

(10.31) Corollary: If $A \in L(X)$ is similar to a triangular matrix $\widehat{A}$, then μ is an eigenvalue for A if and only if $\mu = \widehat{A}_{jj}$ for some j. In a formula,

$$\operatorname{spct}(A) = \{\widehat{A}_{jj} : \text{ all } j\}.$$

More precisely, if $\widehat{A} = V^{-1}AV$ is upper triangular and j is the smallest index for which $\mu = \widehat{A}_{jj}$, then there is an eigenvector for A belonging to μ available in the form $w = V\mathbf{a}$, with $\mathbf{a}$ the element in the standard basis for $\operatorname{null}(\widehat{A} - \mu\,\mathrm{id})$ associated with the (free) jth column of $\widehat{A} - \mu\,\mathrm{id}$, i.e., $\mathbf{a} \in \operatorname{null}(\widehat{A} - \mu\,\mathrm{id})$, $a_j = 1$, and all other entries corresponding to free columns of $\widehat{A} - \mu\,\mathrm{id}$ are 0; cf. (4.14).

The now-standard algorithm for computing the eigenvalues of a given matrix A is the **QR method**. It generates a sequence $B_1, B_2, B_3, \ldots$ of matrices all similar to A that converges to an upper triangular matrix. To the extent that the lower triangular entries of B_k are small (compared to $\|B_k\|$, say), the diagonal entries of B_k are close to eigenvalues of B_k, hence of A. The actual version of the QR method used in `MATLAB` is quite sophisticated, as much care has gone into making the algorithm fast as well as reliable in the presence of round-off.

The `MATLAB` command `eig(`A`)` gives you the list of eigenvalues of A. The more elaborate command `[V,M]=eig(`A`)` gives you, in `V`, a list of corresponding 'eigenvectors', in the sense that, approximately, $A\mathtt{V}(:,j) = \mathtt{V}(:,j)\mathtt{M}(j,j)$, all j.

(10.32) Theorem: Every complex (square) matrix is similar to an upper triangular matrix.

Proof: The proof is by induction on the order, n, of the given matrix A.

If $n = 1$, then A is a 1×1-matrix, hence trivially upper triangular. Assume that we have proved the theorem for all matrices of order $n-1$ and let A be of order n. Since the scalar field is $\mathbb{C}$, we know that A has an eigenvector, $\mathbf{u}_1$, say, with corresponding eigenvalue, μ_1 say. Extend $[\mathbf{u}_1]$ to a basis $U = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n]$ for $\mathbb{C}^n$. Then

$$AU = [A\mathbf{u}_1, \ldots, A\mathbf{u}_n] = [\mathbf{u}_1\mu_1, A\mathbf{u}_2, \ldots, A\mathbf{u}_n].$$

We want to compute $U^{-1}AU$. For this, observe that $U^{-1}\mathbf{u}_1 = U^{-1}U\mathbf{e}_1 = \mathbf{e}_1$. Therefore,

$$U^{-1}AU = [\mathbf{e}_1\mu_1, U^{-1}A\mathbf{u}_2, \ldots, U^{-1}A\mathbf{u}_n].$$

Writing this out in detail, we have

$$U^{-1}AU =: \widehat{A} = \begin{bmatrix} \mu_1 & \times & \cdots & \times \\ 0 & \times & \cdots & \times \\ \vdots & \vdots & \ldots & \vdots \\ 0 & \times & \cdots & \times \end{bmatrix} =: \begin{bmatrix} \mu_1 & C \\ 0 & A_1 \end{bmatrix}.$$

Here, C is some $1 \times (n-1)$ matrix of no further interest, A_1 is a matrix of order $n-1$, hence, by induction hypothesis, there is some invertible W so that $\widehat{A}_1 := W^{-1}A_1W$ is upper triangular. We compute

$$\begin{aligned} \operatorname{diag}(1, W^{-1})\widehat{A}\ \operatorname{diag}(1, W) &= \begin{bmatrix} 1 & 0 \\ 0 & W^{-1} \end{bmatrix} \begin{bmatrix} \mu_1 & C \\ 0 & A_1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & W \end{bmatrix} \\ &= \begin{bmatrix} \mu_1 & CW \\ 0 & W^{-1}A_1W \end{bmatrix}. \end{aligned}$$

The computation uses the fact that multiplication from the left (right) by a block-diagonal matrix multiplies the corresponding rows (columns) from the left (right) by the corresponding diagonal blocks (see Problem 2.17). Since

$$\operatorname{diag}(1, W^{-1}) \operatorname{diag}(1, W) = \operatorname{diag}(1, \operatorname{id}_{n-1}) = \operatorname{id}_n,$$

this shows that $\widehat{A}$ is similar to an upper triangular matrix. Since A is similar to $\widehat{A}$, this finishes the proof. □

Various refinements in this proof are possible (as we will show later, in the discussion of the 'Schur form'), to give more precise information about possible upper triangular matrices similar to a given A. For the present, though, this is sufficient for our needs since it allows us to prove the following:

(10.33) Corollary: Every complex (square) matrix is similar to an 'almost diagonal' matrix. Precisely, for every complex matrix A and every $\varepsilon > 0$, there exists an upper triangular matrix B_ε similar to A whose off-diagonal entries are all $< \varepsilon$ in absolute value.

Proof: By (10.32)Theorem, we know that any such A is similar to an upper triangular matrix. Since similarity is transitive (see (10.29)Proposition), it is therefore sufficient to prove this Corollary in case A is upper triangular, of order n, say.

The proof in this case is a simple trick: Consider the matrix

$$B := W^{-1}AW,$$

with

$$W := \operatorname{diag}(\delta^1, \delta^2, \ldots, \delta^n),$$

and the *scalar* δ to be set in a moment. W is indeed invertible as long as $\delta \neq 0$, since then

$$W^{-1} = \operatorname{diag}(\delta^{-1}, \delta^{-2}, \ldots, \delta^{-n}).$$

Now, multiplying a matrix by a diagonal matrix from the *left* (*right*) multiplies the *rows* (*columns*) of that matrix by the diagonal entries of the diagonal matrix. Therefore,

$$B_{ij} = (W^{-1}AW)_{ij} = A_{ij}\delta^{j-i}, \quad \text{all } i, j.$$

In particular, B is again upper triangular, and its diagonal entries are those of A. However, all its possibly nontrivial off-diagonal entries lie above the diagonal, i.e., are entries B_{ij} with $j > i$, hence are the corresponding entries of A multiplied with some *positive* power of the scalar δ. Thus, if

$$c := \max_{i<j} |A_{ij}|$$

and we choose $\delta := \min\{\varepsilon/c, 1\}$, then, we can be certain that

$$|B_{ij}| \leq \varepsilon, \qquad \text{all } i \neq j,$$

regardless of how small we choose that positive ε. □

10.28 T/F

(a) The only diagonalizable matrix A having just one factorization $A = V\mathrm{M}V^{-1}$ with M diagonal is the empty matrix.

(b) If A is the linear map of multiplication by a scalar, then any basis for its domain is an eigenbasis for A.

(c) A triangular matrix of order n is diagonalizable if and only if it has n different diagonal entries.

(d) Any (square) triangular matrix is diagonalizable.

(e) Any matrix of order 1 is diagonalizable.

(f) A matrix of order n has n eigenvalues.

(g) Similar linear maps have the same spectrum.

(h) The linear map of differentiation on $\Pi_{\le k}$ is nilpotent.

(i) The identity map is idempotent.

(j) If the matrix A has 3 eigenvalues, then it must have at least 3 columns.

(k) If $\mathrm{null}(A - \mu\,\mathrm{id})$ is not trivial, then every one of its elements is an eigenvector for A belonging to the eigenvalue μ.

11 Convergence of the power sequence

Convergence of sequences in a normed vector space

Our discussion of the power sequence $A^0, A^1, A^2, \ldots$ of a linear map naturally involves the *convergence* of such a sequence.

Convergence of a vector sequence or a map sequence is most conveniently described with the aid of a norm, as introduced earlier, starting at page 120.

Suppose $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots$ is an infinite sequence of n-vectors. In order to avoid confusion, I refer to the jth entry of the kth term $\mathbf{x}_k$ in such a vector sequence by $\mathbf{x}_k(j)$. We say that this sequence **converges to the** $\boldsymbol{n}$**-vector** $\mathbf{x}_\infty$ and write

$$\mathbf{x}_\infty = \lim_{k\to\infty} \mathbf{x}_k,$$

in case

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

Such convergence takes place entry-wise, i.e.,

$$\mathbf{x}_\infty = \lim_{k\to\infty} \mathbf{x}_k \quad \Longleftrightarrow \quad \forall i,\ \mathbf{x}_\infty(i) = \lim_{k\to\infty} \mathbf{x}_k(i).$$

Note that $\mathbf{x}_\infty = \lim_{k\to\infty} \mathbf{x}_k$ if and only if, for every $\varepsilon > 0$, there is some k_0 so that, for all $k > k_0$, $\|\mathbf{x}_\infty - \mathbf{x}_k\| < \varepsilon$. This says that, for any given $\varepsilon > 0$ however small, all the terms in the sequence from a certain point on lie in the "ball"

$$B_\varepsilon(\mathbf{x}_\infty) := \{\mathbf{y} \in \mathbb{F}^n : \|\mathbf{y} - \mathbf{x}_\infty\| < \varepsilon\}$$

whose center is $\mathbf{x}_\infty$ and whose radius is ε.

(11.1) Lemma: A convergent sequence is necessarily bounded. More explicitly, if the sequence $(\mathbf{x}_k)$ of n-vectors converges, then $\sup_k \|\mathbf{x}_k\| < \infty$, i.e., there is some c so that, for all k, $\|\mathbf{x}_k\| \le c$.

The proof is a verbatim repeat of the proof of this assertion for scalar sequences, as given on the pages 274ff on scalar sequences.

In the same way, we say that the sequence $A_1, A_2, A_3, \ldots$ of matrices **converges** to the matrix B and write

$$\lim_{k\to\infty} A_k = B,$$

in case

$$\lim_{k\to\infty} \|B - A_k\|_\infty = 0.$$

As in the case of vector sequences, a convergent sequence of matrices is necessarily bounded.

Here, for convenience, we have used the map norm associated with the max-norm since we have the simple and explicit formula (7.17) for it. Yet we know from (7.25)Proposition that any two norms on any finite-dimensional normed vector space are equivalent. In particular, if $\|\ \|'$ is any norm on $L(\mathbb{F}^n) = \mathbb{F}^{n\times n}$, then there is a positive constant c so that

$$\forall A \in \mathbb{F}^{n\times n}, \quad \|A\|_\infty/c \le \|A\|' \le c\|A\|_\infty.$$

This implies that $\lim_{k\to\infty} \|B - A_k\|_\infty = 0$ if and only if

$$\lim_{k\to\infty} \|B - A_k\|' = 0,$$

showing that our definition of what it means for A_k to converge to B is independent of the particular matrix norm we use. We might even have chosen the matrix norm

$$\|A\|' := \max_{i,j} |A_{ij}| = \max_{x\ne 0} \frac{\|Ax\|_\infty}{\|x\|_1},$$

and so explicitly confirmed that convergence of matrices is entry-wise, i.e., $\lim_{k\to\infty} A_k = B$ if and only if

$$\forall i, j, \quad \lim_{k\to\infty} (A_k)_{ij} = B_{ij}.$$

11.1 For each of the following matrices A, work out A^k for arbitrary $k \in \mathbb{N}$ and, from that, determine directly whether or not the power sequence $A^0, A^1, A^2, \ldots$ converges; if it does, also determine that limit. (i) $A := \alpha\,\mathrm{id}_X$; (ii) $A := \begin{bmatrix} 1/2 & 2^{10} \\ 0 & 1/2 \end{bmatrix}$; (iii) $A := [-\mathbf{e}_1, \mathbf{e}_2]$; (iv) $A = \begin{bmatrix} a & b \\ 0 & c \end{bmatrix}$.

Three interesting properties of the power sequence of a linear map

We have already most of the tools in hand needed to analyze the following three interesting properties that the **power sequence of** A, i.e., the sequence

$$(11.2) \qquad A^0, A^1, A^2, \ldots$$

may have.

Let $A \in L(X)$ with $\dim X < \infty$. Then, for any basis V of X,

$$\widehat{A} := V^{-1}AV$$

is a matrix similar to A, and, for any k,

$$A^k = V\widehat{A}^k V^{-1}.$$

Thus, if we understand the sequence (11.2) for any square *matrix* A, then we understand (11.2) for any $A \in L(X)$ with $\dim X < \infty$.

For this reason, we state here the three interesting properties only for a *matrix* A.

We call the matrix A **power-bounded** in case its power sequence is bounded, i.e., $\sup_k \|A^k\|_\infty < \infty$, i.e., there is a constant c so that, for all k, $\|A^k\|_\infty \le c$.

We call the matrix A **convergent** in case its power sequence converges, i.e., in case, for some matrix B, $B = \lim_{k\to\infty} A^k$.

We call the matrix A **convergent to** 0 in case

$$\lim_{k\to\infty} A^k = 0.$$

See the pages 274ff on the convergence of scalar sequences and, in particular, (17.5)Lemma concerning the scalar sequence $(\zeta^0, \zeta^1, \zeta^2, \ldots)$.

The first property is fundamental in the study of evolutionary (i.e., time-dependent) processes, such as weather or fluid flow. In the *simplest* approximation, the state of the system (be it the weather or waves on the ocean or whatever) at time t is described by some vector $y(t)$, and the state $y(t+\Delta t)$ at some slightly later time $t + \Delta t$ is computed as

$$y(t+\Delta t) = Ay(t),$$

with A some time-independent matrix. Such a process is called **stable** if $\|y(t)\|$ remains bounded for all time regardless of the initial state, $y(0)$, of the system. Since $y(k\Delta t) = A^k y(0)$, the requirement of stability is equivalent to the power-boundedness of A.

The third property is fundamental in the study of iterative processes, as discussed earlier, starting on page 150.

The second property is in between the other two. In other words, we have listed the three properties here in the order of increasing strength: if A is convergent to 0, then it is, in particular, convergent. Again, if A is convergent, then it is, in particular, power-bounded.

Suppose now that x is an eigenvector for A, with corresponding eigenvalue μ. Then $Ax = \mu x$, hence $A^k x = \mu^k x$ for $k = 1, 2, 3, \ldots$. Suppose A is power-bounded. Then, in particular, for some c, we should have $c\|x\|_\infty \ge \|A^k\|_\infty \|x\|_\infty \ge \|A^k x\|_\infty = \|\mu^k x\|_\infty = |\mu|^k \|x\|_\infty$. Since $\|x\|_\infty \ne 0$, this implies that the scalar sequence $(|\mu|^k : k = 1, 2, 3, \ldots)$ must be bounded, hence $|\mu| \le 1$. Since we took an arbitrary eigenvector, we conclude that

$$A \text{ power-bounded} \quad \Longrightarrow \quad \rho(A) \le 1. \tag{11.3}$$

Actually, more is true. Suppose that μ is a *defective* eigenvalue for A, which, to recall, means that

$$\operatorname{null}(A - \mu\,\mathrm{id}) \cap \operatorname{ran}(A - \mu\,\mathrm{id}) \ne \{0\}.$$

In other words, there exists an eigenvector for A belonging to μ of the form $x = (A - \mu\,\mathrm{id})y$. This implies that

$$Ay = x + \mu y.$$

Therefore

$$A^2 y = Ax + \mu Ay = \mu x + \mu(x + \mu y) = 2\mu x + \mu^2 y,$$

hence

$$A^3 y = 2\mu Ax + \mu^2 Ay = 2\mu^2 x + \mu^2(x + \mu y) = 3\mu^2 x + \mu^3 y.$$

By now, the pattern is clear:

$$A^k y = k\mu^{k-1} x + \mu^k y.$$

This also makes clear the difficulty: If $|\mu| = 1$, then

$$\|A^k\|_\infty \|y\|_\infty \ge \|A^k y\|_\infty \ge k\|x\|_\infty - \|y\|_\infty.$$

This shows that A cannot be power-bounded.

We conclude:

(11.4) Proposition: If the matrix A is power-bounded, then, for all $\mu \in \text{spct}(A)$, $|\mu| \le 1$, with equality only if μ is a nondefective eigenvalue for A.

Now we consider the case that A is convergent (hence, in particular, power-bounded). If A is convergent, then, for any eigenvector x with associated eigenvalue μ, the sequence $(\mu^k x : k = 0, 1, 2, \ldots)$ must converge. Since x stays fixed, this implies that the scalar sequence $(\mu^k : k = 0, 1, 2, \ldots)$ must converge. This, to recall, implies that $|\mu| \le 1$ with equality only if $\mu = 1$.

Finally, if A is convergent to 0, then, for any eigenvector x with associated eigenvalue μ, the sequence $(\mu^k x)$ must converge to 0. Since x stays fixed (and is nonzero), this implies that the scalar sequence (μ^k) must converge to 0. This, to recall, implies that $|\mu| < 1$.

Remarkably, these simple necessary conditions just derived, for power-boundedness, convergence, and convergence to 0, are also sufficient; see (11.10)Theorem.

For the proof, we need one more piece of information, namely a better understanding of the distinction between defective and nondefective eigenvalues.

11.2 For each of the following four matrices A, determine whether or not it is (a) power-bounded, (b) convergent, (c) convergent to zero. (i) id_n; (ii) $[1, 1; 0, 1]$; (iii) $[8/9, 10^{10}; 0, 8/9]$; (iv) $-\text{id}_n$.

Splitting off the nondefective eigenvalues

Recall that the scalar μ is called a *defective* eigenvalue for $A \in L(X)$ in case

$$\text{null}(A - \mu\,\text{id}) \cap \text{ran}(A - \mu\,\text{id}) \ne \{0\}.$$

(11.5) Proposition: If M is a set of nondefective eigenvalues of $A \in L(X)$, for some finite-dimensional vector space X, then X has a basis $U = [V, W]$, with V consisting entirely of eigenvectors of A belonging to these nondefective eigenvalues, and W any basis for the subspace $Z := \text{ran}\, p(A)$, with $p(t) := \prod_{\mu \in M}(t - \mu)$.

Further, Z is **A-invariant**, meaning that $A(Z) \subset Z$, hence $A|_Z : Z \to Z : z \mapsto Az$ is a well-defined map on Z, and $\text{spct}(A|_Z) = \text{spct}(A) \backslash M$.

Proof: Since $Ap(A) = p(A)A$, we have

$$AZ = A(\operatorname{ran} p(A)) = \operatorname{ran} Ap(A) = p(A) \operatorname{ran} A \subset \operatorname{ran} p(A) = Z,$$

showing Z to be A-invariant. This implies that $A|_Z : Z \to Z : z \mapsto Az$ is a well-defined linear map on Z.

We claim that X is the direct sum of $\operatorname{null} p(A)$ and $\operatorname{ran} p(A)$, i.e.,

$$X = \operatorname{null} p(A) \dot{+} \operatorname{ran} p(A). \tag{11.6}$$

Since, by (3.23)Dimension Formula, $\dim X = \dim \operatorname{null} p(A) + \dim \operatorname{ran} p(A)$, it is, by (3.33)Proposition, sufficient to prove that

$$\operatorname{null} p(A) \cap \operatorname{ran} p(A) = \{0\}. \tag{11.7}$$

For its proof, we show that we can break up each $x \in X$ into a sum of $\#\mathrm{M}$ components x_μ, $\mu \in \mathrm{M}$, with

$$x_\mu \in \begin{cases} \operatorname{null}(A - \mu \operatorname{id}), & \text{if } x \in \operatorname{null} p(A), \\ \operatorname{ran}(A - \mu \operatorname{id}), & \text{if } x \in \operatorname{ran} p(A), \end{cases}$$

which proves (11.7) since $(A - \mu \operatorname{id}) \cap \operatorname{ran}(A - \mu \operatorname{id}) = \{0\}$ for all $\mu \in \mathrm{M}$. For the breakup, let

$$p_\mu : t \mapsto p(t)/(t - \mu), \qquad \mu \in \mathrm{M},$$

and recall from (5.7) that

$$(p_\mu / p_\mu(\mu) : \mu \in \mathrm{M})$$

is a Lagrange basis for the polynomials of degree $< \#\mathrm{M}$. In particular,

$$1 = \sum_{\mu \in \mathrm{M}} p_\mu / p_\mu(\mu).$$

Hence, with (10.21)Lemma, $\operatorname{id} = \sum_{\mu \in \mathrm{M}} p_\mu(A)/p_\mu(\mu)$ and so, for any $x \in X$,

$$x = \sum_{\mu \in \mathrm{M}} x_\mu,$$

with

$$x_\mu := p_\mu(A) x / p_\mu(\mu)$$

in $\operatorname{null}(A - \mu \operatorname{id})$ in case $x \in \operatorname{null} p(A)$ (since $(A - \mu \operatorname{id}) x_\mu = p(A)x/p_\mu(\mu)$), but also in $\operatorname{ran}(A - \mu \operatorname{id})$ in case also $x \in \operatorname{ran} p(A) \subset \operatorname{ran}(A - \mu \operatorname{id})$, hence then $x_\mu = 0$ since we assumed that each $\mu \in \mathrm{M}$ is not defective. This shows (11.7), hence (11.6).

More than that, we just saw that $x \in \operatorname{null} p(A)$ implies that $x = \sum_\mu x_\mu$ with $x_\mu \in \operatorname{null}(A - \mu\,\mathrm{id})$, all $\mu \in \mathrm{M}$, hence, $\operatorname{null} p(A) \subset \operatorname{ran} V$, with

$$V := [V_\mu : \mu \in \mathrm{M}]$$

and V_μ a basis for $\operatorname{null}(A - \mu\,\mathrm{id})$, all μ. On the other hand, each column of V is in $\operatorname{null} p(A)$, hence also $\operatorname{ran} V \subset \operatorname{null} p(A)$, therefore V is onto $\operatorname{null} p(A)$ and, since it is 1-1 by (10.10)Lemma, it is a basis for $\operatorname{null} p(A)$. Therefore, by (11.6), $U := [V, W]$ is a basis for X for any basis W for $Z = \operatorname{ran} p(A)$.

Finally, let $\nu \in \operatorname{spct}(A)$, hence $Ax = \nu x$ for some $x \neq 0$. Then $p(A)x = p(\nu)x$, hence $x \in \operatorname{ran} p(A) = Z$ in case $p(\nu) \neq 0$, i.e., $\nu \notin \mathrm{M}$, and therefore $\nu \in \operatorname{spct}(A|_Z)$. Otherwise, $p(\nu) = 0$, and then $p(A)x = 0$, i.e., $0 \neq x \in \operatorname{null} p(A)$ hence, by (11.7), x cannot be in $\operatorname{ran} p(A) = Z$, i.e., $\nu \notin \operatorname{spct}(A|_Z)$. This proves that $\operatorname{spct}(A|_Z) = \operatorname{spct}(A) \backslash \mathrm{M}$. □

It follows that the matrix representation for A with respect to this basis $U = [V, W]$ has the simple form

$$U^{-1}AU = \begin{bmatrix} \mathrm{M} & 0 \\ 0 & \widehat{B} \end{bmatrix} := \operatorname{diag}(\mu_1, \ldots, \mu_r, \widehat{B}),$$

with $\mu_1, \ldots, \mu_r$ a sequence taken from M, and $\widehat{B}$ some square matrix, namely $\widehat{B} = W^{-1}AW$.

(11.8) Theorem: Let $A \in L(X)$, with X a finite-dimensional vector space.

(i) If A is diagonalizable, then all its eigenvalues are nondefective, and $X = \dot{+}_{\mu \in \operatorname{spct}(A)} \operatorname{null}(A - \mu\,\mathrm{id})$.

(ii) If $\mathbb{F} = \mathbb{C}$ and all of A's eigenvalues are nondefective, then A is diagonalizable.

Proof: (i) The first part is a restatement of (10.15)Corollary; the second part follows from (3.34)Corollary.

(ii) If none of the eigenvalues of A is defective, then we can choose $\mathrm{M} = \operatorname{spct}(A)$ in (11.5)Proposition, leaving $A|_Z$ as a linear map with an empty spectrum. Hence, if also $\mathbb{F} = \mathbb{C}$, then we know from (10.17)Theorem that $\operatorname{ran} W = \operatorname{dom} A|_Z$ must be trivial, hence V is a basis for X. □

Here is a simple *example*. Let $A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$. Then A maps $x := (1, 1)$ to $(3, 3) = 3x$. Hence, $\mu := 3 \in \operatorname{spct}(A)$. We compute

$$\operatorname{ran}(A - \mu\,\mathrm{id}) = \operatorname{ran} \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} = \operatorname{ran} \begin{bmatrix} -1 \\ 1 \end{bmatrix},$$

since the first column of $(A - \mu\,\mathrm{id})$ is bound and the second is free. This also implies that $\mathrm{null}(A - \mu\,\mathrm{id})$ is one-dimensional, with $V := \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ a basis for it.

It follows, by inspection, that $\mathrm{null}(A - \mu\,\mathrm{id}) \cap \mathrm{ran}(A - \mu\,\mathrm{id}) = \{0\}$ since the only vector of the form $(1, 1)\alpha$ *and* of the form $(-1, 1)\beta$ is the zero vector. Equivalently, the matrix $U := \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$ is 1-1, hence a basis for $\mathbb{R}^2$. Consequently, 3 is a nondefective eigenvalue for A.

Now, what about $A|_Z$, with $Z = \mathrm{ran}(A - \mu\,\mathrm{id})$? In this case, things are very simple since Z is one-dimensional. Since $A(Z) \subset Z$, A must map any $z \in Z$ to a scalar multiple of itself! In particular, since $z = (-1, 1) \in \mathrm{ran}(A - \mu\,\mathrm{id})$, A must map this z into a scalar multiple of itself, and this is readily confirmed by the calculation that A maps z to $-(2, 1) + (1, 2) = z$, i.e., to itself. This shows that z is an eigenvector for A belonging to the eigenvalue 1.

Altogether therefore,

$$AU = [Ax, Az] = [3x, z] = U\,\mathrm{diag}(3, 1),$$

showing that A is actually diagonalizable.

This simple example runs rather differently when we change A to $A := \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}$. Since A is upper triangular, its sole eigenvalue is $\mu = 2$. But $(A - \mu\,\mathrm{id}) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$, and we saw earlier that its range and nullspace have the nontrivial vector $\mathbf{e}_1$ in common. Hence, 2 is a defective eigenvalue for this matrix A.

(11.9) Example: Let $A := [x][y]^{\mathrm{t}}$ with $x, y \in \mathbb{R}^n \backslash 0$. Then $\mathrm{rank}\, A = 1$, hence $\mathrm{ran}\, A = \mathrm{ran}[x]$ is one-dimensional, therefore x is an eigenvector for A. Since $Az = x\ (y^{\mathrm{t}} z)$, we have, in particular,

$$Ax = x\ (y^{\mathrm{t}} x),$$

hence x is an eigenvector for A belonging to the eigenvalue $\mu := y^{\mathrm{t}} x$.

Since A is of rank 1, $\dim \mathrm{null}\, A = n - 1$. Let V be a basis for $\mathrm{null}\, A$, i.e., $V \in L(\mathbb{R}^{n-1}, \mathrm{null}\, A)$ invertible. Then $U := [V, x]$ is 1-1 (hence a basis for $\mathbb{R}^n$) if and only if $x \not\in \mathrm{ran}\, V$, i.e., if and only if $x \not\in \mathrm{null}\, A$.

case $x \not\in \mathrm{ran}\, V$: Then $U = [V, x]$ is a basis for $\mathbb{R}^n$. Consider the representation $\widehat{A} = U^{-1} A U$ for A with respect to this basis: With $V =: [v_1, v_2, \ldots, v_{n-1}]$, we have $Au_j = Av_j = 0$ for $j = 1{:}n{-}1$, therefore

$$\widehat{A}\mathbf{e}_j = 0, \qquad j = 1{:}n{-}1.$$

Further, we have $Ax = x\ (y^{\mathrm{t}}x)$, therefore

$$\widehat{A}\mathbf{e}_n = U^{-1}AU\mathbf{e}_n = U^{-1}Ax = (y^{\mathrm{t}}x)\mathbf{e}_n,$$

(recall that, for any $z \in \mathbb{R}^n$, $U^{-1}z$ provides the coordinates of z with respect to the basis U, i.e., $U(U^{-1}z) = z$). Hence, altogether,

$$\widehat{A} = [0, \ldots, 0, (y^{\mathrm{t}}x)\mathbf{e}_n].$$

In particular, A is diagonalizable, with eigenvalues 0 and $y^{\mathrm{t}}x$.

case $x \in \operatorname{ran} V$: Then $U = [V, x]$ is not a basis for $\mathbb{R}^n$. Worse than that, A is now not diagonalizable. This is due to the fact that, in this case, the eigenvalue 0 for A is *defective*: For, $x \neq 0$ while $Ax = 0$, hence

$$\{0\} \neq \operatorname{ran}(A - 0\,\mathrm{id}) = \operatorname{ran} A = \operatorname{ran}[x] \subset \operatorname{null} A = \operatorname{null}(A - 0\,\mathrm{id}).$$

Therefore $\operatorname{null}(A - 0\,\mathrm{id}) \cap \operatorname{ran}(A - 0\,\mathrm{id}) \neq \{0\}$. □

It is hard to tell just by looking at a matrix whether or not it is diagonalizable. There is one exception: If A is hermitian, i.e., equal to its conjugate transpose, then it is not only diagonalizable, but has an orthonormal basis of eigenvectors, as is shown in the next chapter.

11.3* Prove: *If* $A = \begin{bmatrix} B & C \\ 0 & D \end{bmatrix}$, *with* B *and* D *square matrices, then* $\operatorname{spct}(A) = \operatorname{spct}(B) \cup \operatorname{spct}(D)$. (Hint: Prove first that such a matrix A is invertible if and only if both B and D are invertible.)

11.4 Use Problem 11.3 to determine the spectrum of the matrix $A := \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 5 & 6 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 1 & 2 \end{bmatrix}$.

11.5 (a) Use Problem 11.3 to determine the spectrum of the matrix $A := \begin{bmatrix} 1 & 2 & a \\ 2 & 1 & b \\ 0 & 0 & 3 \end{bmatrix}$.
(b) For which choices of a and b is A not diagonalizable?

11.6* (a) Prove: *If* $A = V \operatorname{diag}(\mu, B)V^{-1}$ *for some invertible* B*, scalar* μ *and matrix* B*, then* μ *is an eigenvalue of* A^{t} *with eigenvector the first row of* V^{-1}.

Three interesting properties of the power sequence of a linear map: The sequel

(11.10) Theorem: Let $A \in \mathbb{C}^{n \times n}$. Then:

(i) A is power-bounded iff, for all $\mu \in \operatorname{spct}(A)$, $|\mu| \leq 1$, with $|\mu| = 1$ only if μ is not defective.

(ii) A is convergent iff, for all $\mu \in \operatorname{spct}(A)$, $|\mu| \leq 1$, with $|\mu| = 1$ only if μ is not defective and $\mu = 1$.

(iii) A is convergent to 0 iff $\rho(A) < 1$.

Proof: We only have to prove the implications '$\Longleftarrow$', since we proved all the implications '$\Longrightarrow$' in an earlier section (see pages 172ff).

We begin with (iii). Since A is a matrix over the complex scalars, we know from (10.33)Corollary that, for any $\varepsilon > 0$, we can find an upper triangular matrix B_ε similar to A and with all off-diagonal entries less than ε in absolute value. This means, in particular, that $A = VB^{(\varepsilon)}V^{-1}$ for some (invertible) matrix V, hence, for any k, $A^k = V(B^{(\varepsilon)})^k V^{-1}$, therefore,

$$\|A^k\|_\infty \le \|V\|_\infty \|B^{(\varepsilon)}\|_\infty^k \|V^{-1}\|_\infty.$$

We compute

$$\|B^{(\varepsilon)}\|_\infty = \max_i \sum_j |(B^{(\varepsilon)})_{ij}| \le \max_i |(B^{(\varepsilon)})_{ii}| + (n-1)\varepsilon,$$

since each of those sums involves $n-1$ off-diagonal entries and each such entry is less than ε in absolute value. Further, $B^{(\varepsilon)}$ is upper triangular and similar to A, hence

$$\max_i |(B^{(\varepsilon)})_{ii}| = \max\{|\mu| : \mu \in \operatorname{spct}(A)\} = \rho(A).$$

By assumption, $\rho(A) < 1$. This makes it possible to choose ε positive yet so small that $\rho(A) + (n-1)\varepsilon < 1$. With this choice, $\|B^{(\varepsilon)}\|_\infty < 1$, hence $\lim_{k\to\infty} \|B^{(\varepsilon)}\|_\infty^k = 0$. Therefore, since $\|V\|_\infty$ and $\|V^{-1}\|_\infty$ stay fixed throughout, also $\|A^k\|_\infty \to 0$ as $k \to \infty$. In other words, A is convergent to 0.

With this, we are ready also to handle (i) and (ii). Both assume that all eigenvalues of A of modulus 1 are nondefective. By (11.5)Proposition, this implies the existence of a basis $U = [V, W]$ for $\mathbb{C}^n$ so that V consists of eigenvectors of A belonging to eigenvalues of modulus 1, while $Z := \operatorname{ran} W$ is A-invariant and $A|_Z$ has only eigenvalues of modulus < 1. In particular, $AV = V\mathrm{M}$ for some diagonal matrix M with all diagonal entries of modulus 1, and $AW = WB$ for some matrix B with $\operatorname{spct}(B) = \operatorname{spct}(A|_Z)$, hence $\rho(B) < 1$. Consequently, for any k,

$$A^k U = A^k[V, W] = [A^k V, A^k W] = [V\mathrm{M}^k, WB^k] = U \operatorname{diag}(\mathrm{M}^k, B^k).$$

In other words,

$$A^k = U \operatorname{diag}(\mathrm{M}^k, B^k) U^{-1}.$$

Therefore, $\|A^k\|_\infty \le \|U\|_\infty \max\{\|\mathrm{M}\|_\infty^k, \|B^k\|_\infty\}\|U^{-1}\|_\infty$, and this last expression converges to $\|U\|_\infty \|U^{-1}\|_\infty$ since $\|\mathrm{M}\|_\infty = 1$ while $\|B^k\|_\infty \to 0$, by (iii). Since any convergent sequence is bounded, this implies that also the sequence $(\|A^k\|_\infty)$ must be bounded, hence we have finished the proof of (i).

Assume now, in addition, as in (ii) that all eigenvalues of A of modulus 1 are actually equal to 1. Then $\mathrm{M} = \mathrm{id}$, and so, $\lim_{k\to\infty} A^k = C := U \operatorname{diag}(\mathrm{M}, 0) U^{-1}$ since $A^k - C = U \operatorname{diag}(0, B^k) U^{-1}$, hence

$$\|A^k - C\|_\infty \le \|U\|_\infty \|B^k\|_\infty \|U^{-1}\|_\infty \le \mathrm{const}\|B^k\|_\infty \to 0$$

as $k \to \infty$. □

(11.11) Example: Here is a concrete example, chosen for its simplicity.

Let $A = \begin{bmatrix} 1 & 1 \\ 0 & \alpha \end{bmatrix}$. Then $\operatorname{spct}(A) = \{1, \alpha\}$. In particular, A is diagonalizable if $\alpha \neq 1$ (by (10.11)Corollary) since then A has two eigenvalues. On the other hand, if $\alpha = 1$, then A is not diagonalizable since it then looks like $\mathrm{id}_2 + N$, with $N := [\mathbf{0}, \mathbf{e}_1]$ the simplest example of a non-diagonalizable matrix. Also, in the latter case, the sole eigenvalue, 1, is certainly defective since $\mathbf{e}_1$ is both in $\operatorname{null}(A - \mathrm{id})$ and in $\operatorname{ran}(A - \mathrm{id})$.

Also,

$$A^k = \begin{bmatrix} 1 & 1 + \alpha + \cdots + \alpha^{k-1} \\ 0 & \alpha^k \end{bmatrix} = \begin{cases} \begin{bmatrix} 1 & \frac{1-\alpha^k}{1-\alpha} \\ 0 & \alpha^k \end{bmatrix} & \text{if } \alpha \neq 1; \\ \begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix} & \text{otherwise.} \end{cases}$$

We see that A is power-bounded whenever $|\alpha| \leq 1$ *except* when $\alpha = 1$, i.e., except when there is a defective absolutely largest eigenvalue.

Further, A is convergent iff $|\alpha| < 1$, i.e., if, in addition, the sole eigenvalue of size 1 is equal to 1 and is nondefective. □

The power method

The simple background for the success of the **power method** is the following consequence of (11.10)Theorem (ii).

(11.12) Proposition: If A has just one eigenvalue μ of absolute value $\rho(A)$ and μ is nondefective, then, for almost any x and almost any y, the sequence

$$A^k x / (y^{\mathrm{c}} A^k x), \qquad k = 1, 2, \ldots$$

converges to an eigenvector of A belonging to that absolutely maximal eigenvalue μ. In particular, for almost any vector y, the ratio

$$y^{\mathrm{c}} A^{k+1} x / y^{\mathrm{c}} A^k x$$

converges to μ.

Proof: By assumption, there is (by (11.5)Proposition) a basis $U := [V, W]$, with V a basis for the space $\operatorname{null}(A - \mu\,\mathrm{id})$ comprising all eigenvectors of A belonging to that absolutely largest eigenvalue μ of A, and $B := A|_{\operatorname{ran} W}$ having all its eigenvalues $< |\mu|$ in absolute value. This implies that $\rho(B/\mu) < 1$. Therefore, for any $x =: [V, W](\mathbf{a}, \mathbf{b})$,

$$A^k x = \mu^k V\mathbf{a} + B^k W\mathbf{b} = \mu^k \left(V\mathbf{a} + (B/\mu)^k W\mathbf{b} \right)$$

and $(B/\mu)^k W\mathbf{b} \to \mathbf{0}$ as $k \to \infty$. Consequently, for any y,

$$\frac{y^{\mathrm{c}} A^{k+1} x}{y^{\mathrm{c}} A^k x} = \frac{\mu^{k+1}(y^{\mathrm{c}} V\mathbf{a} + y^{\mathrm{c}}(B/\mu)^{k+1} W\mathbf{b})}{\mu^k (y^{\mathrm{c}} V\mathbf{a} + y^{\mathrm{c}}(B/\mu)^k W\mathbf{b})} = \mu \frac{y^{\mathrm{c}} V\mathbf{a} + y^{\mathrm{c}}(B/\mu)^{k+1} W\mathbf{b}}{y^{\mathrm{c}} V\mathbf{a} + y^{\mathrm{c}}(B/\mu)^k W\mathbf{b}}$$

converges to μ as $k \to \infty$ provided $y^{\mathrm{c}} V\mathbf{a} \neq 0$. □

Note that the speed with which $y^{\mathrm{c}} A^{k+1} x / y^{\mathrm{c}} A^k x$ converges to μ depends on the speed with which $(B/\mu)^k W\mathbf{b} \to 0$ as $k \to \infty$, hence, ultimately, on $\rho(B/\mu)$.

In the **scaled power method**, one would, instead, consider the sequence

$$x_{k+1} := A(x_k/\|x_k\|), \quad k = 0, 1, \ldots,$$

or, more simply, the sequence

$$x_{k+1} := A(x_k/y^{\mathrm{t}} x_k), \quad k = 0, 1, \ldots.$$

The power method is at the heart of good algorithms for the calculation of eigenvalues. In particular, the standard algorithm, i.e., the QR method, can be interpreted as a (very sophisticated) variant of the power method.

11.7 Using `MATLAB` if really necessary, try out the Power method on the following matrices A, starting at the specified vector x, and discuss success or failure. (Note: You can always use `eig(A)` to find out what the absolutely largest eigenvalue of A is (as well as some eigenvector for it), hence can tell whether or not the power method is working for you. If it isn't, identify the source of failure.) (a) $A = \begin{bmatrix} 0 & .2 & .2 & .3 \\ .2 & 0 & .2 & .3 \\ .5 & .4 & 0 & .4 \\ .3 & .4 & .6 & 0 \end{bmatrix}$, $\mathbf{x} = (1, 1, 1, 1)$; (b) $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$, $\mathbf{x} = (1, -1)$; (c) $A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$, $\mathbf{x} = \mathbf{e}_1$; (d) $A = \begin{bmatrix} 4 & 1 & -1 \\ 2 & 5 & -2 \\ 1 & 1 & 2 \end{bmatrix}$, $\mathbf{x} = (1, -2, -1)$.

11.8 T/F

(a) If the matrix A of order n has n eigenvalues, then none of its eigenvalues is defective.

(b) If, for some sequence $(\mathbf{x}_n : n \in \mathbb{N})$ of m-vectors, $\lim_{n\to\infty} \|\mathbf{x}_n\|_2 = 0$, then $\lim_{n\to\infty} \|\mathbf{x}_n\| = 0$ for any norm $\|\cdot\|$ on $\mathbb{F}^m$.

(c) If all the eigenvalues of A are < 1, then $\lim_{k\to\infty} A^k \to 0$.

(d) If all the eigenvalues of A are ≤ 1 in absolute value, then A is power-bounded.

(e) If $p(A)x = 0$ for some polynomial p, $A \in L(X)$ and $x \in X\backslash\{0\}$, then every eigenvalue of A is a zero of p.

12 Canonical forms

Canonical forms exhibit essential aspects of a linear map. Of the three discussed in this chapter, only the Schur form has practical significance. But the mathematics leading up to the other two is too beautiful to be left out.

The only result from this chapter used later in this book is the spectral theorem for hermitian matrices; see (12.2) Corollary.

The Schur form

The discussion of the powers A^k of A used crucially the fact that any square matrix is similar to an upper triangular matrix. The argument we gave there for this fact is due to I. Schur, who used a refinement of it to show that the basis V for which $V^{-1}AV$ is upper triangular can even be chosen to be *unitary* or *orthonormal*, i.e., so that

$$V^{\mathrm{c}}V = \mathrm{id}.$$

(12.1) Schur's theorem: Every $A \in L(\mathbb{C}^n)$ is **unitarily similar** to an upper triangular matrix, i.e., there exists a unitary basis U for $\mathbb{C}^n$ so that $\widehat{A} := U^{-1}AU = U^{\mathrm{c}}AU$ is upper triangular.

Proof: Simply repeat the proof of (10.32)Theorem, with the following modifications: Normalize the eigenvector $\mathbf{u}_1$, i.e., make it have (Euclidean) length 1, then extend it to an o.n. basis for $\mathbb{C}^n$ (as can always be done by applying Gram-Schmidt to an arbitrary basis $[\mathbf{u}_1, \ldots]$ for $\mathbb{C}^n$). Also, observe that unitary similarity is also an equivalence relation since the product of unitary matrices is again unitary. Finally, if W is unitary, then so is $\mathrm{diag}(1, W)$. □

Here is one of the many consequences of Schur's theorem. It concerns **hermitian** matrices, i.e., matrices A for which $A^{\mathrm{c}} = A$. By Schur's theorem, such a matrix, like any other matrix, is unitarily similar to an upper triangular matrix, i.e., for some unitary matrix U , $\widehat{A} := U^{\mathrm{c}}AU$ is upper triangular. On the other hand, for any matrix A and any unitary matrix U,

$$(U^{\mathrm{c}}AU)^{\mathrm{c}} = U^{\mathrm{c}}(A^{\mathrm{c}})U.$$

In other words: if $\widehat{A}$ is the matrix representation for A with respect to a *unitary* basis, then $\widehat{A}^{\mathrm{c}}$ is the matrix representation for A^{c} with respect to the very same basis. For our hermitian matrix A with its upper triangular matrix representation $\widehat{A} = U^{\mathrm{c}}AU$ with respect to the unitary basis U, this means that also $\widehat{A}^{\mathrm{c}} = \widehat{A}$, i.e., that the *upper* triangular matrix $\widehat{A}$ is also *lower triangular* and that its diagonal entries are all real. This proves the hard part of the following remarkable

(12.2) Corollary: A matrix $A \in \mathbb{C}^n$ is hermitian if and only if it is unitarily similar to a real diagonal matrix.

Proof: We still have to prove that if $\widehat{A} := U^{\mathrm{c}}AU$ is real and diagonal for some unitary U, then A is necessarily hermitian. But that follows at once from the fact that then $\widehat{A}^{\mathrm{c}} = \widehat{A}$, therefore $A^{\mathrm{c}} = (U\widehat{A}U^{\mathrm{c}})^{\mathrm{c}} = U\widehat{A}^{\mathrm{c}}U^{\mathrm{c}} = U\widehat{A}U^{\mathrm{c}} = A$. □

12.1 Verify that the symmetric matrix $\begin{bmatrix} 2\mathrm{i} & 1 \\ 1 & 0 \end{bmatrix}$ is not diagonalizable.

A slightly more involved argument makes it possible to characterize all those matrices that are unitarily similar to a diagonal matrix (real or not). Such a matrix has enough eigenvectors to make up an entire orthonormal basis from them. Here are the details.

Start with the observation that diagonal matrices commute with one another. Thus, if $\widehat{A} := U^{\mathrm{c}}AU$ is diagonal, then

$$A^{\mathrm{c}}A = (U\widehat{A}^{\mathrm{c}}U^{\mathrm{c}})(U\widehat{A}U^{\mathrm{c}}) = U\widehat{A}^{\mathrm{c}}\widehat{A}U^{\mathrm{c}} = U\widehat{A}\widehat{A}^{\mathrm{c}}U^{\mathrm{c}} = (U\widehat{A}U^{\mathrm{c}})(U\widehat{A}^{\mathrm{c}}U^{\mathrm{c}}) = AA^{\mathrm{c}},$$

hence having $A^{\mathrm{c}}A = AA^{\mathrm{c}}$ is a necessary condition for A to be unitarily similar to a diagonal matrix. Remarkably, this condition is sufficient as well. Note that this condition can be directly tested by computing the two products and comparing them. It constitutes the only criterion for the diagonalizability of a matrix available that can be tested for by finitely many calculations. Not surprisingly, matrices with this property are very convenient and have, correspondingly, been given a very positive label. They are called **normal**. (Another label might have been **boring**.)

One way to prove that normal matrices are unitarily similar to a diagonal matrix is by way of a refinement of Schur's theorem: It is possible to find a unitary basis that simultaneously upper-triangularizes two matrices A and B provided A and B **commute**, i.e., provided $AB = BA$. This is due to the fact that commuting matrices have some eigenvector in common.

Assuming this refinement of Schur's Theorem (cf. (12.5)Theorem below), one would obtain, for a given normal matrix A, a unitary basis U so that both $U^{\mathrm{c}}AU$ and $U^{\mathrm{c}}A^{\mathrm{c}}U$ are upper triangular. Since one of these is the conjugate transpose of the other, they must both be diagonal. This finishes the proof of

(12.3) Theorem: A matrix $A \in \mathbb{C}^n$ is unitarily similar to a diagonal matrix if and only if $AA^{\mathrm{c}} = A^{\mathrm{c}}A$.

Now for the proof of the Refined Schur's Theorem. Since the proof of Schur's theorem rests on eigenvectors, it is not surprising that a proof of its refinement rests on the following

(12.4) Lemma: If $A, B \in \mathbb{C}^n$ commute, then there exists a vector that is eigenvector for both of them.

Proof: Let $\mathbf{x}$ be an eigenvector for A, $A\mathbf{x} = \mathbf{x}\mu$ say, and let $p = p_{B,\mathbf{x}}$ be the minimal polynomial for B at $\mathbf{x}$. Since $\mathbf{x} \neq 0$, p has zeros. Let ν be one such and set $p =: (\cdot - \nu)q$. Since $\mathbb{F} = \mathbb{C}$, we know that $\mathbf{v} := q(B)\mathbf{x}$ is an eigenvector for B (for the eigenvalue ν). But then, since $AB = BA$, we also have $Aq(B) = q(B)A$, therefore

$$A\mathbf{v} = Aq(B)\mathbf{x} = q(B)A\mathbf{x} = q(B)\mathbf{x}\mu = \mathbf{v}\mu,$$

showing that our eigenvector $\mathbf{v}$ for B is also an eigenvector for A. □

(12.5) Refined Schur's Theorem: For every $A, B \in L(\mathbb{C}^n)$ that commute, there exists a unitary basis U for $\mathbb{C}^n$ so that both $U^{\mathrm{c}}AU$ and $U^{\mathrm{c}}BU$ are upper triangular.

Proof: This is a further refinement of the proof of (10.32)Theorem. The essential step in that proof was to come up with some eigenvector for A which was then extended to a basis, well, to an o.n. basis U for the proof of

Schur's Theorem. Therefore, to have U simultaneously upper-triangularize both A and B, all that's needed is (i) to observe that, by (12.4)Lemma, we may choose $\mathbf{u}_1$ to be a (normalized) eigenvector of A *and* B since, by assumption, $AB = BA$; and (ii) verify that the submatrices A_1 and B_1 obtained in the first step again commute (making it possible to apply the induction hypothesis to them). Here is the verification of this latter fact:

Assuming the eigenvalue of B corresponding to the eigenvector u_1 to be ν, we have

$$U^{\mathrm{c}}AU = \begin{bmatrix} \mu & C \\ 0 & A_1 \end{bmatrix} \qquad U^{\mathrm{c}}BU = \begin{bmatrix} \nu & D \\ 0 & B_1 \end{bmatrix}.$$

Therefore

$$\begin{aligned} \begin{bmatrix} \mu\nu & \mu D + CB_1 \\ 0 & A_1B_1 \end{bmatrix} &= \begin{bmatrix} \mu & C \\ 0 & A_1 \end{bmatrix} \begin{bmatrix} \nu & D \\ 0 & B_1 \end{bmatrix} \\ &= U^{\mathrm{c}}AUU^{\mathrm{c}}BU = U^{\mathrm{c}}ABU = U^{\mathrm{c}}BAU \\ &= U^{\mathrm{c}}BU\, U^{\mathrm{c}}AU = \begin{bmatrix} \nu\mu & \nu C + DA_1 \\ 0 & B_1A_1 \end{bmatrix}, \end{aligned}$$

hence also A_1 and B_1 commute. □

The primary decomposition

The following analysis goes back to Frobenius and could be viewed as a first step toward a finest A-invariant direct sum decomposition, a.k.a. the Jordan form, though the Jordan form is deduced in the next section without any reference to this section. We give the analysis here in the more general situation when the scalar field $\mathbb{F}$ may not be algebraically closed.

The 'primary decomposition' refers to the following facts (taken for granted here). The ring Π of (univariate) polynomials over the field $\mathbb{F}$ is a **unique factorization domain**. This means that each monic polynomial can be written in exactly one way (up to order of the factors) as a product of **irreducible** polynomials, i.e., monic polynomials that have no proper factors. Here, p is called a **proper factor of** q if (i) $0 < \deg p < \deg q$, and (ii) $q = hp$ for some polynomial h.

If $\mathbb{F} = \mathbb{C}$ (or any other algebraically closed field), then each such irreducible polynomial is a monic linear polynomial, i.e., of the form $(\cdot - \mu)$ for some scalar μ. Otherwise, irreducible polynomials may well be of higher than first degree. In particular, if $\mathbb{F} = \mathbb{R}$, then an irreducible polynomial may be of second degree, like the polynomial $()^2 + 1$, but no irreducible polynomial would be of higher degree than that.

The irreducible polynomials are the 'primes' in the 'ring' Π, hence the above-mentioned unique factorization is one into powers of 'primes', or a **prime factorization**.

To obtain the 'primary decomposition' of the vector space X with respect to the linear map $A \in L(X)$, it is convenient to start with the set

$$\mathcal{N}_A := \{p \in \Pi : \operatorname{null} p(A) \neq \{0\}\}$$

of all polynomials p for which $p(A)$ fails to be invertible. This set is not trivial, meaning that it contains more than just the zero polynomial if, as we continue to assume, $\dim X < \infty$, since then

$$\forall x \in X, \quad p_{A,x} \in \mathcal{N}_A, \tag{12.6}$$

with $p_{A,x}$ the *minimal polynomial for A at x*, which, to recall, is the monic polynomial p of smallest degree for which $p(A)x = 0$.

Call an element of $\mathcal{N}_A$ **minimal** if it is monic and none of its proper factors is in $\mathcal{N}_A$, and let

$$\mathcal{Q}_A$$

be the collection of all minimal elements of $\mathcal{N}_A$.

The set $\mathcal{Q}_A$ is not empty since $\mathcal{N}_A$ is not empty, and is closed under multiplication by a scalar, hence contains a monic polynomial of smallest degree. Any $q \in \mathcal{Q}_A$ is necessarily *irreducible*, since, otherwise, it would be the product of certain polynomials p with $p(A)$ 1-1, hence also $q(A)$ would be 1-1.

(12.7) Lemma: Let p be a product of elements of $\mathcal{Q}_A$,

$$p =: \prod_{q \in \mathcal{Q}'_A} q(A)^{d_q}$$

say, with $d_q \in \mathbb{N}$ and $\mathcal{Q}'_A$ a finite subset of $\mathcal{Q}_A$. Then,

$$X_p := \operatorname{null} p(A) = \mathop{\dot{+}}_{q \in \mathcal{Q}'_A} \operatorname{null} q(A)^{d_q}, \tag{12.8}$$

i.e., $X_p = \operatorname{null} p(A)$ is the direct sum of the spaces

$$Y_q := \operatorname{null} q(A)^{d_q}.$$

In other words (by (3.33)Proposition), with V_q a basis for Y_q,

$$V_p := [V_q : q \in \mathcal{Q}'_A]$$

is a basis for X_p.

Proof: There is nothing to prove if $\mathcal{Q}'_A$ has just one element. So, assume that $\#\mathcal{Q}'_A > 1$, and consider the set

$$\mathcal{I} := \sum_{q\in\mathcal{Q}'_A} (p/q^{d_q})\Pi := \{\sum_{q\in\mathcal{Q}'_A} (p/q^{d_q})p_q : p_q \in \Pi\}$$

of all polynomials writable as a weighted sum of the polynomials

$$(12.9) \qquad p/q^{d_q} = \prod_{g\in\mathcal{Q}'_A\backslash q} g^{d_g}, \quad q \in \mathcal{Q}'_A,$$

with polynomial (rather than just scalar) weights. This set is a polynomial **ideal**, meaning that it is closed under addition, as well as under multiplication by polynomials. More than that (see page 279), $\mathcal{I} = \Pi\, q^*$, with q^* the monic polynomial of smallest degree in $\mathcal{I}$. In other words, the monic polynomial q^* is a factor of every $q \in \mathcal{I}$, in particular of every p/q^{d_q} with $q \in \mathcal{Q}'_A$. But these polynomials have no proper factor in common because each $q \in \mathcal{Q}'_A$ is irreducible. Therefore, q^* is necessarily the monic polynomial of degree 0, i.e., $q^* = ()^0$.

It follows that

$$()^0 = \sum_{q\in\mathcal{Q}'_A} (p/q^{d_q})h_q$$

for certain polynomials h_q. This implies that, for the corresponding linear maps

$$(12.10) \qquad P_q : X_p \to X_p : y \mapsto (p/q^{d_q})(A)h_q(A)y, \quad q \in \mathcal{Q}'_A,$$

(well-defined since $X_p = \operatorname{null} p(A)$ is $r(A)$-invariant for any $r \in \Pi$; see Problem 12.2) we have

$$(12.11) \qquad \operatorname{id}_{X_p} = \sum_q P_q.$$

Also, for $q \neq g$, $P_qP_g = s(A)p(A)$ for some $s \in \Pi$, by (12.10) and (12.9). Therefore, $P_qP_g = 0$ for $g \neq q$. Therefore also

$$P_q = P_q \operatorname{id}_{X_p} = P_q(\sum_g P_g) = \sum_g P_qP_g = P_qP_q.$$

This shows that each P_q is a linear projector, hence, by (5.12)Proposition, that X_p is the direct sum of the ranges of the P_q. It remains to show that

$$(12.12) \qquad \operatorname{ran} P_q = Y_q = \operatorname{null} q(A)^{d_q}.$$

It is immediate that $\operatorname{ran} P_q \subset Y_q \subset X_p$. With that, $Y_q \subset \operatorname{null} P_g$ for all $g \in \mathcal{Q}'_A\backslash q$, and this implies (12.12), by (12.11). □

Now let $p = p_A$ be the **minimal (annihilating) polynomial for** A, meaning the monic polynomial p of smallest positive degree for which $p(A) = 0$.

To be sure, there is such a polynomial since X is finite-dimensional, hence so is $L(X)$ (by (3.22)Corollary), therefore $[A^r : r = 0{:}\dim L(X)]$ must fail to be 1-1, i.e., there must be some $\mathbf{a}$ for which

$$p(A) := \sum_{j \le \dim L(X)} a_j A^j = 0,$$

yet $a_j \neq 0$ for some $j > 0$, hence the set of all annihilating polynomials of positive degree is not empty, therefore must have an element of minimal degree, and it will remain annihilating and of that degree if we divide it by its leading coefficient.

By the minimality of p_A, every proper factor of p_A is necessarily in $\mathcal{N}_A$. Hence p_A is of the form

$$p_A = \prod_{q \in \mathcal{Q}'_A} q^{d_q}$$

for some positive integers d_q and some $\mathcal{Q}'_A \subset \mathcal{Q}_A$. (In fact, it is immediate from (12.7)Lemma that necessarily $\mathcal{Q}'_A = \mathcal{Q}_A$, but we don't need that here.) This gives, with (12.7)Lemma, the **primary decomposition for X wrto A:**

$$X = \dot{+}_q \operatorname{null} q(A)^{d_q}, \tag{12.13}$$

and each d_q is the smallest natural number for which

$$\operatorname{null} q(A)^{d_q} = \bigcup_r \operatorname{null} q(A)^r. \tag{12.14}$$

Indeed, from (12.13), every $x \in X$ is uniquely writable as $x = \sum_g x_g$ with $x_g \in \operatorname{null} g(A)^{d_g}$, all $g \in \mathcal{Q}'_A$, and, since each $\operatorname{null} g(A)^{d_g}$ is A-invariant, we therefore have $q(A)^r x = \sum_g q(A)^r x_g = 0$ if and only if $q(A)^r x_g = 0$ for all $g \in \mathcal{Q}'_A$. However, $\operatorname{null} q(A) \subset \operatorname{null} q(A)^{d_q}$ hence, by (12.13), $\operatorname{null} q(A) \cap \operatorname{null} g(A)^{d_g} = \{0\}$ for every $g \in \mathcal{Q}'_A \backslash q$, therefore, for every $g \in \mathcal{Q}'_A \backslash q$, $q(A)$ maps $\operatorname{null} g(A)^{d_g}$ 1-1 into itself, hence $q(A)^r x_g = 0$ if and only if $x_g = 0$. Therefore, altogether, $x \in \operatorname{null} q(A)^r$ if and only if $x = x_q \in \operatorname{null} q(A)^{d_q} \cap \operatorname{null} q(A)^r$, thus proving (12.14). If now $\operatorname{null} q(A)^r = \operatorname{null} q(A)^{d_q}$ for some $r < d_q$, then already $p := p_A / q^{d_q - r}$ would annihilate A, contradicting p_A's minimality.

If $\mathbb{F} = \mathbb{C}$, then each q is of the form $(\cdot - \mu_q)$ for some scalar μ_q and, correspondingly,

$$X = \dot{+}_q \operatorname{null}(A - \mu_q \operatorname{id})^{d_q}.$$

In particular, $A - \mu_q \operatorname{id}$ is nilpotent on

$$Y_q := \operatorname{null}(A - \mu_q \operatorname{id})^{d_q},$$

with degree of nilpotency equal to d_q. Since

$$A = \mu_q \operatorname{id} + (A - \mu_q \operatorname{id}),$$

it follows that

$$\begin{aligned} \exp(tA) &= \exp(t\mu_q \operatorname{id}) \exp(t(A - \mu_q \operatorname{id})) \\ &= \exp(t\mu_q) \sum_{r<d_q} t^r (A - \mu_q \operatorname{id})^r / r! \quad \text{on} \quad Y_q, \end{aligned} \tag{12.15}$$

thus providing a very helpful detailed description of the solution $\mathbf{y} : t \mapsto \exp(tA)\mathbf{c}$ to the first-order ODE $\mathbf{y}'(t) = A\mathbf{y}(t)$, $\mathbf{y}(0) = \mathbf{c}$, introduced in (10.4).

12.2* Prove: *For every $r, p \in \Pi$ and every $A \in L(X)$, $r(A)(\operatorname{null} p(A)) \subset \operatorname{null} p(A)$*, i.e., $\operatorname{null}(A)$ is $r(A)$-invariant.

12.3 Assume that $p, q \in \mathcal{N}_A$ with $\operatorname{null} p(A) \cap \operatorname{null} q(A) \neq \{0\}$. Prove that p and q have a common factor of degree > 0, hence that $p = q$ in case they are both irreducible.

12.4 A subset F of the vector space $X := C^{(1)}(\mathbb{R})$ of continuously differentiable functions is called D**-invariant** if the derivative Df of any $f \in F$ is again in F.

Prove: *Any finite-dimensional D-invariant linear subspace Y of $C^{(1)}(\mathbb{R})$ is necessarily in the nullspace of a constant-coefficient ordinary differential operator*, i.e., an operator of the form $p(D)$ for some polynomial p.

It follows that Y is spanned by certain **exponential polynomials**, i.e., functions of the form $t \mapsto q(t)\exp(\xi t)$ for certain polynomials q and scalars ξ, the latter being the roots of p.

The Jordan form

The Jordan form is the result of the search for the 'simplest' matrix representation for $A \in L(X)$ for some n-dimensional vector space X. It starts off from the following observation.

Suppose X is the direct sum

$$X = Y_1 \dot{+} Y_2 \dot{+} \cdots \dot{+} Y_r \tag{12.16}$$

of r linear subspaces, each of which is A-invariant. Then

$$\operatorname{spct}(A) = \bigcup_j \operatorname{spct}(A|_{Y_j}).$$

More than that, with V_j a basis for Y_j, we have $AV_j \subset \operatorname{ran} V_j$, all j. This implies that the coordinate vector of any column of AV_j with respect to the basis $V := [V_1, \ldots, V_r]$ for X has nonzero entries only corresponding to columns of V_j, and these possibly nonzero entries can be found as the corresponding column in the matrix $V_j^{-1}AV_j$. Consequently, the matrix representation $\widehat{A} = V^{-1}AV$ for A with respect to the basis V is block-diagonal, i.e., of the form

$$\widehat{A} = \operatorname{diag}(V_j^{-1}AV_j : j = 1{:}r) = \begin{bmatrix} V_1^{-1}AV_1 & & \\ & \ddots & \\ & & V_r^{-1}AV_r \end{bmatrix}.$$

The smaller we can make the A-invariant summands Y_j, the simpler and more helpful is our overall description $\widehat{A}$ of the linear map A. Of course, the smallest possible A-invariant subspace of X is the trivial subspace, but it would not contribute any columns to V, hence we will assume from now on that our A-invariant direct sum decomposition (12.16) is **proper**, meaning that none of its summands Y_j is trivial.

With that, each Y_j has dimension ≥ 1, hence is as small as possible if it is 1-dimensional, $Y_j = \operatorname{ran}[v_j]$ say, for some nonzero v_j. In this case, A-invariance says that Av_j must be a scalar multiple of v_j, hence v_j is an eigenvector for A, and the sole entry of the matrix $[v_j]^{-1}A[v_j]$ is the corresponding eigenvalue for A.

Thus, at best, each Y_j is 1-dimensional, hence V consists entirely of eigenvectors for A, i.e., A is diagonalizable. Since we know that not every matrix is diagonalizable, we know that this best situation cannot always be attained. But we can try to make each Y_j as small as possible, in the following way.

```
(12.17) Jordan Algorithm:
input: $X$ $n$-dimensional vector space, $A \in L(X)$.
$\mathcal{Y} \leftarrow \{X\}$
while $\exists Z_1 \dotplus Z_2 \in \mathcal{Y}$ with both $Z_j$ nontrivial and $A$-invariant, do:
        replace $Z_1 \dotplus Z_2$ in $\mathcal{Y}$ by $Z_1$ and $Z_2$.
endwhile
output: the proper $A$-invariant direct sum decomposition $X = \dotplus_{Y \in \mathcal{Y}} Y$.
```

At all times, the elements of $\mathcal{Y}$ form a proper direct sum decomposition for X. Hence

$$\#\mathcal{Y} \le \sum_{Y \in \mathcal{Y}} \dim Y = \dim X = n.$$

Since each pass through the **while**-loop increases $\#\mathcal{Y}$ by 1, the algorithm must terminate after at most $n - 1$ steps.

Now consider any particular Y in the collection $\mathcal{Y}$ output by the algorithm. It is, by construction, not the direct sum of two proper A-invariant spaces, a fact to be used twice in the arguments to follow. However, Y is a nontrivial A-invariant subspace. Hence, with the assumption that $\mathbb{F} = \mathbb{C}$, we know that $A|_Y : Y \to Y : y \mapsto Ay$ is a linear map with some eigenvalue, μ say. This implies that the linear map

$$N : Y \to Y : y \mapsto (A - \mu\,\mathrm{id})y$$

is well-defined and has a nontrivial nullspace.

Claim 1: For some $y \in Y$ and some $q \in \mathbb{N}$, $N^{q-1}y \neq 0 = N^q y$.

Proof: Indeed, since $\operatorname{null} N \neq \{0\}$, this holds, e.g., for $q = 1$ and $y \in \operatorname{null} N\backslash 0$. □

Claim 2: For any y and q as in Claim 1, there is (see Problem 5.4) $\lambda \in Y'$ with $\lambda N^{q-1}y \neq 0$ and, for any such λ, $Y = \operatorname{null}\Lambda^{\mathrm{t}} \dot{+} \operatorname{ran} V$, with $\Lambda := [\lambda N^{i-1} : i = 1{:}q]$, and $V := [N^{q-j}y : j = 1{:}q]$ 1-1.

Proof: The Gramian matrix $\Lambda^{\mathrm{t}}V = (\lambda N^{i-1}N^{q-j}y : i, j = 1{:}q)$ is square and upper triangular, with all diagonal entries equal to $\lambda N^{q-1}y \neq 0$, hence $\Lambda^{\mathrm{t}}V$ is invertible. This implies that V is 1-1 and, by (5.9), that Y is the direct sum of $\operatorname{null}\Lambda^{\mathrm{t}}$ and $\operatorname{ran} V$. □

Claim 3: There is a largest q satisfying Claim 1, and for that q, $Y = \operatorname{null} N^q \dot{+} \operatorname{ran} N^q$.

Proof: The V of Claim 2 is 1-1, hence $q = \#V \leq \dim Y$, therefore there is a largest q satisfying Claim 1. For that q, $\operatorname{null} N^q \cap \operatorname{ran} N^q$ is trivial: indeed, if $x \in \operatorname{null} N^q \cap \operatorname{ran} N^q$, then $x = N^q u$ for some $u \in Y$, and also $N^{2q}u = N^q x = 0$, but if $N^q u \neq 0$, then, for some $r > q$, $N^{r-1}u \neq 0 = N^r u$, which would contradict the maximality of q. Hence $x = N^q u = 0$. But also, by the (3.23)Dimension Formula, $\dim Y = \dim\operatorname{null} N^q + \dim\operatorname{ran} N^q$, therefore, by (3.33)Proposition, Y is the direct sum of $\operatorname{null} N^q$ and $\operatorname{ran} N^q$. □

12.5 Prove: *For every noninvertible* $N \in L(X)$ *with* $\dim X < \infty$, *there exists* $q \in \mathbb{N}$ *so that* $\operatorname{ran} N^q \cap \operatorname{null} N^q = \{0\}$, *hence* $X = \operatorname{ran} N^q \dot{+} \operatorname{null} N^q$. The smallest such q is called the **index** of N. The index of an invertible N is defined to be 0.

12.6 Prove: *The index of a real symmetric matrix is* ≤ 1.

12.7 Prove: *For every* $N \in L(X)$ *with* $\dim X < \infty$, *(i) the sequence* $\operatorname{null} N^j$, $j = 0, 1, 2, \ldots$ *is strictly increasing, and (ii) the sequence* $\operatorname{ran} N^j$, $j = 0, 1, 2, \ldots$ *is strictly decreasing, as long as j is less than the index of N; after that, the sequences become stationary.*

Claim 4: For the largest q, $V_Y := [N^{q-j}y : j = 1{:}q] = V$ of Claim 2 is a basis for Y, hence $q = \dim Y$ and the matrix representation for $A|_Y$ with respect to the basis V_Y for Y has the simple form

$$(12.18)\qquad V_Y^{-1}(A|_Y)V_Y = \begin{bmatrix} \mu & 1 & 0 & \cdots & 0 & 0 \\ 0 & \mu & 1 & \cdots & 0 & 0 \\ 0 & 0 & \mu & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \mu & 1 \\ 0 & 0 & 0 & \cdots & 0 & \mu \end{bmatrix} =: J(\mu, q) \in \mathbb{F}^{q\times q}.$$

Proof: We know from Claim 3 that, for a largest q satisfying Claim 1, Y is the direct sum of $\operatorname{null} N^q$ and $\operatorname{ran} N^q$, and both subspaces are N-invariant, hence A-invariant, therefore necessarily one of them must be trivial, and, as by choice, $\operatorname{null} N^q$ is not trivial, it follows that $\operatorname{ran} N^q = \{0\}$, hence $N^q = 0$. This implies that, for this q, the space $\operatorname{null} \Lambda^{\mathrm{t}}$ of Claim 2 is N-invariant, while $\operatorname{ran} V$ there is N-invariant for any q since $NV = V[\mathbf{0}, \mathbf{e}_1, \ldots, \mathbf{e}_{q-1}]$. Since, by Claim 2, Y is the direct sum of these N-invariant, hence A-invariant, spaces, only one can be nontrivial and, since $0 \neq y \in \operatorname{ran} V$, it follows that $Y = \operatorname{ran} V = \operatorname{ran} V_Y$ and, since V_Y is 1-1 by Claim 2, V_Y is a basis for Y, and $V_Y^{-1}(A - \mu\,\mathrm{id})|_Y V_Y = V_Y^{-1} N V_Y = [\mathbf{0}, \mathbf{e}_1, \ldots, \mathbf{e}_{q-1}]$, hence $V_Y^{-1} A|_Y V_Y = \mu\,\mathrm{id}_q + [0, \mathbf{e}_1, \ldots, \mathbf{e}_{q-1}]$, which proves (12.18). □

It follows that the matrix representation for A with respect to the basis

$$[V_Y : Y \in \mathcal{Y}]$$

for X is block-diagonal, with each diagonal block a **Jordan block**, $J(\mu, q)$, i.e., of the form (12.18) for some scalar μ and some natural number q. Any such matrix representation for A is called a **Jordan (canonical) form** for A.

There is no reason to believe that such a Jordan form is unique. After all, it depends on the particular order we choose for the elements of $\mathcal{Y}$ when we make up the basis $[V_Y : Y \in \mathcal{Y}]$. More than that, there is, in general, nothing unique about $\mathcal{Y}$. For example, if $A = 0$ or, more generally $A = \alpha\,\mathrm{id}$, then any direct sum decomposition for X is A-invariant, hence $[V_Y : Y \in \mathcal{Y}]$ can be any basis for X whatsoever for this particular A.

Nevertheless, the Jordan form is canonical in the following sense.

(12.19) Proposition: Let $\widehat{A} =: \operatorname{diag}(J(\mu_Y, \dim Y) : Y \in \mathcal{Y})$ be a Jordan canonical form for $A \in L(X)$. Then

(i) $\operatorname{spct}(A) = \{\widehat{A}_{jj} : j = 1{:}n\} = \cup_{Y\in\mathcal{Y}}\operatorname{spct}(A|_Y)$.

(ii) For each $\mu \in \operatorname{spct}(A)$ and each $q \in \mathbb{N}$,

$$n_\mu(q) := \dim\operatorname{null}(A - \mu\,\mathrm{id})^q = \sum_{\mu_Y=\mu} \min(q, \dim Y), \tag{12.20}$$

hence $\Delta n_\mu(q) := n_\mu(q+1) - n_\mu(q)$ equals the number of blocks for μ of order $> q$, giving the decomposition-independent expression $-\Delta^2 n_\mu(q-1) = \Delta n_\mu(q-1) - \Delta n_\mu(q)$ for the number of Jordan blocks of order q for μ.

Proof: Since $\widehat{A}$ is a block-diagonal matrix representation for A,

$$\dim\operatorname{null}(A - \mu\,\mathrm{id})^q = \sum_{Y\in\mathcal{Y}} \dim\operatorname{null} J(\mu_y - \mu, \dim Y)^q$$

while $\operatorname{null} J(\sigma, s)^q \ne \{0\}$ only for $\sigma = 0$, and

$$J(0, s)^q \ = \ [\mathbf{0}, \ldots, \mathbf{0}, \mathbf{e}_1, \ldots, \mathbf{e}_{s-q}] \quad \text{for} \quad q \le s,$$

hence $\dim\operatorname{null} J(0, s)^q = \min(q, s)$ for arbitrary $q \in \mathbb{N}$. □

In particular, the Jordan form is unique up to an ordering of its blocks.

Also, (12.20) tells us that $\dim\operatorname{null}(A - \mu\,\mathrm{id})$ equals the number of Jordan blocks associated with μ, while the number of times that μ appears on the diagonal of a Jordan canonical form for A, i.e., $\sum_{\mu_Y=\mu} \dim Y$, equals $\max_q \dim\operatorname{null}(A-\mu\,\mathrm{id})^q = \dim \cup_{q\in\mathbb{N}} \operatorname{null}(A-\mu\,\mathrm{id})^q$, the last equality because $\operatorname{null}(A - \mu\,\mathrm{id})^q$, $q = 1, 2, \ldots$, is an increasing sequence. Correspondingly,

$$\#_g\mu := \dim\operatorname{null}(A - \mu\,\mathrm{id}), \qquad \mu \in \operatorname{spct}(A), \tag{12.21}$$

is called the **geometric multiplicity** of the eigenvalue μ, as it counts the maximum number of columns in a 1-1 column map staffed entirely by eigenvectors for μ, while

(12.22)
$$\#_a\mu := \max_q \dim\operatorname{null}(A - \mu\,\mathrm{id})^q = \dim \bigcup_{q\in\mathbb{N}} \operatorname{null}(A - \mu\,\mathrm{id})^q, \quad \mu \in \operatorname{spct}(A),$$

is called the **algebraic multiplicity** of μ. We will return to these multiplicity notions later, after bringing determinants into play.

While the Jordan form is mathematically quite striking and useful, it is of no practical relevance since it does not depend continuously on the entries of A, hence cannot be determined reliably by floating-point calculations.

12.8 Give an example to show that the Jordan form of a matrix A does not depend continuously on the entries of A.

12.9* Prove that the minimal polynomial p_A for the Jordan block $A := J(\mu, q)$ is $(\cdot - \mu)^q$.

12.10* What is the multiplicity of an eigenvalue of A as the zero of the minimal polynomial p_A for A? What is the relationship, if any, with the geometric or algebraic multiplicity of that eigenvalue? (Feel free to use the primary decomposition; however, these questions can also be answered with the aid of the Jordan form.)

12.11 Prove: *Let A_x be the restriction of $A \in L(X)$ to the Krylov subspace $K_{A,x}$ for some nonzero x in some finite-dimensional vector space X. Then, the matrix representation for A_x with respect to the basis $V := [x, Ax, \ldots, A^{d-1}x]$ of $K_{A,x}$ is the companion matrix for $p_{A,x}$.*

12.12* Prove: *$A \in L(X)$ is diagonalizable if and only if X is the direct sum of* $(\mathrm{null}(A - \mu\,\mathrm{id}) : \mu \in \mathrm{spct}(A))$.

12.13 Prove that $A \in L(X)$ with $\mathbb{F} = \mathbb{C}$ is diagonalizable if and only if all its Jordan blocks are of order 1.

12.14* Prove: *If $A, B \in L(X)$ and $AB = BA$, then, for every $\mu \in \mathrm{spct}(A)$, $\mathrm{null}(A - \mu\,\mathrm{id})$ is B-invariant.* Conclude that, under this condition, and with $\nu \in \mathrm{spct}(B)$, $\mathrm{null}(A - \mu\,\mathrm{id}) \cap \mathrm{null}(B - \nu\,\mathrm{id})$ is both A- and B-invariant.

12.15* Prove: *If $\mathbb{F} = \mathbb{C}$ and $A, B \in L(X)$, and $AB = BA$, then the diagonalizability of B implies that the restriction of B to $Y := \mathrm{null}(A - \mu\,\mathrm{id})$ is diagonalizable for every $\mu \in \mathrm{spct}(A)$.*

12.16* Prove: (a) *If $\mathbb{F} = \mathbb{C}$ and $A, B \in L(X)$ are both diagonalizable and $AB = BA$, then X is the direct sum of $\mathrm{null}(A - \mu\,\mathrm{id}) \cap \mathrm{null}(B - \nu\,\mathrm{id})$, $\mu \in \mathrm{spct}(A)$, $\nu \in \mathrm{spct}(B)$.* (Hint: Problem 3.28) Conclude that (b) there is some basis consisting of eigenvectors for both A and B, i.e., A and B are **simultaneously diagonalizable.**

12.17 Prove: *If $A_1, \ldots, A_r \in L(X)$ are all diagonalizable and commute with each other, then they are simultaneously diagonalizable, i.e., there is some basis for X all of whose columns are eigenvectors for every A_i.*

12.18* Prove: *If $\mathcal{Y}$ provides a proper A-invariant direct sum decomposition for X and $A \in L(X)$ is diagonalizable, then every $B := A\big|_Y$, $Y \in \mathcal{Y}$, is diagonalizable.*

The Weyr form

The Weyr (canonical) form can be obtained from a Jordan normal form by permutation of the columns of the corresponding ‘Jordan’ basis that gives rise to the Jordan form.

Recall that each column $\mathbf{v}$ in a ‘Jordan’ basis can be uniquely identified by three properties: the *eigenvalue* μ to which it belongs, its **degree**, i.e., the smallest integer d for which $\mathbf{v} \in \mathrm{null}(A - \mu\,\mathrm{id})^d$, and the eigenvector $(A - \mu\,\mathrm{id})^{d-1}\mathbf{v}$ in the ‘Jordan’ basis. Moreover, it will be important to associate with each eigenvector $\mathbf{v}$ its **order**, i.e., the largest d for which $\mathbf{v} = (A - \mu\,\mathrm{id})^{d-1}\mathbf{w}$ for some column $\mathbf{w}$ of the ‘Jordan’ basis.

The corresponding ‘Weyr’ basis W groups all columns of the Jordan basis associated with the same eigenvalue together, producing the basis $W =$

$[W_\mu : \mu \in \operatorname{spct}(A)]$. In particular,

$$\operatorname{ran} W_\mu = \cup_q \operatorname{null}(A - \mu \operatorname{id})^q.$$

The resulting matrix representation $W^{-1}AW$ is therefore block-diagonal with as many diagonal blocks as there are eigenvalues.

We now consider the structure of the diagonal block associated with the eigenvalue μ which will depend on the order in which we choose to list in W_μ all the columns of the 'Jordan' basis associated with μ. We choose to list them by degree, first those of degree 1, then those of degree 2, then those of degree 3, etc. . More precisely, those of degree 1 are listed in decreasing order with those of the same order ordered arbitrarily. Those of degree higher than 1 are listed to match the order chosen for their corresponding eigenvectors. In this way, we obtain W_μ in the form

$$W_\mu = [W_{\mu,1} W_{\mu,2}, \ldots],$$

with $W_{\mu,d}$ the columns of degree d, all d. Also, for each d,

$$W_{\mu,d-1} = [(A - \mu \operatorname{id}) W_{\mu,d}, E_{\mu,d}],$$

with $E_{\mu,d}$ comprising the columns associated with μ of degree d that are not the image under $(A - \mu \operatorname{id})$ of some column of the 'Jordan' basis.

It follows that the Weyr block associated with the eigenvalue μ has the following characteristic form:

$$\begin{bmatrix} D_1 & B_1 & & & \\ & D_2 & B_2 & & \\ & & \ddots & \ddots & \\ & & & D_{q-1} & B_{q-1} \\ & & & & D_q \end{bmatrix},$$

with each $D_i = \mu \operatorname{id}_{r_i}$ for a decreasing sequence $r_1 \ge r_2 \ge \cdots$, and B_i an identity matrix of order $r_i \times r_{i+1}$, i.e., of the form

$$B_i = \begin{bmatrix} \operatorname{id}_{r_{i+1}} \\ 0 \end{bmatrix}.$$

Precisely, $[W_{\mu,1}, \ldots, W_{\mu,d}]$ is a basis for $\operatorname{null}(A - \mu \operatorname{id})^d$, hence

$$r_i = \dim \operatorname{null}(A - \mu \operatorname{id})^i - \dim \operatorname{null}(A - \mu \operatorname{id})^{i-1}$$

are numbers depending on A alone. In particular, the Weyr form for A is unique, up to the ordering of the diagonal blocks. Each diagonal block looks like a block version of a Jordan block.

13 Localization of eigenvalues

In this short chapter, we discuss briefly the standard techniques for 'localizing' the spectrum of a given linear map A. Such techniques specify regions in the complex plane that must contain parts or all of the spectrum of A. To give a simple example, we proved (in (12.2)Corollary) that all the eigenvalues of a hermitian matrix must be real, i.e., that $\operatorname{spct}(A) \subset \mathbb{R}$ in case $A^{c} = A$. More precise localization statements for hermitian matrices can be found in the chapter on optimization and quadratic forms.

Since $\mu \in \operatorname{spct}(A)$ iff $(A - \mu\,\mathrm{id})$ is not invertible, it is not surprising that many localization theorems derive from a test for invertibility.

Gershgorin's circles

Let μ be an eigenvalue for A with corresponding eigenvector x. Without loss of generality, we may assume that $\|x\| = 1$ in whatever vector norm on $X = \operatorname{dom} A$ we are interested in at the moment. Then

$$|\mu| = |\mu|\|x\| = \|\mu x\| = \|Ax\| \le \|A\|\|x\| = \|A\|,$$

with $\|A\|$ the corresponding map norm for A. This proves that the spectrum of A must lie in the closed disk $\overline{B}_{\|A\|}$ of radius $\|A\|$ centered at the origin. In other words,

$$\rho(A) \le \|A\| \tag{13.1}$$

for any map norm $\|\cdot\|$.

For example, no eigenvalue of $A = \begin{bmatrix} 1 & 2 \\ -2 & -1 \end{bmatrix}$ can be bigger than 3 in absolute value since $\|A\|_\infty = 3$.

A more refined containment set is obtained by the following more refined analysis.

If $E \in \mathbb{F}^{n\times n}$ has map-norm < 1, then $A := \mathrm{id}_n - E$ is 1-1 since then

$$\|A\mathbf{x}\| = \|\mathbf{x} - E\mathbf{x}\| \ge \|\mathbf{x}\| - \|E\mathbf{x}\| \ge \|\mathbf{x}\| - \|E\|\|\mathbf{x}\| = \|\mathbf{x}\|(1 - \|E\|)$$

with the factor $(1 - \|E\|)$ *positive*, hence $A\mathbf{x} = 0$ implies that $\|\mathbf{x}\| = 0$. Moreover,

$$\|A^{-1}\| = \|\,\mathrm{id}_n - E^{-1}\| \le 1/(1 - \|E\|). \tag{13.2}$$

Now consider a **diagonally dominant** A, i.e., a matrix A with the property that

$$\forall i, \quad |A_{ii}| > \sum_{j\ne i} |A_{ij}|. \tag{13.3}$$

For example, of the three matrices

$$\begin{bmatrix} 2 & -1 \\ 2 & 3 \end{bmatrix}, \quad \begin{bmatrix} -2 & -1 \\ 3 & 3 \end{bmatrix}, \quad \begin{bmatrix} -2 & -1 \\ 4 & 3 \end{bmatrix}, \tag{13.4}$$

only the first is diagonally dominant. Setting

$$D := \operatorname{diag} A = \operatorname{diag}(\ldots, A_{ii}, \ldots),$$

we notice that (i) D is invertible (since all its diagonal entries are nonzero); and (ii) the matrix E defined by $D^{-1}A =: \mathrm{id} - E$ satisfies

$$E_{ij} = \begin{cases} -A_{ij}/A_{ii}, & \text{if } i \ne j; \\ 0, & \text{otherwise,} \end{cases}$$

hence has norm

$$\|E\|_\infty = \max_i \sum_{j\ne i} |A_{ij}/A_{ii}| < 1,$$

by the assumed diagonal dominance of A. This implies that the matrix $\mathrm{id} - E = D^{-1}A$ is invertible, therefore also A is invertible. This proves

(13.5) Proposition: Any diagonally dominant matrix is invertible.

In particular, the first of the three matrices in (13.4) we now know to be invertible. As it turns out, the other two are also invertible; thus, diagonal dominance is only sufficient but not necessary for invertibility. Equivalently, a noninvertible matrix cannot be diagonally dominant.

In particular, for $(A - \mu\,\mathrm{id})$ to be *not* invertible, it must fail to be diagonally dominant, i.e.,

$$\exists i \quad |A_{ii} - \mu| \le \sum_{j\ne i} |A_{ij}|. \tag{13.6}$$

This gives the famous

(13.7) **Gershgorin Circle Theorem**: The spectrum of $A \in \mathbb{C}^{n\times n}$ is contained in the union of the closed disks

$$\overline{B}_{r_i}(A_{ii}) := \{z \in \mathbb{C} : |A_{ii} - z| \le r_i := \sum_{j\ne i} |A_{ij}|\}, \quad i = 1{:}n.$$

For the three matrices in (13.4), this says that

$$\mathrm{spct}(\begin{bmatrix} 2 & -1 \\ 2 & 3 \end{bmatrix}) \subset \overline{B}_1(2) \cup \overline{B}_2(3), \quad \mathrm{spct}(\begin{bmatrix} -2 & -1 \\ 3 & 3 \end{bmatrix}) \subset \overline{B}_1(-2) \cup \overline{B}_3(3),$$

$$\mathrm{spct}(\begin{bmatrix} -2 & -1 \\ 4 & 3 \end{bmatrix}) \subset \overline{B}_1(-2) \cup \overline{B}_4(3).$$

More than that, according to the *refinement of the Gershgorin Circle Theorem* discussed in Problem 13.6, the second matrix must have one eigenvalue in the closed disk $\overline{B}_1(-2)$ and another one in the closed disk $\overline{B}_3(3)$, since these two disks have an empty intersection. By the same refinement, if the third matrix has only one eigenvalue, then it would necessarily have to be the point -1, i.e., the sole point common to the two disks $\overline{B}_1(-2)$ and $\overline{B}_4(3)$.

13.1 Does each of the two Gershgorin disks of the matrix $A := \begin{bmatrix} 5 & -1 \\ 6 & 0 \end{bmatrix}$ contain an eigenvalue of A?

The trace of a linear map

Recall that the *trace* of a square matrix A is given by

$$\mathrm{trace}(A) = \sum_j A_{jj}.$$

Further, as already observed in (6.31), if the product of the two matrices B and C is square, then

$$\mathrm{trace}(BC) = \sum_j \sum_k B_{jk} C_{kj} = \sum_{jk} B_{jk} C_{kj} = \mathrm{trace}(CB). \tag{13.8}$$

Hence, if $A = V\widehat{A}V^{-1}$, then

$$\operatorname{trace}(A) = \operatorname{trace}(V(\widehat{A}V^{-1})) = \operatorname{trace}(\widehat{A}V^{-1}V) = \operatorname{trace}\widehat{A}.$$

This proves

(13.9) Proposition: Any two similar matrices have the same trace.

This permits the definition of the **trace** of an arbitrary linear map A on an arbitrary finite-dimensional vector space X as the trace of the matrices similar to it. In particular, trace(A) equals the sum of the diagonal entries of any Schur form for A, i.e., trace(A) is the sum of the eigenvalues of A, however with some of these eigenvalues possibly repeated.

For example, trace(id_n) $= n$, while spct(id_n) $= \{1\}$.

Offhand, such eigenvalue *multiplicity* seems to depend on the particular Schur form (or any other triangular matrix representation) for A. But, since all of these matrices have the same trace, you will not be surprised to learn that all these triangular matrix representations for A have each eigenvalue appear on its diagonal with exactly the same multiplicity, necessarily its algebraic multiplicity (12.22) as any Jordan canonical form for A is a triangular matrix representation for A. The proof of this claim is most easily given with the aid of yet another tool for testing invertibility, namely determinants, to which we turn next.

Determinants

The **determinant** is, by definition, the unique map

$$\det : \mathbb{F}^{n\times n} \to \mathbb{F} : A \mapsto \det A$$

for which the induced map

$$(\mathbb{F}^n)^n \to \mathbb{F} : (\mathbf{a}_1, \ldots, \mathbf{a}_n) \mapsto \det[\mathbf{a}_1, \ldots, \mathbf{a}_n]$$

is multilinear and alternating, while

$$\det \mathrm{id}_n = 1. \tag{13.10}$$

(All claims about determinants made in this section, including the uniqueness just mentioned, are proved in the chapter entitled "More about determinants" which starts on page 223.) Here, **multilinear** means that $\det[\mathbf{a}_1, \ldots, \mathbf{a}_n]$ is *linear* in each of the n columns $\mathbf{a}_i$, i.e.,

$$\det[\ldots, \mathbf{a} + \alpha\mathbf{b}, \ldots] = \det[\ldots, \mathbf{a}, \ldots] + \alpha \det[\ldots, \mathbf{b}, \ldots]. \tag{13.11}$$

(Here and below, the various ellipses ... indicate the other columns, the ones that are kept fixed.) Further, **alternating** means that the interchange of two columns reverses the sign, i.e.,

$$\det[\ldots,\mathbf{a},\ldots,\mathbf{b},\ldots] = -\det[\ldots,\mathbf{b},\ldots,\mathbf{a},\ldots].$$

In particular, $\det A = 0$ in case two columns of A are the same, i.e.,

$$\det[\ldots,\mathbf{b},\ldots,\mathbf{b},\ldots] = 0.$$

Combining this last with (13.11), we find that

$$\det[\ldots,\mathbf{a}+\alpha\mathbf{b},\ldots,\mathbf{b},\ldots] = \det[\ldots,\mathbf{a},\ldots,\mathbf{b},\ldots],$$

i.e., *addition of a scalar multiple of one column to a different column does not change the determinant.*

In particular, *if* $A = [\mathbf{a}_1,\mathbf{a}_2,\ldots,\mathbf{a}_n]$ *is not invertible, then* $\det A = 0$ since then there must be some column $\mathbf{a}_j$ of A writable as a linear combination of other columns, i.e.,

$$\det A = \det[\ldots,\mathbf{a}_j,\ldots] = \det[\ldots,\mathbf{0},\ldots] = 0,$$

the last equality by the multilinearity of the determinant.

Conversely, *if* A *is invertible, then* $\det A \neq 0$, and this follows from the fundamental determinantal identity (proved on page 229)

$$\det(AB) = \det(A)\det(B) \tag{13.12}$$

since it implies that, for an invertible A,

$$1 = \det \mathrm{id}_n = \det(AA^{-1}) = \det(A)\det(A^{-1}),$$

the first equality by (13.10).

(13.13) Theorem: For all $A \in \mathbb{C}^{n\times n}$,

$$\mathrm{spct}(A) = \{\mu \in \mathbb{C} : \det(A - \mu\,\mathrm{id}) = 0\}.$$

Of course, this theorem is quite useless unless we have in hand an explicit formula for the determinant. Here is the standard formula:

(13.14)

$$\det(A) = \det[\mathbf{a}_1,\mathbf{a}_2,\ldots,\mathbf{a}_n] = \sum_{\sigma\in\mathbf{S_n}} (-1)^\sigma \prod_j \mathbf{a}_j(\sigma_j) = \sum_{\sigma\in\mathbf{S_n}} (-1)^\sigma \prod_j A_{\sigma_j,j}$$

in which the sum is over all permutations σ of degree n, i.e., all 1-1 (hence invertible) maps $\sigma : \{1,\ldots,n\} \to \{1,\ldots,n\}$, and

$$(-1)^{\sigma} \ := \ \operatorname{signum} \Delta(\sigma), \quad \Delta(\sigma) \ := \ \left(\prod_{j<k}(\sigma_k - \sigma_j)\right)$$

is 1 or -1 depending on whether it takes an even or an odd number of interchanges to bring the sequence σ back into increasing order (see Problem 15.1 for a proof of this assertion).

For $n = 1$, we get the trivial fact that, for any scalar a, $\operatorname{spct}([a]) = \{a\}$.

For $n = 2$, (13.13) implies that

$$\operatorname{spct}(\begin{bmatrix} a & b \\ c & d \end{bmatrix}) = \{\mu \in \mathbb{C} : (a-\mu)(d-\mu) = bc\}.$$

For $n = 3$, we get

$$\operatorname{spct}(\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}) = \{\mu \in \mathbb{C} : p(\mu) = 0\},$$

with

$$p(\mu) := (a-\mu)(e-\mu)(i-\mu) + bfg + chd - c(e-\mu)g - (a-\mu)fh - bd(i-\mu).$$

For $n = 4$, (13.14) already involves 24 summands, and, for general n, we have $n! = 1 \cdot 2 \cdots n$ summands. Thus, even with this formula in hand, the theorem is mostly only of theoretical interest since already for modest n, the number of summands involved becomes too large for any practical computation.

In fact, the determinant $\det A$ of a given matrix A is usually computed with the aid of some factorization of A, relying on the fundamental identity (13.12) and on the following

(13.15) Lemma: The determinant of any triangular matrix is the product of its diagonal entries.

Proof: This observation follows at once from (13.14) since any permutation σ other than the identity $(1, 2, \ldots, n)$ must have $\sigma_k < k$ for some k, hence the corresponding product $\prod_j A_{\sigma_j,j}$ in (13.14) will be zero for any lower triangular matrix. Since any such σ must also have $\sigma_h > h$ for some h, the corresponding product will also vanish for any upper triangular matrix. Thus, in either case, only the product $\prod_j A_{\sigma_j,j}$ is possibly nonzero. □

So, with $A = PLU$ as constructed by Gauss-elimination, with L unit lower triangular and U upper triangular, and P a permutation matrix, we have

$$\det A = (-1)^P \prod_j U_{jj},$$

with the number $(-1)^P$ equal to 1 or -1 depending on the parity of the permutation carried out by P, i.e., whether the number of row interchanges made during Gauss elimination is even or odd.

Formula (13.14) is often taken as the definition of $\det A$. It is a simple consequence of the fundamental identity (13.12), and the latter follows readily from the multilinearity and alternation property of the determinant. For these and other details, see the chapter 'More on determinants'.

Annihilating polynomials

Recall that the nontrivial polynomial p is called **annihilating for** $A \in L(X)$ if $p(A) = 0$.

For example, A is *nilpotent* exactly when, for some k, the monomial $()^k$ annihilates A, i.e., $A^k = 0$. As another example, A is a linear projector (or, idempotent) exactly when the polynomial $p : t \mapsto t(t-1)$ annihilates A, i.e., $A^2 = A$.

Annihilating polynomials are of interest because of the following version of the **Spectral Mapping Theorem**:

(13.16) Theorem: For any polynomial p and any linear map $A \in L(X)$ with $\mathbb{F} = \mathbb{C}$,

$$\operatorname{spct}(p(A)) = p(\operatorname{spct}(A)) := \{p(\mu) : \mu \in \operatorname{spct}(A)\}.$$

Proof: If $\mu \in \operatorname{spct}(A)$, then, for some nonzero x, $Ax = \mu x$, therefore also $p(A)x = p(\mu)x$, hence $p(\mu) \in \operatorname{spct}(p(A))$. In other words, $p(\operatorname{spct}(A)) \subset \operatorname{spct}(p(A))$.

Conversely, if $\nu \in \operatorname{spct}(p(A))$, then $p(A) - \nu\,\mathrm{id}$ fails to be 1-1. However, assuming without loss of generality that p is a monic polynomial of degree r, we have $p(t) - \nu = (t - \mu_1)\cdots(t - \mu_r)$ for some scalars $\mu_1, \ldots, \mu_r$, therefore

$$p(A) - \nu\,\mathrm{id} = (A - \mu_1\,\mathrm{id})\cdots(A - \mu_r\,\mathrm{id}),$$

and, since the left-hand side is not 1-1, at least one of the factors on the right must fail to be 1-1. This says that some $\mu_j \in \operatorname{spct}(A)$, while $p(\mu_j) - \nu = 0$. In other words, $\operatorname{spct}(p(A)) \subset p(\operatorname{spct}(A))$. □

In particular, if p annihilates A, then $p(A) = 0$, hence $\operatorname{spct}(p(A)) = \{0\}$, therefore $\operatorname{spct}(A) \subset \{\mu \in \mathbb{C} : p(\mu) = 0\}$.

For example, 0 is the only eigenvalue of a nilpotent linear map. The only possible eigenvalues of an idempotent map are the scalars 0 and 1.

The best-known annihilating polynomial for a given $A \in \mathbb{F}^{n\times n}$ is its **characteristic polynomial**, i.e., the polynomial

$$\chi_A : t \mapsto \det(t \operatorname{id}_n - A).$$

To be sure, by (10.32), we can write any such A as the product $A = V\widehat{A}V^{-1}$ with $\widehat{A}$ upper triangular. Correspondingly,

$$(13.17)\qquad \begin{aligned} \chi_A(t) &= \det V \det(t \operatorname{id}_n - \widehat{A})(\det V)^{-1} = \det(t \operatorname{id}_n - \widehat{A}) \\ &= \chi_{\widehat{A}}(t) \;=\; \prod_j (t - \widehat{A}_{jj}), \end{aligned}$$

the last equation by (13.15)Lemma. Consequently, $\chi_A(A) = V\chi_A(\widehat{A})V^{-1}$, with

$$\chi_A(\widehat{A}) = (\widehat{A} - \mu_1 \operatorname{id}) \cdots (\widehat{A} - \mu_n \operatorname{id}), \qquad \mu_j := \widehat{A}_{jj}, \quad j = 1{:}n,$$

and this, we claim, is necessarily the zero map, for the following reason: The factor $(\widehat{A} - \mu_j \operatorname{id})$ is upper triangular, with the jth diagonal entry equal to zero. This implies that, for each i, $(\widehat{A} - \mu_j \operatorname{id})$ maps

$$T_i := \operatorname{ran}[\mathbf{e}_1, \ldots, \mathbf{e}_i]$$

into itself, but maps T_j into T_{j-1}. Therefore

$$\begin{aligned} \operatorname{ran} \chi_A(\widehat{A}) = \chi_A(\widehat{A})T_n &= (\widehat{A} - \mu_1 \operatorname{id}) \cdots (\widehat{A} - \mu_n \operatorname{id})T_n \\ &\subset (\widehat{A} - \mu_1 \operatorname{id}) \cdots (\widehat{A} - \mu_{n-1} \operatorname{id})T_{n-1} \\ &\subset (\widehat{A} - \mu_1 \operatorname{id}) \cdots (\widehat{A} - \mu_{n-2} \operatorname{id})T_{n-2} \\ &\cdots \\ &\subset (\widehat{A} - \mu_1 \operatorname{id})T_1 \subset T_0 = \{0\}, \end{aligned}$$

or, $\chi_A(\widehat{A}) = 0$, therefore also $\chi_A(A) = 0$. This is known as the **Cayley-Hamilton Theorem**.

Note that the collection $\mathcal{I}_A := \{p \in \Pi : p(A) = 0\}$ of all polynomials that annihilate a given linear map A is an **ideal**, meaning that it is a linear subspace of Π that is also closed under multiplication by polynomials: if $p \in \mathcal{I}_A$ and $q \in \Pi$, then their product $qp : t \mapsto q(t)p(t)$ is also in $\mathcal{I}_A$. Since $\mathcal{I}_A$ is not empty, it contains a monic polynomial of minimal degree. We called this polynomial on page 164 the **minimal (annihilating) polynomial for** A and denoted it by p_A. It generates the ideal in the sense that $\mathcal{I}_A = \Pi p_A$. In technical terms, I_A is a **principal ideal**, more precisely the principal ideal generated by p_A.

In exactly the same way, the collection $\mathcal{I}_{A,x} := \{p \in \Pi : p(A)x = 0\}$ was seen on page 164 to be the principal ideal $\Pi p_{A,x}$, with $p_{A,x}$ the minimal annihilating polynomial for A at x, i.e., the unique monic polynomial of smallest degree in it. Since $\mathcal{I}_A \subset \mathcal{I}_{A,x}$, it follows that $p_{A,x}$ must be a factor for any $p \in \mathcal{I}_A$ and, in particular, for χ_A.

13.2* (a) Prove: *If the minimal polynomial* $p = p_{A,x}$ *of the linear map* $A \in L(X)$ *at some* $x \in X\backslash 0$ *has degree equal to* $\dim X$, *then* $p_{A,x}(A) = 0$. (b) Prove that the spectrum of the companion matrix (see Problem 10.20) of the monic polynomial p equals the zero set of p.

13.3 Recall that a matrix A of order n is *non-derogatory* if it has a *cyclic vector*, i.e., if, for some $\mathbf{x}$, $[\mathbf{x}, A\mathbf{x}, \ldots, A^{n-1}\mathbf{x}]$ is 1-1 (hence a basis).

Prove that the non-derogatory matrices of order n are **dense**, i.e., for every matrix B of order n and every $\varepsilon > 0$, there exists a non-derogatory matrix A so that $\|B - A\|_\infty \le \varepsilon$. (Hint: prove first that there are non-derogatory matrices (e.g., companion matrices (why?)), then consider the function $z \mapsto \det[\mathbf{x}, (B + zA)\mathbf{x}, \ldots, (B + zA)^{n-1}\mathbf{x}]$ with $\mathbf{x}$ a cyclic vector for A.)

13.4 Formulate and prove a 'spectral mapping theorem' for the maps $p : \mathbb{F}^{n\times n} \to \mathbb{F}^{n\times n}$ given by $p(A) := A^{\mathrm{t}}$ and $p(A) := A^{\mathrm{c}}$.

13.5 make one about the coefs of char.pol. being symmetric functions of evs, and one about the ith coeff. being the sum of the $n - i$th principal minors. all of these, including the trace, are invariant under similarity. **still to be done!**

The multiplicities of an eigenvalue

Since χ_A is of degree n in case $A \in \mathbb{C}^n$, χ_A has exactly n zeros *counting multiplicities*. This means that

$$\chi_A(t) = (t - \nu_1)\cdots(t - \nu_n) \tag{13.18}$$

for a certain n-sequence ν. Further,

$$\mathrm{spct}(A) = \{\nu_j : j = 1{:}n\},$$

and this set may well contain only one number, as it does when $A = 0$ or $A = \mathrm{id}$.

Since, by (13.17), (13.18) holds with ν the sequence of diagonal entries of any triangular matrix representation for A, we know that such a sequence contains each eigenvalue μ of A to its algebraic multiplicity $\#_a\mu$ (12.22), i.e., the multiplicity with which μ appears in any Jordan canonical form.

In this way, if $\mathbb{F} = \mathbb{C}$ and $\dim X = n$, then any $A \in L(X)$ has exactly n eigenvalues counting (algebraic) multiplicity.

13.19 Proposition: For any eigenvalue, the algebraic multiplicity is no smaller than the geometric multiplicity, with equality if and only if the eigenvalue is not defective.

Proof: From (12.21) and (12.22),

$$\#_g\mu = \dim \operatorname{null}(A - \mu \operatorname{id}) \le \dim \bigcup_{q \in \mathbb{N}} \operatorname{null}(A - \mu \operatorname{id})^q = \#_a\mu, \quad \mu \in \operatorname{spct}(A),$$

with equality if and only if $\operatorname{null}(A - \mu \operatorname{id}) = \operatorname{null}(A - \mu \operatorname{id})^2$ if and only if $\operatorname{null}(A - \mu \operatorname{id}) \cap \operatorname{ran}(A - \mu \operatorname{id}) = \{0\}$ if and only if μ is not defective. □

An eigenvalue for which algebraic and geometric multiplicity coincide is called **semisimple**, as a generalization of a **simple eigenvalue** which is an eigenvalue for which $\#_a\mu = 1$, hence $\#_a\mu = \#_g\mu$.

For example, the matrix id_n has only the eigenvalue 1, but with algebraic and geometric multiplicity n. In other words, the sole eigenvalue is semisimple as it should be since id_n is trivially diagonalizable.

In contrast, the sole eigenvalue, 0, of $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ has algebraic multiplicity 2 but its geometric multiplicity is only 1. In other words, its sole eigenvalue is defective as it should be since this matrix is not diagonalizable.

13.6* Prove the following generalization of the (13.7)Gershgorin Circle Theoren: *If the union of k of the disks $\overline{B}_{r_i}(A(i,i))$ is disjoint from the union of the remaining disks, then it contains exactly k of the eigenvalues, counting algebraic multiplicities.* (Hint: use the facts that χ_A is a continuous function of A and that the zeros of a polynomial are continuous functions of its coefficients.)

13.7 Using, perhaps, (13.17), determine the algebraic and geometric multiplicities for all the eigenvalues of the following matrix. (Read off the eigenvalues; use elimination to determine geometric multiplicities.)

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

13.8* Let $A \in L(X)$, X finite-dimensional, $\mathbb{F} = \mathbb{C}$. Prove that the multiplicity of $\mu \in \operatorname{spct}(A)$ as a zero of the minimal polynomial p_A of A equals the order of the largest Jordan block belonging to μ, hence is bounded by $\#_a\mu$.

13.9 Give examples to show that the multiplicity of $\mu \in \operatorname{spct}(A)$ as a zero of p_A can be larger than, equal to, or smaller than $\#_g\mu$.

13.10* Prove: *If $\mathbb{F} = \mathbb{C}$, and $\deg p_A = n$ for some $A \in \mathbb{C}^{n \times n}$, then every eigenvalue of A has geometric multiplicity 1.*

Perron-Frobenius

We call the vector $\mathbf{y} \in \mathbb{R}^n$ **positive** (**nonnegative**) and write

$$\mathbf{y} > \mathbf{0} \qquad (\mathbf{y} \ge \mathbf{0})$$

if all its entries are positive (nonnegative). Since a vector has, in general, several entries, there is room for confusion here, since $\mathbf{y}$ can be nonnegative and not zero without being positive.

We will also use the notation

$$|\mathbf{y}| := (|y_i| : i = 1{:}n), \qquad |B| := (|B_{ij}| : i, j = 1{:}n)$$

for the pointwise absolute value of the vector $\mathbf{y}$ and the matrix B, respectively.

Analogously, we call a matrix A **positive** (**nonnegative**) and write $A > 0$ ($A \ge 0$) in case all its entries are positive (nonnegative).

13.11* Prove, *for* $A \in \mathbb{F}^{n\times n}$, $\mathbf{x} \in \mathbb{F}^n$ *that if* $A > 0$ *and* $\mathbf{0} \ne \mathbf{x} \ge \mathbf{0}$, *then* $A\mathbf{x} > 0$, *while* $0 \ne A \ge 0$, $\mathbf{x} > 0$ *does not imply that* $A\mathbf{x} > 0$.

13.12* Let $\mathbf{a}, \mathbf{b}, \mathbf{y} \in \mathbb{R}^n$. Prove: *If* $\mathbf{y} > \mathbf{0}$, *and* $\mathbf{b} \le \mathbf{a}$, *then* $\mathbf{b}^{\mathrm{t}}\mathbf{y} \le \mathbf{a}^{\mathrm{t}}\mathbf{y}$ *with equality if and only if* $\mathbf{b} = \mathbf{a}$.

A positive (nonnegative) matrix A of order n maps the **positive orthant**

$$\mathbb{R}^n_+ := \{\mathbf{y} \in \mathbb{R}^n : \mathbf{y} \ge 0\}$$

into its interior (into itself). Thus the (scaled) power method, applied to a nonnegative A and started at a nonnegative vector, would converge to a nonnegative vector if it converges. This suggests that the absolutely largest eigenvalue for a nonnegative matrix is nonnegative, with a corresponding nonnegative eigenvector. The Perron-Frobenius theorem makes this intuition precise. For its derivation, we assume that A is a nonnegative matrix of order n.

Since A maps $\mathbb{R}^n_+$ into itself, it makes sense to consider, for given $\mathbf{y} \in \mathbb{R}^n_+\backslash\mathbf{0}$, scalars α for which $A\mathbf{y} \ge \alpha\mathbf{y}$ (in the sense that $(Ay)_j \ge \alpha y_j$, all j), i.e., for which $A\mathbf{y} - \alpha\mathbf{y} \ge \mathbf{0}$. The largest such scalar is the nonnegative number

$$r(\mathbf{y}) := \min\{(A\mathbf{y})_j / y_j : y_j > 0\}$$

which is well-defined for every $\mathbf{y} \in \mathbb{R}^n_+\backslash\mathbf{0}$. The basic observation is that

$$(13.20) \qquad A\mathbf{y} - \alpha\mathbf{y} > 0 \quad \Longrightarrow \quad r(\mathbf{y}) > \alpha.$$

The function r so defined is dilation-invariant, i.e., $r(\alpha\mathbf{y}) = r(\mathbf{y})$ for all $\alpha > 0$, hence r takes on all its values already on the set

$$S_+ := \{\mathbf{y} \ge \mathbf{0} : \|\mathbf{y}\| = 1\}.$$

At this point, we need, once again, a result that goes beyond the scope of this book, namely the fact (see (17.6)Theorem) that S_+ is compact, while r is continuous at any $\mathbf{y} > \mathbf{0}$ and upper semicontinuous at any $\mathbf{y} \geq \mathbf{0}$, hence r takes on its supremum over $\mathbb{R}^n_+\backslash\mathbf{0}$ at some point in S_+. I.e., there exists $\mathbf{x} \in S_+$ for which

$$\mu := r(\mathbf{x}) = \sup r(S_+) = \sup r(\mathbb{R}^n_+\backslash\mathbf{0}).$$

Assume now, in addition to $A \geq 0$, that also $p(A) > 0$ for some polynomial p.

Claim 1: $A\mathbf{x} = \mu\mathbf{x}$.

Proof: Assume that $A\mathbf{x} \neq \mu\mathbf{x}$. Since $\mu = r(\mathbf{x})$, we have $\mathbf{0} \neq A\mathbf{x} - \mu\mathbf{x} \geq \mathbf{0}$, therefore $A(p(A)\mathbf{x}) - \mu p(A)\mathbf{x} = p(A)(A\mathbf{x} - \mu\mathbf{x}) > \mathbf{0}$ by Problem 13.11, hence, by (13.20), $r(p(A)\mathbf{x}) > \mu = \sup r(S_+)$, a contradiction. □

Claim 2: $\mathbf{x} > \mathbf{0}$.

Proof: Since $\mathbf{0} \neq \mathbf{x} \geq \mathbf{0}$ and $p(A) > 0$, we have $p(\mu)\mathbf{x} = p(A)\mathbf{x} > \mathbf{0}$, hence $x_i \neq 0$, all i, therefore, $\mathbf{x} > \mathbf{0}$. □

13.13 Prove that, under the same conditions, any nonnegative eigenvector for any eigenvalue ν for such A must be positive, and $p(\nu) > 0$.

Consequence 1: $\mathbf{x}$ *is the unique maximizer for* r *(in* S_+*).*

Proof: If also $r(\mathbf{y}) = \mu$ for some $\mathbf{y} \in S_+$, then, by the same argument, $A\mathbf{y} = \mu\mathbf{y}$, therefore $A\mathbf{z} = \mu\mathbf{z}$ for all $\mathbf{z} = \mathbf{x} + \alpha(\mathbf{y} - \mathbf{x})$ with $\alpha \in \mathbb{R}$, and each of these $\mathbf{z}$ must be positive if it is nonnegative, and this is possible only if $\mathbf{y} - \mathbf{x} = 0$ (since otherwise, since $\mathbf{x} > 0$, there is an absolutely smallest α for which some component of $\mathbf{z}$ is zero, hence $\mathbf{z} \geq \mathbf{0}$ yet not $> \mathbf{0}$). □

Consequence 2: *For any eigenvalue* ν *of any matrix* B *with eigenvector* $\mathbf{y}$*, if* $|B| \leq A$*, then* $|\nu| \leq \mu$*, with equality only if* $|\,\mathbf{y}/\|\mathbf{y}\|\,| = \mathbf{x}$ *and* $|B| = A$*.* More than that, equality implies that $B = (\mu/\nu)DAD^{-1}$, with $D := \operatorname{diag}(\ldots, y_j/|y_j|, \ldots)$.

Proof: Observe that

$$|\nu||\mathbf{y}| = |B\mathbf{y}| \leq |B|\,|\mathbf{y}| \leq A|\mathbf{y}|, \tag{13.21}$$

hence $|\nu| \leq r(|\mathbf{y}|) \leq \mu$. If now $|\nu| = \mu$, then, by the uniqueness of the minimizer $\mathbf{x}$ (and assuming without loss that $\|\mathbf{y}\| = 1$), we must have $|\mathbf{y}| = \mathbf{x}$ and equality throughout (13.21), and this implies $|B| = A$, by Problem 13.12, since we now know that $|\mathbf{y}| > 0$, while $|B| \leq A$ by assumption.
Moreover, the positivity of $|\mathbf{y}|$ implies that $D := \operatorname{diag}(\ldots, y_j/|y_j|, \ldots)$ is well defined and satisfies $\mathbf{y} = D|\mathbf{y}|$, therefore $A|\mathbf{y}| = \mu|\mathbf{y}| = (\mu/\nu)D^{-1}(\nu\mathbf{y})$ while $\nu\mathbf{y} = B\mathbf{y} = BD|\mathbf{y}|$, hence, altogether, $A|\mathbf{y}| = C|\mathbf{y}|$ with $C := (\mu/\nu)D^{-1}BD$, hence $|C| = |B| \leq A$, therefore $A = C$ by Problem 13.12. □

Consequence 3: By choosing $B = A$ in Consequence 2, we get that $\mu = \rho(A) := \max\{|\nu| : \nu \in \sigma(A)\}$, and that μ *has geometric multiplicity 1 (as an eigenvalue of* A*)* since, for any eigenvector $\mathbf{y}$ of A for the eigenvalue μ, we must have $A = DAD^{-1}$ with $D = \mathrm{diag}(\ldots, y_i/|y_i|, \ldots)$, hence also $0 < p(A) = Dp(A)D^{-1}$ which implies that $0 < D$, hence that $\mathbf{y} > 0$, hence $\mathbf{y}/\|\mathbf{y}\| = \mathbf{x}$. We also get that $\rho(A)$ *is strictly monotone in the entries of* A, i.e., that $\rho(\widetilde{A}) > \rho(A)$ in case $\widetilde{A} \ge A \ne \widetilde{A}$ (using the fact that $p(A) > 0$ and $\widetilde{A} \ge A$ implies that also $q(\widetilde{A}) > 0$ for some polynomial q; see Problem 13.14).

13.14* Prove: *(i) If* $0 \le A \le \widetilde{A}$, *then* $0 \le A^k \le \widetilde{A}^k$ *for* $k \in \mathbb{N}$*; (ii) Use (i) to show that if* $p(A) > 0$ *for some polynomial* p, *and* $\widetilde{A} \ge A \ge 0$, *then* $q(A) > 0$ *for some polynomial* q.

As a consequence, we find *computable* upper and lower bounds for the spectral radius of A:

Claim 3:

$$\forall \mathbf{y} > 0, \quad r(\mathbf{y}) \le \rho(A) \le R(\mathbf{y}) := \max_i (A\mathbf{y})_i / y_i,$$

with equality in one or the other if and only if there is equality throughout if and only if $\mathbf{y} = \alpha\mathbf{x}$ *(for some positive* α*). In particular,* $\rho(A)$ *is the only eigenvalue of* A *with nonnegative eigenvector.*

Proof: Assume without loss that $\mathbf{y} \in S_+$. We already know that, for any $\mathbf{0} < \mathbf{y} \in S_+$, $r(\mathbf{y}) \le \rho(A)$ with equality if and only if $\mathbf{y} = \mathbf{x}$. For the other inequality, observe that $R(\mathbf{y}) = \|C^{-1}AC\mathbf{e}\|_\infty$ with $C := \mathrm{diag}(\ldots, y_j, \ldots)$ and $\mathbf{e} := (1, \ldots, 1)$, hence $AC\mathbf{e} = A\mathbf{y}$. Since $C^{-1}AC \ge 0$, it takes on its max-norm at $\mathbf{e}$, by Problem 7.7, hence

$$R(\mathbf{y}) = \|C^{-1}AC\|_\infty \ge \rho(C^{-1}AC) = \rho(A).$$

Now assume that $r(\mathbf{y}) = R(\mathbf{y})$. Then $A\mathbf{y} = r(\mathbf{y})\mathbf{y}$, hence $r(\mathbf{y}) \le r(\mathbf{x}) = \rho(A) \le R(\mathbf{y}) = r(\mathbf{y})$, therefore equality must hold throughout and, in particular, $\mathbf{y} = \mathbf{x}$. Note that $R(\mathbf{y}) = \nu = r(\mathbf{y})$ in case $\mathbf{y}$ is a positive eigenvector for A with eigenvalue ν, hence $\rho(A)$ is, indeed, the only eigenvalue of A with a positive eigenvector.

If, on the other hand, $r(\mathbf{y}) < R(\mathbf{y})$, then we can find $\widetilde{A} \ne A \le \widetilde{A}$ so that $\widetilde{A}\mathbf{y} = R(\mathbf{y})\mathbf{y}$ (indeed, then $\mathbf{z} := R(\mathbf{y})\mathbf{y} - A\mathbf{y}$ is nonnegative but not $\mathbf{0}$, hence $\widetilde{A} := A + y_1^{-1}[\mathbf{z}]\mathbf{e}_1{}^{\mathrm{t}}$ does the job) therefore $r_{\widetilde{A}}(\mathbf{y}) = R(\mathbf{y}) = R_{\widetilde{A}}(\mathbf{y})$, hence $R(\mathbf{y}) = \rho(\widetilde{A}) > \rho(A)$. □

Claim 4: μ *has simple algebraic multiplicity.*

Proof: Since we already know that μ has simple geometric multiplicity, it suffices to show that μ is not a defective eigenvalue, i.e., that $\mathrm{null}(A - \mu\,\mathrm{id}) \cap \mathrm{ran}(A - \mu\,\mathrm{id}) = \{0\}$. So assume to the contrary that $A\mathbf{y} - \mu\mathbf{y}$ is an eigenvector of A belonging to μ. Since A and μ are real, then

$(A-\mu\,\mathrm{id})\,\mathrm{Re}\,\mathbf{y}$ and $(A-\mu\,\mathrm{id})\,\mathrm{Im}\,\mathbf{y}$ are in $\mathrm{null}(A-\mu\,\mathrm{id})$ and they can't both be zero. Therefore, we may assume that $(A-\mu\,\mathrm{id})\mathbf{y}$ is real, hence, by the simple geometric multiplicity of μ, we may assume without loss that $A\mathbf{y}-\mu\mathbf{y}=\mathbf{x}$, or $A\mathbf{y}=\mu\mathbf{y}+\mathbf{x}$, therefore, by induction on k, $A^k\mathbf{y}=\mu^k\mathbf{y}+k\mu^{k-1}\mathbf{x}$, hence, finally,

$$(A/\mu)^k\mathbf{y}=\mathbf{y}+k(\mathbf{x}/\mu).$$

Hence, for large enough k, $\mathbf{z}:=(A/\mu)^k\mathbf{y}\in\mathbb{R}^n_+$, and $A\mathbf{z}=\mu(A/\mu)^{k+1}\mathbf{y}=\mu(\mathbf{z}+\mathbf{x}/\mu)>\mu\mathbf{z}$, therefore $r(\mathbf{z})>\mu$ by (13.20), a contradiction. □

The collection of these claims/consequences constitutes the **Perron-Frobenius Theorem**. Oskar Perron proved all this under the assumption that $A>0$ (i.e., $p=()^1$) but observed that it is sufficient to assume that $A^k>0$ for some k (i.e., $p=()^k$). Frobenius extended it to all $A\ge 0$ that are **irreducible**. While this term has some algebraic and geometric meaning (see below), its most convenient definition for the present purpose is that $p(A)>0$ for some polynomial p. In the contrary case, A is called **reducible**, and not(iv) below best motivates such a definition. Here are some equivalent statements:

Claim 5: *Let $A\ge 0$. Then the following are equivalent:*

(i) *$p(A)>0$ for some polynomial p.*

(ii) *For all (i,j), there exists $k=k(i,j)$ so that $(A^k)_{ij}>0$.*

(iii) *No proper A-invariant subspace is spanned by some $\mathbf{e}_j$'s.*

(iv) *For no permutation matrix P is*

$$PAP^{-1}=\begin{bmatrix}B & C\\ 0 & D\end{bmatrix} \tag{13.22}$$

with B,D square matrices of positive order.

(v) *The directed graph for A is strongly connected.*

Proof: (ii)$\Longrightarrow$(i) since then $p(A):=\sum_{i,j}A^{k(i,j)}>0$.

If (ii) does not hold, then there exists (i,j) so that $(A^k)_{ij}=0$ for all k. But then also $p(A)_{ij}=0$ for all polynomials p; in other words, not(ii) implies not(i) which is equivalent to (i)$\Longrightarrow$(ii). Further, it says that the set $J:=J(j):=\{r:\exists\{k\}\ (A^k)_{rj}\ne 0\}$ is a proper subset of $\{1,\ldots,n\}$ (since it doesn't contain i), but neither is it empty (since it contains j, as $(A^0)_{jj}\ne 0$). Since $(A^{k+\ell})_{rj}=\sum_m(A^k)_{rm}(A^\ell)_{mj}$, it follows that $J(m)\subset J(j)$ for all $m\in J(j)$. This implies, in particular, that $A_{rm}=0$ for all $r\notin J(j), m\in J(j)$, hence that $\mathrm{ran}[\mathbf{e}_m : m\in J(j)]$ is a proper A-invariant subspace, thus implying not(iii). It also implies not(iv), since it shows that the columns $A_{:m}$, $m\in J(j)$, have zero entries in rows r, $r\notin J(j)$, i.e., that (13.22) holds for the permutation $P=[(\mathbf{e}_m)_{m\in J(j)},(\mathbf{e}_r)_{r\notin J(j)}]$, with both B and D of order $<n$.

Conversely, if e.g., (iii) does not hold, and $\mathrm{ran}[\mathbf{e}_m : m\in J(j)]$ is that proper A-invariant subspace, then it is also invariant under any $p(A)$, hence also $p(A)_{rm} = 0$ for every $r \notin J(j)$, $m \in J(j)$, i.e., (i) does not hold, while not(iv) evidently implies not(iii).

The final characterization is explicitly that given by Frobenius, – except that he did not formulate it in terms of graphs; that was done much later, by Rosenblatt (1957) and Varga (1962). Frobenius (1912) observed that, since

$$(A^k)_{ij} = \sum_{\nu_1} \cdots \sum_{\nu_{k-1}} A_{i,\nu_1} \cdots A_{\nu_{k-1},j},$$

therefore, for $i \neq j$, $(A^k)_{ij} \neq 0$ if and only if there exists some sequence $i =: i_0, i_1, \ldots, i_{k-1}, i_k := j$ so that $A_{i_r,i_{r+1}} \neq 0$ for all r. Now, the **directed graph** of A is the graph with n vertices in which the directed edge (i,j) is present iff $A_{ij} \neq 0$. Such a graph is called **strongly connected** in case it contains, for each $i \neq j$, a path connecting vertex i with vertex j, and this, as we just observed, is equivalent to having $(A^k)_{ij} \neq 0$ for some $k > 0$. In short, (ii) and (v) are equivalent. □

There are various refinements of this last claim available. For example, in testing whether the directed graph of A is strongly connected, we only need to check paths involving distinct vertices, and such paths involve at most n vertices. Hence, in condition (ii), we need to check only for $k < n$. But, with that restriction, (ii) is equivalent to having $\mathrm{id}_n + A + \cdots + A^{n-1} > 0$, i.e., to having (i) hold for quite a specific polynomial.

13.15 T/F

(a) A noninvertible matrix cannot be diagonally dominant.

(b) A symmetric matrix has only real eigenvalues.

(c) If $\mu \in \mathrm{spct}(A)$, $\nu \in \mathrm{spct}(B)$ for some $A, B \in L(X)$, then $\mu\nu \in \mathrm{spct}(AB)$.

(d) The geometric multiplicity of an eigenvalue cannot exceed its algebraic multiplicity.

(e) For a simple eigenvalue, its algebraic multiplicity equals its geometric multiplicity.

(f) The collection of all univariate polynomials vanishing at 1 is an ideal.

(g) If $\deg q < \deg p_A$, then $q(A)$ is invertible.

(h) If $p(A) = 0$ for some polynomial p, then every zero of p is an eigenvalue of A.

(i) If the sum $A + B$ of two matrices is defined, then $\det(A + B) = \det A + \det B$.

14 Optimization and quadratic forms

Minimization

We are interested in *minimizing* a given function

$$f : \operatorname{dom} f \subset \mathbb{R}^n \to \mathbb{R},$$

i.e., we are looking for $\mathbf{x} \in \operatorname{dom} f$ so that

$$\forall \mathbf{y} \in \operatorname{dom} f, \quad f(\mathbf{x}) \le f(\mathbf{y}).$$

Any such $\mathbf{x}$ is called a **minimizer for** f; in symbols:

$$\mathbf{x} \in \operatorname{argmin} f.$$

The discussion applies, of course, also to finding some $\mathbf{x} \in \operatorname{argmax} f$, i.e., finding a **maximizer** for f, since $\mathbf{x} \in \operatorname{argmax} f$ iff $\mathbf{x} \in \operatorname{argmin}(-f)$.

Finding minimizers is, in general, an impossible problem since one cannot tell whether or not $\mathbf{x} \in \operatorname{argmin} f$ except by checking *every* $\mathbf{y} \in \operatorname{dom} f$ to make certain that, indeed, $f(\mathbf{x}) \le f(\mathbf{y})$. However, if f is a 'smooth' function, then one can in principle check whether, at least, $\mathbf{x}$ is a **local minimizer**, i.e., whether $f(\mathbf{x}) \le f(\mathbf{y})$ for all 'nearby' $\mathbf{y}$, by checking whether the **gradient**

$$Df(\mathbf{x}) = (D_i f(\mathbf{x}) : i = 1{:}n)$$

of f at $\mathbf{x}$ is zero. Here, $D_i f = \partial f / \partial x_i$ is the derivative of f with respect to its ith argument.

To be sure, the vanishing of the gradient of f at $\mathbf{x}$ is only a *necessary* condition for $\mathbf{x}$ to be a minimizer for f, since the gradient of a (smooth) function must also vanish at any local *maximum*, and may vanish at points that are neither local minima nor local maxima but are, perhaps, only saddle points. By definition, any point $\mathbf{x}$ for which $Df(\mathbf{x}) = 0$ is a **critical point** for f.

At a critical point, f is locally flat. This means that, in the Taylor expansion

$$f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + (Df(\mathbf{x}))^{\mathrm{t}}\mathbf{h} + \mathbf{h}^{\mathrm{t}}(D^2 f(\mathbf{x})/2)\mathbf{h} + \text{h.o.t.}(\mathbf{h})$$

for f at $\mathbf{x}$, the linear term, $(Df(\mathbf{x}))^{\mathrm{t}}\mathbf{h}$, is zero. Thus, if the matrix

$$H := D^2 f(\mathbf{x}) = (D_i D_j f(\mathbf{x}) : i, j = 1{:}n)$$

of second derivatives of f is 1-1, then $\mathbf{x}$ is a local minimizer (maximizer) for f if and only if $\mathbf{0}$ is a minimizer (maximizer) for the **quadratic form**

$$\mathbb{R}^n \to \mathbb{R} : \mathbf{h} \mapsto \mathbf{h}^{\mathrm{t}} H \mathbf{h}$$

associated with the **Hessian** $H = D^2 f(\mathbf{x})$ for f at $\mathbf{x}$.

If all second derivatives of f are continuous, then also $D_i D_j f = D_j D_i f$, hence the Hessian is real symmetric, therefore

$$H^{\mathrm{t}} = H.$$

However, in the contrary case, one simply defines H to be

$$H := (D^2 f(\mathbf{x}) + (D^2 f(\mathbf{x}))^{\mathrm{t}})/2,$$

thus making it real symmetric while, still,

$$\forall \mathbf{h} \in \mathbb{R}^n, \quad \mathbf{h}^{\mathrm{t}} H \mathbf{h} = \mathbf{h}^{\mathrm{t}} D^2 f(\mathbf{x}) \mathbf{h}.$$

In any case, it follows that *quadratic forms model the behavior of a smooth function 'near' a critical point* (and this is true in a trivial sort of way even when the Hessian is 0 at that point). The importance of minimization of real-valued functions is the prime motivation for the study of quadratic forms, to which we now turn.

Quadratic forms

Each $A \in \mathbb{R}^{n\times n}$ gives rise to a quadratic form, denoted by a_A, via

$$q_A : \mathbb{R}^n \to \mathbb{R} : \mathbf{x} \mapsto \mathbf{x}^{\mathrm{t}} A \mathbf{x}.$$

However, as we already observed, the quadratic form 'sees' only the **symmetric part**

$$(A + A^{\mathrm{t}})/2$$

of A, i.e.,

$$\forall \mathbf{x} \in \mathbb{R}^n, \quad \mathbf{x}^{\mathrm{t}} A \mathbf{x} = \mathbf{x}^{\mathrm{t}} \frac{A + A^{\mathrm{t}}}{2} \mathbf{x}.$$

For this reason, in discussions of the quadratic form q_A, we will always assume that A is real symmetric.

The Taylor expansion for q_A is very simple. One computes

$$\begin{aligned} q_A(\mathbf{x}+\mathbf{h}) = (\mathbf{x}+\mathbf{h})^{\mathrm{t}} A(\mathbf{x}+\mathbf{h}) &= \mathbf{x}^{\mathrm{t}} A\mathbf{x} + \mathbf{x}^{\mathrm{t}} A\mathbf{h} + \mathbf{h}^{\mathrm{t}} A\mathbf{x} + \mathbf{h}^{\mathrm{t}} A\mathbf{h} \\ &= q_A(\mathbf{x}) + 2(A\mathbf{x})^{\mathrm{t}}\mathbf{h} + \mathbf{h}^{\mathrm{t}} A\mathbf{h}, \end{aligned}$$

using the fact that $A^{\mathrm{t}} = A$, thus $\mathbf{h}^{\mathrm{t}} A\mathbf{x} = \mathbf{x}^{\mathrm{t}} A\mathbf{h} = (A\mathbf{x})^{\mathrm{t}}\mathbf{h}$, hence

$$Dq_A(\mathbf{x}) = 2A\mathbf{x}, \quad D^2 q_A(\mathbf{x}) = 2A.$$

It follows that, for any 1-1 A, $\mathbf{0}$ is the only critical point of q_A. The sought-for classification of critical points of smooth functions has led to the following classification of quadratic forms:

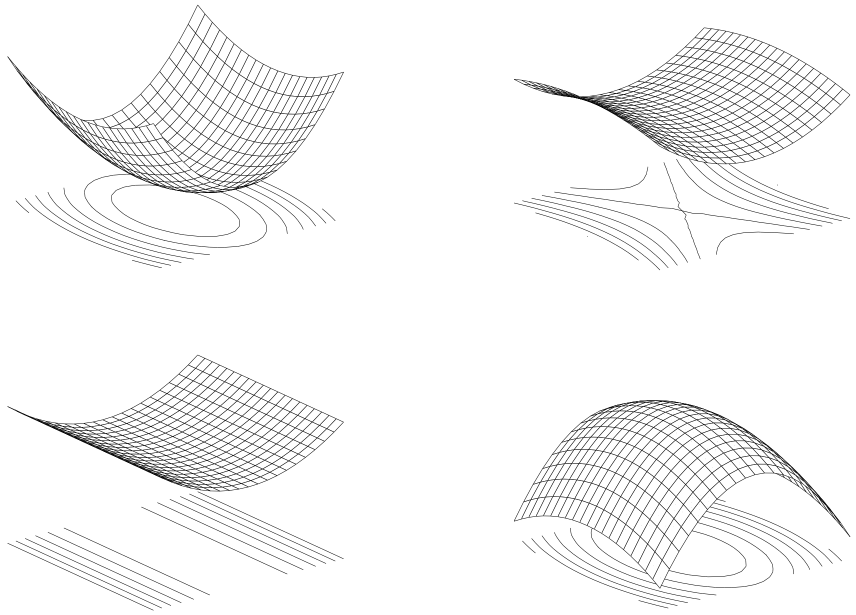
$$A \text{ is } \begin{array}{c} \textbf{positive} \\ \textbf{positive semi-} \\ \textbf{negative semi-} \\ \textbf{negative} \end{array} \textbf{definite} \;:=\; 0 \text{ is } \begin{array}{l} \textbf{the unique minimizer} \\ \textbf{a minimizer} \\ \textbf{a maximizer} \\ \textbf{the unique maximizer} \end{array} \text{ for } q_A.$$

If none of these conditions obtains, i.e., if there exist $\mathbf{x}$ and $\mathbf{y}$ so that $\mathbf{x}^{\mathrm{t}} A\mathbf{x} < 0 < \mathbf{y}^{\mathrm{t}} A\mathbf{y}$, then q_A is called **indefinite** and, in this case, $\mathbf{0}$ is a **saddle point** for q_A.

(14.1)Figure shows three quadratic forms near their unique critical point. One is a minimizer, another is a saddle point, and the last one is a maximizer. Also shown is a quadratic form with a whole straight line of critical points. The figure (generated by the `MATLAB` command `meshc`) also shows some **contour lines** or **level lines**, i.e., lines in the domain $\mathbb{R}^2$ along which the function is constant. The contour plots are characteristic: Near an extreme point, be it a maximum or a minimum, the level lines are ellipses, with the extreme point their center, while near a saddle point, the level lines are hyperbolas, with the extreme point their center and with two level lines actually crossing at the saddle point.

There is an intermediate case between these two, also shown in (14.1)Figure, in which the level lines are parallel lines and, correspondingly, there is a whole line of critical points. In this case, the quadratic form is semidefinite. Note, however, that the *definition* of semidefiniteness does not exclude the possibility that the quadratic form is actually definite.

Since, near any critical point $\mathbf{x}$, a smooth f behaves like its quadratic term $\mathbf{h} \mapsto \mathbf{h}^{\mathrm{t}}(D^2 f(\mathbf{x})/2)\mathbf{h}$, we can be sure that a contour plot for f near an extremum would approximately look like concentric ellipses while, near a saddle point, it would look approximately like concentric hyperbolas.



(14.1) Figure. Local behavior near a critical point.

These two patterns turn out to be the only possible ones for quadratic forms on $\mathbb{R}^2$ with a unique critical point. On $\mathbb{R}^n$, there are only $\lceil (n+1)/2 \rceil$ possible distinct patterns, as follows from the fact about to be proved that, *for every quadratic form q_A, there are o.n. coordinate systems U for which*

$$q_A(\mathbf{x}) = \sum_{i=1}^{n} d_i \; (U^{c}\mathbf{x})_i^2.$$

14.1 For each of the following three functions on $\mathbb{R}^2$, compute the Hessian $D^2 f(\mathbf{0})$ at $\mathbf{0}$ and use it to determine whether $\mathbf{0}$ is a (local) maximum, minimum, or neither. (In an effort to make the derivation of the Hessians simple, I have made the problems so simple that you could tell by inspection what kind of critical point $\mathbf{0} = (0,0) \in \mathbb{R}^2$ is; nevertheless, give your answer based on the spectrum of the Hessian.)

(a) $f(x,y) = (x-y)\sin(x+y)$;

(b) $f(x,y) = (x+y)\sin(x+y)$;

(c) $f(x,y) = (x+y)\cos(x+y)$.

14.2 Construct a bivariate quadratic form q for which the level lines are parabolas and show that it has no critical point.

14.3 Discuss the $3 = \lceil (4+1)/2 \rceil$ essentially different patterns near a unique critical point of a quadratic form of 4 variables.

14.4 Why is $\mathbf{0}$ called a monkey saddle point for $f(\mathbf{x}) := x_1^3 - 3x_1^2 x_2 - 3x_1 x_2^2 + x_2^3$? (Draw a picture.) Why does this behavior, which matches none of the figures in

(16.2)Figure, not contradict the earlier claim that, in the neighborhoof of a unique critical point, a quadratic function has just two possible patterns?

Reduction of a quadratic form to a sum of squares

Consider the effects of a *change of basis*. Let $V \in \mathbb{R}^n$ be a basis for $\mathbb{R}^n$ and consider the map

$$f := q_A \circ V.$$

We have $f(\mathbf{x}) = (V\mathbf{x})^{\mathrm{t}} AV\mathbf{x} = \mathbf{x}^{\mathrm{t}}(V^{\mathrm{t}}AV)\mathbf{x}$, hence

$$q_A \circ V = q_{V^{\mathrm{t}}AV}.$$

This makes it interesting to look for bases V for which $V^{\mathrm{t}}AV$ is as simple as possible. Matrices A and B for which $B = V^{\mathrm{t}}AV$ are said to be **congruent** to each other. Note that congruent matrices are not necessarily similar; in particular, their spectra can be different. However, by Sylvester's Law of Inertia (see (14.9) below), congruent *hermitian* matrices have the same number of positive, of zero, and of negative, eigenvalues. This is not too surprising in view of the following *reduction to a sum of squares* which is possible for any quadratic form.

(14.2) Proposition: Every quadratic form q_A on $\mathbb{R}^n$ can be written in the form

$$q_A(\mathbf{x}) = \sum_{j=1}^{n} d_j \ ({\mathbf{u}_j}^{\mathrm{t}}\mathbf{x})^2,$$

for some suitable o.n. basis $U = [\mathbf{u}_1, \ldots, \mathbf{u}_n]$ for which

$$U^{\mathrm{t}}AU = \mathrm{diag}(d_1, \ldots, d_n) \in \mathbb{R}^{n\times n}.$$

Proof: Since A is hermitian, there exists, by (12.2)Corollary, some o.n. basis $U = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n]$ for $\mathbb{F}^n$ for which $U^{\mathrm{t}}AU = \mathrm{diag}(d_1, d_2, \ldots, d_n) \in \mathbb{R}^{n\times n}$. Now use the fact that $U^{\mathrm{t}}U = \mathrm{id}_n$ and therefore $q_A(\mathbf{x}) = q_{U^{\mathrm{t}}AU}(U^{\mathrm{t}}\mathbf{x})$ to obtain for $q_A(\mathbf{x})$ the displayed expression. □

What about the classification introduced earlier, into positive or negative (semidefinite)? The proposition permits us to visualize $q_A(\mathbf{x})$ as a weighted sum of squares (with real weights $d_1, \ldots, d_n$) and $U^{\mathrm{t}}\mathbf{x}$ an arbitrary n-vector (since U is a basis), hence permits us to conclude that q_A is definite if and only if all the d_j are strictly of one sign, semidefinite if and only if all the d_j are of one sign (with zero possible), and indefinite if and only if there are both positive and negative d_j.

MATLAB readily provides these numbers d_j by the command eig(A).

Consider specifically the case $n = 2$ for which we earlier provided some pictures. Assume without loss that $d_1 \le d_2$. If $0 < d_1$, then A is positive definite and, correspondingly, the contour line

$$c_r := \{\mathbf{x} \in \mathbb{R}^2 : q_A(\mathbf{x}) = r\} = \{\mathbf{x} \in \mathbb{R}^2 : d_1(\mathbf{u}_1{}^{\rm t}\mathbf{x})^2 + d_2(\mathbf{u}_2{}^{\rm t}\mathbf{x})^2 = r\}$$

for $r > 0$ is an ellipse, with axes parallel to $\mathbf{u}_1$ and $\mathbf{u}_2$. If $0 = d_1 < d_2$, then these ellipses turn into parallel straight lines. Similarly, if $d_2 < 0$, then the contour line

$$c_r := \{\mathbf{x} \in \mathbb{R}^2 : q_A(\mathbf{x}) = r\} = \{\mathbf{x} \in \mathbb{R}^2 : d_1(\mathbf{u}_1{}^{\rm t}\mathbf{x})^2 + d_2(\mathbf{u}_2{}^{\rm t}\mathbf{x})^2 = r\}$$

for $r < 0$ is an ellipse, with axes parallel to $\mathbf{u}_1$ and $\mathbf{u}_2$. Finally, if $d_1 < 0 < d_2$, then, for any r, the contour line

$$c_r := \{\mathbf{x} \in \mathbb{R}^2 : q_A(\mathbf{x}) = r\} = \{\mathbf{x} \in \mathbb{R}^2 : d_1(\mathbf{u}_1{}^{\rm t}\mathbf{x})^2 + d_2(\mathbf{u}_2{}^{\rm t}\mathbf{x})^2 = r\}$$

is a hyperbola, with axes parallel to $\mathbf{u}_1$ and $\mathbf{u}_2$.

Note that such an o.n. basis U is Cartesian, i.e., its columns are orthogonal to each other (and are normalized). This means that we can visualize the change of basis, from the natural basis to the o.n. basis U, as a rigid motion, involving nothing more than rotations and reflections.

Rayleigh quotient

This section is devoted to the proof and exploitation of the following remarkable

Fact: The eigenvectors of a hermitian matrix A are the critical points of the corresponding **Rayleigh quotient**

$$R_A(\mathbf{x}) := \langle A\mathbf{x}, \mathbf{x}\rangle / \langle \mathbf{x}, \mathbf{x}\rangle,$$

and $R_A(\mathbf{x}) = \mu$ in case $A\mathbf{x} = \mu\mathbf{x}$.

This fact has many important consequences concerning how the eigenvalues of a hermitian matrix depend on that matrix, i.e., how the eigenvalues change when the entries of the matrix are changed, by round-off or for other reasons.

This perhaps surprising connection has the following intuitive explanation: Suppose that $A\mathbf{x} \notin \operatorname{ran}[\mathbf{x}]$. Then $\mathbf{x} \ne \mathbf{0}$ and the error $\mathbf{h} := A\mathbf{x} - R_A(\mathbf{x})\mathbf{x}$

in the least-squares approximation to $A\mathbf{x}$ from ran$[\mathbf{x}]$ is not zero, and is perpendicular to ran$[\mathbf{x}]$. Consequently, $\langle A\mathbf{x}, \mathbf{h}\rangle = \langle \mathbf{h}, \mathbf{h}\rangle > 0$, and therefore the value

$$\langle A(\mathbf{x}+t\mathbf{h}), \mathbf{x}+t\mathbf{h}\rangle = \langle A\mathbf{x}, \mathbf{x}\rangle + 2t\langle A\mathbf{x}, \mathbf{h}\rangle + t^2\langle A\mathbf{h}, \mathbf{h}\rangle$$

of the numerator of $R_A(\mathbf{x}+t\mathbf{h})$ grows linearly for t near 0, while its denominator

$$\langle \mathbf{x}+t\mathbf{h}, \mathbf{x}+t\mathbf{h}\rangle = \langle \mathbf{x}, \mathbf{x}\rangle + t^2\langle \mathbf{h}, \mathbf{h}\rangle$$

grows only quadratically, i.e., much less fast for t near zero. It follows that, more precisely, the derivative of $f(t) := R_A(\mathbf{x}+t\mathbf{h})$ at $t = 0$ is $\langle \mathbf{h}, \mathbf{h}\rangle/\langle \mathbf{x}, \mathbf{x}\rangle \neq 0$, hence $\mathbf{x}$ cannot be a critical point for R_A. – To put it differently, for any critical point $\mathbf{x}$ for R_A, we necessarily have $A\mathbf{x} \in$ ran$[\mathbf{x}]$, therefore $A\mathbf{x} = R_A(\mathbf{x})\mathbf{x}$. Of course, that makes any such $\mathbf{x}$ an eigenvector with corresponding eigenvalue $R_A(\mathbf{x})$. □

Next, recall from (12.2) that a hermitian matrix is unitarily similar to a real diagonal matrix. This means that we may assume, after some reordering if necessary, that

$$A = U\mathrm{M}U^{\mathrm{c}}$$

with U unitary and with $\mathrm{M} = \mathrm{diag}(\mu_1, \ldots, \mu_n)$ where

$$\mu_1 \leq \mu_2 \leq \cdots \leq \mu_n.$$

At times, we will write, more explicitly,

$$\mu_j(A)$$

to denote the jth eigenvalue of the hermitian matrix A in this ordering. Note that there may be coincidences here, i.e., $\mu_j(A)$ is the jth smallest eigenvalue of A *counting multiplicities*. Note also that, in contrast to the singular values (and in contrast to most books), we have put here the eigenvalues in *increasing* order.

Now recall that a unitary basis has the advantage that it preserves angles and lengths since $\langle U\mathbf{x}, U\mathbf{y}\rangle = \langle \mathbf{x}, \mathbf{y}\rangle$ for any orthonormal U. Thus

$$\langle A\mathbf{x}, \mathbf{x}\rangle = \langle U\mathrm{M}U^{\mathrm{c}}\mathbf{x}, \mathbf{x}\rangle = \langle \mathrm{M}(U^{\mathrm{c}}\mathbf{x}), U^{\mathrm{c}}\mathbf{x}\rangle,$$

and $\langle \mathbf{x}, \mathbf{x}\rangle = \langle U^{\mathrm{c}}\mathbf{x}, U^{\mathrm{c}}\mathbf{x}\rangle$. Therefore

$$R_A(\mathbf{x}) = \langle A\mathbf{x}, \mathbf{x}\rangle/\langle \mathbf{x}, \mathbf{x}\rangle = \langle \mathrm{M}(U^{\mathrm{c}}\mathbf{x}), U^{\mathrm{c}}\mathbf{x}\rangle/\langle U^{\mathrm{c}}\mathbf{x}, U^{\mathrm{c}}\mathbf{x}\rangle = R_{\mathrm{M}}(U^{\mathrm{c}}\mathbf{x}).$$

This implies that

$$\begin{matrix}\max_{\mathbf{x}}\\ \min_{\mathbf{x}}\end{matrix} R_A(\mathbf{x}) = \begin{matrix}\max_{\mathbf{y}}\\ \min_{\mathbf{y}}\end{matrix} R_{\mathrm{M}}(\mathbf{y}).$$

On the other hand, since M is diagonal, $\langle \mathrm{M}\mathbf{y}, \mathbf{y}\rangle = \sum_j \mu_j |y_j|^2$, therefore

$$R_{\mathrm{M}}(\mathbf{y}) = \sum_j \mu_j |y_j|^2 / \sum_j |y_j|^2,$$

and this shows that

$$\min_{\mathbf{x}} R_A(\mathbf{x}) = \min_{\mathbf{y}} R_{\mathrm{M}}(\mathbf{y}) = \mu_1(A), \qquad \max_{\mathbf{x}} R_A(\mathbf{x}) = \max_{\mathbf{y}} R_{\mathrm{M}}(\mathbf{y}) = \mu_n(A).$$

This is **Rayleigh's Principle**. It characterizes the extreme eigenvalues of a hermitian matrix. The intermediate eigenvalues $\mu_j(A)$, $1 < j < n$, are the solution of more subtle extremum problems. This is the content of the **Courant-Fischer Minimax Theorem** and its companion **Maximin Theorem**. It seems most efficient to combine both in the following

(14.3) MMM (or, **Maximinimax**) **Theorem:** Let A be a hermitian matrix of order n, hence $A = U\mathrm{M}U^{\mathrm{c}}$ for some unitary U and some real diagonal matrix $\mathrm{M} = \mathrm{diag}(\cdots, \mu_j(A), \ldots)$ with $\mu_1(A) \le \cdots \le \mu_n(A)$. Then, for $j = 1{:}n$,

$$\max_{\dim G < j} \min_{\mathbf{x}\perp G} R_A(\mathbf{x}) = \mu_j(A) = \min_{j \le \dim H} \max_{\mathbf{x}\in H} R_A(\mathbf{x}),$$

with G and H otherwise arbitrary linear subspaces.

Proof: If $\dim G < j \le \dim H$, then one can find $\mathbf{y} \in H\backslash\mathbf{0}$ with $\mathbf{y} \perp G$ (since, with V a basis for G and W a basis for H, this amounts to finding a nontrivial solution to the equation $V^{\mathrm{c}}W? = 0$, and this system is homogeneous with more unknowns than equations). Therefore

$$\min_{\mathbf{x}\perp G} R_A(\mathbf{x}) \ \le R_A(\mathbf{y}) \le \ \max_{\mathbf{x}\in H} R_A(\mathbf{x}).$$

Hence,

$$\sup_{\dim G < j} \ \min_{\mathbf{x}\perp G} R_A(\mathbf{x}) \ \le \ \inf_{j \le \dim H} \ \max_{\mathbf{x}\in H} R_A(\mathbf{x}).$$

On the other hand, for $G = \mathrm{ran}[\mathbf{u}_1, \ldots, \mathbf{u}_{j-1}]$ and $H = \mathrm{ran}[\mathbf{u}_1, \ldots, \mathbf{u}_j]$,

$$\min_{\mathbf{x}\perp G} R_A(\mathbf{x}) \ = \ \mu_j(A) \ = \ \max_{\mathbf{x}\in H} R_A(\mathbf{x}).$$

□

The MMM theorem has various useful (and immediate) corollaries.

(14.4) Interlacing Theorem: If the matrix B is obtained from the hermitian matrix A by crossing out the kth row and column (i.e., $B = A_{I,I}$ with $I := (1{:}k-1, k+1{:}n)$), then

$$\mu_j(A) \le \mu_j(B) \le \mu_{j+1}(A), \quad j < n.$$

Proof: It is sufficient to consider the case $k = n$, since we can always achieve this situation by interchanging rows k and n, and columns k and n, of A, and this will not change $\operatorname{spct}(A)$. Let $J : \mathbb{F}^{n-1} \to \mathbb{F}^n : \mathbf{x} \mapsto (\mathbf{x}, 0)$. Then $R_B(\mathbf{x}) = R_A(J\mathbf{x})$ and $\operatorname{ran} J = \operatorname{ran}[\mathbf{e}_n]\perp$, therefore also $J(G\perp) = (JG + \operatorname{ran}[\mathbf{e}_n])\perp$ and $\{JG + \operatorname{ran}[\mathbf{e}_n] : \dim G < j, G \subset \mathbb{F}^{n-1}\} \subset \{\widetilde{G} : \dim \widetilde{G} < j+1, \widetilde{G} \subset \mathbb{F}^n\}$. Hence

$$\begin{aligned}\mu_j(B) = \max_{\dim G<j}\min_{\mathbf{x}\perp G} R_A(J\mathbf{x}) &= \max_{\dim G<j}\ \min_{\mathbf{y}\perp JG+\operatorname{ran}[\mathbf{e}_n]} R_A(\mathbf{y})\\ &\le \max_{\dim \widetilde{G}<j+1}\ \min_{\mathbf{y}\perp\widetilde{G}} R_A(\mathbf{y}) = \mu_{j+1}(A).\end{aligned}$$

Also, since $\{JH : j \le \dim H, H \subset \mathbb{F}^{n-1}\} \subset \{\widetilde{H} : j \le \dim \widetilde{H}, \widetilde{H} \subset \mathbb{F}^n\}$,

$$\begin{aligned}\mu_j(B) &= \min_{j\le\dim H}\ \max_{\mathbf{x}\in H} R_A(J\mathbf{x}) = \min_{j\le\dim H}\ \max_{\mathbf{y}\in JH} R_A(\mathbf{y})\\ &\ge \min_{j\le\dim\widetilde{H}}\ \max_{\mathbf{y}\in\widetilde{H}} R_A(\mathbf{y}) = \mu_j(A).\end{aligned}$$

□

(14.5) Corollary: If $A = \begin{bmatrix} B & C \\ D & E \end{bmatrix} \in \mathbb{F}^{n\times n}$ is hermitian, and $B \in \mathbb{F}^{r\times r}$, then at least r eigenvalues of A must be $\le \max\operatorname{spct}(B)$ and at least r eigenvalues of A must be $\ge \min\operatorname{spct}(B)$.

In particular, if the spectrum of B is negative and the spectrum of E is positive, then A has exactly r negative, and $n-r$ positive, eigenvalues.

A different, simpler, application of the MMM theorem is based on the following observation: If

$$\forall t, \quad f(t) \le g(t),$$

then this inequality persists if we take on both sides the maximum or minimum over the same set T, i.e., then

$$\max_{t\in T} f(t) \le \max_{t\in T} g(t), \qquad \min_{t\in T} f(t) \le \min_{t\in T} g(t).$$

It even persists if we further take the minimum or maximum over the same family $\mathbf{T}$ of subsets T, e.g., then also

$$\max_{T\in\mathbf{T}} \min_{t\in T} f(t) \le \max_{T\in\mathbf{T}} \min_{t\in T} g(t).$$

Consequently,

(14.6) Corollary: If A, B are hermitian, and $R_A(\mathbf{x}) \le R_B(\mathbf{x}) + c$ for some constant c and all $\mathbf{x}$, then

$$\forall j, \quad \mu_j(A) \le \mu_j(B) + c.$$

This gives

(14.7) Weyl's Inequalities: If $A = B + C$, with A, B, C hermitian, then

$$\forall j, \quad \mu_j(B) + \mu_1(C) \le \mu_j(A) \le \mu_j(B) + \mu_n(C).$$

Proof: Since $\mu_1(C) \le R_C(\mathbf{x}) \le \mu_n(C)$ (by Rayleigh's principle), while $R_B(\mathbf{x}) + R_C(\mathbf{x}) = R_A(\mathbf{x})$, the preceding corollary provides the proof. □

A typical *application of Weyl's Inequalities* is the observation that, for $A = BB^{c} + C \in \mathbb{F}^{n\times n}$ with $B \in \mathbb{F}^{n\times k}$ and A hermitian (hence also C hermitian), $\mu_1(C) \le \mu_j(A) \le \mu_n(C)$ for all $j < (n-k)$, since $\operatorname{rank} BB^{c} \le \operatorname{rank} B \le k$, hence $\mu_j(BB^{c})$ must be zero for $j < (n-k)$.

Since $C = A - B$, Weyl's Inequalities imply that

$$|\mu_j(A) - \mu_j(B)| \le \max\{|\mu_1(A-B)|, |\mu_n(A-B)|\} = \rho(A-B).$$

Therefore, with the substitutions $A \leftarrow A+E$, $B \leftarrow A$, we obtain

(14.8) max-norm Wielandt-Hoffman: If A and E are both hermitian, then

$$\max_j |\mu_j(A+E) - \mu_j(A)| \le \max_j |\mu_j(E)|.$$

A corresponding statement involving 2-norms is valid but much harder to prove.

Finally, a totally different application of the MMM Theorem is

(14.9) Sylvester's Law of Inertia: Any two *congruent* hermitian matrices have the same number of positive, zero, and negative eigenvalues.

Proof: It is sufficient to prove that if $B = V^{\mathrm{c}}AV$ for some hermitian A and some invertible V, then $\mu_j(A) > 0$ implies $\mu_j(B) > 0$. For this, we observe that, by the MMM Theorem, $\mu_j(A) > 0$ implies that R_A is positive somewhere on every j-dimensional subspace, while (also by the MMM Theorem), for some j-dimensional subspace H,

$$\mu_j(B) \;=\; \max_{\mathbf{x}\in H} R_B(\mathbf{x}) = \max_{\mathbf{x}\in H} R_A(V\mathbf{x})R_{V^{\mathrm{c}}V}(\mathbf{x}),$$

and this is necessarily positive, since $\dim VH = j$ and

$$R_{V^{\mathrm{c}}V}(\mathbf{x}) = \|V\mathbf{x}\|^2/\|\mathbf{x}\|^2$$

is positive for any $\mathbf{x} \neq \mathbf{0}$. □

It follows that we don't have to diagonalize the real symmetric matrix A (as we did in the proof of (14.2)Proposition) in order to find out whether or not A or the corresponding quadratic form q_A is definite. Assuming that A is invertible, hence has no zero eigenvalue, it is sufficient to use Gauss elimination without pivoting to obtain the factorization $A = LDL^{\mathrm{c}}$, with L unit lower triangular. By Sylvester's Law of Inertia, the number of positive (negative) eigenvalues of A equals the number of positive (negative) diagonal entries of D.

This fact can be used to locate the eigenvalues of a real symmetric matrix by *bisection*. For, the number of positive (negative) diagonal entries in the diagonal matrix D_μ obtained in the factorization $L_\mu D_\mu {L_\mu}^{\mathrm{c}}$ for $(A - \mu\,\mathrm{id})$ tells us the number of eigenvalues of A to the right (left) of μ, hence makes it easy to locate and refine intervals that contain just one eigenvalue of A.

15 More on determinants

Determinants are often brought into courses such as this quite unnecessarily. But when they are useful, they are remarkably so. The use of determinants is a bit bewildering to the beginner, particularly if confronted with the classical definition as a sum of signed products of matrix entries.

I find it more intuitive to follow Weierstrass and begin with a few important properties of the determinant, from which all else follows, including that classical definition (which is practically useless anyway).

As to the many determinant identities available, in the end I have almost always managed with just one nontrivial one, viz. *Sylvester's determinant identity*, and this is a direct consequence of Gauss elimination The only other one I have used at times is the *Binet-Cauchy Formula.* Both are stated and derived at the end of this chapter.

Definition and basic properties

The determinant is a map,

$$\det : \mathbb{F}^{n\times n} \to \mathbb{F} : A \mapsto \det A,$$

with various properties. The first one in the following list is perhaps the most important one; the second one serves as a normalization and, along with properties (iv) and (v), uniquely defines the map, as we will show by, eventually, deriving all the properties listed here, including the property (i), from the three properties (ii), (iv) and (v).

(i) $\det(AB) = \det(A)\det(B)$.

(ii) $\det \mathrm{id} = 1$.

Consequently, for any invertible A,

$$1 = \det \mathrm{id} = \det(AA^{-1}) = \det(A)\det(A^{-1}).$$

Hence,

(iii) *If A is invertible, then* $\det A \neq 0$ *and,* $\det(A^{-1}) = 1/\det A$.

While the determinant is defined as a map on matrices, it is very useful to think of $\det A = \det[\mathbf{a}_1, \ldots, \mathbf{a}_n]$ as a function of the columns $\mathbf{a}_1, \ldots, \mathbf{a}_n$ of A. The next two properties are in those terms:

(iv) *The determinant is a* **multilinear** *form*, i.e., for every j, the map $\mathbf{x} \mapsto \det[\ldots, \mathbf{a}_{j-1}, \mathbf{x}, \mathbf{a}_{j+1}, \ldots]$ *is linear*, meaning that, for any n-vectors $\mathbf{x}$ and $\mathbf{y}$ and any scalar α (and arbitrary n-vectors $\mathbf{a}_i$),

$$\begin{aligned}\det[\ldots,&\mathbf{a}_{j-1}, \mathbf{x} + \alpha\mathbf{y}, \mathbf{a}_{j+1}, \ldots]\\ &= \det[\ldots, \mathbf{a}_{j-1}, \mathbf{x}, \mathbf{a}_{j+1}, \ldots] + \alpha \det[\ldots, \mathbf{a}_{j-1}, \mathbf{y}, \mathbf{a}_{j+1}, \ldots].\end{aligned}$$

(v) *The determinant is an* **alternating** *form*, i.e.,

$$\det[\ldots, \mathbf{a}_i, \ldots, \mathbf{a}_j, \ldots] = -\det[\ldots, \mathbf{a}_j, \ldots, \mathbf{a}_i, \ldots].$$

In words: Interchanging two columns changes the sign of the determinant but not its absolute value.

It can be shown (see page 229) that (ii) + (iv) + (v) implies (i) (and anything else you may wish to prove about determinants). Here are some basic consequences first.

Since 0 is the only scalar α with the property that $\alpha = -\alpha$, it follows from (v) that

(vi) $\det A = 0$ *if two columns of A are the same.*

Using first (iv) and then the consequence (vi) of (v), we compute

$$\begin{aligned}\det[\ldots, \mathbf{a}_i, \ldots, \mathbf{a}_j + \alpha\mathbf{a}_i, \ldots]&\\ &= \det[\ldots, \mathbf{a}_i, \ldots, \mathbf{a}_j, \ldots] + \alpha \det[\ldots, \mathbf{a}_i, \ldots, \mathbf{a}_i, \ldots]\\ &= \det[\ldots, \mathbf{a}_i, \ldots, \mathbf{a}_j, \ldots].\end{aligned}$$

This proves

(vii) *Adding a multiple of one column of A to another column of A doesn't change the determinant.*

Here comes a very important use of (vii): Assume that $\mathbf{y} = A\mathbf{x}$ and consider

$$\det[\ldots, \mathbf{a}_{j-1}, \mathbf{y}, \mathbf{a}_{j+1}, \ldots].$$

Since $\mathbf{y} = x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n$, subtraction of x_i times column i from column j, i.e., subtraction of $x_i\mathbf{a}_i$ from $\mathbf{y}$ here, for each $i \neq j$ is, by (vii), guaranteed not to change the determinant, yet changes the jth column to $x_j\mathbf{a}_j$; then, pulling out that scalar factor x_j (permitted by (iv)), leaves us finally with $x_j \det A$. This proves

(viii) *If* $\mathbf{y} = A\mathbf{x}$, *then* $\det[\ldots, \mathbf{a}_{j-1}, \mathbf{y}, \mathbf{a}_{j+1}, \ldots] = x_j \det A$.

Hence, if $\det A \neq 0$, then $\mathbf{y} = A\mathbf{x}$ implies

$$x_j = \det[\ldots, \mathbf{a}_{j-1}, \mathbf{y}, \mathbf{a}_{j+1}, \ldots]/\det A, \qquad j = 1, \ldots, n.$$

This is **Cramer's rule**.

In particular, if $\det A \neq 0$, then $A\mathbf{x} = \mathbf{0}$ implies that $x_j = 0$ for all j, i.e., then A is 1-1, hence invertible (since A is square). This gives the converse to (iii), i.e.,

(ix) *If* $\det A \neq 0$*, then* A *is invertible.*

In old-fashioned mathematics, a matrix was called **singular** if its determinant is 0. So, (iii) and (ix) combined say that *a matrix is nonsingular iff it is invertible.*

The suggestion that one actually construct the solution to $A? = \mathbf{y}$ by Cramer's rule is ridiculous under ordinary circumstances since, even for a linear system with just two unknowns, it is more efficient to use Gauss elimination. On the other hand, if the solution is to be constructed *symbolically* (in a symbol-manipulating system such as `Maple` or `Mathematica`), then Cramer's rule is preferred to Gauss elimination since it treats all unknowns equally. In particular, the number of operations needed to obtain a particular unknown is the same for all unknowns.

We have proved all these facts (except (i)) about determinants from certain postulates (namely (ii), (iv), (v)) without ever saying how to *compute* $\det A$. Now, it is the actual formula for $\det A$ that has given determinants such a bad name. Here is the standard one, which (see page 229) can be derived from (ii), (iv), (v), in the process of proving (i):

(x)

$$\det A = \sum_{\sigma \in \mathbb{S}_n} (-1)^\sigma \prod_{j=1}^{n} A_{\sigma(j),j}.$$

Once we have proved this formula, we have also proved uniqueness of the map satisfying (ii), (iv), and (v).

In the formula, $\sigma \in \mathbb{S}_n$ is shorthand for: σ is a **permutation of the first** n **integers**, i.e.,

$$\sigma = (\sigma(1), \sigma(2), \ldots, \sigma(n)),$$

where $\sigma(j) \in \{1, 2, \ldots, n\}$ for all j, and $\sigma(i) \neq \sigma(j)$ if $i \neq j$. In other words, σ is a 1-1 and onto map from $\{1, \ldots, n\}$ to $\{1, \ldots, n\}$. Also,

$$(15.1) \qquad (-1)^\sigma := \operatorname{signum} \Delta(\sigma), \quad \text{with} \quad \Delta(\sigma) := \prod_{i<j} (\sigma(j) - \sigma(i)),$$

is the **sign of the permutation** σ. It equals 1 or -1 depending on whether the number of out-of-order pairs, i.e., $(\sigma(i), \sigma(j))$ with $i < j$ yet $\sigma(i) > \sigma(j)$, is even or odd, and the parity of this number is therefore called the **parity** of σ. This parity can also be determined as the parity of the number of interchanges needed, starting with $\sigma = (\sigma(1), \ldots, \sigma(n))$, to end up with the sequence $(1, 2, \ldots, n)$. To be sure, if you and I both try to bring the entries of σ into increasing order by interchanges, the number of steps taken may differ, but their parity never will; if it takes me an even number of steps, it will take you an even number of steps, due to the fact that any one interchange will change $\Delta(\sigma)$ to its negative (see Problem 15.1) while $\Delta((1, 2, \ldots, n))$ is positive.

15.1* Let $\sigma^\#$ denote the number of out-of-order pairs in the permutation σ (hence $\Delta(\sigma) = (-1)^{\sigma^\#}$), and let τ be the permutation obtained from σ by interchange of the ith and jth entry. (a) Prove: *If $\sigma(i)$ and $\sigma(j)$ are out of order, then $\sigma^\# - \tau^\#$ is positive and odd.* (b) Conclude that $\sigma^\# - \tau^\#$ is negative and odd in case $\sigma(i)$ and $\sigma(j)$ are in order. (c) Conclude that any permutation σ can be brought into order by at most $\sigma^\#$ interchanges, and give an example of a permutation for which fewer than $\sigma^\#$ interchanges suffice.

Here is a simple example: $\sigma = (3, 1, 4, 2)$ has the pairs $(3, 1)$, $(3, 2)$, and $(4, 2)$ out of order, hence $(-1)^\sigma = -1$. Equivalently, the following sequence of 3 interchanges gets me from σ to $(1, 2, 3, 4)$:

$$\begin{gathered}(3,1,4,2)\\(3,1,2,4)\\(1,3,2,4)\\(1,2,3,4)\end{gathered}$$

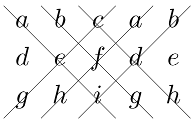
Therefore, again, $(-1)^\sigma = -1$.

Now, fortunately, we don't really ever have to use this stunning formula (x) in calculations, nor is it physically possible to use it for n much larger than 8 or 10. For $n = 1, 2, 3$, one can derive from it explicit rules for computing $\det A$:

$$\det [\,a\,] = a, \quad \det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc,$$

$$\det \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} = aei + bfg + cdh - (ceg + afh + bdi);$$

the last one can be remembered easily by the following mnemonic:



For $n > 3$, this mnemonic *does not work*, and one would not usually make use of (x), but use instead (i) and the following immediate consequence of (x):

(xi) *The determinant of a triangular matrix equals the product of its diagonal entries.*

Indeed, when A is upper triangular, then $A_{ij} = 0$ whenever $i > j$. Now, if $\sigma(j) > j$ for some j, then the factor $A_{\sigma(j),j}$ in the corresponding summand $(-1)^\sigma \prod_{j=1}^n A_{\sigma(j),j}$ is zero. This means that the only possibly nonzero summands correspond to σ with $\sigma(j) \le j$ for all j, and there is only one permutation that manages that, the **identity permutation** $(1, 2, \ldots, n)$, and its parity is even (since it takes no interchanges to bring it into increasing order). Therefore, the formula in (x) gives $\det A = A_{11} \cdots A_{nn}$ in this case. – The proof for a lower triangular matrix is analogous; else, use (xiii) below.

Consequently, if $A = LU$ with L unit triangular and U upper triangular, then

$$\det A = \det U = U_{11} \cdots U_{nn}.$$

If, more generally, $A = PLU$, with P some permutation matrix, then

$$\det A = \det(P) U_{11} \cdots U_{nn},$$

i.e.,

(xii) $\det A$ *is the product of the pivots used in elimination, times* $(-1)^i$, *with* i *the number of row interchanges made.*

Since, by elimination, any $A \in \mathbb{F}^{n\times n}$ can be factored as $A = PLU$, with P a permutation matrix, L unit lower triangular, and U upper triangular, (xii) provides the standard way to compute determinants.

Note that, then, $A^{\rm t} = U^{\rm t} L^{\rm t} P^{\rm t}$, with $U^{\rm t}$ lower triangular, $L^{\rm t}$ unit upper triangular, and $P^{\rm t}$ the inverse of P, hence

(xiii) $\det A^{\rm t} = \det A$.

This can also be proved directly from (x). Note that this converts all our statements about the determinant in terms of *columns* to the corresponding statements in terms of *rows*.

(xiv) “expansion by minors”:

Since, by (iv), the determinant is slotwise linear, and $\mathbf{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \cdots + x_n\mathbf{e}_n$, we obtain, for any $j \in \underline{n}$,

$$\det[\ldots, \mathbf{a}_{j-1}, \mathbf{x}, \mathbf{a}_{j+1}, \ldots] = x_1 C_{1j} + x_2 C_{2j} + \cdots + x_n C_{nj}, \tag{15.2}$$

with

$$C_{ij} := \det[\ldots, \mathbf{a}_{j-1}, \mathbf{e}_i, \mathbf{a}_{j+1}, \ldots]$$

the socalled **cofactor** of A_{ij}. With the choice $\mathbf{x} = \mathbf{a}_k$, this implies

$$\begin{aligned} A_{1k}C_{1j} + A_{2k}C_{2j} + \cdots + A_{nk}C_{nj} &= \det[\ldots, \mathbf{a}_{j-1}, \mathbf{a}_k, \mathbf{a}_{j+1}, \ldots] \\ &= \begin{cases} \det A & \text{if } k = j; \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The case $k = j$ gives the **expansion by minors** for $\det A$ (and justifies the name 'cofactor' for C_{ij}). The case $k \neq j$ is justified by (vi). In other words, with

$$\mathrm{adj}A := \begin{bmatrix} C_{11} & C_{21} & \cdots & C_{n1} \\ C_{12} & C_{22} & \cdots & C_{n2} \\ \vdots & \vdots & \cdots & \vdots \\ C_{1n} & C_{2n} & \cdots & C_{nn} \end{bmatrix}$$

the socalled **adjugate** of A (note that the subscripts appear reversed), we have

$$\mathrm{adj}(A)\, A = (\det A)\ \mathrm{id}. \tag{15.3}$$

This provides another proof of (ix), since it shows that, for a *nonsingular* A,

$$A^{-1} = (\mathrm{adj}A)/\det A.$$

The expansion by minors is useful since, as follows from (x), the cofactor C_{ij} equals $(-1)^{i+j}$ times the determinant of the matrix $A(\mathbf{n}\backslash i \mid \mathbf{n}\backslash j)$ obtained from A by removing row i and column j, i.e.,

$$C_{ij} = (-1)^{i+j} \det \begin{bmatrix} \cdots & \cdots & \cdots & \cdots \\ \cdots & a_{i-1,j-1} & a_{i-1,j+1} & \cdots \\ \cdots & a_{i+1,j-1} & a_{i+1,j+1} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix},$$

and this is a determinant of order $n-1$, and so, if $n-1 > 1$, can itself be expanded along some column (or row).

As a practical matter, for $[\mathbf{a}, \mathbf{b}, \mathbf{c}] := A \in \mathbb{R}^3$, the formula $\mathrm{adj}(A)\, A = (\det A)\ \mathrm{id}$ implies that

$$(\mathbf{a} \times \mathbf{b})^{\mathrm{t}}\mathbf{c} = \det[\mathbf{a}, \mathbf{b}, \mathbf{c}],$$

with

$$\mathbf{a} \times \mathbf{b} := (a_2b_3 - a_3b_2, a_3b_1 - a_1b_3, a_1b_2 - a_2b_1)$$

the **cross product** of $\mathbf{a}$ with $\mathbf{b}$. In particular, $\mathbf{a}\times\mathbf{b}$ is perpendicular to both $\mathbf{a}$ and $\mathbf{b}$. Also, if $[\mathbf{a}, \mathbf{b}]$ is o.n., then so is $[\mathbf{a}, \mathbf{b}, \mathbf{a}\times\mathbf{b}]$ but, in addition, $\det[\mathbf{a}, \mathbf{b}, \mathbf{a}\times\mathbf{b}] = 1$, i.e., $[\mathbf{a}, \mathbf{b}, \mathbf{a}\times\mathbf{b}]$ provides a right-handed cartesian coordinate system for $\mathbb{R}^3$.

(xv) $\det A$ *is the* n*-dimensional (signed) volume of the parallelepiped*

$$\{A\mathbf{x} : 0 \le x_i \le 1, \text{all } i\}$$

spanned by the columns of A.

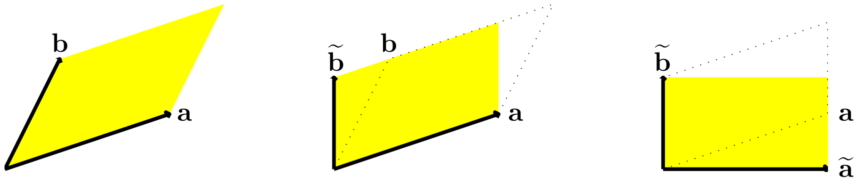
For $n > 3$, this is a *definition*, while, for $n \le 3$, one works it out (see below). This is a very useful *geometric* way of thinking about determinants.

Also, it has made determinants indispensable in the *definition* of multivariate integration and the handling therein of changes of variable.

Since $\det(AB) = \det(A)\det(B)$, it follows that *the linear map* $T : \mathbb{F}^n \to \mathbb{F}^n : \mathbf{x} \mapsto A\mathbf{x}$ *changes volumes by a factor of* $\det A$, meaning that, for any set M in the domain of T,

$$\mathrm{vol}_n(T(M)) = \det(A)\,\mathrm{vol}_n(M).$$

As an example, consider $\det[\mathbf{a}, \mathbf{b}]$, with $\mathbf{a}$, $\mathbf{b}$ vectors in the plane linearly independent, and assume, wlog, that $a_1 \neq 0$. By (iv), $\det[\mathbf{a}, \mathbf{b}] = \det[\mathbf{a}, \widetilde{\mathbf{b}}]$, with $\widetilde{\mathbf{b}} := \mathbf{b} - (b_1/a_1)\mathbf{a}$ having its first component equal to zero, and so, again by (iv), $\det[\mathbf{a}, \mathbf{b}] = \det[\widetilde{\mathbf{a}}, \widetilde{\mathbf{b}}]$, with $\widetilde{\mathbf{a}} := \mathbf{a} - (a_2/\widetilde{b}_2)\widetilde{\mathbf{b}}$ having its second component equal to zero. Therefore, $\det[\mathbf{a}, \mathbf{b}] = \widetilde{a}_1\widetilde{b}_2 = \pm\|\widetilde{\mathbf{a}}\|\,\|\widetilde{\mathbf{b}}\|$ equals ± the area of the rectangle spanned by $\widetilde{\mathbf{a}}$ and $\widetilde{\mathbf{b}}$. However, following the derivation of $\widetilde{\mathbf{a}}$ and $\widetilde{\mathbf{b}}$ graphically, we see, by matching congruent triangles, that the rectangle spanned by $\widetilde{\mathbf{a}}$ and $\widetilde{\mathbf{b}}$ has the same area as the parallelepiped spanned by $\mathbf{a}$ and $\widetilde{\mathbf{b}}$, and, therefore, as the parallelepiped spanned by $\mathbf{a}$ and $\mathbf{b}$. Thus, up to sign, $\det[\mathbf{a}, \mathbf{b}]$ is the area of the parallelepiped spanned by $\mathbf{a}$ and $\mathbf{b}$.



Here, finally, for the record, is a *proof* that (ii) + (iv) + (v) implies (i), hence everything else we have been deriving so far. Let A and B be arbitrary matrices (of order n). Then the multilinearity (iv) implies that

$$\begin{aligned}\det(BA) &= \det[B\mathbf{a}_1, \ldots, B\mathbf{a}_n] \\ &= \det[\ldots, \sum_i \mathbf{b}_i A_{ij}, \ldots] \\ &= \sum_{\sigma\in\{1,\ldots,n\}^n} \det[\mathbf{b}_{\sigma(1)}, \ldots, \mathbf{b}_{\sigma(n)}] \prod_j A_{\sigma(j),j}.\end{aligned}$$

By the consequence (vi) of the alternation property (v), most of these summands are zero. Only those determinants $\det[\mathbf{b}_{\sigma(1)}, \ldots, \mathbf{b}_{\sigma(n)}]$ for which all the entries of σ are different are *not* automatically zero. But that are exactly all the $\sigma \in \mathbb{S}_\mathbf{n}$, i.e., the permutations of the first n integers. Further, for such σ,

$$\det[\mathbf{b}_{\sigma(1)}, \ldots, \mathbf{b}_{\sigma(n)}] = (-1)^\sigma \det B$$

by the alternation property (v), with $(-1)^\sigma = 1$ or -1 depending on whether it takes an even or an odd number of interchanges to change σ into a strictly increasing sequence. Thus

$$\det(BA) = \det(B) \sum_{\sigma\in \mathrm{S}_\mathbf{n}} (-1)^\sigma \prod_j A_{\sigma(j),j}.$$

Choosing, in particular, $B = \mathrm{id}$, we obtain formula (x) since $\mathrm{id}A = A$ while, by the defining property (ii), $\det \mathrm{id} = 1$, and, with that, $\det(BA) = \det(B)\det(A)$ for arbitrary B and A.

On the other hand, starting with the formula in (x) as a definition, one may verify (see Problem 15.2) that det so defined satisfies the three properties (ii) ($\det(\mathrm{id}) = 1$), (iv) (multilinear), and (v) (alternating) claimed for it. In other words, there actually is such a function (necessarily given by (x)).

Sylvester

Here, for the record, is the statement and a proof of Sylvester's Determinant Identity. For it, the following notation will be useful: If $\mathbf{i} = (i_1, \ldots, i_r)$ and $\mathbf{j} = (j_1, \ldots, j_s)$ are suitable integer sequences, then $A_{\mathbf{ij}} = A(\mathbf{i} \mid \mathbf{j})$ is the $r \times s$-matrix whose (p, q) entry is A_{i_p,j_q}, $p = 1, \ldots, r$, $q = 1, \ldots, s$. This is just as in `MATLAB` except for the vertical bar used here at times, for emphasis and in order to list, on either side of it, a sequence without having to encase it in parentheses. Also, $A(\mathbf{i}) := A(\mathbf{i} \mid \mathbf{i})$.

With

$$\underline{k} := 1{:}k$$

now the *sequence* $(1, 2, \ldots, k)$, consider the matrix C with entries

$$C_{ij} := \det A(\underline{k}, i \mid \underline{k}, j).$$

Note that $C_{ij} = 0$ whenever $i \in \underline{k}$ or $j \in \underline{k}$ since then $A(\underline{k}, i \mid \underline{k}, j)$ has two rows the same or two columns the same. If $A(\underline{k})$ is invertible, then the nontrivial part of C, i.e., the submatrix $C(\backslash\underline{k})$, gives rise to the matrix

$$A/A(\underline{k}) \;:=\; C(\backslash\underline{k})/\det A(\underline{k})$$

which is the *Schur complement* in A of the **pivot block** $A(\underline{k})$. This terminology, already mentioned in Problem 4.33, derives from the fact (to be proved in a moment) that $C(\backslash\underline{k})/\det A(\underline{k})$ can be viewed as having been obtained by block elimination, as the (2,2)-block in the 2-by-2 block matrix

$$\begin{bmatrix} A(\underline{k}) & A(\underline{k}, \backslash\underline{k}) \\ 0 & A(\backslash\underline{k}) - A(\backslash\underline{k}, \underline{k})A(\underline{k})^{-1}A(\underline{k}, \backslash\underline{k}) \end{bmatrix}$$

that results when we use the first k rows of A to zero out the first k entries in every row $i > k$.

To prove this claim, expand $C_{ij} = \det A(\underline{k}, i \mid \underline{k}, j)$ by entries of its last column (using property (xiv)) to get

$$C_{ij} = A_{ij} \det A(\underline{k}) - \sum_{r \le k} A_{rj}\ (-1)^{k-r} \det A((\underline{k}\backslash r), i \mid \underline{k}).$$

This shows that, for $i > k$,

$$C_{i:} \in A_{i:} \det A(\underline{k}) - \operatorname{ran} A(\underline{k} \mid :),$$

while $C_{ij} = 0$ for $j \in \underline{k}$ as already observed earlier. Hence, if $\det A(\underline{k}) \neq 0$, then $C_{i:}/\det A(\underline{k})$ is the result of subtracting a weighted sum of the first k rows of A from the ith row of A in such a way that the first k entries of the resulting row are zero.

In other words, for $i > k$, $C_{i:}/\det A(\underline{k})$ is the ith row of the work-array B after k steps of elimination without pivoting, or even with row pivoting as long as only rows with index $\le k$ are interchanged.

This provides the following useful

(15.4) Determinantal Expressions For LDU Factors: Assume that elimination without pivoting can be carried out on the matrix A, resulting in the factorization $A = LDU$, with L unit lower triangular, D diagonal and invertible, and U unit upper triangular. Then, with B the work-array after k steps of elimination without pivoting applied to A,

$$D_{k+1,k+1} = B_{k+1,k+1} = \det A(\underline{k+1})/\det A(\underline{k})$$

is the pivot for the $k+1$st elimination step, hence the $k+1$st diagonal entry of the diagonal factor D, therefore, for $i > k$,

$$L_{i,k+1} = B_{i,k+1}/B_{k+1,k+1} = \det A(\underline{k}, i \mid \underline{k+1})/\det A(\underline{k+1})$$

is the $(i, k+1)$ entry of the resulting unit lower triangular left factor of A and, correspondingly,

$$U_{k+1,i} = B_{k+1,i}/B_{k+1,k+1} = \det A(\underline{k+1} \mid \underline{k}, i)/\det A(\underline{k+1})$$

is the $(k+1, i)$ entry of the resulting unit upper triangular right factor of A.

Since such row elimination is done by elementary matrices with determinant equal to 1, we get

$$\det A = \det \begin{bmatrix} A(\underline{k}) & A(\underline{k} \mid \backslash\underline{k}) \\ 0 & A/A(\underline{k}) \end{bmatrix} = \det A(\underline{k}) \det(A/A(\backslash\underline{k})),$$

hence

$$\det(A/A(\underline{k})) = \det A / \det A(\underline{k})$$

which is **Schur's determinant identity** and is the reason for the somewhat unusual notation $A/A(\underline{k})$ for the Schur complement.

Now note that we are free to replace A in Schur's determinant identity by $A(\underline{k}, \mathbf{i} \mid \underline{k}, \mathbf{j})$ for any $\#\mathbf{i} = \#\mathbf{j}$, hence obtain

(15.5) **Sylvester's determinant identity**. If

$$\forall i, j, \quad B_{ij} := \det A(\underline{k}, i \mid \underline{k}, j) / \det A(\underline{k}),$$

then

$$\det B(\mathbf{i} \mid \mathbf{j}) = \det A(\underline{k}, \mathbf{i} \mid \underline{k}, \mathbf{j}) / \det A(\underline{k}).$$

Binet-Cauchy

(15.6) **Binet-Cauchy Formula**. If BA is defined, then, for $\#\mathbf{i} = \#\mathbf{j}$,

$$\det(BA)(\mathbf{i} \mid \mathbf{j}) = \sum_{\#\mathbf{h} = \#\mathbf{i}} \det B(\mathbf{i} \mid \mathbf{h}) \ \det A(\mathbf{h} \mid \mathbf{j}).$$

Even the special case $\#\mathbf{i} = \#A$ of this, i.e., the most important determinant property (i),

$$\det(BA) = \det B \det A,$$

Binet and Cauchy were the first to prove. Not surprisingly, the proof of the formula follows our earlier proof of that identity.

Proof: Since $(BA)(\mathbf{i} \mid \mathbf{j}) = B(\mathbf{i} \mid :) A(: \mid \mathbf{j})$, it is sufficient to consider the case $B, A^{\mathrm{t}} \in \mathbb{F}^{m \times n}$ for some m and n. If $m > n$, then B cannot be onto, hence BA must fail to be invertible, while the sum is empty, hence has value 0. It is therefore sufficient to consider the case $m \le n$.

For this, using the linearity of the determinant in each slot,

$$\begin{aligned}\det(BA) &= \det[BA_{:,1},\ldots,BA_{:,m}]\\ &= \sum_{h_1}\cdots\sum_{h_m}\det[B_{:,h_1}A_{h_1,1},\ldots,B_{:,h_m}A_{h_m,m}]\\ &= \sum_{h_1}\cdots\sum_{h_m}\det[B_{:,h_1},\ldots,B_{:,h_m}]\,A_{h_1,1}\cdots A_{h_m,m}\\ &= \sum_{h_1<\cdots<h_m}\det B(:\mid \mathbf{h})\sum_{\sigma\in\mathbb{S}_m}(-1)^\sigma A_{h_{\sigma(1)},1}\cdots A_{h_{\sigma(m)},m}\\ &= \sum_{h_1<\cdots<h_m}\det B(:\mid \mathbf{h})\det A(\mathbf{h}\mid :).\end{aligned}$$

□

15.2* Prove that the function det : $\mathbb{F}^{n\times n}\to\mathbb{F}$ given by the formula in (x) necessarily satisfies (ii), (iv), and (v).

15.3 Prove: *For any $A\in\mathbb{F}^{n\times n+1}$, the vector $((-1)^k\det A(\,:\,,\backslash k) : k = 1{:}n+1)$ is in* null A.

15.4 Let $A\in\mathbb{Z}^{n\times n}$, i.e., a matrix of order n with integer entries, and assume that A is invertible. Prove: $A^{-1}\in\mathbb{Z}^{n\times n}$ *if and only if* $|\det A| = 1$. (Hint: Use Cramer's Rule to prove that such A maps $\mathbb{Z}^n$ onto itself in case $\det A = \pm 1$.)

15.5 Prove: $|\det A| = \sqrt{\det(A^{c}A))}$.

15.6 Prove **Hadamard's inequality**: $|\det[\mathbf{a}_1,\ldots,\mathbf{a}_n]| \le \|\mathbf{a}_1\|\cdots\|\mathbf{a}_n\|$.

15.7 Let R be a ring (see page 278). Prove the following claim, of use in ideal theory: *If $A\mathbf{x} = 0$ for $A\in R^{n\times n}$ and $\mathbf{x}\in R^n$, then $x_i\det A = 0$ for all i.*

15.8 Use the previous homework to prove the following (see page 278 for background on rings): *If R is a commutative ring with identity, $s_1,\ldots,s_n\in R$,*

$$F := [s_1,\ldots,s_n](R^n) = \{\sum_j s_j r_j : (r_j)\in R^n\}$$

and H is an ideal in R for which $F\subset HF := \{hf : h\in H, f\in F\}$, then, for some $h\in H$, $(1-h)F = 0$.

15.9 Prove that the elementary matrix $A := \mathrm{id}_n - qr^{\mathrm t}$ has a factorization $A = LDU$ with L unit lower triangular, D diagonal, and U unit upper triangular provided the numbers

$$p_i := 1 - \sum_{j\le i} q_j r_j$$

are nonzero for $i < n$, and verify that then $D = \operatorname{diag}(p_i/p_{i-1} : i = 1{:}n)$ and

$$L(i,j) = -q_i r_j/p_j = U(j,i),\quad i > j.$$

15.10 T/F

(a) If A and B are matrices for which both AB and BA are defined, then $\det(AB) = \det(BA)$.

(b) If $A\in\mathbb{Z}^{n\times n}$ then $\det A\in\mathbb{Z}$.

(c) $\det A^{c} = \det A^{-1}$ if A is invertible.

16 Some applications

The cross product in 3-space

In the vector space $X = \mathbb{R}^3$, the standard inner product is also called the **dot product**, because of the customary notation

$$\mathbf{y}^{\mathrm{t}}\mathbf{x} = \langle \mathbf{x}, \mathbf{y} \rangle =: \mathbf{x} \cdot \mathbf{y}, \qquad \mathbf{x}, \mathbf{y} \in \mathbb{R}^3.$$

In this most familiar vector space, another vector 'product' is of great use, the so-called **cross product** $\mathbf{x} \times \mathbf{y}$. It is most efficiently defined implicitly, i.e., by

$$(16.1) \qquad \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^3, \quad (\mathbf{x} \times \mathbf{y})^{\mathrm{t}}\mathbf{z} := \det[\mathbf{x}, \mathbf{y}, \mathbf{z}].$$

From (13.14) (see also page 226), we work out that

$$\det[\mathbf{x}, \mathbf{y}, \mathbf{z}] = (x_2y_3 - x_3y_2)z_1 + (x_3y_1 - x_1y_3)z_2 + (x_1y_2 - x_2y_1)z_3,$$

hence

$$\mathbf{x} \times \mathbf{y} = (x_2y_3 - x_3y_2, x_3y_1 - x_1y_3, x_1y_2 - x_2y_1).$$

Given what you already know about determinants, the definition (16.1), though implicit, makes all the basic facts about the cross product immediate:

(i) *The cross product* $\mathbf{x} \times \mathbf{y}$ *is linear in its two arguments,* $\mathbf{x}$ *and* $\mathbf{y}$.

(ii) *The cross product* $\mathbf{x} \times \mathbf{y}$ *is* **alternating**, *meaning that* $\mathbf{y} \times \mathbf{x} = -(\mathbf{x} \times \mathbf{y})$.

(iii) Perhaps most importantly, $\mathbf{x} \times \mathbf{y}$ *is a vector perpendicular to both* $\mathbf{x}$ *and* $\mathbf{y}$.

(iv) $\mathbf{x} \times \mathbf{y} = 0$ *if and only if* $[\mathbf{x}, \mathbf{y}]$ *is not 1-1.*

Indeed, if $[\mathbf{x}, \mathbf{y}]$ is 1-1, then we can always extend it to a basis $[\mathbf{x}, \mathbf{y}, \mathbf{z}]$ for $\mathbb{R}^3$, and then $(\mathbf{x} \times \mathbf{y})^{\mathrm{t}}\mathbf{z}$ is not zero, hence then $\mathbf{x} \times \mathbf{y} \neq \mathbf{0}$. If $[\mathbf{x}, \mathbf{y}]$ fails to be

1-1, then $[\mathbf{x},\mathbf{y},\mathbf{x}{\times}\mathbf{y}]$ fails to be 1-1, hence then $\|\mathbf{x}{\times}\mathbf{y}\|^2 = \det[\mathbf{x},\mathbf{y},\mathbf{x}{\times}\mathbf{y}] = 0$, therefore $\mathbf{x} \times \mathbf{y} = 0$. In particular, $\mathbf{x} \times \mathbf{x} = \mathbf{0}$.

So, assuming that $[\mathbf{x},\mathbf{y}]$ is 1-1, we can compute the *direction* vector

$$\mathbf{u} := (\mathbf{x} \times \mathbf{y})/\|\mathbf{x} \times \mathbf{y}\|,$$

and so conclude that

$$\|\mathbf{x} \times \mathbf{y}\|_2^2 = \det[\mathbf{x},\mathbf{y},\mathbf{x} \times \mathbf{y}] = \|\mathbf{x} \times \mathbf{y}\| \det[\mathbf{x},\mathbf{y},\mathbf{u}].$$

In other words,

(v) *the Euclidean length of* $\mathbf{x} \times \mathbf{y}$ *gives the (unsigned) area of the parallelepiped spanned by* $\mathbf{x}$ *and* $\mathbf{y}$.

This also holds when $[\mathbf{x},\mathbf{y}]$ fails to be 1-1 since then that area is zero.

When $[\mathbf{x},\mathbf{y}]$ is 1-1, then there are exactly two directions perpendicular to the plane $\operatorname{ran}[\mathbf{x},\mathbf{y}]$ spanned by $\mathbf{x}$ and $\mathbf{y}$, namely $\mathbf{u} := (\mathbf{x} \times \mathbf{y})/\|\mathbf{x} \times \mathbf{y}\|$ and $(\mathbf{y} \times \mathbf{x})/\|\mathbf{y} \times \mathbf{x}\| = -\mathbf{u}$, with $\mathbf{u}$ the choice that makes $\det[\mathbf{x},\mathbf{y},\mathbf{u}]$ positive. If you imagine the thumb of your *right* hand to be $\mathbf{x}$, and the pointer of that hand to be $\mathbf{y}$, then the middle finger, bent to be perpendicular to both thumb and pointer, would be pointing in the direction of $\mathbf{x} \times \mathbf{y}$. For that reason, any basis $[\mathbf{x},\mathbf{y},\mathbf{z}]$ for $\mathbb{R}^3$ with $\det[\mathbf{x},\mathbf{y},\mathbf{z}] > 0$ is said to be **right-handed**.

16.1 Relate the standard choice $(x_2, -x_1)$ for a vector perpendicular to the 2-vector $\mathbf{x}$ to the above construction.

16.2 Give a formula for an n-vector $\mathbf{x}_1 \times \cdots \times \mathbf{x}_{n-1}$ that is perpendicular to the $n-1$ n-vectors $\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}$ and whose Euclidean length equals the (unsigned) volume of the parallelepiped spanned by the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}$.

Rotation in 3-space

A particularly useful transformation of 3-space is counter-clockwise rotation by some angle θ around some given axis-vector $\mathbf{a}$. Let $R = R_{\theta,\mathbf{a}}$ be this rotation. We are looking for a computationally efficient way to represent this map.

This rotation leaves its **axis**, i.e., $\operatorname{ran}[\mathbf{a}]$, pointwise fixed, and rotates any vector in the plane $H := \mathbf{a}^\perp$ by an angle of θ radians counterclockwise with respect to the direction of $\mathbf{a}$; see (16.2)Figure. In other words, with

$$\mathbf{p} = \mathbf{q} + \mathbf{r}, \quad \text{where} \quad \mathbf{q} := P_{\operatorname{ran}[\mathbf{a}]}\mathbf{p}, \quad \text{hence} \quad \mathbf{r} = \mathbf{p} - \mathbf{q},$$

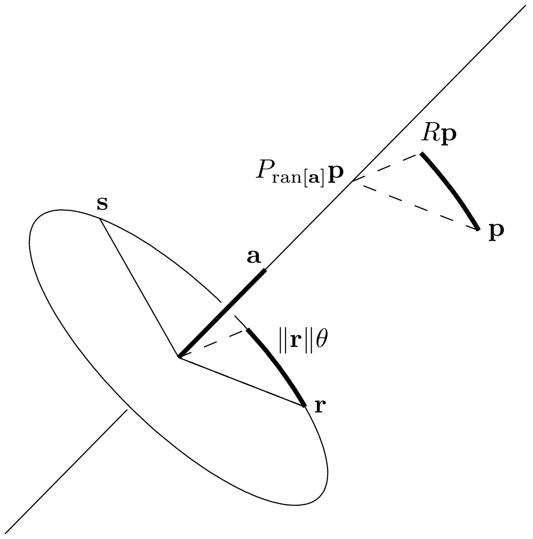
we have

$$R\mathbf{p} = \mathbf{q} + R\mathbf{r},$$

by the linearity of the rotation. To compute $R\mathbf{r}$, let $\mathbf{s}$ be the vector in H obtained by rotating $\mathbf{r}$ counterclockwise $\pi/2$ radians. Then

$$R\mathbf{r} = \cos(\theta)\mathbf{r} + \sin(\theta)\mathbf{s},$$

and that's it.



(16.2) Figure. Rotation of the point $\mathbf{p}$ counterclockwise θ radians around the axis spanned by the vector $\mathbf{a}$. The orthogonal projection $\mathbf{r}$ of $\mathbf{p}$ into the plane H with normal $\mathbf{a}$, together with its rotation $\mathbf{s}$ counterclockwise $\pi/2$ radians around that axis, serve as a convenient orthogonal coordinate system in H.

It remains to construct $\mathbf{s}$, and this is traditionally done with the aid of the cross product $\mathbf{a} \times \mathbf{r}$ since (see (16.1)) it is a vector perpendicular to $\mathbf{a}$ and $\mathbf{r}$. Hence, assuming without loss that $\mathbf{a}$ is normalized, we now know that $\mathbf{a} \times \mathbf{r}$ is in the plane H and perpendicular to $\mathbf{r}$ and of the same length as $\mathbf{r}$. Of the two vectors in H that have this property, it also happens to be the one obtained from $\mathbf{r}$ by a $(\pi/2)$-rotation that appears counterclockwise when looking down on H from the side that the vector $\mathbf{a}$ points into. (Just try it out.)

The calculations can be further simplified. The map

$$\mathbf{r} \mapsto \mathbf{a} \times \mathbf{r}$$

is linear and, by inspection, $\mathbf{a} \times \mathbf{a} = 0$. Since $\mathbf{a}$ is normalized by assumption, we compute

$$\mathbf{r} = \mathbf{p} - (\mathbf{a}^{\mathrm{t}}\mathbf{p})\mathbf{a},$$

hence

$$\mathbf{a} \times \mathbf{r} = \mathbf{a} \times \mathbf{p}.$$

So, altogether

$$\begin{aligned} R\mathbf{p} &= (\mathbf{a}^{\mathrm{t}}\mathbf{p})\mathbf{a} + \cos(\theta)(\mathbf{p} - (\mathbf{a}^{\mathrm{t}}\mathbf{p})\mathbf{a}) + \sin(\theta)(\mathbf{a} \times \mathbf{p}) \\ &= \cos(\theta)\mathbf{p} + (1 - \cos(\theta))(\mathbf{a}^{\mathrm{t}}\mathbf{p})\mathbf{a} + \sin(\theta)(\mathbf{a} \times \mathbf{p}). \end{aligned}$$

This is the formula that is most efficient for the calculation of $R\mathbf{p}$. However, if the matrix for $R = R\,\mathrm{id}_3$ (with respect to the natural basis) is wanted, we read it off as

$$R = \cos(\theta)\,\mathrm{id}_3 + (1 - \cos(\theta))[\mathbf{a}][\mathbf{a}]^{\mathrm{t}} + \sin(\theta)(\mathbf{a}\times),$$

with

$$\mathbf{a}\times := \begin{bmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{bmatrix}$$

the matrix for the linear map $\mathbf{r} \mapsto \mathbf{a} \times \mathbf{r}$.

Flats: points, vectors, barycentric coordinates, differentiation

In CAGD and Computer Graphics, Linear Algebra is mainly used to change one's point of view, that is, to change coordinate systems. In this, even the familiar 3-space, $\mathbb{R}^3$, is often treated as an 'affine space' or 'flat' rather than a vector space, in order to deal simply with useful maps other than linear maps, namely the affine maps.

For example, the **translation**

$$\tau_{\mathbf{v}} : \mathbb{R}^3 \to \mathbb{R}^3 : \mathbf{p} \mapsto \mathbf{p} + \mathbf{v}$$

of $\mathbb{R}^3$ by the vector $\mathbf{v}$ is not a linear map (since it fails to map 0 to 0). Nevertheless, it can be represented by a matrix, using the following trick. Embed $\mathbb{R}^3$ into $\mathbb{R}^4$ by the 1-1 map

$$\mathbb{R}^3 \to \mathbb{R}^4 : \mathbf{x} \mapsto (\mathbf{x}, 1).$$

The image of $\mathbb{R}^3$ under this map is the 'flat'

$$F := \mathbb{R}^3 \times \{1\} = \{(\mathbf{x}, 1) : \mathbf{x} \in \mathbb{R}^3\} \ \subset \ \mathbb{R}^4.$$

Consider the linear map on $\mathbb{R}^4$ given by

$$T_{\mathbf{v}} := \begin{bmatrix} \mathrm{id}_3 & \mathbf{v} \\ 0 & 1 \end{bmatrix}.$$

Then, for any $\mathbf{x} \in \mathbb{R}^3$,

$$T_{\mathbf{v}}(\mathbf{x}, 1) = (\ \mathrm{id}_3\mathbf{x}\ + \mathbf{v}, 0^{\mathrm{t}}\mathbf{x} + 1) = (\mathbf{x} + \mathbf{v}, 1).$$

In other words, the linear map $T_{\mathbf{v}}$ carries F into itself in such a way that the point $\mathbf{p} = (\mathbf{x}, 1)$ is carried to its 'translate' $(\mathbf{x} + \mathbf{v}, 1) = \mathbf{p} + (\mathbf{v}, 0)$.

Let, now, $A \in \mathbb{R}^{4\times 4}$ be an arbitrary linear map on $\mathbb{R}^4$ subject only to the condition that it map F into itself. Breaking up A in the same way as we did $T_{\mathbf{v}}$, i.e.,

$$A =: \begin{bmatrix} A_0 & \mathbf{w} \\ [\mathbf{u}]^{\mathrm{t}} & t \end{bmatrix},$$

we get

$$A\,(\mathbf{x}, 1) = (A_0\mathbf{x} + \mathbf{w}, \mathbf{u}^{\mathrm{t}}\mathbf{x} + t),$$

hence want $\mathbf{u}^{\mathrm{t}}\mathbf{x} + t = 1$ for all $\mathbf{x} \in \mathbb{R}^3$, and this holds if and only if $\mathbf{u} = 0$ and $t = 1$, i.e.,

$$A = \begin{bmatrix} A_0 & \mathbf{w} \\ 0 & 1 \end{bmatrix}$$

is the most general such map. Its action on $\mathbb{R}^3$ is an arbitrary linear transformation, A_0, followed by translation by an arbitrary $\mathbf{w}$.

Such a description of an affine map on $\mathbb{R}^3$ is used in `MATLAB` graphics, as follows. Three-dimensional plots in `MATLAB` plot, in fact, the orthogonal projection onto the (x,y)-plane *after* an affine transformation of $\mathbb{R}^3$ that makes the center of the plotting volume the origin and a rotation that moves a line through that center, specified by azimuth and elevation, to the z-axis. This affine map is recorded in a matrix of order 4, obtainable by the command `view`, and also changeable by that command, but, fortunately, in down-to-earth terms like *azimuth* and *elevation*, or *viewing angle*.

As an example, let us so plot the unit cube. Its edges are traversed by the piecewise linear path through the points specified by the columns of the following matrix:

```
cube = [0 1 1 0 0 0 1 1 0 0 1 1 1 1 0 0;
        0 0 1 1 0 0 0 1 1 0 0 0 1 1 1 1;
        0 0 0 0 0 1 1 1 1 1 1 0 0 1 1 0];
```

Its center is the point (1,1,1)/2, and translating that point to the origin is handled by the affine map described by

$$\tau = \begin{bmatrix} 1 & 0 & 0 & -1/2 \\ 0 & 1 & 0 & -1/2 \\ 0 & 0 & 1 & -1/2 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

The default azimuth angle is $\alpha := -37.5\pi/180$ radians, and the corresponding affine map (i.e., the rotation in the (x, y)-plane that carries the default viewpoint to a point with x-coordinate 0 and a nonpositive y-coordinate) is given by

$$A = \begin{bmatrix} -\sin(a) & \cos(a) & 0 & 0 \\ -\cos(a) & -\sin(a) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

with $a := 3\pi/2 + \alpha$. As a check, note that this map is the identity in case $\alpha = 0$, i.e., $a = 3\pi/2$, and, in general, carries the vector $(\cos(a), \sin(a), z, 1)$ to the vector $(0, -1, z, 1)$. Finally, the default elevation angle is $\varepsilon := 30\pi/180$ radians, and the corresponding map (i.e., the rotation in the (y, z)-plane that further carries the view point to a point with y-coordinate 0 and nonnegative z-coordinate) is given by

$$E = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \sin(e) & -\cos(e) & 0 \\ 0 & \cos(e) & \sin(e) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

with $e := \pi - \varepsilon$. As a check, note that this map is the identity in case $\varepsilon = \pi/2$, i.e, $e = \pi/2$, and, in general, carries the vector $(x, \cos(e), \sin(e), 1)$ to the vector $(x, 0, 1, 1)$.

To check these formulæ, we compare the resulting product for this case

$$EAT =$$

with the transformation matrix

obtained in MATLAB via

```
plot3(cube(1,:), cube(2,:), cube(3,:))
view
```

which produces the following plot

The "flat" F is not a vector subspace of $\mathbb{R}^4$, i.e., it is not closed under vector addition or scalar multiplication. However, it is closed under formation of so-called **affine combinations**, i.e., weighted sums of the form

$$\sum_{j=0}^{r} \mathbf{p}_j \alpha_j$$

with $\mathbf{p}_0, \ldots, \mathbf{p}_r \in F$, and $\alpha_0, \ldots, \alpha_r \in \mathbb{R}$ such that

$$\sum_{j=0}^{r} \alpha_j = 1.$$

Thus, as far as the set F is concerned, these are the only weighted sums allowed. Note that such an affine sum can always be rewritten as

$$\mathbf{p}_0 + \sum_{j=1}^{r} (\mathbf{p}_j - \mathbf{p}_0)\beta_j,$$

where now the weights β_j, $j = 1{:}r$, are *arbitrary.* In other words, an affine sum in F is obtained by adding to some point in F an arbitrary weighted sum of elements in the vector space $F{-}F$.

An **affine map** on F is any map A from F to F that preserves affine combinations, i.e., for which

$$A(\mathbf{p}_0 + \sum_j (\mathbf{p}_j - \mathbf{p}_0)\alpha_j) = A\mathbf{p}_0 + \sum_j (A\mathbf{p}_j - A\mathbf{p}_0)\alpha_j$$

for all $\mathbf{p}_j \in F$, $\alpha_j \in \mathbb{R}$. It follows that the map on $F{-}F$ defined by

$$A_0 : F{-}F \to F{-}F : \mathbf{p} - \mathbf{q} \mapsto A\mathbf{p} - A\mathbf{q}$$

must be well-defined and linear, hence A is necessarily the restriction to F of some linear map $\widetilde{A}$ on $\mathbb{R}^4$ that carries F into itself and therefore also carries the linear subspace $F{-}F$ into itself.

The main pay-off, in CAGD and in Computer Graphics, of these considerations is the fact that one can represent the composition of affine maps by the product of their corresponding matrices.

This concrete example has led to the following abstract definition of a flat, whose notational conventions strongly reflect the concrete example. You should verify that the standard example is, indeed, a flat in the sense of this abstract definition.

(16.3) Definition: A **flat** or **affine space** or **linear manifold** is a nonempty set F of **point**s, a vector space T of **translation**s, and a map

(16.4) $$\varphi : F \times \mathrm{T} \to F : (p, \tau) \mapsto \tau(p) =: p + \tau$$

satisfying the following:

(a) $\forall (p, \tau) \in F \times \mathrm{T}, \quad p + \tau = p \quad \iff \quad \tau = 0.$

(b) $\forall \tau, \sigma \in \mathrm{T}, \quad (\cdot + \tau) + \sigma = \cdot + (\tau + \sigma).$

(c) $\exists p_0 \in F, \quad \varphi(p_0, \cdot)$ is onto.

Translations are also called **vector**s since (like ‘vehicles’ or ‘conveyors’, words that have the same Latin root as ‘vector’) they carry points to points.

Condition (a) ensures the uniqueness of the solution of the equation $p{+}? = q$ whose existence (see the proof of (3) below) is guaranteed by (c).

Condition (b) by itself is already satisfied, for arbitrary F and T, by, e.g., $\varphi : (p, \tau) \mapsto p$.

Condition (c) is needed to be certain that T is rich enough. (a)&(b) is already satisfied, e.g., by $\mathrm{T} = \{0\}$, $\varphi(\cdot, 0) = \mathrm{id}$. As we will see in a moment, (a)&(b)&(c) implies that $\varphi(p, \cdot)$ is onto for every $p \in F$. In other words, there is nothing special about the p_0 that appears in (c). In fact, the notion of a flat was developed explicitly as a set that, in contrast to a vector space which has an origin, does not have a distinguished point.

Consequences

(1) $\varphi(\cdot, 0) = \mathrm{id}$ (by (a)).

(2) For any $\tau \in \mathrm{T}$, $\varphi(\cdot, \tau)$ is invertible; its inverse is $\varphi(\cdot, -\tau)$ (by (1) and (b)). The corresponding abbreviation

$$p - \tau := p + (-\tau)$$

is helpful and standard.

(3) $\forall p, q \in F, \quad \exists \tau \in \mathrm{T}, \quad p + \tau = q$. This *unique* τ is correspondingly denoted

$$q - p.$$

Proof: If $p + \tau = q = p + \sigma$, then, by (2) and (b), $p = q + (-\sigma) = (p + \tau) + (-\sigma) = p + (\tau - \sigma)$, therefore, by (1), $\tau - \sigma = 0$, showing the *uniqueness* of the solution to $p+? = q$, regardless of p and q. The *existence* of a solution is, offhand, only guaranteed, by (c), for $p = p_0$. However, with the invertibility of $\varphi(p_0, \cdot) : \mathrm{T} \to F$ thus established, hence with $p - p_0$ and $q - p_0$ well-defined, we have $q = p_0 + (q - p_0)$ and $p = p_0 + (p - p_0)$, hence $p_0 = p - (p - p_0)$, therefore

$$q = p - (p - p_0) + (q - p_0),$$

showing that the equation $p+? = q$ has a solution (namely the vector $(q - p_0) - (p - p_0)$). □

16.3 Prove that the map $\Phi : \mathrm{T} \to F^F$, given by the rule $\Phi(\tau) : p \mapsto \varphi(p, \tau)$, is a semi-homomorphism, from the additive group of the vector space T into the semi-group F^F with composition the semi-group action.

(4) Note that (3) provides a 1-1 correspondence (in many different ways) between F and T. Specifically, for any particular $o \in F$,

$$F \to \mathrm{T} : p \mapsto p - o$$

is an invertible map, as is its inverse,

$$\mathrm{T} \to F : \tau \mapsto o + \tau.$$

However, the wish to avoid such an arbitrary choice of an 'origin' o in F provided the impetus to define the concept of flat in the first place. The

dimension of a flat is, by definition, the dimension of the associated vector space of translations. Also, since the primary focus is usually the flat, F, it is very convenient to write its vector space of translations as

$$F{-}F.$$

(5) The discussion so far has only made use of the additive structure of T. Multiplication by scalars provides additional structure. Thus, for arbitrary $Q \subset F$, the **affine hull** of Q, or the **flat spanned by** Q is, by definition,

$$\flat(Q) := q + \operatorname{span}(Q - q),$$

with the right side certainly independent of the choice of $q \in Q$, by (4). The affine hull of Q is, itself, a flat, with $\operatorname{span}(Q - q)$ the vector space of its translations.

(6) In particular, the affine hull of a finite subset Q of F is

$$\flat(Q) = q_0 + \operatorname{ran}[q - q_0 : q \in Q\backslash q_0], \quad q_0 \in Q.$$

Let

$$q_0 + \sum_{q \neq q_0} (q - q_0)\alpha_q$$

be one of its elements. In order to avoid singling out $q_0 \in Q$, it is customary to write instead

$$\sum_q q\alpha_q, \quad \text{with} \quad \alpha_{q_0} := 1 - \sum_{q \neq q_0} \alpha_q.$$

This makes $\flat(Q)$ the set of all **affine combination**s

$$\sum_{q \in Q} q\alpha_q, \quad \sum_q \alpha_q = 1,$$

of the elements of Q. The affine hull $\flat(q_0, \ldots, q_r)$ of a sequence $q_0, \ldots, q_r$ in F is defined analogously. But I prefer to work here with the set Q in order to stress the point of view that, in a flat, all points are of equal importance.

A special case is the straight line through $p \neq q$, i.e.,

$$\flat(p, q) = p + \mathbb{R}(q - p) = q + \mathbb{R}(p - q) = \{(1 - \alpha)p + \alpha q : \alpha \in \mathbb{R}\}.$$

(7) The finite set $Q \subset F$ is called **affinely independent** in case, for some (hence for every) $o \in Q$, $[q - o : q \in Q\backslash o]$ is 1-1. In that case, each $p \in \flat(Q)$ can be written in exactly one way as an affine combination

$$p =: \sum_q q\ell_q(p), \quad \sum_q \ell_q(p) = 1,$$

of the $q \in Q$. Indeed, in that case, for any particular $o \in Q$, $V_o := [q - o : q \in Q \backslash o]$ is a basis for the vector space of translations on $\flat(Q)$, hence, for all $p \in \flat(Q)$,

$$p = o + (p - o) = o + V_o V_o^{-1}(p - o) = \sum_{q \in Q} q \ell_q(p),$$

with

$$(\ell_q(p) : q \in Q \backslash o) := V_o^{-1}(p - o), \quad \ell_o(p) := 1 - \sum_{q \neq o} \ell_q(p).$$

The 'affine' vector $\ell(p) = (\ell_q(p) : q \in Q) \in \mathbb{R}^Q$ constitutes the **barycentric coordinates of p with respect to Q**.

It follows that, for arbitrary $p_i \in \flat(Q)$ and arbitrary $\alpha_i \in \mathbb{R}$ with $\sum_i \alpha_i = 1$, we have

$$\sum_i \alpha_i p_i = \sum_i \alpha_i \sum_q \lambda_q(p_i) q = \sum_q (\sum_i \alpha_i \lambda_q(p_i)) q,$$

with

$$\sum_i \alpha_i (\sum_q \lambda_q(p_i)) = \sum_i \alpha_i = 1.$$

Hence, by the uniqueness of the barycentric coordinates, the map

$$\lambda : \flat(Q) \to \mathbb{R}^Q : p \mapsto (\lambda_q(p) : q \in Q)$$

is **affine**, meaning that

$$\lambda(\sum_i \alpha_i p_i) = \sum_i \alpha_i \lambda(p_i).$$

It is also 1-1, of course, and so is, for our flat $\flat(Q)$, what a coordinate map is for a vector space, namely a convenient structure-preserving numerical representation of the flat.

It follows that, with $f_0 : Q \to G$ an arbitrary map on Q into some flat G, the map

$$f : \flat(Q) \to G : \sum_{q \in Q} \lambda_q(p) q \mapsto \sum_{q \in Q} \lambda_q(p) f_0(q)$$

is affine. Hence, if $A : f \to G$ is an affine map that agrees with f_0 on Q, then it must equal f.

(8) Let the $r + 1$-subset Q of the r-dimensional flat F be affinely independent. Then, for any $o \in Q$, $[q - o : q \in Q \backslash o]$ is a basis for $F - F$, and the scalar-valued map

$$\ell_o : F \to \mathbb{R} : p \mapsto \ell_o(p)$$

is a **linear polynomial** on F. Some people prefer to call it an **affine polynomial** since, after all, it is not a *linear* map. However, the adjective 'linear' is used here in the sense of 'degree ≤ 1', in distinction to quadratic, cubic, and higher-degree polynomials. A description for the latter can be obtained directly from the ℓ_q, $q \in Q$, as follows. The column map

$$[\ell_\alpha := \prod_{q\in Q} (\ell_q)^{\alpha(q)} : \alpha \in \mathbb{Z}_+^Q, |\alpha| = k]$$

into $\mathbb{R}^F$ is a basis for the (scalar-valued) polynomials of degree $\le k$ on F.

(9) An affine combination with nonnegative weights is called a **convex combination**. The weights being affine, hence summing to 1, they must also be no bigger than 1. The set

$$[p\,..\,q] := \{(1-\alpha)p + \alpha q : \alpha \in [0\,..\,1]\}$$

of all convex combinations of the two points p, q is called the **interval with endpoints** p**,** q. The set

$$\sigma_Q := \{\sum_{q\in Q} q\alpha_q : \alpha \in [0\,..\,1]^Q, \sum_q \alpha_q = 1\}$$

of all convex combinations of points in the finite set Q is called the **simplex with vertex set** Q in case Q is affinely independent.

(10) Flats are the proper setting for *differentiation*. Assume that the flat F is finite-dimensional. Then there are many ways to introduce a vector norm on the corresponding vector space $F{-}F$ of translations, hence a notion of convergence, but which vector sequences converge and which don't is independent of the choice of that norm. This leads in a natural way to convergence on F: *The point sequence* $(p_n : n \in \mathbb{N})$ *in* F **converges** *to* $p \in F$ *exactly when* $\lim_{n\to\infty} \|p_n - p\| = 0$. Again, this characterization of convergence does not depend on the particular vector norm on $F{-}F$ chosen.

With this, the function $f : F \to G$, on the finite-dimensional flat F to the finite-dimensional flat G, is **differentiable at** $p \in F$ in case the limit

$$D_\tau f(p) := \lim_{h\searrow 0} (f(p + h\tau) - f(p))/h$$

exists for every $\tau \in (F{-}F)\backslash 0$. In that case, $D_\tau f(p)$ is called the **derivative of** f **at** p **in the direction** τ.

Notice that $D_\tau f(p)$ is a *vector*, in $G{-}G$. It tells us the direction into which $f(p)$ gets translated as we translate p to $p+\tau$. Further, its magnitude gives an indication of the size of the change as a function of the size of the change in p. Exactly,

$$f(p + h\tau) = f(p) + hD_\tau f(p) + o(\|\tau\|h), \quad h \ge 0.$$

In particular, if f is differentiable at p, then

$$Df(p) : F{-}F \to G{-}G : \tau \mapsto D_\tau f(p)$$

is a well-defined map, from $F{-}F$ to $G{-}G$. This map is positively homogeneous, i.e.,

$$D_{h\tau} f(p) = h D_\tau f(p), \quad h \ge 0.$$

If this map $Df(p)$ is linear, it is called the **derivative of f at p**. Note that then

$$(16.5) \qquad f(p+\tau) = f(p) + Df(p)\tau + o(\|\tau\|), \quad \tau \in F{-}F.$$

If V is any particular basis for $F{-}F$ and W is any particular basis for $G{-}G$, then the matrix

$$Jf(p) := W^{-1} Df(p) V$$

is the **Jacobian** of f at p. Its (i,j) entry tells us how much $f(p+\tau)$ moves in the direction of w_i because of a unit change in τ in the direction of v_j. More precisely, if $\tau = V\alpha$, then $Df(p)\tau = W\, Jf(p)\alpha$.

A practical high-point of these considerations is the **chain rule**, i.e., the observation that if $g : G \to H$ is a ‘uniformly’ differentiable map, then their composition, gf, is differentiable, and

$$D(gf)(p) = Dg(f(p))Df(p).$$

In most applications, both F and G are coordinate spaces and, correspondingly, the bases V and W are the standard ones.

If, in particular, $F = \mathbb{R}^n$ and $G = \mathbb{R}$, i.e., if f is a scalar-valued function of n real variables, then the Jacobian Df is a 1-row matrix or vector, called the **gradient** of f, and denoted

$$\mathrm{grad} f = \nabla f = (D_1 f, \ldots, D_n f),$$

with $D_i f$ the directional derivative of f in the direction of $\mathbf{e}_i$. Then, directly from (16.5), the gradient $\nabla f(p)$ gives the direction of steepest ascent at p.

An example from CAGD

In Computer-Aided Geometric Design, one uses repeated corner-cutting to refine a given polygon into a smooth curve of approximately the same shape. The best-known example is the **Chaikin algorithm**. This algorithm consists in applying repeatedly, until satisfied, the following step:

```
    input: the vertices $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \mathbf{x}_{n+1} := \mathbf{x}_1 \in \mathbb{R}^2$ of a closed polygon.
for $j = 1{:}n$, do: $\mathbf{y}_{2j-1} \leftarrow (3\mathbf{x}_j + \mathbf{x}_{j+1})/4$; $\mathbf{y}_{2j} \leftarrow (\mathbf{x}_j + 3\mathbf{x}_{j+1})/4$; enddo
    output: the vertices $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{2n}, \mathbf{y}_{2n+1} := \mathbf{y}_1 \in \mathbb{R}^2$ of a closed polygon that is inscribed into the input polygon.
```

In other words,

$$[\mathbf{y}_1, \ldots, \mathbf{y}_{2n}] = [\mathbf{x}_1, \ldots, \mathbf{x}_n] C_n,$$

with C_n the $n \times (2n)$-matrix

$$C_n := \begin{bmatrix} 3 & 1 & 0 & 0 & 0 & 0 & \cdots & 1 & 3 \\ 1 & 3 & 3 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & 3 & 3 & 1 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 3 & 1 \end{bmatrix} /4.$$

It is possible to show that, as $k \to \infty$, the polygon with vertex sequence

$$[\mathbf{x}_1^{(k)}, \ldots, \mathbf{x}_{2^k n}^{(k)}] := [\mathbf{x}_1, \ldots, \mathbf{x}_n] C_n C_{2n} \cdots C_{2^k n}$$

converges to a smooth curve, namely the curve

$$t \mapsto \sum_j \mathbf{x}_j B_2(t-j),$$

with B_2 a certain smooth piecewise quadratic function, a so-called quadratic B-spline (see, e.g., page “pageBspline).

Here, we consider the following much simpler and more radical corner-cutting:

$$[\mathbf{y}_1, \ldots, \mathbf{y}_n] = [\mathbf{x}_1, \ldots, \mathbf{x}_n] A,$$

with

$$A := \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 1 \\ 1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 1 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 1 \end{bmatrix} /2. \qquad (16.6)$$

In other words, the new polygon is obtained from the old by choosing as the new vertices the midpoints of the edges of the old.

Simple examples, hand-drawn, quickly indicate that the sequence of polygons, with vertex sequence

$$[\mathbf{x}_1^{(k)}, \ldots, \mathbf{x}_n^{(k)}] := [\mathbf{x}_1, \ldots, \mathbf{x}_n] A^k$$

seem to shrink eventually into a point. Here is the analysis that this is, in fact, the case, with that limiting point equal to the average, $\sum_j \mathbf{x}_j/n$, of the original vertices.

(i) The matrix A, defined in (16.6), is a **circulant**, meaning that each row is obtainable from its predecessor by shifting everything one to the right, with the right-most entry in the previous row becoming the left-most entry of the next row. All such matrices have (see Problem 16.4) eigenvectors of the form

$$\mathbf{u}_\lambda := (\lambda^1, \lambda^2, \ldots, \lambda^n), \tag{16.7}$$

with the scalar λ chosen so that $\lambda^n = 1$, hence $\lambda^{n+1} = \lambda$.

For our A, we compute

$$A\mathbf{u}_\lambda = (\lambda^n + \lambda^1, \lambda^1 + \lambda^2, \ldots, \lambda^{n-1} + \lambda^n)/2.$$

Hence, if $\lambda^n = 1$, therefore $\lambda^{-1} = \lambda^{n-1}$, we get

$$A\mathbf{u}_\lambda = \frac{\lambda^{n-1}+1}{2}\mathbf{u}_\lambda.$$

(ii) The equation $\lambda^n = 1$ has exactly n distinct solutions, namely the n **roots of unity**

$$\lambda_j := \exp(2\pi \mathrm{i} j/n) = \omega^j, \quad j = 1{:}n.$$

Here,

$$\omega := \omega_n := \exp(2\pi\mathrm{i}/n)$$

is a **primitive nth root of unity**. Note that

$$\overline{\omega} = 1/\omega.$$

Let

$$V = [\mathbf{v}_1, \ldots, \mathbf{v}_n] := [\mathbf{u}_{\lambda_1}, \ldots, \mathbf{u}_{\lambda_n}]$$

be the column map whose jth column is the eigenvector

$$\mathbf{v}_j := (\omega^j, \omega^{2j}, \ldots, \omega^{nj})$$

of A, with corresponding eigenvalue

$$\mu_j := (\lambda_j^{n-1} + 1)/2 = (\omega^{-j} + 1)/2, \quad j = 1{:}n.$$

Since these eigenvalues are distinct, V is 1-1 (by (10.10)Lemma), hence V is a basis for $\mathbb{C}^n$. In particular,

$$A = V \operatorname{diag}(\ldots, \mu_j, \ldots)V^{-1}.$$

(iii) It follows that

$$A^k = V\,\mathrm{diag}(\ldots,\mu_j^k,\ldots)V^{-1} \xrightarrow[k\to\infty]{} V\,\mathrm{diag}(0,\ldots,0,1)V^{-1}$$

since $|\mu_j| < 1$ for $j < n$, while $\mu_n = 1$. Hence

$$\lim_{k\to\infty} A^k = [\mathbf{v}_n]V^{-1}(n,:).$$

(iv) In order to compute $(V^{-1})_{n\mathbf{:}}$, we compute $V^{\mathrm{c}}V$ (recalling that $\overline{\omega} = \omega^{-1}$):

$$(V^{\mathrm{c}}V)_{jk} = {\mathbf{v}_j}^{\mathrm{c}}\mathbf{v}_k = \sum_{r=1}^{n} \omega^{-rj}\,\omega^{rk} = \sum_{r=1}^{n} \omega^{(k-j)r}.$$

That last sum is a geometric series, of the form $\sum_{r=1}^{n} \nu^r$ with $\nu := \omega^{k-j}$, hence equals n in case $k = j$, and otherwise $\nu \neq 1$ and the sum equals $(\nu^{n+1} - \nu)/(\nu - 1) = 0$ since $\nu^n = 1$, hence $\nu^{n+1} - \nu = 0$. It follows that

$$V^{\mathrm{c}}V = n\,\mathrm{id}_n,$$

i.e., $V/\sqrt{n}$ is *unitary*, i.e., an o.n. basis for $\mathbb{C}^n$. In particular, $V^{-1} = V^{\mathrm{c}}/n$, therefore

$$(V^{-1})_{n\mathbf{:}} = {\mathbf{v}_n}^{\mathrm{c}}/n.$$

(v) It follows that

$$\lim_{k\to\infty} A^k = \mathbf{v}_n{\mathbf{v}_n}^{\mathrm{c}}/n,$$

with

$$\mathbf{v}_n = (1,1,\ldots,1).$$

Consequently,

$$\lim_{k\to\infty}[\ldots,x_j^{(k)},\ldots] \;=\; \sum_j x_j/n\,{\mathbf{v}_n}^{\mathrm{c}} \;=\; [\ldots,\sum_j x_j/n,\ldots],$$

i.e., the rank-one matrix all of whose columns equal the average $\sum_j x_j/n$ of the vertices of the polygon we started out with.

16.4* Prove that *any circulant matrix of order n has eigenvectors in $\mathbb{C}^n$ of the form (16.7) with $\lambda^n = 1$.* Conclude that any circulant is diagonalizable, with an orthogonal basis of eigenvectors.

Tridiagonal Toeplitz matrix

Circulants are a special case of *Toeplitz* matrices, i.e., of matrices that are constant along diagonals. Precisely, the matrix A is **Toeplitz** if

$$\forall i, j, \quad A_{ij} = a_{i-j}$$

for some sequence $(\ldots, a_{-2}, a_{-1}, a_0, a_1, a_2, \ldots)$ of appropriate domain. Circulants are special in that the determining sequence $\mathbf{a}$ for them is periodic, i.e., $a_{i+n} = a_i$, all i, if A is of order n.

Consider now the case of a **tridiagonal** Toeplitz matrix A. For such a matrix, only the (main) diagonal and the two next-to-main diagonals are (perhaps) nonzero; all other entries are zero. This means that only a_{-1}, a_0, a_1 are, perhaps, nonzero, while $a_i = 0$ for $|i| > 1$. If also $a_{-1} = a_1 \neq 0$, then the circulant trick, employed in the preceding section, still works, i.e., we can fashion some eigenvectors from vectors of the form $\mathbf{u}_\lambda = (\lambda^1, \ldots, \lambda^n)$. Indeed, now

$$(A\mathbf{u}_\lambda)_j = \begin{cases} a_0\lambda + a_1\lambda^2 & \text{for } j = 1; \\ a_1\lambda^{j-1} + a_0\lambda^j + a_1\lambda^{j+1} & \text{for } j = 2{:}n{-}1; \\ a_1\lambda^{n-1} + a_0\lambda^n & \text{for } j = n. \end{cases}$$

Hence,

$$A\mathbf{u}_\lambda = (a_1/\lambda + a_0 + a_1\lambda)\mathbf{u}_\lambda - a_1(\mathbf{e}_1 + \lambda^{n+1}\mathbf{e}_n).$$

At first glance, this doesn't look too hopeful since we are after eigenvectors. However, recall that, for a unimodular λ, i.e., for $\lambda = \exp(\mathrm{i}\varphi)$ for some real φ, we have $1/\lambda = \overline{\lambda}$, hence $a_1/\overline{\lambda} + a_0 + a_1\overline{\lambda} = a_1/\lambda + a_0 + a_1\lambda$ and therefore

$$A\mathbf{u}_{\overline{\lambda}} = (a_1/\lambda + a_0 + a_1\lambda)\mathbf{u}_{\overline{\lambda}} - a_1(\mathbf{e}_1 + \overline{\lambda^{n+1}}\mathbf{e}_n).$$

It follows that, with $\lambda =: \exp(\mathrm{i}\varphi)$ and

$$\mathbf{v}_\lambda := (\mathbf{u}_\lambda - \mathbf{u}_{\overline{\lambda}})/(2\mathrm{i}) = (\sin(k\varphi) : k = 1{:}n),$$

we obtain

$$A\mathbf{v}_\lambda = \mu_\lambda \mathbf{v}_\lambda + a_1 \sin((n+1)\varphi)\mathbf{e}_n$$

where

$$\mu_\lambda := a_0 + a_1(\lambda + \overline{\lambda}) = a_0 + a_1 2\cos(\varphi).$$

Thus, with $\lambda_k := \exp(\mathrm{i}k/(n+1))$, $k = 1{:}n$, $\mathbf{v}_k := \mathbf{v}_{\lambda_k}$ is an eigenvector of A with corresponding eigenvalue $\mu_k := \mu_{\lambda_k}$ and, since we assumed that $a_1 \neq 0$, these n numbers μ_k are pairwise distinct, hence $V =: [\mathbf{v}_1, \ldots, \mathbf{v}_n]$ is 1-1 by (10.10)Lemma, hence a basis for $\mathbb{C}^n$. In fact, since V maps $\mathbb{R}^n$ to $\mathbb{R}^n$, V is a basis for $\mathbb{R}^n$. Hence if both a_0 and a_1 are real, then also each μ_j is real and then, A is diagonalizable even over $\mathbb{F} = \mathbb{R}$.

Markov Chains

Recall from page 151 our example of a random walk on some graph. There we were interested in the matrices M^k, $k = 1, 2, 3, \ldots$, with the entries of the matrix M of order n all nonnegative and all entries in any particular row adding up to 1. In other words, $M \geq 0$ and $M\mathbf{e} = \mathbf{e}$, with

$$\mathbf{e} := (1, 1, \ldots, 1) \in \mathbb{R}^n.$$

In particular, $1 \in \operatorname{spct}(M)$. Further, since $\|M\|_\infty = 1$, we conclude from (13.1) that $\rho(M) \leq 1$. Hence, 1 is an absolutely largest eigenvalue for M. Assume, in addition, that M is irreducible. This is certainly guaranteed if $M > 0$. Then, by the Perron-Frobenius theory, 1 is a nondefective eigenvalue of M, and is the unique absolutely largest eigenvalue. By (11.10)Theorem, this implies that, with $Y := \operatorname{ran}(M - \operatorname{id})$, $\rho(M|_Y) < 1$, hence, for any basis W for Y, $V := [\mathbf{e}, W]$ is a basis for $\mathbb{R}^n$, and

$$MV = [\mathbf{e}, MW] = V \operatorname{diag}(1, B)$$

with $B := W^{-1}MW$ the matrix representation of $M|_Y$ with respect to W, hence $\rho(B) < 1$. Therefore,

$$M^k = V \operatorname{diag}(1, B^k) V^{-1} \xrightarrow[k\to\infty]{} V \operatorname{diag}(1, 0) V^{-1}.$$

In other words,

$$\lim_{k\to\infty} M^k = \mathbf{e}\mathbf{u}^{\mathrm{t}},$$

with $\mathbf{u}$ the first row of V^{-1}, hence (see, e.g., Problem 11.6) $M^{\mathrm{t}}\mathbf{u} = \mathbf{u}$, i.e., $\mathbf{u}$ is an eigenvector of M^{t} belonging to the eigenvalue 1. In particular, all rows of M^k converge to this particular nonnegative vector whose entries sum to 1 since $\mathbf{u}^{\mathrm{t}}\mathbf{e} = (V^{-1}V)_{1,1} = 1$.

Polynomial interpolation and divided differences

Polynomial interpolation is a fundamental topic, particularly for Numerical Analysis and Scientific computing. We discussed it briefly in (3.37)Example in Chapter 3, to illustrate the use of the fact that triangular matrices with nonzero diagonal entries are invertible.

To recall, polynomial interpolation involves the construction of a polynomial of degree $< k$ that matches given scalar values at a given k-sequence $\boldsymbol{\tau} := (\tau_h : h = 1{:}k)$.

In this section, we look in some detail at the **Newton form** wrto $\boldsymbol{\tau}$, i.e., the expansion

$$p =: \sum_j w_{j-1,\tau} a_j \tag{16.8}$$

of the resulting interpolating polynomial in terms of the so-called *Newton polynomials*

$$w_{j,\boldsymbol{\tau}} : t \mapsto \prod_{0<h\le j} (t - \tau_h), \quad j = 0, 1, 2, \ldots$$

introduced in (3.37)Example for an arbitrary sequence $\boldsymbol{\tau} = (\tau_h : h \in \mathbb{N})$ of scalars.

Consider the linear map

$$W_{\boldsymbol{\tau}} : \mathbb{F}_0^{\mathbb{N}} \to \Pi : \mathbf{a} \mapsto \sum_j w_{j-1,\boldsymbol{\tau}} a_j$$

from infinite scalar sequences with only finitely many nonzero entries into the space

$$\Pi = \Pi(\mathbb{F})$$

of (univariate) scalar-valued polynomials.

(16.9) Proposition. *$W_{\boldsymbol{\tau}}$ is 1-1 and onto, hence a basis for Π.*

Proof: By Problem 3.31, for any $k \in \mathbb{N}$,

$$W_{k,\boldsymbol{\tau}} := [w_{0,\boldsymbol{\tau}}, \ldots, w_{k-1,\boldsymbol{\tau}}]$$

is a basis for $\Pi_{<k}$.

Since $\Pi = \cup_k \Pi_{<k}$, this proves that $W_{\boldsymbol{\tau}}$ is invertible. □

It follows that each coefficient a_j in (16.8) is a linear function of p, i.e.,

$$a_j = (W_{\boldsymbol{\tau}}^{-1} p)_j, \quad p \in \Pi, \quad j \in \mathbb{N}, \tag{16.10}$$

and depends, in particular, on $\boldsymbol{\tau}$. We now investigate that dependence in some detail, based on the simple observation that, for $k < j$, $w_{k,\boldsymbol{\tau}}$ is a factor of $w_{j,\boldsymbol{\tau}}$. This permits us to write, for $p \in \Pi$ and $k \in \mathbb{N}$,

$$p = \sum_{j\in\mathbb{N}} w_{j-1,\boldsymbol{\tau}} a_j =: p_{k,\boldsymbol{\tau}} + w_{k,\boldsymbol{\tau}} q_{k,\boldsymbol{\tau}},$$

with

$$p_{k,\boldsymbol{\tau}} := \sum_{j=1}^{k} w_{j-1,\boldsymbol{\tau}} a_j$$

the sum of the first k terms, and with $q_{k,\boldsymbol{\tau}}$ some polynomial which we will look at more closely in a moment.

Since $\deg p_{k,\boldsymbol{\tau}} < k = \deg w_{k,\boldsymbol{\tau}}$, it follows (see page 281) from

$$p = p_{k,\boldsymbol{\tau}} + w_{k,\boldsymbol{\tau}} q_{k,\boldsymbol{\tau}} \tag{16.11}$$

that $p_{k,\boldsymbol{\tau}}$ *is the remainder after division of* p *by* $w_{k,\boldsymbol{\tau}} = (\cdot - \tau_1)\cdots(\cdot - \tau_k)$, *hence depends only on* p *and* $\tau_{\underline{k}} := (\tau_1, \ldots, \tau_k)$. This is reflected in the notation

$$\Delta(\tau_{\underline{k}})p := (W_{\boldsymbol{\tau}}^{-1}p)_k, \quad p \in \Pi, \quad k \in \mathbb{N} \tag{16.12}$$

we adopt from now on for the coefficient a_k of $w_{k-1,\boldsymbol{\tau}}$ in the Newton form (16.8) for $p \in \Pi$. In effect, we use the definition

$$p =: \sum_k w_{k-1,\boldsymbol{\tau}} \Delta(\tau_{\underline{k}})p, \quad p \in \Pi. \tag{16.13}$$

This definition is quite powerful. For example, since the $w_{j,\boldsymbol{\tau}}$ depend continuously on $\boldsymbol{\tau}$, so does $W_{\boldsymbol{\tau}}$, hence so does $(W_{\boldsymbol{\tau}})^{-1}$ (see Problem 7.11), therefore $\Delta(\tau_{\underline{k}})p$ *is a continuous function of* $\tau_{\underline{k}} = (\tau_1, \ldots, \tau_k)$.

This continuity is particularly useful to know since

$$p = \sum_k (\cdot - t)^k (D^k p)(t)/k!$$

(the Taylor expansion for $p \in \Pi$ at t), while, for $\boldsymbol{\tau}$ the constant sequence $(t, t, \ldots)$, $w_{k,\boldsymbol{\tau}} = (\cdot - t)^k$, all k, hence we conclude that

$$\Delta([t^{[k+1]})p := \Delta(\underbrace{t, \ldots, t}_{k+1 \text{ terms}})p = D^k p(t)/k!, \tag{16.14}$$

and therefore

$$\lim_{(\tau_0, \ldots, \tau_k) \to t^{[k+1]}} \Delta(\tau_0, \ldots, \tau_k)p = D^k p(t)/k!.$$

It follows that $\Delta(t)$ is a colorful symbol for the linear functional $p \mapsto p(t)$ of evaluation at t.

The symbol $\Delta(...)$ is read **divided difference at** ..., and the reason for this terminology is the following.

For $k = 2$, (16.13) says that

$$p(\tau_2) = p(\tau_1) + (\tau_2 - \tau_1)\Delta(\tau_1, \tau_2)p,$$

hence, in case $\tau_2 \neq \tau_1$,

$$\Delta(\tau_1, \tau_2)p = \frac{p(\tau_2) - p(\tau_1)}{\tau_2 - \tau_1},$$

a ratio of differences, i.e., a **divided difference**, and this is true for every k, as the following argument shows.

Since $p_{k+1,\boldsymbol{\tau}} = p_{k,\boldsymbol{\tau}} + w_{k,\boldsymbol{\tau}}\Delta(\tau_1,\ldots,\tau_{k+1})p$, while $p = p_{k,\boldsymbol{\tau}} + w_{k,\boldsymbol{\tau}}q_{k,\boldsymbol{\tau}}$ and $p_{k+1,\boldsymbol{\tau}}(\tau_{k+1}) = p(\tau_{k+1})$, we have

$$w_{k,\boldsymbol{\tau}}(\tau_{k+1})\Delta(\tau_1,\ldots,\tau_{k+1})p = w_{k,\boldsymbol{\tau}}(\tau_{k+1})q_{k,\boldsymbol{\tau}}(\tau_{k+1}),$$

therefore

$$q_{k,\boldsymbol{\tau}}(\tau_{k+1}) = \Delta(\tau_1,\ldots,\tau_{k+1})p,$$

at least for any τ_{k+1} for which $w_{k,\boldsymbol{\tau}}(\tau_{k+1}) \neq 0$, hence for every $\tau_{k+1} \in \mathbb{F}$, by the continuity of $q_{k,\boldsymbol{\tau}}$ and the continuous dependence of $\Delta(\tau_1,\ldots,\tau_{k+1})$ on τ_{k+1}. In short,

$$q_{k,\boldsymbol{\tau}} = \Delta(\tau_{\underline{k}},\cdot)p.$$

More than that, since $w_{k,\boldsymbol{\tau}}$ is symmetric in the τ_i, $i \in \underline{k}$, $p_{k,\boldsymbol{\tau}}$ also does not depend on the order of the τ_i, $i \in \underline{k}$

We describe this by saying that $p_{k,\boldsymbol{\tau}}$ **agrees with** p **at** $\tau_{\underline{k}} := (\tau_h : h \in \underline{k})$ and observe that this means (see Problem 16.6) that

$$D^r p(z) = D^r p_{k,\boldsymbol{\tau}}(z), \quad 0 \le r < m_i := \#\{i \in \underline{k} : \tau_i = z\}, \quad z \in \mathbb{F}.$$

This is called **Hermite interpolation** in case some of the τ_i coincide, and is called **Lagrange interpolation** otherwise.

16.5* Prove that, *for fixed* $p \in \Pi$, $(W_{\boldsymbol{\tau}})^{-1}p$ *depends continuously on* $\boldsymbol{\tau}$.

16.6* Prove that,for $f \in \Pi$,

$$D^i f(z) = 0, \quad 0 \le i < m \quad \iff \quad f \in (\cdot - z)^m\Pi,$$

i.e, iff f has an m-fold zero at z.

Linear Programming

This application can serve as a reinforcement of the discussion of Elimination in Chapter 3.

In Linear Programming, one seeks a minimizer for a *linear* **cost function**

$$\mathbf{x} \mapsto \mathbf{c}^{\mathrm{t}}\mathbf{x}$$

on the set

$$F := \{\mathbf{x} \in \mathbb{R}^n : A\mathbf{x} \le \mathbf{y}\}$$

of all n-vectors $\mathbf{x}$ that satisfy the m **linear constraint**s

$$A(i,:)\mathbf{x} \le y_i, \quad i = 1{:}m,$$

with $\mathbf{c} \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$, $\mathbf{y} \in \mathbb{R}^m$ given. Here and below, for $\mathbf{y},\mathbf{z} \in \mathbb{R}^m$,

$$\mathbf{y} \le \mathbf{z} \ := \ \mathbf{z} - \mathbf{y} \in \mathbb{R}^m_+ := \{\mathbf{u} \in \mathbb{R}^m : 0 \le u_j,\ j = 1{:}m\},$$

i.e., the inequality is to hold pointwise (or, entry-wise).

The set F, also called the **feasible set**, is the intersection of m halfspaces, i.e., sets of the form

$$H(\mathbf{a}, y) := \{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}^{\mathrm{t}}\mathbf{x} \le y\}.$$

Such a **halfspace** consists of all the points that lie on that side of the corresponding **hyperplane**

$$h(\mathbf{a}, y) := \{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}^{\mathrm{t}}\mathbf{x} = y\}$$

that the **normal a** of the hyperplane points away from; see (2.9)Figure, or (16.16)Figure.

Here is a simple example: Minimize

$$2x_1 + x_2$$

over all $\mathbf{x} \in \mathbb{R}^2$ for which

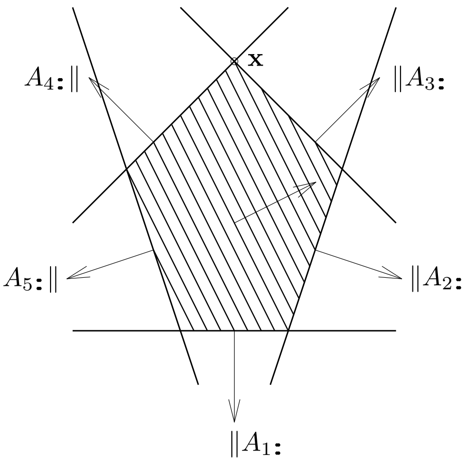
$$x_2 \ge -2, \quad 3x_1 - x_2 \le 5, \quad x_1 + x_2 \le 3,$$

$$x_1 - x_2 \ge -3, \quad 3x_1 + x_2 \ge -5.$$

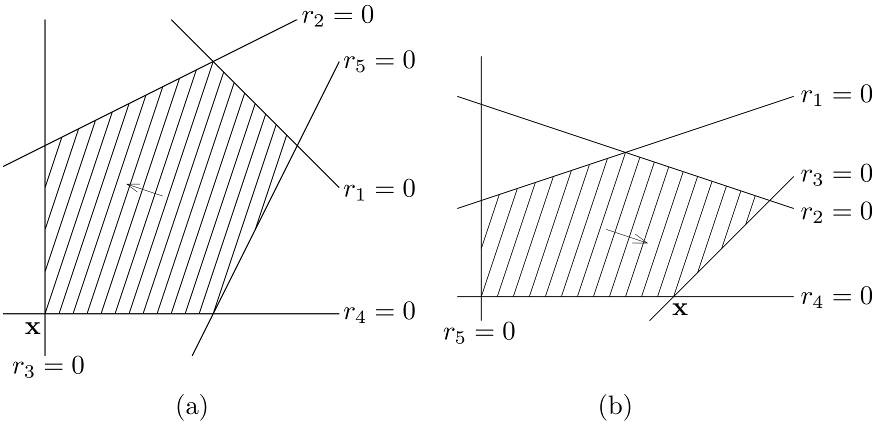
In matrix notation, and more uniformly written, this is the set of all $\mathbf{x} \in \mathbb{R}^2$ for which $A\mathbf{x} \le \mathbf{y}$ with

$$(16.15) \qquad A := \begin{bmatrix} 0 & -1 \\ 3 & -1 \\ 1 & 1 \\ -1 & 1 \\ -3 & -1 \end{bmatrix}, \quad [\mathbf{y}] := \begin{bmatrix} 2 \\ 5 \\ 3 \\ 3 \\ 5 \end{bmatrix}.$$

In this simple setting, you can visualize the set F by drawing each of the hyperplanes $h(A_{i:}, y_i)$ along with a vector $\|A_{i:}$ parallel to $A_{i:}$, i.e., pointing in the same direction as its normal vector, $A_{i:}$; the set F lies on the side that the normal vector points away from; see (16.16)Figure. □



(16.16) Figure. The feasible set for five linear constraints in the plane, as filled out by some level lines of the cost function. Since the gradient of the cost function is shown as well, the location of the minimizer is clear.



(16.17) Figure. The same setting as in (16.16)Figure but viewed in terms of the (nonbasic) variables (a) r_3, r_4; (b) r_5, r_4.

In order to provide a handier description for F, one introduces the so-called **slack variables**

$$\mathbf{r} := \mathbf{y} - A\mathbf{x};$$

earlier, we called this the *residual.* With their aid, we can describe F as

$$F = \{\mathbf{x} \in \mathbb{R}^n : \exists \mathbf{r} \in \mathbb{R}^m_+ \text{ s.t. } (\mathbf{x}, \mathbf{r}, -1) \in \text{null}[A,\ \text{id}_m, \mathbf{y}]\},$$

and use elimination to obtain a concise description of null$[A,\ \text{id}_m, \mathbf{y}]$.

For this, assume that A is 1-1. Then, each column of A is bound, hence is also bound in $[A,\ \mathrm{id}_m, \mathbf{y}]$. Therefore, after n steps of the (4.2)Elimination Algorithm applied to $[A,\ \mathrm{id}_m, \mathbf{y}]$, we will arrive at a matrix B, with the same nullspace as $[A,\ \mathrm{id}_m, \mathbf{y}]$, and an n-vector $\mathtt{f}$ (with f_k the row used as pivot row for the kth unknown or column for $k = 1{:}n$), for which

$$B(\mathtt{f}, 1{:}n)$$

is upper triangular with nonzero diagonals while, with $\mathtt{b}$ the $m - n$ rows not yet used as pivot rows,

$$B(\mathtt{b}, 1{:}n) = 0.$$

Further, since the columns $(n+1){:}m$ of $[A,\ \mathrm{id}_m, \mathbf{y}]$ have nonzero entries in these pivot rows $\mathtt{f}$ only in columns $n + \mathtt{f}$, the other columns, i.e., columns $n + \mathtt{b}$, will remain entirely unchanged. It follows that if we choose, as we may, $\mathtt{b}$ to be increasing, then

$$B(\mathtt{b}, n + \mathtt{b}) = \mathrm{id}_{m-n}.$$

Therefore, after dividing each of the n pivot rows by their pivot element and then using each pivot row to eliminate its unknown also from all other pivot rows, we will arrive at a matrix, still called B and with $\operatorname{null} B = \operatorname{null}[A,\ \mathrm{id}_m, \mathbf{y}]$, for which now also

$$B(\mathtt{f}, 1{:}n) = \mathrm{id}_n.$$

For our particular example, $n = 2$, hence this matrix B will be reached after just two steps (in which I chose the pivot rows capriciously):

$$[A,\ \mathrm{id}_m, \mathbf{y}] = \begin{bmatrix} 0 & -1 & 1 & 0 & 0 & 0 & 0 & 2 \\ 3 & -1 & 0 & 1 & 0 & 0 & 0 & 5 \\ \underline{\mathbf{1}} & 1 & 0 & 0 & 1 & 0 & 0 & 3 \\ -1 & 1 & 0 & 0 & 0 & 1 & 0 & 3 \\ -3 & -1 & 0 & 0 & 0 & 0 & 1 & 5 \end{bmatrix} \to$$

$$\begin{bmatrix} 0 & -1 & 1 & 0 & 0 & 0 & 0 & 2 \\ 0 & -4 & 0 & 1 & -3 & 0 & 0 & -4 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 3 \\ 0 & \underline{\mathbf{2}} & 0 & 0 & 1 & 1 & 0 & 6 \\ 0 & 2 & 0 & 0 & 3 & 0 & 1 & 14 \end{bmatrix} \to$$

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 1/2 & 1/2 & 0 & 5 \\ 0 & 0 & 0 & 1 & -1 & 2 & 0 & 8 \\ 1 & 0 & 0 & 0 & 1/2 & -1/2 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1/2 & 1/2 & 0 & 3 \\ 0 & 0 & 0 & 0 & 2 & -1 & 1 & 8 \end{bmatrix} =: B, \qquad \begin{matrix} \mathtt{f} = (3, 4), \\ \mathtt{b} = (1, 2, 5). \end{matrix}$$

The columns $n+\mathtt{f}$ of B are free in the sense that we can freely choose $r_{\mathtt{f}}$, i.e., the slack variables associated with the n pivot rows (in Linear Programming, they are called the **nonbasic** variables), and, once they are chosen, then $\mathbf{x}$ as well as the bound slack variables, $r_{\mathtt{b}}$ (called the **basic** variables in Linear Programmming), are uniquely determined by the requirement that $(\mathbf{x}, \mathbf{r}, -1) \in \operatorname{null} B$.

This suggests eliminating $\mathbf{x}$ altogether, i.e., using the pivot rows $B(\mathtt{f}, :)$ to give

$$\mathbf{x} = B_{\mathtt{f},\mathtt{end}} - B(\mathtt{f}, n+\mathtt{f})\, r_{\mathtt{f}}, \tag{16.18}$$

(with `end` being `MATLAB`'s convenient notation for the final row or column index) and, with that, rewrite the cost function $\mathbf{x} \mapsto \mathbf{c}^{\mathrm{t}}\mathbf{x}$ in terms of $r_{\mathtt{f}}$:

$$r_{\mathtt{f}} \mapsto \mathbf{c}^{\mathrm{t}} B_{\mathtt{f},\mathtt{end}} - \mathbf{c}^{\mathrm{t}} B(\mathtt{f}, n+\mathtt{f})\, r_{\mathtt{f}}. \tag{16.19}$$

This formulation of the cost function does not involve $\mathbf{x}$, hence we can now ignore the pivot rows $\mathtt{f}$ (which are the only rows involving $\mathbf{x}$) and concentrate on finding an $r_{\mathtt{f}}$ that minimizes the cost function (16.19) subject only to the restriction that $B(\mathtt{b}, n+(1{:}m{+}1))\,(\mathbf{r}, -1) = 0$. These equations have a unique solution $\mathbf{r}$ for given $r_{\mathtt{f}}$, and the condition that $\mathbf{r} \ge 0$ is the only restriction on the possible choices of $r_{\mathtt{f}}$.

Correspondingly, we simplify the work-array B in the following two ways:

(i) We append the row $B(m{+}1, :) := \mathbf{c}^{\mathrm{t}} B(\mathtt{f}, :)$ which then permits us to write the cost function in the form (16.22).

(ii) Then, we drop entirely the n rows $\mathtt{f}$ (storing those rows perhaps in some other place against the possibility that we need to compute $\mathbf{x}$ from $r_{\mathtt{f}}$ at some later date), and also drop the first n columns.

In our example, the equation (16.18) becomes

$$\mathbf{x} = (0, 3) - \begin{bmatrix} 1/2 & -1/2 \\ 1/2 & 1/2 \end{bmatrix} (r_3, r_4) \tag{16.20}$$

and the changes to B leave us with the following, smaller, array B:

$$B = \begin{bmatrix} 1 & 0 & 1/2 & 1/2 & 0 & 5 \\ 0 & 1 & -1 & 2 & 0 & 8 \\ 0 & 0 & 2 & -1 & 1 & 8 \\ 0 & 0 & 3/2 & -1/2 & 0 & 3 \end{bmatrix}, \quad \mathtt{b} = (1, 2, 5), \quad \mathtt{f} = (3, 4), \tag{16.21}$$

in which the columns $\mathtt{f}$ are what is left of the former columns $n+\mathtt{f}$ after deletion of the rows $\mathtt{f}$. Also, $B(1{:}\#\mathtt{b}, \mathtt{b})$ is an identity matrix. In particular, each column b_i, associated with the basic variable r_{b_i}, has only one nonzero entry and that entry is 1 and sits in row i.

This change of independent variables, from $\mathbf{x}$ to the nonbasic (slack) variables $r_{\mathtt{f}}$, turns the n hyperplanes $h(A_{k\mathtt{:}}, y_k)$, $k \in \mathtt{f}$, into coordinate planes; see (16.17)Figure. In particular, the choice $r_{\mathtt{f}} = \mathbf{0}$ places us at a point of intersection of these n hyperplanes. In our example, $r_3 = 0 = r_4$, that point is $\mathbf{x} = (0, 3)$ (see (16.18)), and it is marked in (16.16)Figure and (16.17)Figure, and functions as the origin in (16.17)Figure(a).

In terms of the reduced work-array B (reproduced here

$$B = \begin{bmatrix} 1 & 0 & 1/2 & 1/2 & 0 & 5 \\ 0 & 1 & -1 & 2 & 0 & 8 \\ 0 & 0 & 2 & -1 & 1 & 8 \\ 0 & 0 & 3/2 & -1/2 & 0 & 3 \end{bmatrix}, \quad \mathtt{b} = (1, 2, 5), \quad \mathtt{f} = (3, 4),$$

for our numerical example from (16.21) for ready reference), and with

$$m' := m - n = \#\mathtt{b},$$

our minimization problem now reads: *Minimize the cost function*

$$(16.22) \qquad r_{\mathtt{f}} \mapsto B_{\mathtt{end},\mathtt{end}} - B(\mathtt{end}, \mathtt{f})\, r_{\mathtt{f}}$$

over all $r_{\mathtt{f}} \in \mathbb{R}^n_+$ *for which*

$$B_{1:m',\mathtt{end}} - B(1{:}m', \mathtt{f})\, r_{\mathtt{f}} \geq \mathbf{0}.$$

This is the **canonical form** in which linear programming problems are usually stated, and from which most textbooks start their discussion of such problems.

Note how easily accessible various relevant information now is.

(i) $B_{\mathtt{end},\mathtt{end}}$ tells us the value of the cost function at the **current point**, $r_{\mathtt{f}} = \mathbf{0}$.

(ii) Our current point is (in) the intersection of the n hyperplanes $r_k = 0$, $k \in \mathtt{f}$.

(ii) Our current point is feasible if and only if the other slack variables are nonnegative, i.e., $r_{\mathtt{b}} = B_{1:m',\mathtt{end}} \geq \mathbf{0}$. Assume that this is so. Then the only way we can further decrease the cost function is to move one of the nonbasic variables r_k, $k \in \mathtt{f}$, from its present value 0 to something positive.

(iii) For any $k \in \mathtt{f}$, r_k enters the cost function (16.22) as $-B_{\mathtt{end},k}\, r_k$. Since feasibility requires us to move such r_k from 0 to something positive, such a move would therefore lower the cost function if and only if $B_{\mathtt{end},k} > 0$.

(iv) In particular, if $B_{\mathtt{end},\mathtt{f}} \leq 0$, then no permissible change will decrease the cost function, i.e., we are at the minimum.

(v) If we were to change the nonbasic variable r_k from zero to something positive, then the basic variable r_{b_i} would change, from $B_{i,\mathtt{end}}$ to $B_{i,\mathtt{end}} -$

$B_{ik}\, r_k$. Hence, assuming $B_{i,\mathtt{end}} > 0$ and $B_{ik} > 0$, we could change r_k only to $B_{i,\mathtt{end}}/B_{ik}$ before the b_ith constraint would be violated.

(vi) Thus, the maximal feasible change for r_k would be the minimum of $B_{i,\mathtt{end}}/B_{ik}$ over all i with B_{ik} and $B_{i,\mathtt{end}}$ positive. Assume that, for the k we chose, i is chosen that way.

(vii) Such a maximal feasible change of r_k, from 0 to $B_{i,\mathtt{end}}/B_{ik}$, would make r_{b_i} zero, hence move our current point off the constraint hyperplane $r_k = 0$ and onto the constraint hyperplane $r_{b_i} = 0$, while leaving it on all the other constraint hyperplanes $r_{b_j} = 0$, $j \neq i$, making $r_{b_i} = 0$ nonbasic and r_k basic. This change can be achieved by dividing row i by B_{ik}, then using that row to eliminate r_k from all the other rows of B, and correspondingly, interchanging b_i (in $\mathtt{b}$) with k in $\mathtt{f}$. By also, in this way, eliminating r_k from the last row, we make certain that we have in hand a description of the cost function in which only the (new) $r_{\mathtt{f}}$ occurs explicitly, hence $B_{\mathtt{end},\mathtt{end}}$ gives the value of the cost function at the new current point $r_{\mathtt{f}} = \mathbf{0}$.

In our example (have a look at (16.17)Figure(a)), we already observed that our current point, $r_{\mathtt{f}} = 0$, is, indeed, feasible. But we notice that $B_{\mathtt{end},4} < 0$, hence any feasible change of r_4 would only increase the value of the cost function (16.22). This is also evident from the gradient of the cost function indicated by that arrow in the figure. On the other hand, $B_{\mathtt{end},3}$ is positive, hence we can further decrease the cost function (16.22) by increasing r_3. Such a change is limited by concerns for the positivity of

$$r_{\mathtt{b}} = B_{1:\#\mathtt{b},\mathtt{end}} - B_{1:\#\mathtt{b},3}r_3.$$

As for $r_{b_1} = r_1$, we would reach $r_1 = 0$ when r_3 equals $B_{1,\mathtt{end}}/B_{13} = 5/(1/2) = 10$, while any positive change of r_3 would make $r_2 = r_{b_2}$ only more positive since $B_{23} < 0$, and finally, $r_{b_3} = r_5$ we would reach 0 when r_3 equals $B_{3,\mathtt{end}}/B_{33} = 8/2 = 4$. We take the smaller change and thereby end up at a new vector $\mathbf{r}$, with $r_4 = 0 = r_5$, i.e., are now at the intersection of the constraint hyperplanes 4 and 5, with the cost further reduced by $B_{\mathtt{end},3} * 4 = (3/2)4 = 6$, to -3. This is the basic step: we exchange one nonbasic variable with a basic variable.

In other words, after this change, r_4 and r_5 are now the nonbasic variables. In order to have our B tell us about this new situation, and since $5 = b_3$, we merely divide its 3rd row by B_{33} then use that row to eliminate r_3 from all other rows of B. This ensures that r_3 is now basic (i.e., column 3 is a coordinate direction), and the array B has changed to

$$B = \begin{bmatrix} 1 & 0 & 0 & 3/4 & -1/4 & 3 \\ 0 & 1 & 0 & 3/2 & 1/2 & 12 \\ 0 & 0 & 1 & -1/2 & 1/2 & 4 \\ 0 & 0 & 0 & 1/2 & -3/4 & -3 \end{bmatrix}, \quad \mathtt{b} = (1,2,3), \quad \mathtt{f} = (5,4).$$

In particular, we see that the cost at $r_4 = 0 = r_5$ is, indeed, -3. We also see that $r_3 = 4$ and that $B_{\mathtt{end},4} > 0 > B_{\mathtt{end},5}$, hence know that it is possible to reduce the cost feasibly only by changing r_4 from 0 to something positive. Such a change would only increase r_3, but would reduce r_1 to zero by the choice $r_3 = B_{1,\mathtt{end}}/B_{14} = 3/(3/4) = 4$ and would reduce r_2 to zero by the choice $r_3 = B_{2,\mathtt{end}}/B_{24} = 12/(3/2) = 8$. Hence, this change is limited to the smaller one, i.e., to the change $r_4 = 4$ that makes $r_1 = 0$.

We carry out this exchange, thus making r_4 basic and r_1 nonbasic, by dividing row 1 by B_{14} and then using that row to eliminate r_4 from all other rows, to get the following B:

$$(16.23)\quad B = \begin{bmatrix} 4/3 & 0 & 0 & 1 & -1/3 & 4 \\ -2 & 1 & 0 & 0 & 1 & 6 \\ 2/3 & 0 & 1 & 0 & 1/3 & 6 \\ -1/3 & 0 & 0 & 0 & -2/3 & -4 \end{bmatrix}, \quad \mathtt{b} = (4, 2, 3), \quad \mathtt{f} = (1, 5).$$

In particular, now $B_{\mathtt{end},\mathtt{f}} \le 0$, showing that no further improvement is possible, hence -4 is the minimum of the cost function on the feasible set. At this point, $r_{3:4} = B_{[3,1],\mathtt{end}} = (6, 4)$ (since $b_{[3,1]} = (3, 4)$), hence, from the rows used as pivot rows to eliminate $\mathbf{x}$ (and saved earlier, see (16.20)), we find that, in terms of $\mathbf{x}$, our optimal point is $\mathbf{x} = (0, 3) - (1/2)\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}(6, 4) = -(1, 2)$, and, indeed, $\mathbf{c}^{\mathrm{t}}\mathbf{x} = (2, 1)^{\mathrm{t}}(-1, -2) = -4$.

The steps just carried out for our example are the standard steps of the **Simplex Method**. In this method (as proposed by Dantzig), one examines the value of the cost function only at a **vertex**, i.e., at the intersection of n of the constraint hyperplanes, i.e., at a point corresponding to $r_{\mathtt{f}} = 0$ for some choice of the n-sequence $\mathtt{f}$ in $\{1, \ldots, m\}$. Assuming the corresponding vertex feasible, i.e., that

$$r_{\mathtt{b}} = B_{1:m',\mathtt{end}} \ge \mathbf{0}$$

for the array B corresponding to this choice for $\mathtt{f}$, one checks whether

$$B_{\mathtt{end},\mathtt{f}} \le \mathbf{0}.$$

If it is, then one knows that one is at the minimum since one knows that, at any feasible point, the cost function is $B_{\mathtt{end},\mathtt{end}} - B(\mathtt{end}, \mathtt{f})\, r_{\mathtt{f}}$ for some nonnegative $r_{\mathtt{f}}$. Otherwise, one moves to a neighboring vertex at which the cost is less by exchanging a r_k for which $B_{\mathtt{end},k} > 0$ (usually the one for which $B_{\mathtt{end},k}$ is as large as possible) with some r_{b_i} with i chosen as the minimizer for $B_{i,\mathtt{end}}/B_{ik}$ over all i with $B_{ik} > 0$. This exchange is carried out by just one full elimination step applied to B, by dividing row i by B_{ik} and then using this row to eliminate r_k from all other rows, and then updating the sequences $\mathtt{b}$ and $\mathtt{f}$.

This update step is one full elimination step. It is sometimes called a **(Gauss-)Jordan** step in order to distinguish it from the **Gauss** step, the step we used in the (4.2)Elimination Algorithm in which the unknown is eliminated only from the rows not yet used as pivot rows.

Since all the information contained in the columns $B(:,\mathtt{b})$ is readily derivable from $\mathtt{b}$ and $\mathtt{f}$, one usually doesn't bother to carry these columns along. This makes the updating of the matrix $B(:,[\mathtt{f},m{+}1])$ a bit more mysterious.

Finally, there are the following points to consider:

unbounded feasible set If, for some $k \in \mathtt{f}$, $B_{\mathtt{end},k}$ is the only positive entry in its column, then increasing r_k will strictly decrease the cost *and* increase all basic variables. Hence, if $r_{\mathtt{f}} = \mathbf{0}$ is a feasible point, then we can make the cost function on the feasible set as close to $-\infty$ as we wish. In our example, this would be the case if we dropped constraints 1 and 5. Without these constraints, in our very first Simplex Method step, we could have increased r_3 without bound and so driven the cost to $-\infty$.

finding a feasible point In our example, we were fortunate in that the very first vertex we focused on was feasible. In other words, we had $B_{i,\mathtt{end}} \ge 0$ for all $i \in \mathtt{b}$, hence $r_{\mathtt{b}} \ge \mathbf{0}$ at the point $r_{\mathtt{f}} = 0$. However, if this does not hold, then $r_{\mathtt{f}} = \mathbf{0}$ is not a feasible point. Yet, we can then use the very Simplex Method to try to find a feasible point. The idea is quite simple. Suppose $B_{i,\mathtt{end}} < 0$, hence $r_{b_i} < 0$ at $r_{\mathtt{f}} = 0$. If, for some $k \in \mathtt{f}$, B_{ik} is negative, then increasing r_k will increase $B_{i,\mathtt{end}} = r_{b_i}$. In the contrary case, increasing any r_k, $k \in \mathtt{f}$, will only make $B_{i,\mathtt{end}} = r_{b_i}$ more negative. In other words, then there are no feasible points.

For example, in the following situation, the point $r_{\mathtt{f}} = 0$ is not feasible since $r_2 = B_{2,\mathtt{end}} = -12 < 0$:

$$B = \begin{bmatrix} 2 & 0 & 1 & 1 & 0 & 10 \\ -4 & 1 & -3 & 0 & 0 & -12 \\ 2 & 0 & 3 & 0 & 1 & 18 \\ 1 & 0 & 2 & 0 & 0 & 8 \end{bmatrix}, \quad \mathtt{b} = (4,2,5), \quad \mathtt{f} = (1,3).$$

Fortunately, both B_{21} and B_{23} are negative, hence a positive change in either r_1 or r_3 will increase $B_{2,\mathtt{end}}$. Since B_{21} is more negative than B_{23}, we choose to change r_1 (rather than r_3) to something positive. The limit on r_1 from the first row is $B_{1,\mathtt{end}}/B_{11} = 10/2 = 5$ and from the third row is $B_{3,\mathtt{end}}/B_{31} = 18/2 = 9$. So, we divide row 1 by $B_{11} = 2$, then use it to eliminate r_1 from all the other rows, and obtain

$$B = \begin{bmatrix} 1 & 0 & 1/2 & 1/2 & 0 & 5 \\ 0 & 1 & -1 & 2 & 0 & 8 \\ 0 & 0 & 2 & -1 & 1 & 8 \\ 0 & 0 & 3/2 & -1/2 & 0 & 3 \end{bmatrix}, \quad \mathtt{b} = (1,2,5), \quad \mathtt{f} = (3,4),$$

for which $r_{\mathtt{f}} = 0$ is a feasible point.

In general, one may have to repeat this step until one either reaches a B with $B_{1:m',\mathtt{end}} \ge \mathbf{0}$, hence the corresponding point $r_{\mathtt{f}}$ is feasible, or else, for some i, $B_{i,\mathtt{f}} \ge \mathbf{0}$ while $B_{i,\mathtt{end}} < 0$ in which case there are no feasible points.

Note that, in this way, the Simplex Method can be used to solve any finite set of linear inequalities in the sense of either providing a point satisfying them all or else proving that none exists.

degeneracies We have behaved as if the m constraint hyperplanes were **in general position**, meaning that any n of them have exactly one point in common while any $n+1$ of them have no point in common. The contrary case is called **degenerate**. In a degenerate situation, it can happen that, at the current point $r_{\mathtt{f}} = \mathbf{0}$, $r_{b_i} = B_{i,\mathtt{end}} = 0$ for some $i \in \mathtt{b}$, hence we could not move any r_k for which $B_{ik} > 0$. Instead, we could exchange r_k with r_{b_i} but since that would not change the cost function, we might not be certain of a finite termination.

convergence in finitely many steps If we can guarantee that, at each step, we strictly decrease the cost, then we must reach the vertex with minimal cost in finitely many steps since, after all, there are only finitely many vertices. A complete argument has to deal with the fact that the cost may not always strictly decrease because the current point may lie on more than just n of the constraint hyperplanes.

16.7 Find the maximum of the cost function $\mathbf{x} \mapsto 2x_1 + x_2$ over $F := \{\mathbf{x} \in \mathbb{R}^2 : A\mathbf{x} \le \mathbf{y}\}$ with A and $\mathbf{y}$ given by (16.15).

16.8 How would you modify the algorithm outlined above if the constraint set was $\{\mathbf{x} \in \mathbb{R}^n : A\mathbf{x} \ge \mathbf{y}\}$ (rather than $\{\mathbf{x} \in \mathbb{R}^n : A\mathbf{x} \le \mathbf{y}\}$)?

16.9* Show that any canonical form reached during the simplex method does not depend on the sequence of simplex method steps carried out to reach it from the starting canonical form, but only on the current sequence $\mathtt{b}$ and the starting canonical form. (Hint: remember the $\mathtt{b}$-form and Problem 4.33).

16.10 Show that the constraints $x_1 - x_2 \le -1,\ x_1 + x_2 \ge 1,\ x_1 - 2x_2 \ge -1$ are infeasible.

Total positivity

This application highlights the power of the determinant concept, albeit in a very special context.

A matrix is **totally positive** (or, **tp** for short) if all its (square) submatrices have a nonnegative determinant. It would have been better to have called such a matrix **totally nonnegative** in order to distinguish it from a matrix all of whose (square) submatrices have positive determinant which are called **strictly totally positive** (or, **stp** for short).

Any zero matrix is trivially totally positive, as is any identity matrix. Here are some more interesting examples:

(i) any **Cauchy matrix**

(ii)

(iii) all which which are actually stp.

By the multilinearity of determinants, *any positive weighted sum of tp matrices is tp.*

By the (15.6)Binet-Cauchy Formula, *the product of tp matrices is tp.*

If A is an invertible matrix and tp, then its inverse is **checkerboard** in the sense that

$$(-1)^{i+j}(A^{-1})_{ij} \geq 0.$$

This is a direct consequence of the formula (15.3)

$$(A^{-1})_{ij} = (-1)^{i+j} \det A(\backslash j \mid \backslash i)/\det A\,.$$

Consequently, $\|A\mathbf{x}\|_\infty/\|\mathbf{x}\|_\infty$ is minimized by the choice $\mathbf{x} = ((-1)^i : i = 1, 2, \ldots)$.

The following proposition offers a very striking property, a kind of monotonicity in the entries of the inverses of principal submatrices of an invertible tp matrix.

(16.24) Proposition. *If $A \in \mathbb{R}^{n\times n}$ is tp and invertible, and I is any integer subinterval in $1{:}n$, then $A(I,I)$ is also invertible and*

$$(-1)^{i+j}(A(I,I)^{-1})_{ij} \leq (-1)^{i+j}(A^{-1})_{ij}, \quad i, j \in I,$$

with the convention $A(I,I)^{-1} \in \mathbb{R}^{I\times I}$.

Least-squares approximation by broken lines

For a given strictly increasing sequence

$$\boldsymbol{\xi} = (a = \xi_1 < \cdots < \xi_{\ell+1} = b),$$

consider the collection

$$\mathrm{BL}_{\boldsymbol{\xi}} := \{f \in C([a\,..\,b]) : D^2 f(x) = 0, x \notin \boldsymbol{\xi}\}$$

of all continuous functions on the closed interval $[a\,..\,b]$ that are straight lines on each of the ℓ intervals $[\xi_i\,..\,\xi_{i+1}]$, $i = 1{:}\ell$. We call each element of $\mathrm{BL}_{\boldsymbol{\xi}}$ a **broken line with break sequence $\boldsymbol{\xi}$**.

The conditions imposed on the elements of $\mathrm{BL}_{\boldsymbol{\xi}}$, i.e., continuity and the vanishing of the second derivative at all points in $[a\,..\,b]$ other than the **interior breaks** ξ_i, $i = 2{:}\ell$, are linear, hence $\mathrm{BL}_{\boldsymbol{\xi}}$ is a linear subspace of $C([a\,..\,b])$. What is its dimension?

Since $f \in \mathrm{BL}_{\boldsymbol{\xi}}$ is a straight line on $[\xi_i\,..\,\xi_{i+1}]$, we know, e.g., from (5.7) that

$$f(x) = \frac{f(\xi_i)(\xi_{i+1} - x) + f(\xi_{i+1})(x - x_i)}{\xi_{i+1} - \xi_i} \quad \text{for} \quad x \in [\xi_i\,..\,\xi_{i+1}], \quad i = 1{:}\ell.$$

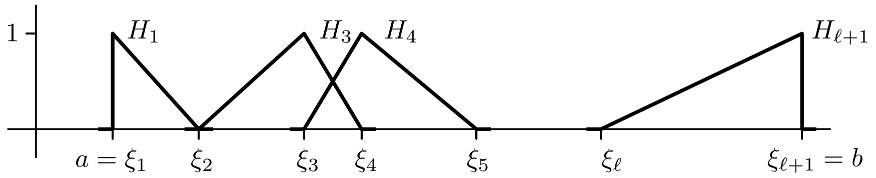
This shows that

$$f = \sum_{i=1}^{\ell+1} H_i f(\xi_i),$$

with

$$H_i(x) := \left\{ \begin{array}{ll} \frac{x-\xi_{i-1}}{\xi_i-\xi_{i-1}}, & \xi_{i-1} \le x \le \xi_i \\ \frac{\xi_{i+1}-x}{\xi_{i+1}-\xi_i}, & \xi_i \le x \le x_{i+1} \\ 0 & \text{otherwise} \end{array} \right\}, \quad a \le x \le b,$$

the unique element of $\mathrm{BL}_{\boldsymbol{\xi}}$ that is zero at all the breaks but at ξ_i where it has the value 1. Because of their characteristic shape (see (16.25) Figure), the H_i are known as **hat function**s or **chapeau function**s.



(16.25) Figure. Some of the elements of the standard basis for $\mathrm{BL}_{\boldsymbol{\xi}}$.

Consequently, $H := [H_1, \ldots, H_{\ell+1}]$ is onto $\mathrm{BL}_{\boldsymbol{\xi}}$, while the evident fact that $\delta_{\boldsymbol{\xi}} : f \mapsto f(\boldsymbol{\xi})$ is a left inverse for H implies that H is also 1-1, hence a basis for $\mathrm{BL}_{\boldsymbol{\xi}}$, with $\delta_{\boldsymbol{\xi}}$ (restricted to $\mathrm{BL}_{\boldsymbol{\xi}}$) the corresponding coordinate map. Also,

$$\|H\mathbf{a}\|_\infty = \|\mathbf{a}\|_\infty, \tag{16.26}$$

with

$$\|g\|_\infty := \max g([a \,.\,.\, b]), \quad g \in C([a \,.\,.\, b]).$$

It follows with (6.17)Theorem that we can construct, for any $g \in C[a..b]$, its least-squares approximation by broken lines with break sequence $\boldsymbol{\xi}$ as

$$P_H g := H(H^{\mathrm{c}}H)^{-1}H^{\mathrm{c}}g,$$

with $H^{\mathrm{c}}H$ the tridiagonal matrix with nonzero entries ${H_i}^{\mathrm{c}}H_j$, $|i-j| \le 1$, which for the standard inner product

$$\langle g, f\rangle = f^{\mathrm{c}}g := \int_a^b g(x)f(x)\,\mathrm{d}x$$

work out to be, with $h_i := \Delta\xi_i := \xi_{i+1} - \xi_i$ and $s := (x - \xi_i)/h_i$,

$${H_i}^{\mathrm{c}}H_{i+1} = \int_{\xi_i}^{\xi_{i+1}} \frac{x-\xi_i}{h_i}\frac{\xi_{i+1}-x}{h_i}\,\mathrm{d}x = h_i \int_0^1 s(1-s)\,\mathrm{d}s = h_i/6;$$

hence

$$H_i{}^{\mathrm{c}}H_{i-1} = H_{i-1}{}^{\mathrm{c}}H_i = h_{i-1}/6;$$

and, finally, with $t := (x - \xi_{i-1})/h_{i-1}$,

$$H_i{}^{\mathrm{c}}H_i = h_{i-1}\int_0^1 t^2\,\mathrm{d}t + h_i\int_0^1 (1-s)^2\,\mathrm{d}s = (h_{i-1}+h_i)/3.$$

It is helpful to divide the ith equation in the resulting linear system for the coordinate vector $\mathbf{a} := H^c P_H g$ of the least-squares approximation from $\mathrm{BL}_{\boldsymbol{\xi}}$ to g by

$$\int H_i = (h_{i-1}+h_i)/2,$$

and so obtain the linear system

$$(16.27)\qquad \frac{h_{i-1}/3}{h_{i-1}+h_i}a_{i-1} + (2/3)a_i + \frac{h_i/3}{h_{i-1}+h_i}a_{i+1} = M_i^c g, \quad i = 1{:}(\ell{+}1),$$

with the choice $h_0 := 0 =: h_{\ell+1}$, and with

$$M_i := 2H_i/(h_{i-1}+h_i)$$

a hat function normalized to have integral 1.

Since each M_i is nonnegative, this implies that the right-hand sides of the linear system are bounded by $\|g\|_\infty := \max(|g([a\,..\,b])|)$. Hence, if i is such that $|a_i| = \|\mathbf{a}\|_\infty$, then, by (16.27),

$$\|g\|_\infty \ge (2/3)|a_i| - (1/3)\|\mathbf{a}\|_\infty = \|\mathbf{a}\|_\infty/3,$$

or, with (16.26),

$$\|P_H g\|_\infty = \|\mathbf{a}\|_\infty \le 3\|g\|_\infty.$$

This shows that P_H, as a linear map on the normed vector space $C([a\,..\,b])$, has $\|P_H\| \le 3$, hence, by Lebesgue's inequality (7.30), the error $g - P_H g$ in the least-squares approximation to $g \in C([a\,..\,b])$ is at most 4 times the smallest possible error achievable (in the max norm) by any element of $\mathrm{BL}_{\boldsymbol{\xi}}$.

The B-spline basis for a spline space

Frames

An interesting and useful generalization of bases are **frame**s, i.e., column maps that are onto but not necessarily 1-1. The previous section offered the well-known and popular example of a **barycentric frame**, i.e., an (n+1)-column map $V = [v_0, \ldots, v_n]$ onto an n-dimensional vector space X, whose 1-dimensional nullspace has only the zero vector in common with the subspace

$$\mathrm{null}(\mathbf{e}^{\mathrm{t}}) = \{\mathbf{a} \in \mathbb{R}^{0:n} : \sum_i a_j = 0\},$$

hence V maps the flat

$$\{\mathbf{a} \in \mathbb{R}^{0:n} : \mathbf{e}^{\mathrm{t}}\mathbf{a} = 1\}$$

1-1 onto X. Such a barycentric representation of X has the advantage that it provides a stable description of the neighborhoods of $n+1$ more or less arbitrary elements of X. It also provides the means for discussing polynomials defined on an arbitrary finite-dimensional vector space.

A multivariate polynomial interpolant of minimal degree

This application makes essential use of the concept of bound and free columns introduced during the discussion of elimination; it also provides a striking example of the use of determinants.

Consider polynomial interpolation to data at a finite set T of sites in $\mathbb{F}^d$. When $d = 1$, and $k := \#\mathrm{T}$, we know that $\Pi_{<k}$ contains, for each $\mathbb{F}$-valued function g defined at least on T, exactly one p that matches g on T in the sense that $p(\boldsymbol{\tau}) = g(\boldsymbol{\tau})$ for all $\boldsymbol{\tau} \in \mathrm{T}$.

What should we do when $d > 1$?

Here are several ways to choose a polynomial subspace F that is **correct** for T in the sense that it contains, for each $a \in \mathbb{F}^{\mathrm{T}}$ exactly one p that matches a on T. This will happen exactly when $\delta_{\mathrm{T}} : f \mapsto f(\mathrm{T}) := (f(\boldsymbol{\tau}) : \boldsymbol{\tau} \in \mathrm{T}) \in \mathbb{F}^{\mathrm{T}}$ maps F *onto* $\mathbb{F}^{\mathrm{T}}$.

One idea is to use the Lagrange form

$$Pg := [\ell_{\boldsymbol{\tau}} : \boldsymbol{\tau} \in \mathrm{T}]g(\mathrm{T})$$

with the $\ell_{\boldsymbol{\tau}}$ a suitable multivariate generalization of the elements of the (univariate) Lagrange basis (5.7). Here is one such generalization:

$$\ell_{\boldsymbol{\tau}}(\mathbf{x}) := \prod_{\boldsymbol{\sigma} \in \mathrm{T}\backslash\boldsymbol{\tau}} \frac{(\boldsymbol{\tau} - \boldsymbol{\sigma})^{\mathrm{t}}(\mathbf{x} - \boldsymbol{\sigma})}{(\boldsymbol{\tau} - \boldsymbol{\sigma})^{\mathrm{t}}(\boldsymbol{\tau} - \boldsymbol{\sigma})}, \quad \boldsymbol{\tau} \in \mathrm{T}.$$

Note that, as a product of $< \#\mathrm{T}$ linear factors, each $\ell_{\boldsymbol{\tau}}$ is a polynomial of degree $< \#\mathrm{T}$ and that

$$\delta_{\mathrm{T}}[\ell_{\boldsymbol{\tau}} : \boldsymbol{\tau} \in \mathrm{T}] = \mathrm{id}. \tag{16.28}$$

Hence,

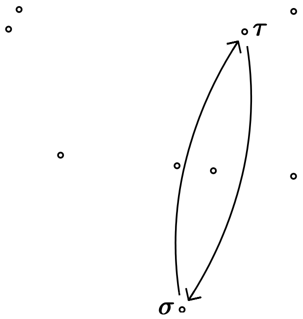
$$P_{\mathrm{T}} g := [\ell_{\boldsymbol{\tau}} : \boldsymbol{\tau} \in \mathrm{T}] g(\mathrm{T})$$

is a polynomial of degree $< \#\mathrm{T}$ that matches g on T and is the unique such in $F := \mathrm{ran}[\ell_{\boldsymbol{\tau}} : \boldsymbol{\tau} \in \mathrm{T}]$. In other words, this F is correct for T.

There are at least two objections to this choice when compared with the standard choice in the univariate case: (a) the space F depends not only on $\#\mathrm{T}$ but on T itself; and (b) the degree of the interpolating polynomial may be unnecessarily large.

Objection (a) turns out to be unreasonable because of

(16.29) Mairhuber's Theorem ([M]). For any n-dimensional subspace F of $\Pi(\mathbb{F}^d)$ with $n, d > 1$, there exists $\mathrm{T} \subset \mathbb{F}^d$ with $\#\mathrm{T} = n$ for which F is not correct.



(16.30) Figure. Interchanging two sites in the plane by a continuous move while keeping all sites distinct (something that cannot be done when all sites are restricted to lie on a straight line).

Proof: The proof is a nice demonstration of the power of determinants. We assume that, to the contrary, a particular F is correct for every choice of $\mathrm{T} \subset \mathbb{F}^d$ with $\#\mathrm{T} = \dim F > 1$, and derive from this a contradiction, as follows. Pick a basis V for F. Then the Gram matrix $\delta_{\mathrm{T}} V$ must be invertible for all choices of such T, hence

$$\det(\delta_{\mathrm{T}} V) \neq 0, \quad \mathrm{T} \subset \mathbb{F}^d, \ \#\mathrm{T} = \dim F.$$

This determinant is linear in the entries $v(\boldsymbol{\tau})$, $v \in V$, $\boldsymbol{\tau} \in \mathrm{T}$ of the Gram matrix, hence a continuous function of the $\boldsymbol{\tau}$. Since $n > 1$, T contains at least two elements, $\boldsymbol{\tau}$ and $\boldsymbol{\sigma}$, say. Since $d > 1$, we can carry out a continuous motion of these two sites which will move each to the former location of the other without ever coinciding with any other site in T. This will change the determinant continuously but, by assumption, it will never be 0, yet, at the end, we will in effect have interchanged two rows of the Gram matrix, hence will have changed the value of the determinant to its negative in a continuous manner without ever crossing 0, a contradiction. □

Objection (b), on the other hand, is well-founded. For example, since $\dim \Pi_{<2}(\mathbb{F}^d) = d+1$, we would expect to be able to find an interpolant of degree < 2 to arbitrary data $a(\mathrm{T})$ at 'most' $(d{+}1)$-sets T in $\mathbb{F}^d$ while the interpolant $[\ell_{\boldsymbol{\tau}} : \boldsymbol{\tau} \in \mathrm{T}]a(\mathrm{T})$ constructed above would usually be of degree d.

One way to deal with Objection (b) is to construct, for given T, a polynomial interpolant of minimal degree, as follows.

We know from (16.28) that δ_{T} maps

$$Y := \Pi_{<\#\mathrm{T}}(\mathbb{F}^d)$$

onto $\mathbb{F}^{\mathrm{T}}$. Therefore, in looking for interpolants of minimal degrees, it is sufficient to look for them in Y. Moreover, we know that, for any basis V of Y, the Gramian

$$\delta_{\mathrm{T}} V$$

has rank $\#\mathrm{T}$, therefore has exactly $\#\mathrm{T}$ bound columns, and we can determine the sequence $\mathtt{b}$ of the indices of these bound columns by Gauss elimination. With that, we know that, with

$$V(\,:\,,\mathtt{b}) := [v_j : j \in \mathtt{b}]$$

the column map formed from the bound columns of V, the square matrix $\delta_{\mathrm{T}} V(\,:\,,\mathtt{b})$ is 1-1, hence invertible. Therefore, $F := \operatorname{ran} V(\,:\,,\mathtt{b})$ is correct for T.

Now choose the basis V for Y to have monomials as its columns, i.e.,

$$V = [()^{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \in (\mathbb{Z}_+)^d,\ |\boldsymbol{\alpha}| < \#\mathrm{T}],$$

with

$$()^{\boldsymbol{\alpha}} : \mathbb{F}^d \to \mathbb{F} : \mathbf{x} \mapsto \mathbf{x}^{\boldsymbol{\alpha}} := x_1^{\alpha_1} \cdots x_d^{\alpha_d}$$

an unusual but handy notation for the **monomial with exponent $\boldsymbol{\alpha}$**, and

$$|\boldsymbol{\alpha}| := \sum_j \alpha_j$$

its **degree**. The resulting polynomial subspace $F = \operatorname{ran} V(\,:\,,\mathtt{b})$ will depend strongly on the order in which the monomials appear in the column map V. I will use the notation

$$\boldsymbol{\alpha} \prec \boldsymbol{\beta}$$

to indicate that $()^{\boldsymbol{\alpha}}$ appears in V to the left of $()^{\boldsymbol{\beta}}$. We define the corresponding **$\prec$-degree** of

$$r =: \sum_{\boldsymbol{\alpha}} ()^{\boldsymbol{\alpha}} \widehat{r}_{\boldsymbol{\alpha}} \in \operatorname{ran} V = \Pi_{\le k}(\mathbb{F}^d)$$

as the exponent of the right-most column of V needed in any power form of r, i.e.,

$$\deg_{\prec}(r) := \max{}_{\prec}\{\boldsymbol{\alpha} : \widehat{r}_{\boldsymbol{\alpha}} \ne 0\}.$$

(16.31) Proposition. With $k := \#\mathrm{T}$, let

$$\boldsymbol{\beta}_1 \prec \cdots \prec \boldsymbol{\beta}_{k+1}$$

be the exponents associated with the bound columns of the Gram matrix $\delta_{\mathrm{T}} V$ and set

$$W := [()^{\boldsymbol{\beta}_1}, \ldots, ()^{\boldsymbol{\beta}_{k+1}}].$$

Then, $F := \operatorname{ran} W$ is correct for T, and, for arbitrary $a \in \mathbb{F}^{\mathrm{T}}$,

$$p_a := W(\delta_{\mathrm{T}} W)^{-1} a \in F$$

is an interpolant for a on T of **minimal $\prec$-degree** in the sense that $\deg_{\prec}(p_a)$ is minimal for all polynomials that interpolate a at T.

Proof: The fact that $\delta_{\mathrm{T}} p_a = a$ is evident. For the minimality of the $\prec$-degree of p_a, we introduce the linear projector

$$P := W(\delta_{\mathrm{T}} W)^{-1} \delta_{\mathrm{T}}$$

on $\Pi(\mathbb{F}^d)$ with range $F = \operatorname{ran} W$ and nullspace $\operatorname{null} \delta_{\mathrm{T}}$, and prove the following stronger statement:

$$\deg_{\prec}(Pr) \preceq \deg_{\prec}(r), \quad r \in \operatorname{ran} V, \tag{16.32}$$

which claims that P is **$\prec$-degree reducing**. Since P is a linear map, it is sufficient to prove this claim for r a monomial, $r = ()^{\gamma}$ say. If γ is the exponent associated with a bound column of $\delta_{\mathrm{T}} V$, then $Pr = r$, and we are done. Otherwise, γ is associated with a free column of $\delta_{\mathrm{T}} V$. But this says that $\delta_{\mathrm{T}} r$ is in the range of $[\delta_{\mathrm{T}} ()^{\boldsymbol{\beta}_j} : \boldsymbol{\beta}_j \prec \gamma]$, hence $Pr \in \operatorname{ran}[()^{\boldsymbol{\beta}_j} : \boldsymbol{\beta}_j \prec \gamma]$ and therefore $\deg_{\prec}(Pr) \prec \deg_{\prec}(r)$.

Since $Pr = p_a$ for all polynomial interpolants r to a on T, (16.32) implies the claimed minimality. □

Now, by choosing, as we may, the order $\prec$ to respect the ordinary polynomial degree in the sense that

$$\boldsymbol{\alpha} \prec \boldsymbol{\beta} \quad \Longrightarrow \quad |\boldsymbol{\alpha}| \le |\boldsymbol{\beta}|,$$

we are certain that the resulting linear projector P is also **degree-reducing** in the sense that

$$\deg(Pr) \le \deg(r), \quad r \in \Pi(\mathbb{F}^d),$$

hence that p_a is an interpolant of minimal degree.

16.11 A subspace F of the space $C(R)$ of continuous functions on the subset R of $\mathbb{F}^d$ is called a **Haar space** if it is correct for every $\mathrm{T} \subset R$ with $\#\mathrm{T} = \dim F$. Prove that no subspace F of $C(R)$ can be a Haar space if $\dim F > 1$ and R contains a set shaped like a ”Y”. (Hint: marshalling yard.)

The reduced monic Gröbner basis for a zero-dimensional ideal

17 Background

A nonempty finite subset of $\mathbb{R}$ contains a maximal element

Let m be an arbitrary element of the set M in question; there is at least one, by assumption. Then the algorithm

$$\textbf{for } r \in M \textbf{ do: if } r > m\textbf{, } m \leftarrow r\textbf{, od}$$

produces the maximal element, m, after finitely many steps.

Since a bounded subset of $\mathbb{Z}$ necessarily has only finitely many elements, it follows that *a nonempty bounded subset of $\mathbb{Z}$ contains a maximal element*. This latter claim is used several times in this book.

Also, note the corollary that *a bounded function into the integers takes on its maximal value*: its range then contains a maximal element and any preimage of that maximal element will do.

A nonempty bounded subset of $\mathbb{R}$ has a least upper bound

Let M be a subset of $\mathbb{R}$. Then, as the example of the open interval $(0\,..\,1)$ shows, such M need not have a maximal (or, rightmost) element. However, if the set

$$\{r \in \mathbb{R} : m \le r, \forall m \in M\}$$

of **upper bound**s for M is not empty, then this set has a smallest (or, leftmost) element. This smallest element is called the **least upper bound**, or the **supremum**, for M and is correspondingly denoted

$$\sup M.$$

The existence of a least upper bound for any real set M that has an upper bound is part of our understanding or definition of the set $\mathbb{R}$. What if M has no upper bound? Then some would say that $\sup M = \infty$. What if

M is empty? Then, offhand, $\sup M$ is not defined. On the other hand, since $M \subset N \implies \sup M \le \sup N$, some would, consistent with this, define $\sup\{\} := -\infty$.

One also considers the set

$$\{r \in \mathbb{R} : r \le m, \forall m \in M\}$$

of all **lower bound**s of the set M and understands that this set, if nonempty, has a largest (or, right-most) element. This element is called the **greatest lower bound**, or **infimum**, of M, and is denoted

$$\inf M.$$

What if M has no lower bound? Then some would say that $\inf M = -\infty$. In particular, $\inf \mathbb{R} = -\infty$. Also, some would set $\inf\{\} := \infty = \sup \mathbb{R}$.

Note that

$$-\sup M = \inf(-M).$$

Complex numbers

A complex number is of the form

$$z = a + \mathrm{i}b,$$

with a and b real numbers, called, respectively, the **real part of** z and the **imaginary part of** z, and i the **imaginary unit**, i.e.,

$$\mathrm{i} := \sqrt{-1}.$$

Actually, there are two complex numbers whose square is -1. We denote the other one by $-\mathrm{i}$. Be aware that, in parts of Engineering, the symbol j is used instead of i.

`MATLAB` works internally with (double precision) complex numbers. Both variables `i` and `j` in `MATLAB` are initialized to the value i.

One adds complex numbers by adding separately their real and imaginary parts. One multiplies two complex numbers by multiplying out and rearranging, mindful of the fact that $\mathrm{i}^2 = -1$. Thus,

$$(a + \mathrm{i}b)(c + \mathrm{i}d) = ac + a\mathrm{i}d + b\mathrm{i}c - bd = (ac - bd) + \mathrm{i}(ad + bc).$$

Note that both addition and multiplication of complex numbers is commutative. Further, the product of $z = a + \mathrm{i}b$ with its **complex conjugate**

$$\overline{z} := a - \mathrm{i}b$$

is the nonnegative number

$$z\overline{z} = a^2 + b^2,$$

and its (nonnegative) squareroot is called the **absolute value** or **modulus** of z and denoted by

$$|z| := \sqrt{z\overline{z}}.$$

For $z \neq 0$, we have $|z| \neq 0$, hence $\overline{z}/|z|^2 = a/|z|^2 - \mathrm{i}b/|z|^2$ is a well-defined complex number. It is the **reciprocal** of z since $z\overline{z}/|z|^2 = 1$, of use for *division* by z. Note that, for any two complex numbers z and ζ,

$$|z\zeta| = |z||\zeta|.$$

It is very useful to visualize complex numbers as points in the so called **complex plane**, i.e., to identify the complex number $a + \mathrm{i}b$ with the point (a, b) in $\mathbb{R}^2$. With this identification, its absolute value corresponds to the (Euclidean) distance of the corresponding point from the origin, and its direction $z/|z|$ is called its **sign**, and is denoted

$$\operatorname{signum} z := \begin{cases} z/|z|, & z \neq 0, \\ 0, & z = 0. \end{cases}$$

The sum of two complex numbers corresponds to the vector sum of their corresponding points. The product of two complex numbers is most easily visualized in terms of the **polar form**

$$z = a + \mathrm{i}b = r\exp(\mathrm{i}\varphi),$$

with $r \geq 0$, hence $r = |z|$ its *modulus*, $\exp(\mathrm{i}\varphi) = \operatorname{signum} z$ its *sign*, and $\varphi \in \mathbb{R}$ is called its **argument**. Indeed, for any real φ, $\exp(\mathrm{i}\varphi) = \cos(\varphi) + \mathrm{i}\sin(\varphi)$ has absolute value 1, and φ is the angle (in radians) that the vector (a, b) makes with the positive real axis. Note that, for $z \neq 0$, the argument, φ, is only defined up to a multiple of 2π, while, for $z = 0$, the argument is arbitrary. If now also $\zeta = \alpha + \mathrm{i}\beta = |\zeta|\exp(\mathrm{i}\psi)$, then, by the law of exponents,

$$z\zeta = |z|\exp(\mathrm{i}\varphi)|\zeta|\exp(\mathrm{i}\psi) = |z||\zeta|\exp(\mathrm{i}(\varphi + \psi)).$$

Thus, as already noted, the absolute value of the product is the product of the absolute values of the factors, while the argument of a product is the sum of the arguments of the factors.

For example, in as much as the argument of $\overline{z}$ is the negative of the argument of z, the argument of the product $z\overline{z}$ is necessarily 0. As another example, if $z = a + \mathrm{i}b$ is of modulus 1, then z lies on the unit circle in the complex plane, and so does any power z^k of z. In fact, then $z = \exp(\mathrm{i}\varphi)$ for some real number φ, and therefore $z^k = \exp(\mathrm{i}(k\varphi))$. Hence, the sequence $z^0, z^1, z^2, \ldots$ appears as a sequence of points on the unit circle, equally spaced around that circle, hence fails to converge to one point unless it is a constant sequence, i.e., unless $z = 1$, hence $\varphi = 0$.

(17.1) Lemma: Let z be a complex number of modulus 1. Then the sequence $z^0, z^1, z^2, \ldots$ of powers of z lies on the unit circle, but fails to converge except when $z = 1$.

Convergence of a scalar sequence

A subset Z of $\mathbb{C}$ is said to be **bounded** if it lies in some ball

$$B_r := \{z \in \mathbb{C} : |z| < r\}$$

of (finite) radius r. Equivalently, Z is bounded if, for some r, $|\zeta| < r$ for all $\zeta \in$ Z. In either case, the number r is called a **bound for** Z.

In particular, we say that the scalar sequence $(\zeta_1, \zeta_2, \ldots)$ is **bounded** if the set $\{\zeta_m : m \in \mathbb{N}\}$ is bounded. For example, the sequence $(1, 2, 3, \ldots)$ is not bounded.

(17.2) Lemma: The sequence $(\zeta^1, \zeta^2, \zeta^3, \ldots)$ is bounded if and only if $|\zeta| \le 1$. Here, ζ^k denotes the kth power of the scalar ζ.

Proof: Assume that $|\zeta| > 1$. I claim that, for all m,

$$|\zeta^m| - 1 > (|\zeta| - 1)m. \tag{17.3}$$

This is certainly true for $m = 1$. Assume it correct for $m = k$. Then

$$|\zeta^{k+1}| - 1 = (|\zeta^{k+1}| - |\zeta^k|) + (|\zeta^k| - 1).$$

The first term on the right-hand side gives

$$|\zeta^{k+1}| - |\zeta^k| = (|\zeta| - 1)|\zeta|^{k-1} > |\zeta| - 1,$$

since $|\zeta| > 1$, while, for the second term, $|\zeta^k| - 1 > (|\zeta| - 1)k$ by induction hypothesis. Consequently,

$$|\zeta^{k+1}| - 1 > (|\zeta| - 1) + (|\zeta| - 1)k = (|\zeta| - 1)(k + 1),$$

showing that (17.3) also holds for $m = k + 1$.

In particular, for any given c, choosing m to be any natural number bigger than $c/(|\zeta|-1)$, we have $|\zeta^m| > c$. We conclude that the sequence $(\zeta^1, \zeta^2, \zeta^3, \ldots)$ is unbounded when $|\zeta| > 1$.

Assume that $|\zeta| \le 1$. Then, for any m, $|\zeta^m| = |\zeta|^m \le 1^m = 1$, hence the sequence $(\zeta^1, \zeta^2, \zeta^3, \ldots)$ is not only bounded, it lies entirely in the **unit disk**

$$B_1^- := \{z \in \mathbb{C} : |z| \le 1\}.$$

A sequence $(\zeta_1, \zeta_2, \zeta_3, \ldots)$ of (real or complex) scalars is said to **converge to the scalar** ζ, in symbols:

$$\lim_{m\to\infty} \zeta_m = \zeta,$$

if, for all $\varepsilon > 0$, there is some m_ε so that, for all $m > m_\varepsilon$, $|\zeta - \zeta_m| < \varepsilon$.

Assuming without loss the scalars to be complex, we can profitably visualize this definition as saying the following: Whatever small circle $\{z \in \mathbb{C} : |z-\zeta| = \varepsilon\}$ of radius ε we draw around the point ζ, *all* the terms of the sequence except finitely many are inside that circle.

(17.4) Lemma: A convergent sequence is bounded.

Proof: If $\lim_{m\to\infty} \zeta_m = \zeta$, then there is some m_0 so that, for all $m > m_0$, $|\zeta - \zeta_m| < 1$. Therefore, for all m,

$$|\zeta_m| \le r := |\zeta| + 1 + \max\{|\zeta_k| : k = 1{:}m_0\}.$$

Note that r is indeed a well-defined nonnegative number, since a *finite* set of real numbers always has a largest element. □

(17.5) Lemma: The sequence $(\zeta^1, \zeta^2, \zeta^3, \ldots)$ is convergent if and only if either $|\zeta| < 1$ or else $\zeta = 1$. In the former case, $\lim_{m\to\infty} \zeta^m = 0$, while in the latter case $\lim_{m\to\infty} \zeta^m = 1$.

Proof: Since the sequence is not even bounded when $|\zeta| > 1$, it cannot be convergent in that case. We already noted that it cannot be convergent when $|\zeta| = 1$ unless $\zeta = 1$, and in that case $\zeta^m = 1$ for all m, hence also $\lim_{m\to\infty} \zeta^m = 1$.

This leaves the case $|\zeta| < 1$. Then either $|\zeta| = 0$, in which case $\zeta^m = 0$ for all m, hence also $\lim_{m\to\infty} \zeta^m = 0$. Else, $0 < |\zeta| < 1$, therefore $1/\zeta$ is a well-defined complex number of modulus greater than one, hence, as we showed earlier, $1/|\zeta^m| = |(1/\zeta)^m|$ grows monotonely to infinity as $m \to \infty$. But this says that $|\zeta^m|$ decreases monotonely to 0. In other words, $\lim_{m\to\infty} \zeta^m = 0$. □

A real continuous function on a compact set in $\mathbb{R}^n$ has a maximum

This basic result of Analysis is referred to in this book several times. Its proof goes beyond the scope of this book.

Here is the phrasing of this result that is most suited for this book.

(17.6) Theorem: An upper semicontinuous real-valued function f on a closed and bounded set M in $X := \mathbb{R}^n$ has a maximum, i.e.,

$$\sup f(M) = f(\mathbf{m})$$

for some $\mathbf{m} \in M$.

In particular, $\sup f(M) < \infty$.

A subset M of X is **closed** if $\mathbf{m} = \lim_n \mathbf{x}_n$ and $\mathbf{x}_n \in M$, all n, implies that $\mathbf{m} \in M$.

A subset M of X is **bounded** if $\sup \|M\| < \infty$.

A subset M of X is **compact** if it is closed and bounded.

A function $f : M \subset X \to \mathbb{R}$ is **continuous at m** if $\lim_n \mathbf{x}_n = \mathbf{m}$ implies that $\lim_n f(\mathbf{x}_n) = f(\mathbf{m})$. The function is **continuous** if it is continuous at every point of its domain.

A function $f : M \subset X \to \mathbb{R}$ is **upper semicontinuous at m** if $\lim_n \mathbf{x}_n = \mathbf{m}$ implies that $\lim_n f(\mathbf{x}_{\mu(n)}) \ge f(\mathbf{m})$ for every strictly increasing $\mu : \mathbb{N} \to \mathbb{N}$ for which the corresponding subsequence $n \mapsto f(\mathbf{x}_{\mu(n)})$ of $n \mapsto f(\mathbf{x}_n)$ is convergent.

Let $b := \sup f(M)$. Then, for each $r < b$, the set

$$M_r := \{\mathbf{m} \in M : f(\mathbf{m}) \ge r\}$$

is not empty. Also, M_r is closed, by the upper semicontinuity of f, and bounded. Also, M_r is decreasing as r increases. This implies (by the Heine-Borel Theorem) that $\cap_r M_r$ is not empty. But, for any $\mathbf{m} \in \cap_r M_r$, $f(\mathbf{m}) \ge r$ for all $r < b$, hence $f(\mathbf{m}) \ge b = \sup f(M)$, therefore $f(\mathbf{m}) = \sup f(M)$.

The theorem is also valid if X is any finite-dimensional normed vector space. For, with V any basis for X, we can write $f = gV^{-1}$ with $g := fV$ upper semicontinuous on $V^{-1}M$ and $\sup f(M) = \sup g(V^{-1}M) = g(h)$ for some $h \in V^{-1}M$, and so $m := Vh$ does the job for f.

Groups, rings, and fields

(17.7 A **semigroup** (F, op) is a set F and an **operation** op on F, i.e., a map $op : F \times F \to F : (f, g) \mapsto fg$ that is **associative**, meaning that

$$\forall f, g, h \in F, \quad (fg)h = f(gh).$$

The semigroup is **commutative** if

$$\forall f, g \in F, \quad fg = gf.$$

The prime (and essentially only) example of a semigroup is the set M^M of all maps on some set M, with map composition as the operation, or any of its subsets H that are **closed under** the operation, i.e., satisfy $HH := \{gh : g, h \in H\} \subset H$. M^M is commutative only if $\#M = 1$.

17.1 Prove that (F, op) is a semigroup if and only if the map $\Phi : F \to F^F : f \mapsto (g \mapsto fg)$ is a **semigroup homomorphism**, i.e., $\Phi(fg) = \Phi(f)\Phi(g)$, all $f, g \in F$. Give an example in which Φ fails to be 1-1.

(17.8) A **group** (G, op) is a semigroup (necessarily nonempty) whose operation is a **group operation**, meaning that, in addition to associativity, it has the following properties:

(g.1) there exists a **left neutral element** and a **right neutral element**, i.e., an $e_l, e_r \in G$ (necessarily $e_l = e_r$, hence unique, denoted by e and called the **neutral element**) such that

$$\forall g \in G, \quad e_l g = g = g e_r;$$

(g.2) every $g \in G$ has a **left inverse** and a **right inverse**, i.e., $f, h \in G$ so that $fg = e = gh$ (and, necessarily, these are unique and coincide, leading to the notation $f = g^{-1} = h$).

G is said to be 'a group under the operation op'.

If also H is a group, then a **homomorphism** from G to H is any map $\varphi : G \to H$ that 'respects the group structure', i.e., for which

$$\forall f, g \in G, \quad \varphi(fg) = \varphi(f)\varphi(g).$$

A group G is called **Abelian** if it is commutative, in which case the group operation is denoted $(f, g) \mapsto f + g$, the group inverse of $g \in G$ is denoted $-g$, and $f - g$ is short-hand for $f + (-g)$.

The prime example of a group is the collection of all invertible maps on some set, with map composition the group operation. The most important special case of these is $\mathbb{S}_n$, called the **symmetric group of order** n and consisting of all permutations of order n, i.e., of all invertible maps on $\underline{n} = \{1, 2, \ldots, n\}$. Any finite group G can be **represented by** a subgroup of $\mathbb{S}_n$ for some n in the sense that there is a **group monomorphism** $\varphi : G \to \mathbb{S}_n$, i.e., a 1-1 homomorphism from G to $\mathbb{S}_n$.

Here are some specific examples:

(i) $(\mathbb{Z}, +)$, i.e., the integers under addition; note that, for each $n \in \mathbb{Z}$, the map $n : \mathbb{Z} \to \mathbb{Z} : m \mapsto m + n$ is, indeed, invertible, with $-n : \mathbb{Z} \to \mathbb{Z} : m \mapsto m - n$ its inverse.

(ii) $(\mathbb{Q}\backslash 0, *)$, i.e., the nonzero rationals under multiplication; note that, for each $q \in \mathbb{Q}\backslash 0$, the map $q : \mathbb{Q}\backslash 0 \to \mathbb{Q}\backslash 0 : p \mapsto pq$ is, indeed, invertible, with $q^{-1} : \mathbb{Q}\backslash 0 \to \mathbb{Q}\backslash 0 : p \mapsto p/q$ its inverse.

(iii) The collection of all rigid motions that carry an equilateral triangle to itself. It can be thought of as $\mathbb{S}_3$ since each such motion, being rigid, must permute the vertices and is completely determined once we know what it does to the vertices.

17.2 Prove that, for $M = \{1, 2\}$, the semigroup M^M is not commutative.

17.3 Verify all the parenthetical claims made in the above definition of a group.

17.4 Give an example of a nonabelian group.

(17.9) A **ring** $R = (R, +, *)$ is a set R (necessarily nonempty) with two operations, $(f, g) \mapsto f + g$ and $(f, g) \mapsto f * g =: fg$, called addition and multiplication respectively, such that

(r.1) $(R, +)$ is an Abelian group, with neutral element usually denoted 0;

(r.2) $(R, *)$ is a semigroup;

(r.3) (distributive laws): for every $f \in R$, the maps $R \to R : g \mapsto fg$ and $R \to R : g \mapsto gf$ are homomorphisms of the group $(R, +)$, i.e., $f(g + h) = fg + fh$ and $(g + h)f = gf + hf$.

A **field** is a ring $(R, +, *)$ for which $(R\backslash 0, *)$ is a group.

If multiplication in the ring R is commutative, i.e., $fg = gf$ for all $f, g \in R$, then R is called commutative.

If the ring R has a neutral element for multiplication, i.e., an element e so that $eg = g = ge$ for all $g \neq 0$, then it has exactly one such, and it is usually denoted by 1. In that case, R is called a **ring with identity**. Any field is a ring with identity.

Both $\mathbb{R}$ and $\mathbb{C}$ are commutative fields. The prime example of a ring is the set $\Pi(\mathbb{F}^d)$ of all polynomials in d (real or complex) variables with (real or complex) coefficients, with pointwise addition and multiplication the ring operations. It is a commutative ring with identity. It has given the major impetus to the study of (two-sided) **ideals**, i.e., of nonempty subsets S of a ring R closed under addition, and containing both SR and RS, i.e., closed also under left or right multiplication by any element of the ring. This makes S a subring of R, i.e., a ring in its own right, but not all subrings are ideals. Both $\{0\}$ and R are trivially ideals. Any other ideal in R is called **proper**.

Let R be a commutative ring. Then the set

$$[s_1, \ldots, s_r](R^r) = \{s_1 g_1 + \cdots + s_r g_r : (g_1, \ldots, g_r) \in R^r)\}$$

is an ideal, the ideal **generated by** $(s_1, \ldots, s_r)$. Such an ideal is called **finitely generated**. A ring R is called **Noetherian** if all its ideals are finitely generated. **Hilbert's Basis Theorem** famously states that $\Pi(\mathbb{F}^d)$ is Noetherian.

17.5 Verify that, for any $s_1, \ldots, s_n$ in the commutative ring R, $[s_1, \ldots, s_n](R^n)$ is an ideal.

The ring of univariate polynomials

$\Pi = \Pi(\mathbb{F})$ is, by definition, the set of univariate polynomials, i.e., the collection of all maps

$$p : \mathbb{F} \to \mathbb{F} : z \mapsto \widehat{p}_0 + \widehat{p}_1 z + \widehat{p}_2 z^2 + \cdots + \widehat{p}_d z^d,$$

with $\widehat{p}_0, \ldots, \widehat{p}_d \in \mathbb{F}$ and some $d \in \mathbb{Z}_+$. If $\widehat{p}_d \neq 0$, then d is the **degree** of p, i.e.,

$$d = \deg p := \max\{j : \widehat{p}_j \neq 0\}.$$

This leaves the degree of the zero polynomial, $0 : \mathbb{F} \to \mathbb{F} : z \mapsto 0$, undefined. It is customary to set

$$\deg 0 := -1.$$

As already mentioned, Π is a ring under pointwise addition and multiplication. More than that, Π is a **principal ideal domain**, meaning that any of its proper ideals is generated by just one element. Indeed, if I is a proper ideal, then it contains an element p of smallest possible nonnegative degree and, since $I \neq \Pi$, this degree is positive. If f is any element of Π, then, by the Euclidean algorithm (see page 281), we can find $q, r \in \Pi$ so that $f = qp + r$ and $\deg r < \deg p$. If now $f \in I$, then also $r = f - qp \in I$ and $\deg r < \deg p$ hence, by the minimality of $\deg p$, r must be 0. In other words,

$$I = \Pi p := \{qp : q \in \Pi\}.$$

It follows that Π is a **unique factorization domain**, which, in the case of the ring Π, means that each factorization of a polynomial into a product of irreducible polynomials is unique, up to reordering of the factors and multiplication by a scalar.

17.6* Prove that the ideal generated by the univariate polynomials $p_1, \ldots, p_r$ is generated by their greatest common divisor.

To be sure, already $\Pi(\mathbb{F}^2)$ fails to be a principal ideal domain.

It is simple algebra (see, e.g., the discussion of Horner's method on page 280) that the set

$$Z(p) := \{z \in \mathbb{F} : p(z) = 0\}$$

of zeros of $p \in \Pi$ contains at most $\deg p$ elements. It is the **Fundamental Theorem of Algebra** that $\#Z(p) = \deg p$, counting multiplicities, in case $\mathbb{F} = \mathbb{C}$. More explicitly, this theorem says that, with $d := \deg p$,

$$p = c(\cdot - z_1) \cdots (\cdot - z_d)$$

for some nonzero constant c and some $\mathbf{z} \in \mathbb{C}^d$.

It is in this sense that $\mathbb{C}$ is said to be **algebraically closed** while $\mathbb{R}$ is not. E.g., the real polynomial $()^2 + 1$ has no real zeros. It is remarkable that, by adjoining one, hence the other, of the 'imaginary' zeros of $()^2 + 1$, i.e., $\mathrm{i} = \sqrt{-1}$, appropriately to $\mathbb{R}$, i.e., by forming $\mathbb{C} = \mathbb{R} + \mathrm{i}\mathbb{R}$, we obtain enough additional scalars so that now, even if we consider polynomials with complex coefficients, all nonconstant polynomials have a full complement of zeros (counting multiplicities).

Horner, or: How to divide a polynomial by a linear factor

Recall that, given the polynomial p and one of its roots, μ, the polynomial $q := p/(\cdot - \mu)$ can be constructed by **synthetic division**. This process is also known as **nested multiplication** or **Horner's scheme** as it is used, more generally, to evaluate a polynomial efficiently. Here are the details, for a polynomial of degree ≤ 3.

If $p(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3$, and z is any scalar, then

$$p(z) = \underbrace{a_0 + z\,(\underbrace{a_1 + z\,(\underbrace{a_2 + z\,\underbrace{a_3}_{=:b_3}}_{=:b_2}))}_{=a_1+zb_2=:b_1}}_{a_0+zb_1=:b_0}\,.$$

In other words, we write such a polynomial in **nested form** and then evaluate from the inside out. Each step is of the form

$$b_j := a_j + z b_{j+1}; \qquad (17.10)$$

it involves one multiplication and one addition. The last number calculated is b_0; it is the value of p at z. There are 3 such steps for our cubic polynomial (the definition $b_3 := a_3$ requires no calculation!). So, for a polynomial of degree n, we would use n multiplications and n additions.

Now, not only is b_0 of interest, since it equals $p(z)$; the other b_j are also useful since

$$p(t) = b_0 + (t - z)(b_1 + b_2 t + b_3 t^2).$$

We verify this by multiplying out and rearranging terms according to powers of t. This gives

$$\begin{array}{rcccccccc} b_0 + (t-z)(b_1 + b_2 t + b_3 t^2) & = & b_0 & + & b_1 t & + & b_2 t^2 & + & b_3 t^3 \\ & & -zb_1 & - & zb_2 t & - & zb_3 t^2 & & \\ & = & b_0 - zb_1 & + & (b_1 - zb_2)t & + & (b_2 - zb_3)t^2 & + & b_3 t^3 \\ & = & a_0 & + & a_1 t & + & a_2 t^2 & + & a_3 t^3 \end{array}$$

The last equality holds since, by (17.10),

$$b_j - zb_{j+1} = a_j$$

for $j < 3$ while $b_3 = a_3$ by definition.

(17.11) Nested Multiplication (a.k.a. Horner): To evaluate the polynomial $p(t) = a_0 + a_1 t + \cdots + a_k t^k$ at the point z, compute the sequence $(b_0, b_1, \ldots, b_k)$ by the prescription

$$b_j := \begin{cases} a_j & \text{if } j = k; \\ a_j + zb_{j+1} & \text{if } j < k. \end{cases}$$

Then $p(t) = b_0 + (t - z)q(t)$, with

$$q(t) := b_1 + b_2 t + \cdots + b_k t^{k-1}.$$

In particular, if z is a root of p (hence $b_0 = 0$), then

$$q(t) = p(t)/(t - z).$$

Since $p(t) = (t - z)q(t)$, it follows that $\deg q < \deg p$. This provides another proof (see (3.38)) for the *easy* part of the *Fundamental Theorem of Algebra*, namely that a polynomial of degree k has at most k roots.

The Euclidean Algorithm

Horner's method is a special case of the **Euclidean Algorithm** which constructs, for given polynomials f and p with $\deg p > 0$, (unique) polynomials q and r with $\deg r < \deg p$ so that

$$f = pq + r.$$

For variety, here is a nonstandard discussion of this algorithm which uses the fact that square triangular matrices with nonzero diagonal entries are invertible.

Assume that

$$p(t) = \widehat{p}_0 + \widehat{p}_1 t + \cdots + \widehat{p}_d t^d, \quad \widehat{p}_d \neq 0, \quad d > 0,$$

and

$$f(t) = \widehat{f}_0 + \widehat{f}_1 t + \cdots + \widehat{f}_n t^n$$

for some $n \geq d$, there being nothing to prove otherwise. Then we seek a polynomial

$$q(t) = \widehat{q}_0 + \widehat{q}_1 t + \cdots + \widehat{q}_{n-d} t^{n-d}$$

for which

$$r := f - pq$$

has degree $< d$. With $r(t) =: \widehat{r}_0 + \widehat{r}_1 t + \cdots + \widehat{r}_n t^n$, this requires $\widehat{r}_j = 0$ for $j \geq d$. Since $r = f - pq$, we compute $\widehat{r}_j = \widehat{f}_j - \sum_{i=j-d}^{n-d} \widehat{p}_{j-i}\widehat{q}_i$. Therefore, we require that $\sum_{i=j-d}^{n-d} \widehat{p}_{j-i}\widehat{q}_i = \widehat{f}_j$ for $j = d, \ldots, n$, and so obtain the square upper triangular linear system

$$\begin{array}{ccccccccccc}
\widehat{p}_d\widehat{q}_0 & + & \widehat{p}_{d-1}\widehat{q}_1 & + & \widehat{p}_{d-2}\widehat{q}_2 & + & \cdots & + & \widehat{p}_0\widehat{q}_{n-d} & = & \widehat{f}_d \\
 & & \widehat{p}_d\widehat{q}_1 & + & \widehat{p}_{d-1}\widehat{q}_2 & + & \cdots & + & \widehat{p}_1\widehat{q}_{n-d} & = & \widehat{f}_{d+1} \\
 & & & \ddots & & & & & & & \vdots \\
 & & & & \ddots & & & & & & \vdots \\
 & & & & & & \widehat{p}_d\widehat{q}_{n-d-1} & + & \widehat{p}_{d-1}\widehat{q}_{n-d} & = & \widehat{f}_{n-1} \\
 & & & & & & & & \widehat{p}_d\widehat{q}_{n-d} & = & \widehat{f}_n
\end{array}$$

for the unknown coefficients $\widehat{q}_0, \ldots, \widehat{q}_{n-d}$ which can be uniquely solved by back substitution since the diagonal entries of its coefficient matrix all equal $\widehat{p}_d \neq 0$.

18 List of Notation

try to build up a list of definitions by listing up all lines with :=

Rough index for this book