# **Applied Linear Algebra**

Carl de Boor

draft 20dec94

table of contents

## 0. lists, assignments, 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 relative 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.

### Lists, especially $n$-vectors

#### Lists.

**Definition:** A **list** or, more precisely, an $n$**-list**

$$f = (f_i)_{i=1}^n = (f_i : i = 1, \ldots, n) = (f_1, f_2, \ldots, f_n)$$

associates with each number $i$ in its **domain**

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

some **term** or **item** $f_i$.

The set

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

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

The number

$$\#f$$

denotes the **length** of the list; in particular,

$$\operatorname{dom} f = \{1, \ldots, \#f\},$$

i.e., $\#f = n$ if and only if $f$ is an $n$-list.

Here is an example of a list, the shopping list from last Saturday:

milk
cream
cheese
tomatoes
Roquefort
sweet butter
shoe laces

A list has on it a first item, then a second item, then a third item, and so on all the way down to the last item. There are other names used for an $n$-list: (finite) **sequence** or, more precisely, $n$**-sequence**; also, $n$**-tupel** (though that raises the question of exactly what a 'tupel' might be); also, **one-dimensional array** (of length $n$). In these notes, '$n$-list' and '$n$-sequence' are used interchangeably.

There is often no good reason to write such a list in a vertical column; in fact, that takes up valuable space on the printed page. Writing the items one after the other on a line is more economical with space and is also ok as long as we indicate somehow, e.g., by a comma, where one item ends and the next begins. For the earlier example, writing milk cream cheese tomatoes Roquefort sweet butter shoe laces

leaves some doubt, while

milk, cream, cheese, tomatoes, Roquefort, sweet butter, shoe laces

does not. We will usually specify a list $f$ in this way, i.e., in the form

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

with the individual items $f_i$ of the list written out explicitly. However, if we have some **rule** or formula for calculating $f_i$ from $i$, then we will write

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

For example,

$$f : \underline{n} : i \mapsto n - i$$

is the list $(n, n-1, \ldots, 1)$ in which the first $n$ natural numbers appear in reverse order.

We will have use for the **empty list**, i.e., the list

$$()$$

To be sure, $\operatorname{dom}() = \{\}$, i.e., the **empty set**, and, correspondingly, $\#() = 0$.

As another example, the **cartesian product**

$$X \times Y := \{(x, y) : x \in X, y \in Y\}$$

of the set $X$ with the set $Y$ is, by definition, the collection of all 2-lists with the first item from $X$ and the second item from $Y$.

If both $f$ and $g$ are lists, then

$$(f, g)$$

is the list in which we first list the items of $f$ and then the items of $g$. It is called the **concatenation of** $f$ **with** $g$. To be sure,

$$\#(f, g) = \#f + \#g = \#(g, f),$$

but, usually, $(f, g) \neq (g, f)$.

The shopping list given above is special in that no two items on it are the same. Such a list is called **1-1** or **one-one**. One would hope that the weekly shopping list is 1-1. Also, the standard way to determine the **cardinality**

$$\#S$$

of a (finite) set $S$ is to make up a 1-1 list $s$ with the property that $\operatorname{ran} s = S$, since for any such list $s$, $\#s = \#S$.

On the other hand, many lists in practical life are not 1-1. For example, a hotel will keep an occupancy list $o$ in which $o_i$ is, say, `yes` or `no`, depending on whether or not room $i$ is occupied. It makes little sense to wonder whether the occupancy list is 1-1.

In these notes, we will deal with *vectors*, i.e., lists of *numbers*, such as $(1, 3.14, -14)$ or $(3, 3, 1)$ .

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

In `MATLAB`, there are (at least) two ways to specify an $n$-vector, namely as a one-row matrix, or as a one-column matrix. For example, we can record the 3-vector $x = (1.3, 3.14, -14)$ as the one-row matrix

```
x_as_row = [1.3,3.14,-14];
```

or as the one-column matrix

```
x_as_col = [1.3;3.14;-14];
```

Which way is preferable? For various reasons, it is customary to represent $n$-vectors by one-*column* matrices, and this custom is followed throughout these notes.

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

```
x_as_col = [1.3
     3.14
   -14];
```

`MATLAB` can also handle lists of *characters*. A list of characters is called a **string**. It is specified in `MATLAB` by writing the items of the list *without separating commas* and enclosed in single quotes. For example, the list

s := (I, ,a,m, ,n,o,t, ,a, ,l,i,s,t,!)

would be specified by the `MATLAB` statement

```
s = 'I am not a list!';
```

after which the statement `s(10)` would get the response

```
ans = a
```

since the letter `a` is indeed the 10th item in the list `s`. Note that, now, blanks are significant, i.e., the two lists

`' tabula rasa'` and `'tabula  rasa'`

are quite different.

The empty string is specified as expected, i.e., by

```
empty_s = '';
```

If both `s` and `t` are strings, then

```
[s,t]
```

is their concatenation.

If `s` is any of these lists in `MATLAB`, whether a vector or a string, then its length is obtained by `length(s)`. For example, `length('')` gets the answer `0`.

Incidentally, while the two *lists*, $(\ ,t,a,b,u,l,a,\ ,r,a,s,a)$ and $(t,a,b,u,l,a,\ ,\ ,r,a,s,a)$, are indeed different, they have the same range, namely the *set*

$$\{t,a,b,u,l,a,\ ,r,a,s\} = \{a,b,l,r,s,t,u,\ \} = \{l,a,s,t,\ ,r,u,b\}.$$

Also,

$$\#(\ ,t,a,b,u,l,a,\ ,r,a,s,a) = 12 \neq 8 = \#\,\mathrm{ran}(\ ,t,a,b,u,l,a,\ ,r,a,s,a).$$

## Assignments

**Definition:** An **assignment** or, more precisely, an **$I$-assignment**

$$f = (f_i)_{i\in I} = (f_i : i \in I)$$

associates with each element $i$ in its **domain**

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

some **term** or **item** 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 $i \in \operatorname{dom} f = \operatorname{dom} g$.

An assignment is just like a list, except that its domain is not restricted to be of the form $\underline{n}$ for some $n$, but can be an arbitrary set. Its domain is also called its **index set**. Very confusingly, many mathematicians call an assignment an *indexed set*, even though it is most certainly not a set. The term **family** is also used; however it, too, smacks too much of a set or collection.

An **assignment on** $I$ is, by definition, an assignment with domain or index set $I$. We call the assignment $f$ **1-1** if $f_i = f_j \Longrightarrow i = j$.

The simplest assignment is the **empty assignment**, (), which, not surprisingly, looks just like the empty list since, after all, its index set is the empty set. Note that the empty assignment is 1-1 (why??).

If $\operatorname{dom} f$ is finite, say $\#\operatorname{dom} f = n$, then we could always describe $f$ by listing the $n$ pairs $(i, f_i)$, $i \in I$, 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 right on the map (not bothering to record the value 0 of no death). 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, 'next' to each element of the domain, its assigned item or value.

This is routinely done for matrices, another important object in these notes.

**Matrices**

**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 $n$.

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

In other words, a matrix has a rectangular domain. Correspondingly, it is customary to display such an $m \times n$-matrix $A$ as a rectangle of items:

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

rather than as a list of pairs. 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.

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

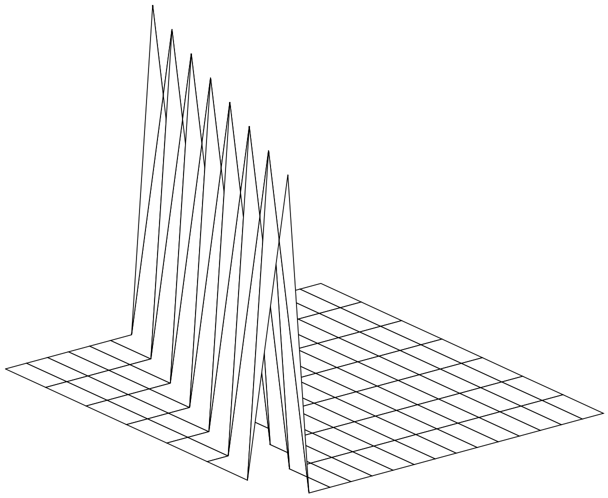
$$A_{i,j} =: A(i,j)$$

corresponding to the index $(i,j)$ is also called the $(i,j)$**-entry** of $A$. The list $A(i,:) := (A(i,j) : j \in \underline{n})$ is called the $i$**th row of** $A$, the list $A(:,j) := (A(i,j) : i \in \underline{m})$ is called the $j$**th column of** $A$ , and the list $(A(i,i) : 1 \le i \le \min\{m,n\})$ is called the **(main) diagonal of** $A$. By definition, $A^T$ denotes the **transpose** of the matrix $A$, i.e., the $n \times m$-matrix whose $(i,j)$-entry is $A(j,i)$, all $i,j$. Because of its importance in the later parts of these notes, we usually use the **conjugate transpose** $A' := \overline{A}^T$ whose $(i,j)$-entry is the scalar $\overline{A(j,i)}$, with $\overline{\alpha}$ the complex conjugate of the scalar $\alpha$.

The notation $A(i,:)$ for the $i$th row and $A(:,j)$ for the $j$th column of the matrix $A$ is taken from `MATLAB`. 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 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* $\mathtt{A}^T$ in the contrary case, you must use the notation `A.'` which is strange since there is nothing *pointwise* about this operation.

The above-mentioned need to look at displays of matrices sideways is evident when we use `MATLAB` to plot a matrix. Here, for example, is the 'picture' of the $8 \times 16$-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.



The rectangular identity matrix `eye(8,16)` as plotted in `MATLAB`

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'

```
[[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 `MATLAB` only knows matrices.

□

## 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 from left to right. Nevertheless, in these notes, it will always be done by columns, i.e., the sequence $(f_1, f_2, \ldots, f_n)$ of $m$-vectors will be associated with the $m \times n$-matrix $A$ whose $j$th column is $f_j$, all $j$. We write this fact in this way:

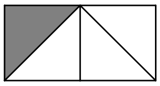
$$A = [f_1, f_2, \ldots, f_n].$$

While this makes it ok to denote by $\#A$ the number of columns of $A$, we will be usually be cautious and use

$$\#_c A, \quad \#_r A$$

to denote the number of columns, respectively, rows of the matrix $A$. Thus

$$\#_r A' = \#_c A = \#A = \#_r A^T.$$



A simple triangulation.

For example, a **triangulation** in the 'plane', i.e., in $\mathbb{R}^2$, is specified by a list $p = (p_1, p_2, \ldots, p_n)$ of points and a list $t = (t_1, t_2, \ldots, t_r)$ of triangles. To be precise, each $p_j$ is an element of $\mathbb{R}^2$, i.e., a 2-vector. Each $t_j$, on the other hand, is a list of three items, specifying the three points which make up this particular triangle. The list $t_j$ could, of course, list the three points explicitly. It is more efficient, though, simply to list the three indices under which these three points can be found in the point list $p$. Thus, the triangulation shown in the above figure is described by the following two matrices, the first a list of the vertices, and the second a list of its triangles (the second triangle in this list is shaded in the figure):

$$p := \begin{bmatrix} 0 & 1 & 2 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \qquad t := \begin{bmatrix} 1 & 1 & 2 & 3 \\ 2 & 5 & 3 & 6 \\ 5 & 4 & 5 & 5 \end{bmatrix}$$

If we wanted to enter the list of points in MATLAB, then it is certainly easier to type each point in a row by itself. Thus one might type in the above matrix $p$ in this way:

```
p = [0 0
1 0
2 0
0 1
1 1
2 1]';
```

i.e., type one point per row, then change to the transpose. On the other hand, for this example, one could also do it in this systematic way:

```
p = [0:2 0:2; zeros(1,3) ones(1,3)];
```

or in this systematic way:

```
p = [ [0:2; zeros(1,3)] [0:2; ones(1,3)] ];
```

□

A more picturesque example is provided by a **polyhedron**. Take, for example, the regular **octahedron**. Its vertex set consists of the three unit vectors and their negatives, i.e.:

```
vs = [1 0 0; -1 0 0; 0 1 0; 0 -1 0; 0 0 1; 0 0 -1]';
```

Each face is a triangle, and we specify it here by giving the index, in the vertex array `vs`, of each of its three vertices:

```
ff = [2 4 5; 2 5 3; 4 1 5; 2 6 4]';
bf = [2 3 6; 6 1 4; 6 3 1; 5 1 3]';
```

The faces have been organized into front faces and back faces, in anticipation of the plotting about to be done, in which we want to plot the front faces strongly, but only lightly indicate the back faces. Be sure to look for specific faces in the figure below, in which the six vertices are numbered as in `vs`. E.g., the first front face, specified by the first column of `ff`, involves the vertices numbered `2, 4, 5`; it is the face we are viewing head-on.

First, we set the frame:

```
plot3([-1 1], [0 0], [0 0],'.')
hold on; axis off
```

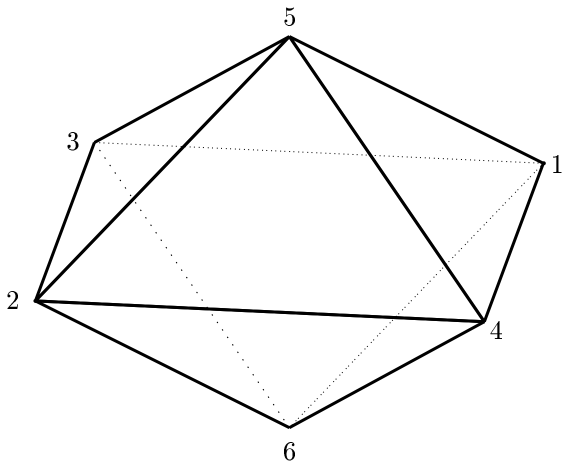
Then we plot the back-faces first (using `r=[1 2 3 1]` to make sure that we plot closed triangles):

```
r = [1 2 3 1];
for j=1:4
   line(vs(bf(j,r),1),vs(bf(j,r),2),vs(bf(j,r),3),'lines',':');
end
```

Then, finally, we plot the front faces and finish the picture:

```
for j=1:4
   line(vs(ff(j,r),1),vs(ff(j,r),2),vs(ff(j,r),3),'linew',1.5);
end
hold off
```

Here is the resulting figure (obtained by the command `print figoctah.eps -deps2` which generates a `postscript` file). I have labelled all the vertices by their index in the vertex list `vs`.




The regular octahedron.

□

## Maps

**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 $X$,

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

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

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

Also, for any $U \subseteq Y$, 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

$$Y^X \quad \text{or} \quad (X \to Y).$$

Names other than **map** are in use, such as **mapping**, **operator**, **morphism**, **transformation** etc., all longer than “map”. A scalar-valued map is often called a **function**. Somewhat confusingly, many mathematicians use the term ‘range’ for what we have called here ‘target’; the same mathematicians use the term **image** for what we 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 on $X$ gives rise to *many* maps, one for each $Y$ that contains $\operatorname{ran} f$, by the prescription $X \to Y : x \mapsto f_x$. We call this the **map into $Y$ given by the assignment** $f$ and denote it by

$$f|^Y$$

if need be.

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?

### 1-1 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 question of whether range and target coincide. This situation is so important that it has been given a special name:

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

We call a map 1-1 if its corresponding assignment is 1-1:

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

These two map properties will play a major role throughout these notes. – 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)**.

We usually don't think of $n$-vectors as giving rise to a map, except when their range is again $\underline{n}$. Any map from $\underline{n}$ to $\underline{n}$ which is 1-1 and onto is called a **permutation of order** $n$, since its list is a reordering of the first $n$ integers. Thus $(3, 2, 1)$ or $(3, 1, 2)$ are permutations of order 3 while the map into $\mathbf{3}$ given by the 3-vector $(3, 3, 1)$ is not a permutation, as it is neither 1-1 nor onto.

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 $\underline{n}$). But we only have to check one of these, and if one holds, then the other one must hold, too. This is also called the **pigeonhole principle**.

**(0.1) Proposition (pigeon hole):** If $f : X \to Y$, with $X$ a finite set, and $Y$ has the same number of elements as $X$, then $f$ is 1-1 if and only if $f$ is onto.

**Proof:** The proof relies on the fact (already mentioned earlier) that, for two *finite* sets $M$ and $N$, $M \subset N$ implies $\#M \le \#N$ with equality if and only if $M = N$. In particular, since $f(X) \subset Y$, we have $\#f(X) \le \#Y$ with equality if and only if $f(X) = Y$, i.e., if and only if $f$ is onto. On the other hand, $f$ is 1-1 if and only if $\#X = \#f(X)$, i.e., if and only if $\#Y = \#f(X)$ (since we assumed that $\#X = \#Y$). □

If $X$ is not finite, then, even if $f : X \to X$, $f$ may be 1-1 without being onto, or onto without being 1-1. For example, consider the **right shift**

$$r : \mathbb{N} \to \mathbb{N} : n \mapsto n + 1. \tag{0.2}$$

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**

$$l : \mathbb{N} \to \mathbb{N} : n \mapsto \max\{n - 1, 1\} \tag{0.3}$$

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 pigeon hole principle continues to hold for certain infinite $X$ and $Y$ and certain special maps $f : X \to Y$. Specifically, according to (3.16)Corollary, if $X$ and $Y$ are *linear spaces* of 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 (2.13)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 0.*

## Some examples

The very act of *counting* is a matter of setting up a map $c : S \to \mathbb{N}$, from the set $S$ to be counted into the natural numbers, in such a way that every natural number appears at most once (**1-1**) and so that, for every natural number $n$ in the range, also every natural number smaller than $n$ is in the range. The biggest integer in the range is the **cardinality** of the set to be counted, i.e., the number $\#S$. As mentioned earlier, we usually set up this map in the inverse way, by making in effect a list, pointing to a first item and saying 'one', then to another item and saying 'two', then to another item and saying 'three', and so forth, until we have exhausted the set; the last natural number mentioned is the cardinality of the set. It is a nice exercise to prove to yourself the 'obvious' fact that this cardinality is independent of how we set up the map, as long as it has those two properties mentioned.

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

$$\text{name} \quad | \quad \text{grade}$$

Here at UW, the target of the grade map is the set

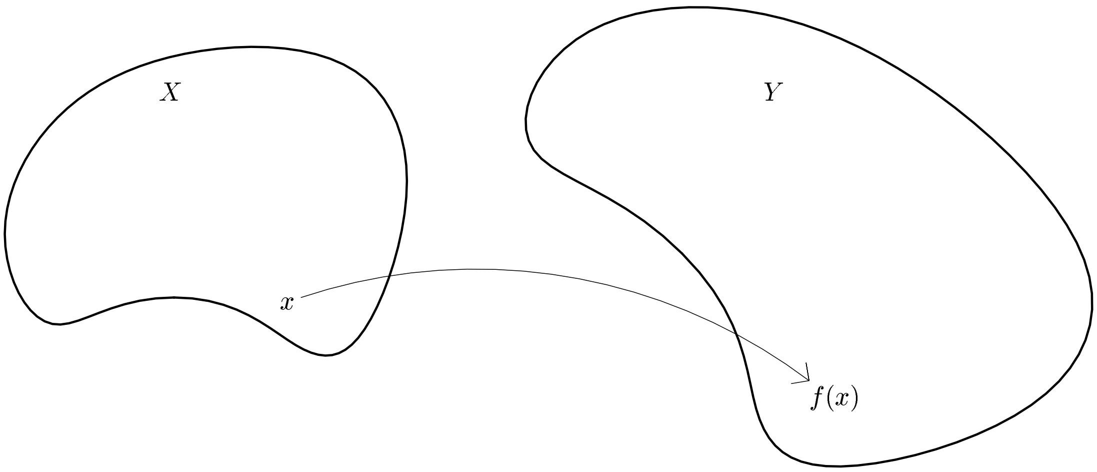
$$\{\text{A, AB, B, BC, C, D, F, I}\},$$

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 which is onto is automatically also 1-1, and a grade map which is 1-1 is automatically also onto.

There are many maps in your life which 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.

## Maps and their graphs

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.

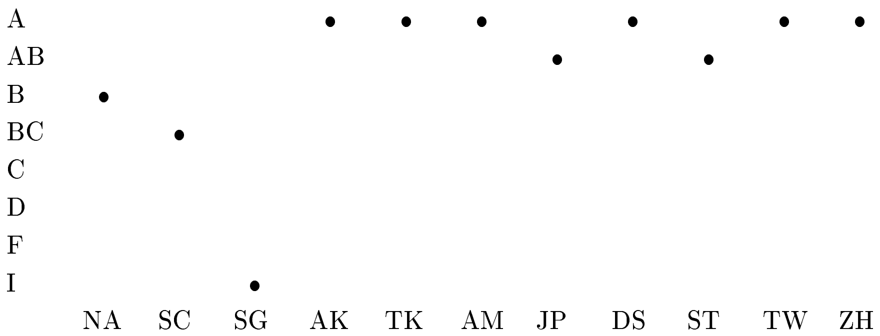



One way to visualize the map $f : X \to Y : x \mapsto f(x)$.

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

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

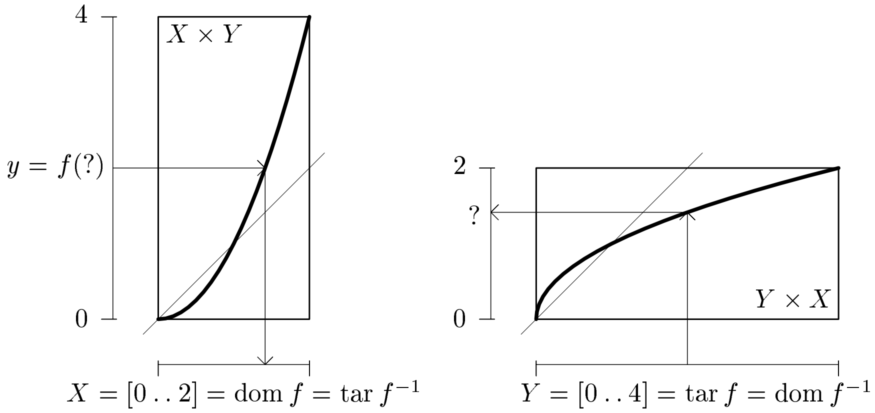
Here, one draws the cartesian product $X \times Y$ of its domain and target, and marks each pair $(x, f(x))$. Here, for example, is the grade map $G$ for a graduate course I taught recently. I abbreviated the students' names, to protect the innocent.




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

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

whose graph is shown in the next figure.




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

## Invertibility

The graph of a map $f$ helps us solve the standard 'computational' problem involving maps, namely the problem of finding an $x \in X$ which 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 $G^{-1}\{\mathrm{AB}\} = \{\mathrm{JP, ST}\}$, 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 $f(x) = y$.

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

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 $f^{-1}$.

For any 1-1 assignment $f$, the map $f|^{\operatorname{ran} f}$ 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 earlier figure 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 (0.2) is 1-1, but not onto, while the left shift (0.3) 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.

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 $f : X \to X$ which is 1-1 *or* onto is automatically invertible.

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**:

**Definition:** The **composition** $f \circ g$ of two maps $f : X \to Y$ and $g : U \to W \subseteq X$ 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 $x$ is an $n$-vector and $y$ is an $m$-vector with $\operatorname{ran} y \subset \underline{n} = \{1, \ldots, n\}$, then

$$z := x_y := x \circ y = (x_{y_1}, \ldots, x_{y_m}).$$

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

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

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

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

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

For example, `A(m:-1:1,:)` is the matrix obtained from `A` by reversing the order of the rows. `A(:,2:2:n)` 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`.

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 $(\mathtt{i},\mathtt{j}) \in \operatorname{dom}\mathtt{D}$, 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 2 3 4]` and `b=[2 2 2]`, then the statement `a(b)=[1 2 3]` changes `a` to `[1 3 3 4]`.

□

As a first use of map composition, here is the standared test for a map being onto or 1-1.

If $fg$ is onto, then $f$ is onto; if $fg$ is 1-1, then $g$ is 1-1.

**Proof:** Since $\operatorname{ran}(fg) \subseteq \operatorname{ran} f \subseteq \operatorname{tar} f = \operatorname{tar} fg$, $fg$ 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 (0.3) with the right shift (0.2) is the identity, hence $l$ is onto and $r$ is 1-1 (as observed earlier).

Remark. The only practical way to check whether a given $g$ is 1-1 is to come up with an $f$ so that $fg$ is 'obviously' 1-1, e.g., invertible. The only practical way to check whether a given $f$ is onto, is to come up with a $g$ so that $fg$ is 'obviously' onto, e.g., invertible.

**Definition:** If $f \in Y^X$ and $g \in X^Y$ and $fg = \mathrm{id}$, then $f$ 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 $fg = \mathrm{id}$, keep in mind that being onto is a property that "happens" in the target of a map while being 1-1 is a property that "happens" in the domain of a map.

Now we consider the converse statements.

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, $gf = \mathrm{id}$ follows. □

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".

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 $f^{-1}$ is also invertible, and $(f^{-1})^{-1} = f$.

**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 and a right inverse, then it must be both 1-1 and onto, hence invertible. Also (using the associativity of map composition),

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

□

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 $\operatorname{ran} gf = g(\operatorname{ran} f) \subseteq \operatorname{ran} g$. 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**.

**(0.4) Proposition:** If $f : X \to Y$ and $fg = \mathrm{id}_Y$, then $gf$ is a projector, with range $\operatorname{ran} g$.

For example, the composition $lr$ of the left shift (0.3) with the right shift (0.2) 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 and onto, hence invertible, then the $n$-vector `cinv` giving its inverse can be obtained with the command

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

Here, the whole purpose of the first command is to make certain that we start with 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 way to construct `cinv` is with the help of the command `sort`, as follows:

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

For (check the `MATLAB` manual!), 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:

`d(i)=c(cinv(i))`, `i=` $1, \ldots, n$.

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.

□

### The inversion of maps and the minimization of functions

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: for $f : X \to Y$ and $y \in Y$, solve

$$f(?) = y. \tag{0.5}$$

When we try to solve (0.5), 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 (0.5) *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 $A? = b$, with $A$ a *linear* map. This is so because linear 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. 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:

**(1.1) Definition.** To say that $X$ is a **linear space** (=:**ls**) (of **vector**s), or a **vector space**, over the field $\mathbb{F}$ (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., addition

- (a.1) is **associative**: $x + (y + z) = (x + y) + z$;
- (a.2) is **commutative**: $x + y = y + x$;
- (a.3) has **neutral** element: $\exists\{0\}\ \forall\{x\}\ x + 0 = x$;
- (a.4) has **inverse**: $\forall\{x\}\ \exists\{y\}\ x + y = 0$.

(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**: $1x = x$.

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 $x + y = 0$. I will denote the neutral element in $X$ by the same symbol, 0, used for the zero scalar.

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 linear 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:

**(1.2) 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** $\alpha f$ 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 linear 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** $\mathbb{F}^n$ whose elements (vectors) we call $n$-vectors.

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

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

(iv) $T = \mathbb{R}^n$, $\mathbb{F} = \mathbb{R}$, 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*:

**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, for all $f,g \in Y$, also $f+g \in Y$, and for all $f \in Y$, $\alpha \in \mathbb{F}$, also $\alpha f \in Y$.

Standard examples include:

(i) $\Pi_k :=$ the set of all **polynomials of degree** $\le k$ as a subset of $\mathbb{F}^{\mathbb{F}}$.

(ii) The set $C[a\,..\,b]$ of all **continuous function**s on the interval $[a\,..\,b]$.

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

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

(v) The set $\mathrm{BL}_\xi \subset C[\xi_1\,..\,\xi_{\ell+1}]$ of all **broken lines** *broken lines* with (interior) breaks at $\xi_2 < \cdots < \xi_\ell$.

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.

**Proposition.** A subset $Y$ of a vector space $X$ is a vector space (with respect to the same addition and multiplication by scalars) if and only if $Y$ is a linear subspace (of $X$), i.e., $Y$ is nonempty and is closed under addition and multiplication by scalars.

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 $\alpha f$ 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 $X$ a whole bunch of vector spaces, namely all the linear subspaces of the vector space $X^T$, with $T$ an arbitrary set.

## Linear maps

**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.,

$$\forall \{x, y \in X\}\ f(x+y) = f(x) + f(y);$$

and

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

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

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

$$L(X, Y).$$

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

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

The map $C(a\,..\,b) \to \mathbb{R} : g \mapsto \lim_{s\to t} g(s)$ (with $t$ some point in $(a\,..\,b)$) is linear.

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

**Example:** If $a \in \mathbb{R}^n$, then

$$a* : \mathbb{R}^n \to \mathbb{R} : x \mapsto a*x := a_1x_1 + a_2x_2 + \cdots + a_nx_n \tag{1.3}$$

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

$$a*? = b$$

for some **coefficient vector** $a \in \mathbb{R}^n$ and some **right side** $b \in \mathbb{R}$. Such an equation has solutions for arbitrary $b$ if and only if $a \neq 0$. You 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

$$a*? = 0.$$

In particular, the map $a*$ cannot be 1-1 unless $n = 1$.

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

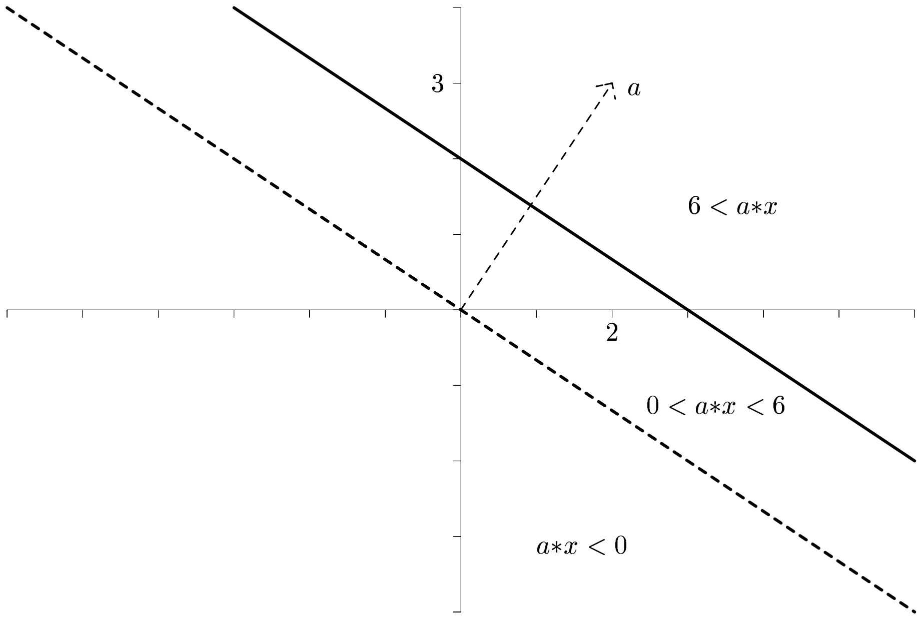
$$\operatorname{null} a* := \{x \in \mathbb{R}^2 : a*x = 0\}$$

through the origin formed by all the solutions to the corresponding homogeneous problem, and perpendicular to the coefficient vector $a$. Note that the 'nullspace' $\operatorname{null} a*$ splits $\mathbb{R}^2$ into the two **half-spaces**

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

one of which contains $a$. Here is such a figure, for the particular equation

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




One way to visualize all the parts of the equation $2x_1 + 3x_2 = 6$.

□

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 $\alpha A$ of $A \in L(X,Y)$ and $\alpha \in \mathbb{F}$ 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 $Y$.

> $L(X,Y)$ 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*.

> 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 $x, y \in X$,

$$(BA)(x+y) = B(A(x+y)) = B(Ax+Ay) = B(Ax) + B(Ay) = (BA)(x) + (BA)(y).$$

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

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

### Linear maps and matrices

**Standard example.** *Take any sequence $v_1, v_2, \ldots, v_n$ in the vector space $X$. Then the map*

$$f : \mathbb{F}^n \to X : a \mapsto v_1 a_1 + v_2 a_2 + \cdots + v_n a_n$$

*is linear.*

**Proof:** The proof is a boring but necessary verification.

(a) additivity:

$$\begin{aligned} f(a+b) &= v_1\,(a+b)_1 + v_2\,(a+b)_2 + \cdots + v_n\,(a+b)_n && \text{(definition of } f) \\ &= v_1\,(a_1+b_1) + v_2\,(a_2+b_2) + \cdots + v_n\,(a_n+b_n) && \text{(addition of } n\text{-vectors)} \\ &= v_1 a_1 + v_1 b_1 \ + \ v_2 a_2 + v_2 b_2 \ + \ \cdots \ + \ v_n a_n + v_n b_n && \text{(multipl. by scalar distributes)} \\ &= v_1 a_1 + v_2 a_2 + \cdots + v_n a_n \quad + \quad v_1 b_1 + v_2 b_2 + \cdots + v_n b_n && \text{(vector addition commutes)} \\ &= f(a) \ + \ f(b) && \text{(definition of } f) \end{aligned}$$

(s) homogeneity:

$$\begin{aligned} f(\alpha a) &= v_1\,(\alpha a)_1 + v_2\,(\alpha a)_2 + \cdots + v_n\,(\alpha a)_n && \text{(definition of } f) \\ &= v_1 \alpha a_1 + v_2 \alpha a_2 + \cdots + v_n \alpha a_n && \text{(multipl. of scalar with } n\text{-vectors)} \\ &= \alpha(v_1 a_1 + v_2 a_2 + \cdots + v_n a_n) && \text{(multipl. by scalar distributes)} \\ &= \alpha f(a) && \text{(definition of } f) \end{aligned}$$

□

**Definition:** The weighted 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** $a_1, a_2, \ldots, a_n$. I will use the suggestive abbreviation

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

hence use
$$[v_1, v_2, \ldots, v_n]$$
for the map $\mathbb{F}^n \to X : a \mapsto v_1a_1 + v_2a_2 + \cdots + v_na_n$. 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 of the $v_j$ are $m$-vectors, and

$$v_1a_1 + v_2a_2 + \cdots + v_na_n = Va,$$

with $V$ the $m \times n$-matrix with columns $v_1, v_2, \ldots, v_n$. This explains why I chose to write the weights in the linear combination $v_1a_1 + v_2a_2 + \cdots + v_na_n$ 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 $v_1, v_2, \ldots, v_n$.

Note that `MATLAB` uses the notation $[v_1, v_2, \ldots, v_n]$ for the matrix with columns $v_1, v_2, \ldots, v_n$, 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 $c_1, c_2, \ldots, c_n$ as the *linear map* $[c_1, c_2, \ldots, c_n] : \mathbb{F}^n \to \mathbb{F}^m : x \mapsto c_1x_1 + c_2x_2 + \cdots + c_nx_n$.

□

**(1.4) Agreement:** For any sequence $v_1, v_2, \ldots, v_n$ of $m$-vectors,

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

denotes both the $m \times n$-matrix $V$ with columns $v_1, v_2, \ldots, v_n$ and the linear map

$$V : \mathbb{F}^n \to \mathbb{F}^m : a \mapsto [v_1, v_2, \ldots, v_n]a = v_1a_1 + v_2a_2 + \cdots + v_na_n.$$

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

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 $\mathbb{F}^m$ which carries the $n$-vector $a$ to the $m$-vector

$Va = v_1a_1 + v_2a_2 + \cdots + v_na_n$. 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 $\operatorname{ran} V \subseteq \mathbb{F}^m$. – If we want the first interpretation, we call $V \in \mathbb{F}^{m\times n}$ a (two-dimensional) **array**.

Next, we prove that there is nothing special about the linear maps of the form $[v_1, v_2, \ldots, v_n]$ from $\mathbb{F}^n$ into the vector space $X$, i.e., *every* linear map from $\mathbb{F}^n$ to $X$ is necessarily of that form. This is due to the fact that the identity map

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

is of this form, i.e.,

$$\mathrm{id}_n = [e_1, e_2, \ldots, e_n]$$

with $e_j$ the $j$**th unit vector**, i.e.,

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

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

$$a = e_1a_1 + e_2a_2 + \cdots + e_na_n, \qquad \forall a \in \mathbb{F}^n.$$

Further,

**(1.5) 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 $a \in \mathbb{F}^n$,

$$(fV)a = f(Va) = f(v_1a_1 + v_2a_2 + \cdots + v_na_n) = f(v_1)a_1 + f(v_2)a_2 + \cdots + f(v_n)a_n = [f(v_1), \ldots, f(v_n)]a.$$

□

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

$$f = f\ \mathrm{id}_n = f\,[e_1, e_2, \ldots, e_n] = [f(e_1), \ldots, f(e_n)].$$

This proves:

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

$$f = [f(e_1), f(e_2), \ldots, f(e_n)].$$

In other words,

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

As a simple example, recall from (1.3) the linear map $a* : \mathbb{R}^n \to \mathbb{R} : x \mapsto a_1x_1 + a_2x_2 + \cdots + a_nx_n$. In this case, $a * e_j = a_j$, all $j$. Hence

$$a* = [a_1, \ldots, a_n] = [a]'. \tag{1.7}$$

**Notation:** I follow `MATLAB` notation. E.g., $[V, W]$ denotes the column map in which first all the columns of $V$ are used and then all the columns of $W$. Also, if $V$ and $W$ are column maps, then I write

$$V \subseteq W$$

to mean that $V$ is obtained by omitting (zero or more) columns from $W$; i.e., $V = W(:, a)$ for some 1-1 list $a$.

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

$$W \subseteq M$$

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

**(1.8) Proposition.** If $Z$ is a linear subspace of $Y$ and $W \in L(\mathbb{F}^m, Y)$, then $W \subseteq Z \implies \operatorname{ran} W \subseteq Z$.

The important (1.5)Proposition is the reason we define the **product of matrices** the way we do. If $A \in \mathbb{F}^{m\times n} = L(\mathbb{F}^n, \mathbb{F}^m)$ and $B = [b_1, b_2, \ldots, b_r] \in \mathbb{F}^{n\times r} = L(\mathbb{F}^r, \mathbb{F}^n)$, then $AB \in L(\mathbb{F}^r, \mathbb{F}^m) = \mathbb{F}^{m\times r}$, and

$$AB = A[b_1, b_2, \ldots, b_r] = [Ab_1, \ldots, Ab_r].$$

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

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

For this reason, most books on elementary linear algebra and most users of linear algebra *identify* the $n$-vector $x$ with the $n \times 1$-matrix $[x]$, hence write simply $x$ for what I have denoted here by $[x]$. 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 $x$ by writing down its entries in a list, like $x = (x_1, x_2, \ldots)$, since that uses much less space than does the writing of

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

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

$$\langle x, y\rangle := \sum_j \overline{y_j} x_j = x{*}\overline{y}.$$

On the other hand,

$$xy' = [x][y]' \ = \ (x_i\overline{y_j} : i, j = 1, \ldots, n)$$

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

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 $x$ or $x'$ is meant to be a matrix, I will write $[x]$ and $[x]'$, respectively.

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

**(1.9) Example.** Here is an example, of help later, for the use of (1.6). Consider the socalled **elementary row operation**

$$E_{i,k}(\alpha)$$

on $n$-vectors, in which one adds $\alpha$ times the $k$-th entry to the $i$th entry. This operation leaves $e_j$ unchanged for every $j \neq k$, and changes $e_k$ to $e_k + \alpha e_i$. Thus, *if* $E_{i,k}(\alpha)$ is linear, then necessarily

$$E_{i,k}(\alpha) = [\ldots, e_{k-1}, e_k + \alpha e_i, e_{k+1}, \ldots] = [e_1, \ldots, e_n] + \alpha[\underbrace{0, \ldots, 0}_{k-1}, e_i, 0, \ldots, 0],$$

or

$$(1.10) \qquad E_{i,k}(\alpha) = \mathrm{id}_n + \alpha e_i e_k'.$$

On the other hand, since

$$\alpha e_i e_k' x = \alpha e_i x_k = (\underbrace{0, \ldots, 0}_{i-1}, \alpha x_k, 0, \ldots, 0),$$

we see that

$$(\mathrm{id}_n + \alpha e_i e_k')x = (\ldots, x_{i-1}, x_i + \alpha x_k, x_{i+1}, \ldots),$$

which is exactly the operation we had in mind. We conclude that (1.10) does indeed hold.

Matrices of the form

$$(1.11) \qquad E_{y,z}(\alpha) := \mathrm{id} + \alpha y z'$$

are called **elementary**. They are very useful since, if invertible, their inverse has the same simple form; see (1.16)Proposition below. □

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

We are ready for 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 $\operatorname{ran} A = Y$. Now, the range of $A$ is the linear image of a vector space, hence itself a vector space. Indeed, if $v_1, v_2, \ldots, v_m$ are elements of $\operatorname{ran} A$, then there must be a sequence $w_1, w_2, \ldots, w_m$ in $X$ with $Aw_j = v_j$, all $j$. Since $X$ is a vector space, it contains $Wa$ for arbitrary $a \in \mathbb{F}^m$, therefore the corresponding linear combination $Va = [Aw_1, Aw_2, \ldots, Aw_m]a = (AW)a = A(Wa)$ must be in $\operatorname{ran} A$. In other words, if $V \subset \operatorname{ran} A$, then $\operatorname{ran} V \subset \operatorname{ran} A$.

Hence, if we wonder whether $A$ is onto, and we happen to know an *onto* column map $[v_1, v_2, \ldots, v_m] = V \in L(\mathbb{F}^m, Y)$, then we only have to check that the *finitely many* columns, $v_1, v_2, \ldots, v_m$, of $V$ are in $\operatorname{ran} A$. For, if some are not in $\operatorname{ran} A$, then, surely, $A$ is not onto. However, if they all are in $\operatorname{ran} A$, then $Y = \operatorname{ran} V \subseteq \operatorname{ran} A \subseteq \operatorname{tar} A = Y$, hence $\operatorname{ran} A = Y$ and $A$ is onto.

**(1.12)Proposition:** The range of a linear map $A \in L(X,Y)$ is a linear subspace, i.e., is closed under vector addition and multiplication by a scalar.

If $Y$ is the range of the column map $V$, then $A$ 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

$$(1.13) \qquad A^{-1}\{y\} = x + \{z \in X : Az = 0\}.$$

In particular, $A$ is 1-1 if and only if $\{z \in X : Az = 0\} = \{0\}$. In other words, 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 $\{z \in X : Az = 0\}$ of all elements of $X$ mapped by $A$ to 0 is singled out.

**Definition:** The set
$$\operatorname{null} A := \{z \in \operatorname{dom} A : Az = 0\}$$
is called the **nullspace** or **kernel** of the linear map $A$.
The 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.

Almost all linear subspaces you'll meet will be of the form $\operatorname{ran} A$ or $\operatorname{null} A$ for some linear map $A$. These two ways of specifying a linear subspace are very different in character. **linear subspace, specification of**

If we are told that our linear subspace $Z$ of $X$ is of the form $\operatorname{null} A$, 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 $\operatorname{ran} A$ for some linear map from some $U$ into $X$, then we can 'write down' explicitly every element of $\operatorname{ran} A$: 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 $\Pi_k$ 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 notation for the **monomial of degree** $j$, i.e., as the collection of all real-valued functions which are of the form
$$t \mapsto a_0 + a_1 t + \cdots + a_k t^k$$
for some coefficient-vector $a$. On the other hand, $\Pi_k$ can also be defined as $\operatorname{null} D^{k+1}$, i.e., as the collection of all real-valued functions which are $k+1$-times continuously differentiable and have their $(k+1)$st derivative identically zero.

## 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 $\mathbb{F}^n \to \mathbb{F}^m : a \mapsto Aa := \sum_j A(:,j)a_j$. For this reason, it is also customary to refer to the range $\operatorname{ran} A$ of a matrix $A$ as the **column space** $\operatorname{col} A$ of that matrix. Further, we have found that, in these terms, the *matrix product* $AB$ is also the *composition* $A \circ B$, i.e.,
$$(A \circ B)a = A(B(a)) = (AB)a = \sum_j (AB)(:,j)a_j.$$
In these terms, the identity map $\mathrm{id}_n$ on $\mathbb{F}^n$ corresponds to the **identity matrix** $[e_1, e_2, \ldots, e_n]$, hence the name for the latter.

**(1.14) 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. (2.14)) that this can only happen when $m = n$.

We will also soon prove (cf. (2.13)Theorem below) the *pigeon hole principle for square matrices*, i.e., that a linear map from $\mathbb{F}^n$ to $\mathbb{F}^n$ is 1-1 if and only if it is onto. In other words, 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:

**(1.15) 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 numerical process called *elimination*, for solving a homogeneous linear system $A? = 0$, i.e., for constructing $\operatorname{null} A$.

In preparation, and as in exercise in invertible matrices, we verify the following useful fact about elementary matrices.

**(1.16) Proposition.** For $x, y \in \mathbb{F}^n$ and $\alpha \in \mathbb{F}$, the elementary matrix

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

is invertible if and only if $1 + \alpha z' y \neq 0$, and, in that case

$$E_{y,z}(\alpha)^{-1} = E_{y,z}\Big(\frac{-\alpha}{1+\alpha z' y}\Big). \tag{1.17}$$

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

$$\alpha y z' \; \beta y z' = \alpha\beta \; (z'y) \; yz',$$

we conclude that

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

In particular, since the factor $(\alpha + \beta + \alpha\beta(z'y))$ is symmetric in $\alpha$ and $\beta$, we conclude that

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

Further, if $1 + \alpha z' y \neq 0$, then the choice

$$\beta = \frac{-\alpha}{1+\alpha z' y}$$

will give $\alpha + \beta + \alpha\beta(z'y) = 0$, hence $E_{y,z}(\beta)E_{y,z}(\alpha) = E_{y,z}(\alpha)E_{y,z}(\beta) = \mathrm{id}_n$. This proves that $E_{y,z}(\alpha)$ is invertible, with its inverse given by (1.17).

Conversely, assume that $1 + \alpha z' y = 0$. Then $y \neq 0$, yet

$$E_{y,z}(\alpha) y = y + \alpha(z'y)y = 0,$$

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

## 2. Elimination, or: The determination of $\operatorname{null} A$ and $\operatorname{ran} A$

### Elimination and Backsubstitution

Elimination has as its goal an efficient description of the solution set for the *homogeneous* linear system $A? = 0$, i.e., of the nullspace of the matrix $A$. It is based on the following observation:

**(2.1) 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 $\alpha$ times row $k$ from row $i$, for some $k \neq i$. Then, by (1.9)Example,

$$B = E_{i,k}(-\alpha)\, A,$$

with $E_{i,k}(-\alpha) = \mathrm{id}_m - \alpha e_i e_k'$. Consequently, $\operatorname{null} A \subseteq \operatorname{null} B$.

However, since $i \neq k$, we have, for any $\alpha$, that $-\alpha(e_k{}' e_i) = 0$. Therefore, by (1.16), also

$$A = E_{i,k}(\alpha)\, B,$$

hence also $\operatorname{null} B \subseteq \operatorname{null} A$. □

One solves the homogeneous linear system $A? = 0$ by **elimination**. This is an *inductive* process which 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. The classification proceeds inductively, from the first to the last unknown, i.e., for $k = 1, 2, \ldots, n$, with the $k$th step as follows:.

At the beginning of the $k$th **elimination step**, we have in hand a matrix $B$, called the **working-array**, which is **equivalent** to our initial matrix $A$ in that $\operatorname{null} B = \operatorname{null} A$. Further, we have already classified 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 $B$ are nonpivot rows; they 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 $k$th unknown and, correspondingly, change $B$, as follows:

**bound case:** We call the $k$th unknown **bound** in case we can find some nonpivot row $B(h,:)$ for which $B(h,k) \neq 0$. We pick one such row and call it the **pivot row** for the $k$th unknown. Further, we use it to eliminate the $k$th unknown from all the remaining nonpivot rows $B(i,:)$ by the calculation

$$B(i,:) := B(i,:) - \frac{B(i,k)}{B(h,k)} B(h,:).$$

**free case:** In the contrary case, we call the $k$th unknown **free**. No action is required in this case, since none of the nonpivot rows involves the $k$th unknown.

By (2.1)Lemma, the changes (if any) made in $B$ will not change $\operatorname{null} B$. This finishes the $k$th elimination step.

For future reference, here is a formal description of the entire algorithm.

**(2.2) Elimination Algorithm.**
**input**: $A \in \mathbb{F}^{m\times n}$.
$B \leftarrow A$, $p \leftarrow (0,\ldots,0) \in \mathbb{R}^n$.
**for** $k = 1,\ldots,n$, **do**:
  **for some** $h \notin \operatorname{ran} p$ with $B(h,k) \neq 0$, **do**:
    $p(k) \leftarrow h$
    **for all** $i \notin \operatorname{ran} p$, **do**:

$$B(i,:) \leftarrow B(i,:) - \frac{B(i,k)}{B(h,k)}\, B(h,:)$$

    **enddo**
  **enddo**
**enddo**
**output**: $B$, $p$, and, possibly, `free` $\leftarrow$ `find(`$p$`==0)`, `bound` $\leftarrow$ `find(`$p$`>0)`.

Note that nothing is done at the $k$th step if there is no $h \notin \operatorname{ran} p$ with $B(h,k) \neq 0$, i.e., if $B(h,k) = 0$ for all $h \notin \operatorname{ran} p$.

**A numerical example** We start with

$$A := \begin{bmatrix} 0 & 2 & 0 & 2 & 5 & 4 & 0 & 6 \\ 0 & 1 & 0 & 1 & 2 & 2 & 0 & 3 \\ 0 & 2 & 0 & 2 & 5 & 4 & -1 & 7 \\ 0 & 1 & 0 & 1 & 3 & 2 & -1 & 4 \end{bmatrix}, \qquad p = (0,0,0,0,0,0,0,0).$$

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

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & \underline{\mathbf{1}} & 0 & 1 & 2 & 2 & 0 & 3 \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 & 1 \end{bmatrix}, \qquad p = (0,2,0,0,0,0,0,0).$$

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 nonpivotal 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 nonpivot rows, i.e., from rows 3 and 4. This gives

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 & \underline{\mathbf{1}} & 0 & 0 & 0 \\ 0 & \mathbf{1} & 0 & 1 & 2 & 2 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}, \qquad p = (0,2,0,0,1,0,0,0).$$

The sixth unknown is free, but there are nonzero entries in the seventh column of the remaining nonpivot rows, 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 nonpivot row. This gives

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 \\ 0 & \mathbf{1} & 0 & 1 & 2 & 2 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \underline{\mathbf{-1}} & 1 \end{bmatrix}, \qquad p = (0,2,0,0,1,0,4,0).$$

With that, there are no nontrivial nonpivot rows left. In particular, the eighth unknown is free, hence we have already in hand the final array.

Hence, 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 nonpivot rows of $B$ (if any) are entirely zero. Thus, in determining $x \in \operatorname{null} B$, we only have to pay attention to the pivot rows. This means that we can determine a particular element $x$ 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 $k$th unknown is bound, i.e., $k \in$ `bound`, determine $x_k$ from its pivot equation (since that equation only involves $x_k, \cdots, x_n$); else, pick $x_k$ arbitrarily (as then the $k$th unknown is free, i.e., $k \in$ `free`).

Here is a more formal description, for future reference.

**(2.3) Backsubstitution Algorithm.**
**input**: $B \in \mathbb{F}^{m \times n}$ and $p$ as output from (2.2), $y \in \mathbb{F}^n$.
$x \leftarrow y$
**for** $k = n, n-1, \ldots, 1$, **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 $A? = 0$ satisfying $x_i = y_i$ for all $i$ with $p(i) = 0$.

Notice that the value of the free unknowns is arbitrary and that, once these are chosen somehow, 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 general element of $\operatorname{null} B$ has exactly as many degrees of freedom as there are free unknowns. Since there are `#free` unknowns, $\operatorname{null} B$ is said to be 'of dimension `#free`'.

In particular, for any $k$, the $k$th entry, $x_k$, of an $x \in \operatorname{null} B$ can be nonzero only in one of two ways: (a) the $k$th unknown is free, i.e., $k \in$ `free`; (b) the $k$th unknown is bound, but $x_j \neq 0$ for some $j > k$. It follows that $x_k$ can be the rightmost nonzero entry of such an $x$ only if the $k$th unknown is free. Further, if the $k$th unknown is free, and $x$ is the element of $\operatorname{null} B = \operatorname{null} A$ computed by setting $x_k = 1$ and setting all other free entries equal to 0, then $x_k$ is necessarily the rightmost nonzero entry of $x$ (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 proves

**(2.4) Observation.** There exists $x \in \operatorname{null} A$ with rightmost nonzero entry $x_k$ if and only if the $k$th unknown is free. In particular, $\operatorname{null} A$ is nontrivial if and only if there are free unknowns.

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 is independent of all the details of the 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.

Here is another consequence of (2.4). If there are more unknowns than equations, then there are not enough equations to go around, i.e., some unknowns must be free, therefore there are nontrivial solutions to our homogeneous equation $A? = 0$. We remember this fundamental result of elimination in the following form:

**(2.5) Theorem:** Any matrix with more columns than rows has a nontrivial nullspace.

### The really reduced row echelon form

The construction of the *really reduced row echelon form* takes elimination four steps further:

(i) When the $h$th pivot row is found, and it is not the $h$th row, then it is exchanged with the current $h$th row to make it the $h$th row. (This keeps things neat; all the rows not yet used as pivot rows lie below all the rows already picked as pivot rows.)

(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 $B(h,k)$ is the pivot element in question (i.e., $\texttt{bound}(h) = k$, i.e., $x_k$ is the $h$th bound unknown), then, after this normalization, one merely subtracts $B(i,k)$ times $B(h,:)$ from $B(i,:)$ to eliminate the $k$th unknown from row $i$.)

(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:

$$B = \mathrm{rref}(A).$$

However, it turns out to be very neat to add the following final step:

(iv) Remove all rows which are entirely zero, thus getting the matrix

$$\Lambda := B(1{:}\#\texttt{bound},:) =: \mathrm{rrref}(A)$$

called the *really reduced row echelon form* of $A$.

Here is a formal description (in which we talk about *the* rref for $A$ even though we prove its *uniqueness* only later, in (2.9)):

**(2.6) Definition.** We say that $\Lambda$ is the **really reduced row echelon form for** $A \in \mathbb{F}^{m\times n}$ and write $\Lambda = \mathrm{rrref}(A)$, in case $\Lambda \in \mathbb{F}^{r\times n}$ for some $r$ and there is a strictly increasing sequence `bound` (provided by certain versions of the `MATLAB` function `rref` along with $\Lambda$) so that the following is true:

1. $\Lambda$ is a **row echelon form for** $A$: This means that (i) $\mathrm{null}\,\Lambda = \mathrm{null}\,A$; and (ii) for each $k = \texttt{bound}(i)$, $\Lambda(i,:)$ is the pivot row for the $k$th unknown, i.e., $\Lambda(i,:)$ is the unique row in $\Lambda$ for which $\Lambda(i,k)$ is the first (or, leftmost) nonzero entry.
2. $\Lambda$ is **really reduced** or normalized, in the sense that $\Lambda(:,\texttt{bound})$ is the identity matrix, i.e., for each $i$, the pivot element $\Lambda(i,\texttt{bound}(i))$ equals 1 and is the only nonzero entry in its column.

**A numerical example, continued** For the earlier schematic example, the rref and the rrref would look like this:

$$B = \begin{bmatrix} 0 & \mathbf{1} & 0 & 1 & 0 & 2 & 0 & 3 \\ 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \underline{\mathbf{1}} & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \qquad \Lambda = \begin{bmatrix} 0 & \mathbf{1} & 0 & 1 & 0 & 2 & 0 & 3 \\ 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \underline{\mathbf{1}} & -1 \end{bmatrix}.$$

Recall that, for this example, $\texttt{bound} = (2,5,7)$ and $\texttt{free} = (1,3,4,6,8)$. □

**A complete description for** $\operatorname{null} A$ **obtained from** $\operatorname{rrref}(A)$

The rrref for $A$ makes it easy to determine all solutions of the homogeneous linear system $A? = 0$, i.e., all the elements of $\operatorname{null} A$.

To recall, we call the $k$th unknown for $A? = 0$ *bound* or *free* depending on whether $k = \mathtt{bound}(j)$ for some $j$ or $k = \mathtt{free}(j)$ for some $j$.

We now obtain from $\Lambda = \operatorname{rrref}(A)$ a 1-1 matrix $C$ with the property that $\operatorname{null} A = \operatorname{ran} C$. Since any such $C$ is 1-1 onto $\operatorname{null} A$, this implies that every $x \in \operatorname{null} A$ can be written in exactly one way in the form $x = Ca$. We call such a $C$ a 'basis' for the vector space $\operatorname{null} A$.

In the discussion, we use the following notation introduced earlier: If $x$ is an $n$-vector and $p$ is a list of length $r$ with range in $\underline{n}$, then $x_p$ is the $r$-vector

$$x_p = (x_{p(i)} : i = 1, \ldots, r).$$

With this,

$$\begin{aligned} x \in \operatorname{null} A \quad \iff \quad 0 = \Lambda x &= \sum_j \Lambda(:,j) x_j \\ &= \sum_{j \in \mathtt{bound}} \Lambda(:,j) x_j \; + \sum_{j \in \mathtt{free}} \Lambda(:,j) x_j \\ &= \Lambda(:,\mathtt{bound}) x_{\mathtt{bound}} \; + \; \Lambda(:,\mathtt{free}) x_{\mathtt{free}} \end{aligned}$$

Since $\Lambda(:,\mathtt{bound}) = \mathrm{id}$ by 2. of (2.6), we conclude that

$$x \in \operatorname{null} A \quad \iff \quad x_{\mathtt{bound}} = -\Lambda(:,\mathtt{free}) x_{\mathtt{free}}.$$

We can write this even more succinctly in matrix form as follows:

$$\operatorname{null} A = \operatorname{ran} C,$$

with $C$ the $(n \times \#\mathtt{free})$-matrix whose 'free' rows form an identity matrix, and whose 'bound' rows are formed by the 'free' columns of $-\Lambda$:

$$C(\mathtt{free},:) = \mathrm{id}, \qquad C(\mathtt{bound},:) = -\Lambda(:,\mathtt{free}). \tag{2.7}$$

E.g., for the schematic example,

$$C = \begin{bmatrix} \mathbf{1} & 0 & 0 & 0 & 0 \\ -0 & -\times & -\times & -\times & -\times \\ 0 & \mathbf{1} & 0 & 0 & 0 \\ 0 & 0 & \mathbf{1} & 0 & 0 \\ -0 & -0 & -0 & -\times & -\times \\ 0 & 0 & 0 & \mathbf{1} & 0 \\ -0 & -0 & -0 & -0 & -\times \\ 0 & 0 & 0 & 0 & \mathbf{1} \end{bmatrix} = \begin{bmatrix} \mathbf{1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & \mathbf{1} & 0 & 0 & 0 \\ 0 & 0 & \mathbf{1} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{1} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathbf{1} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ -0 & -\times & -\times & -\times & -\times \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -0 & -0 & -0 & -\times & -\times \\ 0 & 0 & 0 & 0 & 0 \\ -0 & -0 & -0 & -0 & -\times \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Note that $C$ is 1-1, since $x = Ca = 0$ implies that $0 = x_{\mathtt{free}} = C(\mathtt{free},:)a = a$. Therefore, $C$ is (or, the columns of $C$ form) a basis for $\operatorname{null} A$, in the sense that $C$ is a 1-1 onto column map to $\operatorname{null} A$.

Finally, note the fact that $C$ is 'upper triangular' in the sense that

$$i > \mathtt{free}(j) \quad \Longrightarrow \quad C(i,j) = 0.$$

### The factorisation $A = A(:,\texttt{bound})\mathrm{rrref}(A)$

With $C$ as constructed from $\mathrm{rrref}(A)$ by (2.7), we have $\operatorname{ran} C = \operatorname{null} A$, therefore $AC = 0$. Written out more fully, this says that

$$A(:,\texttt{free})C(\texttt{free},:) + A(:,\texttt{bound})C(\texttt{bound},:) = 0.$$

Substituting (2.7) into this, we get

$$A(:,\texttt{free}) + A(:,\texttt{bound})(-\Lambda(:,\texttt{free})) = 0,$$

or

$$A(:,\texttt{free}) = A(:,\texttt{bound})\Lambda(:,\texttt{free}).$$

On the other hand, $\Lambda(:,\texttt{bound}) = \mathrm{id}$ by 2. of (2.6), hence

$$A(:,\texttt{bound}) = A(:,\texttt{bound})\Lambda(:,\texttt{bound})$$

trivially. We conclude:

**(2.8)** $$A = A(:,\texttt{bound})\,\mathrm{rrref}(A).$$

### A 'basis' for $\operatorname{ran} A$

Here is a first consequence of the factorization: Since $A = A(:,\texttt{bound})\Lambda$, we have $\operatorname{ran} A \subseteq \operatorname{ran} A(:,\texttt{bound})$, while certainly $\operatorname{ran} A(:,\texttt{bound}) \subseteq \operatorname{ran} A$. Hence

$$\operatorname{ran} A = \operatorname{ran} A(:,\texttt{bound}).$$

Also, $A(:,\texttt{bound})$ is 1-1: For, if $A(:,\texttt{bound})a = 0$, then the $n$-vector $x$ with $x_{\texttt{bound}} = a$ and with $x_{\texttt{free}} = 0$ is in $\operatorname{null} A$, hence must have $a = x_{\texttt{bound}} = C(\texttt{bound},:)x_{\texttt{free}} = 0$. Consequently, $A(:,\texttt{bound})$ is (or, the columns of $A(:,\texttt{bound})$ form) a basis for $\operatorname{ran} A$.

### Uniqueness of the $\mathrm{rrref}(A)$

A second consequence of (2.8) is the fact that, since $A(:,\texttt{bound})$ is 1-1, the matrix $\mathrm{rrref}(A)$ is uniquely determined: For, we already observed that, by (2.4), the sequence $\texttt{bound}$ only depends on $\operatorname{null} A$. Further, with $\Lambda := \mathrm{rrref}(A)$, (2.8) says that, for $j = 1,\ldots,n$, $\Lambda(:,j) = \Lambda e_j$ is a solution of the equation $A(:,\texttt{bound})? = A(:,j)$, and this solution is unique since $A(:,\texttt{bound})$ is 1-1.

Further, since $\mathrm{rref}(A)$ differs from $\mathrm{rrref}(A)$ only by those additional $m - \#\texttt{bound}$ zero rows, it follows that also each $A$ has a *unique* rref.

This finishes the proof of the following summarizing theorem.

**(2.9) Theorem:** For given $A \in \mathbb{F}^{m\times n}$, there is exactly one matrix $\Lambda$ having the properties 1. and 2. (listed in (2.6)) of a rrref for $A$. Further, with $\texttt{bound}$ and $\texttt{free}$ the indices of bound and free unknowns, $V := A(:,\texttt{bound})$ is 1-1 onto $\operatorname{ran} A$, and $C \in \mathbb{F}^{n\times\#\texttt{free}}$, given by $C(\texttt{free},:) = \mathrm{id}$, $C(\texttt{bound},:) = -\Lambda(:,\texttt{free})$, is 1-1 onto $\operatorname{null} A$, and $C$ is 'upper triangular' in the sense that $C(i,j) = 0$ for $i > \texttt{free}(j)$.

Finally, since $A(:,k) \in \operatorname{ran} A(:,[1{:}k{-}1])$ if and only if $\operatorname{null} A$ contains some $x$ for which $x_k$ is the rightmost nonzero entry, we have, as a consequence of (2.4)Observation, the following

**(2.10) Proposition:** The $k$th unknown is free if and only if $A(:,k) \in \operatorname{ran} A(:,[1{:}k{-}1])$, i.e., iff the $k$th column of $A$ is a linear combination of the columns of $A$ strictly to the left of it.

### rrref$(A)$ **and the solving of** $A? = b$

(2.10)Proposition is exactly what we need when considering the linear system

$$A? = b \tag{2.11}$$

for given $A \in \mathbb{F}^{m\times n}$ and given $b \in \mathbb{F}^m$. For, here we are hoping to write $b$ as a linear combination of the columns of $A$, and (2.10) tells us that this is possible exactly when the last unknown in the *homogeneous* system

$$[A, b]? = 0 \tag{2.12}$$

is free. Further, the factorization (2.8), applied to the **augmented** matrix $[A, b]$, tells us how to write $b$ as a linear combination of the columns of $A$ in case that can be done. For, with $\Lambda = \mathrm{rrref}([A, b])$, and assuming that the $(n+1)$st unknown is free, it says that

$$b = A(:, \mathtt{bound})\Lambda(:, n+1).$$

More generally, if $\Lambda = \mathrm{rrref}([A, U])$ for some arbitrary matrix $U \in \mathbb{F}^{m\times s}$ and all the unknowns corresponding to columns of $U$ are free, then, by (2.8), applied to $[A, U]$ rather than $A$, we have

$$U = A(:, \mathtt{bound})\Lambda(:, n + [1{:}s]).$$

**A numerical example, continued** Recall our earlier example in which we used elimination to convert a given matrix to a kind of row-echelon form, as follows:

$$\begin{bmatrix} 0 & 2 & 0 & 2 & 5 & 4 & 0 & 6 \\ 0 & 1 & 0 & 1 & 2 & 2 & 0 & 3 \\ 0 & 2 & 0 & 2 & 5 & 4 & -1 & 7 \\ 0 & 1 & 0 & 1 & 3 & 2 & -1 & 4 \end{bmatrix} \to \begin{bmatrix} 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 \\ 0 & \mathbf{1} & 0 & 1 & 2 & 2 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{-1} & 1 \end{bmatrix} \to \begin{bmatrix} 0 & \mathbf{1} & 0 & 1 & 0 & 2 & 0 & 3 \\ 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & -1 \end{bmatrix},$$

hence $\mathtt{bound} = (2, 5, 7)$, $\mathtt{free} = (1, 3, 4, 6, 8)$. 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 $[A, U]$. As soon as we specify the number of columns, in $A$ or $U$, we know $A$ and $U$ exactly.

First, choose $U$ to be a one-column matrix. Then, since the last unknown is free, we conclude that

$$(6, 3, 7, 4) = A(:, \mathtt{bound})\Lambda(:, 8) = \begin{bmatrix} 2 & 5 & 0 \\ 1 & 2 & 0 \\ 2 & 5 & -1 \\ 1 & 3 & -1 \end{bmatrix} (3, 0, -1).$$

If we choose $U$ to be a two-column matrix instead, then the linear system $A? = U$ is unsolvable since now one of the columns of $U$ (the first one) corresponds to a bound unknown. In fact, also the second column of $U$ fails to be in the range of this $A$ since the corresponding unknown turned out to be free only because of the presence of the seventh column, and this seventh column is *not* in the span of the columns to the left of it, hence neither is the eighth column. □

## The pigeon hole principle for square matrices

Here is a surefire way to guarantee that, for the homogeneous system (2.12), the last unknown is free: *Make sure that $A$ is 1-1 and square.*

Indeed, by making $A$ 1-1, we guarantee that the first $n$ unknowns must be bound (since having the $k$th unknown free would entitle us, by (2.4), to some $x$ in the nullspace of $[A, b]$ whose rightmost nonzero entry is $x_k$, hence if $k \le n$, $x_{[1:n]}$ would be a nontrivial element in $\operatorname{null} A$, contrary to the assumption that $A$ is 1-1). With that, making $A$ square guarantees that the first $n$ unknowns have used up all the available rows as pivot rows. Therefore, the $(n+1)$st unknown must be free, regardless of what $b$ might be.

In other words, if $A$ is 1-1 and square, then we can solve (2.11) no matter what the right-hand side $b$ is, i.e., then $A$ is onto. This proves half of the following basic Theorem.

**(2.13) Theorem (pigeon hole principle for square matrices).** A square matrix is 1-1 if and only if it is onto.

**Proof:** We still have left to prove that an onto square matrix $A$ is 1-1. So, assume that $A \in \mathbb{F}^{n \times n}$ is onto. Then, for each $j$, the equation $A? = e_j$ has a solution, $b_j$ say. Set $B := [b_1, b_2, \ldots, b_n]$. Then $AB = A[b_1, b_2, \ldots, b_n] = [Ab_1, Ab_2, \ldots, Ab_n] = [e_1, e_2, \ldots, e_n] = \mathrm{id}_n$. In particular, $B$ is 1-1, therefore, by what we already know, also onto, hence invertible. Since $A$ is a left inverse for $B$, $A$ must be $B^{-1}$. In particular, $A$ must be 1-1. □

In other words, when dealing with a *square* matrix, 1-1 *or* onto is already enough to have 1-1 *and* onto, i.e., to have invertibility.

Note that the proof just given provides the proof of the 'Amazing Fact' (1.15), since it concludes from $AB = \mathrm{id}$ (with $A$, $B$ square) that $A$ must be the inverse of $B$, and this implies, in particular, that also $BA = \mathrm{id}$.

Why did we restrict ourselves so far to *square* matrices? The same simple comparison of number of unknowns with number of available rows makes clear that

**(2.14) Proposition.** An invertible matrix is necessarily square. More precisely, if $A \in \mathbb{F}^{m \times n}$, then (i) $A$ 1-1 implies that $m \ge n$; and (ii) $A$ onto implies that $m \le n$.

**Proof:** If $A$ is 1-1, then, as we already observed, the first $n$ unknowns in (2.12) must be bound. In particular, there must be at least $n$ rows to 'bind' them, i.e., to supply pivot rows for them.

If $A$ is onto, then, by (2.10), the $(n+1)$st unknown in (2.12) must be free, regardless of what $b$ is. This can happen only if all $m$ rows are used as pivot rows by the time we reach the $(n+1)$st unknowns, and this is possible only if $m \le n$. □

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 $\Lambda = \mathrm{rrref}([A, \mathrm{id}_n])$, then, with (2.8), necessarily $[A, \mathrm{id}_n] = A\Lambda$, hence $\Lambda = [\,\mathrm{id}_n, A^{-1}]$.

**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 $A? = b$ in practice by computing $\mathrm{rref}([A, b])$ or by computing `inv(A)*b`. Rather, in `MATLAB` you would compute the solution of `A?=b` as `A\b`. 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.

□

**(2.15) Example** If $V \in L(\mathbb{F}^n, X)$ and $\Lambda \in L(X, \mathbb{F}^n)$, then $\Lambda V$ is a square matrix, of order $n$. If $\Lambda V$ is 1-1, then the Theorem tells us that $\Lambda V$ is invertible. In particular, $V$ is 1-1 and $\Lambda$ is onto, and, for every $x \in X$, there exists exactly one $v \in \operatorname{ran} V$ for which $\Lambda v = \Lambda x$. This is the essence of **interpolation**.

For example, take $X = C(\mathbb{R})$, $V = [()^0, ()^1, \ldots, ()^k]$, hence $\operatorname{ran} V$ equals $\Pi_k$, the collection of all the polynomials of degree $\le k$. Further, take $\Lambda : X \to \mathbb{R}^{k+1} : f \mapsto (f(t_0), f(t_1), \ldots, f(t_k))$ for some fixed sequence $t_0 < t_1 < \cdots < t_k$ of points. Then the equation

$$\Lambda V ? = \Lambda f$$

asks for the (power) coefficients of a polynomial of degree $\le k$ which agrees with the continuous function $f$ at the $k+1$ distinct points $t_j$.

We investigate whether $\Lambda V$ is 1-1. If $\Lambda V a = 0$, then the polynomial $p := Va$ vanishes at the $k+1$ distinct points $t_j$. However, $p$ is a polynomial of degree $\le k$, hence cannot vanish at more than $k$ points without vanishing identically (how do you know that???). Therefore, $p = 0$, i.e., $Va = 0$. So, *assuming* that $V$ is 1-1 (see (3.18)Example), it follows that $a = 0$.

We have 'proved' that $\Lambda V$ is 1-1, hence know now that, for every $f \in C(\mathbb{R})$ and every $k+1$ distinct points $t_1, \ldots, t_k$, there is exactly one polynomial of degree $\le k$ which agrees with $f$ at these $t_j$. □

## 3. The dimension of a vector space

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, one sets up the invertible linear map

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

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 $\mathbb{F}^n$ which 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 which 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 $a \in \mathbb{F}^n$ with $x = Va$. 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** $\mathbb{F}^n$.

The bound part, $A(:, \mathtt{bound})$ of $A \in \mathbb{F}^{m\times n}$ is a basis for $\operatorname{ran} A$. You also know how to construct a basis for the nullspace of any $A \in \mathbb{F}^{m\times n}$ from its $\operatorname{rrref}(A)$.

**(3.1) Proposition.** If $V \in L(\mathbb{F}^n, X)$ and $W \in L(\mathbb{F}^m, X)$ are both bases for $X$, then $m = n$.

**Proof:** If both $V \in L(\mathbb{F}^n, X)$ and $W \in L(\mathbb{F}^m, X)$ are invertible, then $W^{-1}V$ is an invertible linear map from $\mathbb{F}^n$ to $\mathbb{F}^m$, hence an invertible matrix, and therefore, by (2.14)Proposition, it must be a square matrix; in particular, $m = n$. □

In other words, all bases for a given vector space $X$ have the same order, i.e., the same number of columns. This number is called the **dimension** of $X$; it is written

$$\dim X.$$

In particular, since $\mathrm{id}_n$ 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 $X$.

Here is a small difficulty with this (and any other) definition of dimension: What is the dimension of the **trivial space**, i.e., the vector space which consists of the zero vector alone? It is a perfectly well-behaved vector space (though a bit limited).

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 a number 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. Correspondingly, there is exactly one *linear* map from $\mathbb{F}^0$ into $X$, namely the linear map $\mathbb{F}^0 \to X : 0 \mapsto 0$. Since this is a linear map from $\mathbb{F}^0$, we call it the column map into $X$ with *no* columns, and denote it by $[\,]$. Thus,

$$[\,] : \mathbb{F}^0 \to X : 0 \mapsto 0. \tag{3.2}$$

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 $\{0\}$, hence

$$\dim\{0\} = 0.$$

Isn't Mathematics wonderful! - As it turns out, the column map [] will also be very helpful below.

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

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

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

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

$v_1, v_2, \ldots, v_n$ is said to be **linearly independent** in case $V$ is 1-1, i.e., in case $Va = 0$ implies $a = 0$ (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).

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

$v_1, v_2, \ldots, v_n$ 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* $V$ 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.

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

**(3.3) Proposition:** Let $V = [v_1, v_2, \ldots, v_n]$ 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 on 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 $V^{-1}v_j = e_j$, 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). □

## 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 $V \subseteq W$ 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.

For the proof of these assertions, we use the following lemma.

**(3.4) Lemma:** If $V \in L(\mathbb{F}^n, X)$ is 1-1, and $w \in X$, then $[V, w]$ is 1-1 iff $w \notin \operatorname{ran} V$.

**Proof:** If $w \in \operatorname{ran} V$, then $w = Va$ for some $a$, hence $[V,w](a,-1) = Va - w = 0$, showing that $[V,w]$ fails to be 1-1. Conversely, if $w \notin \operatorname{ran} V$, and $0 = [V,w](a,b) = Va + wb$, then $b \neq 0$ would imply that $w = V(-a/b) \in \operatorname{ran} V$, a contradiction; therefore, then necessarily $b = 0$, thus already $Va = 0$, hence also $a = 0$ (since $V$ is 1-1, by assumption). □

Call the column map $V$ into $X$ **maximally 1-1** if $V$ is 1-1 while any $W \supsetneq V$ fails to be 1-1.

**(3.5) Corollary:** Any maximally 1-1 column map into a vector space $X$ is a basis for $X$.

**Proof:** If $V$ is a maximally 1-1 column map into $X$, then, for any $w \in X$, $[V,w]$ fails to be 1-1, therefore, by (3.4)Lemma, $w \in \operatorname{ran} V$, showing that $V$ is also onto $X$, hence invertible. □

**(3.6) Corollary:** If $V \in L(\mathbb{F}^n, X)$ is 1-1, and $W \in L(\mathbb{F}^m, X)$ is onto, then there exists $U \subseteq W$ so that $[V,U]$ is a basis (for $X$).

**Proof:** (3.4)Lemma shows that, at every step of the algorithm:

$U := [\,]$; **for** $w \in W$ **do: if** $w \notin \operatorname{ran}[V,U]$, **then** $U \leftarrow [U,w]$; **endif; enddo**

the column map $[V,U]$ is 1-1. Further, for the final $U$, all the columns of $W$ are contained in $\operatorname{ran}[V,U]$, therefore $X = \operatorname{ran} W \subseteq \operatorname{ran}[V,U] \subseteq X$, i.e., $[V,U]$ is also onto. □

**(3.7) Corollary:** Any 1-1 column map into a finitely generated vector space $X$ can be extended to a basis for $X$.

In particular, if $V \in L(\mathbb{F}^n, X)$ is 1-1, then $n \leq \dim X$.

With the choice $V = [\,]$, (3.6)Corollary states the following.

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

In particular, if $W \in L(\mathbb{F}^m, X)$ is onto, then $\dim X \leq m$.

Call the column map $V$ **minimally onto** $X$ if $\operatorname{ran} V = X$ and any $W \subsetneq V$ fails to be onto $X$.

**(3.9) Corollary:** Any minimally onto column map to a vector space $X$ is a basis for $X$.

**Proof:** If $V$ is onto $X$, then it can be thinned to a basis for $X$, by (3.8)Corollary. If this basis had fewer columns than $V$, then $V$ would not be minimally onto. Hence the basis must be $V$ itself. □

A slight variation on the argument for (3.4)Lemma gives the following useful result.

**(3.10) Proposition:** Let $V$ and $W$ be 1-1 column maps into the vector space $X$. Then, $[V, W]$ is 1-1 if and only if $(\operatorname{ran} V) \cap (\operatorname{ran} W) = \{0\}$.

**Proof:** '$\Longrightarrow$': Let $x \in \operatorname{ran} V \cap \operatorname{ran} W$. Then, for some $a$, $b$, $Va = x = Wb$, therefore $[V, W](a, -b) = 0$. Hence, if $[V, W]$ is 1-1, this implies that both $a = 0$ and $-b = 0$, hence $x = 0$. In other words, if $[V, W]$ is 1-1, then $\operatorname{ran} V \cap \operatorname{ran} W = \{0\}$.

'$\Longleftarrow$': Let $[V, W](a, c) = 0$. Then $x := Va = -Wc$, therefore $x \in \operatorname{ran} V \cap \operatorname{ran} W$. Hence, if $\operatorname{ran} V \cap \operatorname{ran} W = \{0\}$, this implies that $Va = 0 = -Wc$, and, as both $V$ and $W$ are 1-1 by assumption, this implies that $a = 0$ and $c = 0$. In other words, if $\operatorname{ran} V \cap \operatorname{ran} W = \{0\}$ (and both $V$ and $W$ are 1-1), then $[V, W]$ is 1-1. $\square$

### Some uses of the dimension concept

Here is the major use of the dimension concept.

**(3.11) Proposition:** If $X$, $Y$ are vector spaces with $X \subseteq Y$ 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 $X$ (namely the unique linear map from $\mathbb{F}^0$ into $X$), while $\dim Y$ is an upper bound on the number of columns of any 1-1 column map into $X \subseteq Y$ (by (3.7)Corollary), there exists a maximally 1-1 column map $V$ into $X$. By (3.5)Corollary, any such $V$ is necessarily a basis for $X$, hence $X$ is finitely generated and $\dim X \le \dim Y$. If now $X = Y$, then $\dim X = \dim Y$, while if $X \ne Y$, then there exists $y \in Y \backslash \operatorname{ran} V$, hence $[V, y]$ is a 1-1 column map (by (3.4)Lemma) into $Y$, therefore $\dim X < \#V + 1 \le \dim Y$. $\square$

Note the following important (nontrivial) part of (3.11)Proposition:

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

Here is an important refinement of the above (3.11)Proposition (to which it reverts if $Z$ is trivial).

**(3.13) Corollary:** If $Y$ and $Z$ are linear subspaces of the finite-dimensional vector space $X$, and $Y \cap Z = \{0\}$, then
$$\dim Y + \dim Z \le \dim X,$$
with equality if and only if
$$X = Y + Z := \{y + z : y \in Y, z \in Z\}.$$

**Proof:** Let $V$ and $W$ be bases for $Y$ and $Z$, respectively (which exist since $Y$ and $Z$, as linear subspaces of a finite-dimensional vector space, are themselves finite-dimensional, by (3.12)Corollary). By (3.10)Proposition, $[V, W]$ is 1-1 (since $\operatorname{ran} V \cap \operatorname{ran} W = \{0\}$ by assumption). It follows, with (3.7)Corollary, that
$$\dim Y + \dim Z = \#V + \#W = \#[V, W] \le \dim X,$$
with equality if and only if $[V, W]$ is onto $X$, i.e., if and only if
$$Y + Z = \operatorname{ran} V + \operatorname{ran} W = \operatorname{ran}[V, W] = X.$$
$\square$

We will use this corollary most often in the following form.

**(3.14)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; in particular, $Y \cap Z$ contains nonzero elements.

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

**(3.15) 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 (3.12)Corollary), hence $\operatorname{null} A$ has a basis, $V \in L(\mathbb{F}^n, \operatorname{null} A)$ say. By (3.6)Corollary, we can extend this to a basis $[V, U]$ for $\operatorname{dom} A$, with $U \in L(\mathbb{F}^r, \operatorname{dom} A)$ for some $r$. In particular, $[V, U]$ 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 $A[V, U]$ 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 $AUa = 0$, then $Ua \in \operatorname{null} A$, hence, since $V$ maps onto $\operatorname{null} A$, there is some $b$ so that $Ua = Vb$. This implies that $[V, U](b, -a) = 0$ and, since $[V, U]$ is 1-1, this shows that, in particular, $a = 0$. □

**Corollary:** If $Y$ and $Z$ are linear subspaces of the finite-dimensional vector space $X$, then

$$\dim(Y + Z) = \dim Y + \dim Z - \dim(Y \cap Z).$$

**Proof:** $Y \cap Z$ is a linear subspace of $X$, hence is finite-dimensional (by (3.12)Corollary), hence $Y \cap Z$ has a basis, $V$ say. Extend it, as we may (by (3.6)Corollary), to a basis $[U, V]$ of $Y$ and to a basis $[V, W]$ of $Z$, and consider the column map $A := [U, V, V, W]$. It has range equal to $\operatorname{ran}[U, V] + \operatorname{ran}[V, W] = Y + Z$ and has domain $\mathbb{F}^n$ with $n := \#[U, V, V, W] = \dim Y + \dim Z$. By the Dimension Formula, it is therefore sufficient to prove that $\dim \operatorname{null} A = \dim(Y \cap Z)$.

For this, assume that $[U, V, V, W](a, b, c, d) = 0$. Then $Z \supseteq \operatorname{ran} W \ni Wd = -[U, V, V](a, b, c) \in \operatorname{ran}[U, V] = Y$, hence $Wd \in Y \cap Z = \operatorname{ran} V$, i.e., $Wd = Ve$ for some $e$, or $[V, W](d, -e) = 0$, hence, in particular, $d = 0$. Since the situation is symmetric in $Y$ and $Z$, we also conclude that $a = 0$. It follows that $Vb + Vc = 0$ and, since $V$ is 1-1, this is possible if and only if $c = -b$. Consequently, $\operatorname{null} A = \operatorname{ran} B$, with $B : \mathbb{F}^m \to \mathbb{F}^n : b \mapsto (0, b, -b, 0)$ and with $m := \#V$. Since $B$ is 1-1 and $V$ is a basis for $Y \cap Z$, it follows that $\dim \operatorname{null} A = m = \dim(Y \cap Z)$, which is exactly what we needed. □

**(3.16) 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.

In particular, (iii) implies (1.15).

**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. □

**(3.17) Lemma:** Let $X, Y$ be vector spaces, and assume that $X$ is finite-dimensional. Then $\dim X = \dim Y$ if and only if there exists an invertible $A \in L(X, Y)$.

**Proof:** Let $n := \dim X$. Since $n < \infty$, there exists an invertible $V \in L(\mathbb{F}^n, X)$ (, 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 $Y$. □

**(3.18) Example: The dimension of $\Pi_k$.** Recall that $\Pi_k$ denotes the collection of all polynomials of degree $\le k$ in one real argument. This means, by definition, that $\Pi_k$ is the range of the column map

$$V := [()^0, ()^1, \ldots, ()^k] : \mathbb{R}^{k+1} \to \mathbb{R} : a \mapsto \sum_{j=0}^{k} ()^j a_j.$$

In particular, $\dim \Pi_k \le k + 1$, by (3.8)Corollary.

On the other hand, with

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

we have $\Lambda V = \operatorname{id}$, hence $V$ is 1-1, in addition to being onto $\Pi_k$, therefore $V$ is a basis for $\Pi_k$. This says that

$$\dim \Pi_k = k + 1.$$

An analogous argument shows that

$$\dim \Pi_k(\mathbb{R}^d) = \#\{\alpha \in \mathbb{Z}_+^d : |\alpha| := \sum_{j=1}^{d} \alpha_j \le k\},$$

and the latter number can be shown to equal $\binom{k+d}{d}$. □

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

Recall from (1.2) 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 $T = \underline{m} \times \underline{n} := \{(i, j) : i = 1, \ldots, m; j = 1, \ldots, n\}$.

**(3.19) Proposition:** If $\#T :=$ number of elements of $T$ is finite, then $\dim \mathbb{F}^T = \#T$.

**Proof:** Since $T$ is finite, we can order its elements,

$$T =: \{t_1, \ldots, t_n\},$$

say. Correspondingly, for $j = 1, \ldots, n$, let

$$f_j : T \to \mathbb{F} : t \mapsto \begin{cases} 1, & t = t_j; \\ 0, & \text{otherwise.} \end{cases}$$

Let $F := [f_1, f_2, \ldots, f_n]$. Then $F \in L(\mathbb{F}^n, \mathbb{F}^T)$, and

$$(3.20) \qquad (Fa)(t_i) = \sum_{j=1}^{n} f_j(t_i) a_j = a_i, \qquad i = 1, \ldots, n.$$

This implies that $RF = \mathrm{id}$, with

$$R : \mathbb{F}^T \to \mathbb{F}^n : f \mapsto (f(t_1), f(t_2), \ldots, f(t_n)).$$

Further, (3.20) also implies that, for any $f \in \mathbb{F}^T$, $(FRf)(t_i) = f(t_i)$ for all $i$, i.e., $FRf = f$. Consequently, also $FR = \mathrm{id}$. In particular, $F$ is invertible, hence a basis for $\mathbb{F}^T$, and therefore $\dim \mathbb{F}^T = n = \#T$. □

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

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

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

of dimension equal to $\dim \mathbb{F}^S = \#S$. 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 (3.11)Proposition. □

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

**Proof:** In this case, $\mathbb{F}^{m\times n} = \mathbb{F}^T$ with $T = \underline{m} \times \underline{n} := \{(i,j) : i = 1, \ldots, m; j = 1, \ldots, n\}$, hence $\#T = mn$. □

**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 $\widehat{A} := W A V^{-1} \in \mathbb{F}^{m\times n}$, with $V$ a basis for $X$ and $W$ a basis for $Y$. This sets up a map

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

and this map is linear and invertible (indeed, its inverse is the map $\mathbb{F}^{m\times n} \to L(X, Y) : B \mapsto V B W^{-1}$). Consequently, by (3.17)Lemma, $L(X, Y)$ and $\mathbb{F}^{m\times n}$ have the same dimension. □

## Elimination in vector spaces

We know from (3.8)Corollary that any onto column map $W \in L(\mathbb{F}^m, X)$ can be thinned to a basis. The proof relies on the following algorithm (in the proof of (3.6)Corollary, but taken here with $V = [\,]$):

**Basis Selection Algorithm:** $U := [\,]$; **for** $w \in W$ **do: if** $w \notin \operatorname{ran}[V, U]$, **then** $U \leftarrow [U, w]$; **endif; enddo**

whose output, $U$, is a basis for $X$ since it is 1-1, and $W \subset \operatorname{ran} U$, therefore $X = \operatorname{ran} W \subseteq \operatorname{ran} U \subseteq X$, so $U$ is onto $X$.

The efficient way to answer the question whether or not $w \in \operatorname{ran} U$ is via elimination, applied to the column map $W$. For, by (2.10)Proposition, the $k$th unknown for $A? = 0$ is free if and only if $A(:,k) \in \operatorname{ran} A(:,[1{:}k-1])$. Thus, the above algorithm would adjoin a column of $W$ to $U$ exactly when the corresponding unknown in $W? = 0$ is bound. In particular, the column map $U$ output by the above algorithm is exactly the column map $W(:,\texttt{bound})$, with `bound` as determined by elimination applied to $W? = 0$.

All this is immediate in case $X \subset \mathbb{F}^r$ for some $r$, for, then $W$ is just an $r \times m$-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.

If $T$ is not finite, then we may not be able to determine in finite time whether or not a given unknown 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

$$W = VA,$$

with $V$ some 1-1 column map, hence $A$ a matrix. Under these circumstances, the $k$th unknown for $W? = 0$ is free if and only if the $k$th unknown for $A? = 0$ is free, and the latter we can determine by elimination applied to $A$.

Indeed, if $A = A(:\,\texttt{bound})\Lambda$ is the factorization for $A$ produced by elimination, then $W = VA(:,\texttt{bound})\Lambda$. Now, $W = VA$ implies that $W(:,\texttt{bound}) = VA(:,\texttt{bound})$ and this, as a product of 1-1 maps, must be 1-1. Consequently,

$$W = W(:\,\texttt{bound})\Lambda,$$

with $W(:,\texttt{bound})$ 1-1, hence $\Lambda$ is the interesting part of $\operatorname{rref}(W)$.

As an example, assume that $W = [w_1, w_2, w_3]$, with $w_j : \mathbb{F} \to \mathbb{F} : t \mapsto (t - j)$, $j = 1, 2, 3$. Hence,

$$W = VA, \quad \text{with } V := [()^1, ()^0], \ \ A := \begin{bmatrix} 1 & 1 & 1 \\ -1 & -2 & -3 \end{bmatrix},$$

and we know (e.g., from (3.18)) that $V$ is 1-1.

We compute

$$\operatorname{rref}(\begin{bmatrix} 1 & 1 & 1 \\ -1 & -2 & -3 \end{bmatrix}) = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \end{bmatrix}.$$

Therefore,

$$W = W(:,\texttt{bound})\Lambda = [w_1, w_2] \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 2 \end{bmatrix}.$$

## Elimination as factorization

The description (2.2) 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 workarray $B$ 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) which are never going to be used as pivot rows. With that done, we have, at the beginning of the algorithm,

$$A = PB$$

for some permutation matrix $P$, and all the work in the algorithm consists of repeatedly subtracting some multiple $\alpha$ 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

$$B \leftarrow E_{i,h}(-\alpha)B$$

with $i > h$. Since, by (1.16), $E_{i,h}(-\alpha)^{-1} = E_{i,h}(\alpha)$, this implies that

$$A = PLU,$$

with $L$ the product of all those elementary matrices $E_{i,h}(\alpha)$ (in the appropriate order), and $U$ the final state of the workarray $B$. Specifically, $U$ is in row-echelon form (as defined in (2.6)); in particular, $U$ is upper triangular.

Each $E_{i,h}(\alpha)$ is **unit lower triangular**, i.e., of the form $\mathrm{id} + N$ with $N$ **strictly lower triangular**, i.e.,

$$N(r,s) \neq 0 \quad \Longrightarrow \quad r > s.$$

For, because of the initial ordering of the rows in $B$, only $E_{i,h}(\alpha)$ 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 $m$ unknowns are bound, hence the remaining unknowns 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 $A? = C$, there is no need to subject all of $[A, C]$ to the elimination algorithm. If elimination just applied to $A$ gives the factorization

$$A = PLU$$

for an invertible $A$, and $C = [b]$ for some $b \in \mathbb{F}^m$, then we can find the unique solution $x$ to the equation $A? = b$ by the two-step process:

$$c \leftarrow L^{-1}P^{-1}b$$
$$x \leftarrow U^{-1}c$$

and these two steps are easily carried out. The first step amounts to subjecting the rows of the matrix $[b]$ to all the row operations (including reordering) used during elimination applied to $A$. The second step is handled by the Backsubstitution Algorithm (2.3), with input $B = [U, c]$, $p = (1, 2, \ldots, m, 0)$, and $y = (0, \ldots, , 0, -1)$.

Once it is understood that the purpose of elimination for solving $A? = b$ is the factorization of $A$ into a product of "easily" invertible factors, then it is possible to seek other factorizations which 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 $Q$ 'orthonormal', i.e., $Q'Q = \mathrm{id}$. 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

$$H_w := E_{w,w}(-2/w'w) = \mathrm{id} - \frac{2}{w'w}ww',$$

and are easily seen to be **self-inverse** or **involutory** (i.e., $H_wH_w = \mathrm{id}$), **hermitian** (i.e., $H'_w = H_w$), hence **unitary** (i.e., $H'_wH_w = \mathrm{id} = H_wH'_w$).

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 both a PLU factorization $A = PLU$ and in a QR factorization $A = QR$ can be shown to be proportional to the 'condition numbers' of the factors. Since $Q$ is o.n., $\kappa(Q) = 1$ and $\kappa(R) = \kappa(A)$, while, for a PLU factorization $A = PLU$, only the permutation matrix, $P$, is o.n., and $\kappa(L)$ and $\kappa(U)$ can be quite large.

## 4. The inverse of a basis

### A formula for the coordinate map

Let $V \in L(\mathbb{F}^n, Y)$ be a basis for the vector space $Y$. How do we find the coordinates

$$a = V^{-1}y \tag{4.1}$$

for given $y \in Y$?

Offhand, we solve the linear system $V? = y$ for $a$. Since $V$ is a basis, we know that this linear system has exactly one solution. But that is not the same thing as having a concrete formula for $a$ in terms of $y$.

If $Y = \mathbb{F}^n$, then $V^{-1}$ is a matrix; in this case, (4.1) is an explicit formula. However, if $Y$ is only a linear subspace of $\mathbb{F}^n$, then (4.1) is merely a formal expression.

We now give a recipe for getting an explicit formula. It does require you to know *some* onto linear map $\Lambda$ from $Y$ to $\mathbb{F}^n$. However, the recipe works with any such linear map.

So, assume that $y = Va$ and let $\Lambda \in L(Y, \mathbb{F}^n)$ be onto. Then $\Lambda y = \Lambda V a$, and $\Lambda V$ is a linear map from $\mathbb{F}^n$ to $\mathbb{F}^n$, i.e., a square matrix, and is also onto (since $V$ is a basis, hence onto), therefore invertible, by (2.13)Theorem. Consequently,

$$a = (\Lambda V)^{-1} \Lambda y,$$

the promised formula.

> **(4.2) Proposition.** If $V \in L(\mathbb{F}^n, Y)$ is a basis for the vector space $Y$, and $\Lambda \in L(Y, \mathbb{F}^n)$ is onto, then
>
> $$V^{-1} = (\Lambda V)^{-1} \Lambda.$$
>
> In particular, the coordinates $a = V^{-1}y$ for any $y \in Y$ are given by the formula
>
> $$a = (\Lambda V)^{-1} \Lambda y.$$

**(4.3) Example: Polynomial interpolation** You have learned to think of polynomials as linear combinations of monomials. You are (unfortunately) not happy until you have written a polynomial $p \in \Pi_{<n}$ in the form $Va$, with $V$ the **power basis**, i.e.,

$$V := [()^0, ()^1, \ldots, ()^{n-1}] := [\ \ , t, \ldots, t^{n-1}].$$

If all you know about your polynomial $p$ are its values $(p(t_i) : i = 1, \ldots, n)$ at $n$ distinct points, $t_1 < \cdots < t_n$ say, then you do know $\Lambda p$, with

$$\Lambda : \Pi_{<n} \to \mathbb{R}^n : f \mapsto (f(t_i) : i = 1, \ldots, n).$$

Now, for $i = 1, \ldots, n$, $\Pi_{<n}$ contains the **Lagrange polynomial**

$$\ell_i : t \mapsto \prod_{j \neq i} \frac{t - t_j}{t_i - t_j}.$$

Notice that

$$\ell_i(t_r) = \begin{cases} 1 & i = r; \\ 0 & \text{otherwise.} \end{cases}$$

In other words, the column map

$$\ell := [\ell_1, \ell_2, \ldots, \ell_n]$$

into $\Pi_{<n}$ satisfies

$$\Lambda\ell = \mathrm{id}_n.$$

This implies that $\Lambda$ is onto, hence, by (4.2), the polynomial coefficients for $p$ are given by

$$a = (\Lambda V)^{-1}\Lambda p.$$

The matrix $\Lambda V = (t_i^j : i = 1,\ldots,n; j = 0,\ldots,n-1)$ is called the **Vandermonde matrix** for the point sequence $(t_i : i = 1,\ldots,n)$. The Vandermonde matrix is notorious for giving trouble if one actually attempts to compute the coordinate vector $a$ as the solution to the linear system

$$(\Lambda V)? = \Lambda p.$$

This trouble is due to the fact that the power basis is 'illconditioned'; see xxx .

Incidentally, we have just proved the *easy part* of the Fundamental Theorem of Algebra; we have proved that a polynomial of degree $< n$ cannot have more than $n-1$ distinct zeros without being identically zero. For, we have proved that the linear map $p \mapsto (p(t_i) : i = 1,\ldots,n)$ is 1-1 on $\Pi_{<n}$ in case $t_i \neq t_j$ for $i \neq j$. This implies that any polynomial $p \in \Pi_{<n}$ which vanishes at more than $n-1$ distinct points must be the zero polynomial. □

### Interpolation and linear projectors

As the example (4.3) already intimated, our formula in (4.2) for the inverse of a basis $V \in L(\mathbb{F}^n, Y)$ can be much more than that. It is useful for *interpolation*, in the following sense. Assume that, in fact, $\Lambda$ is a **data map**, i.e., a linear map *to* $\mathbb{F}^n$ from some vector space $X$ containing $Y$, and that $\Lambda$ maps $Y$ *onto* $\mathbb{F}^n$. Note that, by (3.7)Corollary(iii), this last assumption is equivalent to having $\Lambda$ be 1-1 on $Y$ since $\dim Y = n = \dim \operatorname{tar} \Lambda$. In other words, under this assumption, $\Lambda$ maps $Y = \operatorname{ran} V$ *1-1* onto $\mathbb{F}^n$.

What does this say? It says that, for each $b \in \mathbb{F}^n$, there is exactly one $y \in Y$ for which $\Lambda y = b$. Equivalently, it says that, *for each $x \in X$, there is exactly one $y \in Y$ which* **agrees with** $x$ **at** $\Lambda$ *in the sense that*

$$\Lambda y = \Lambda x.$$

For example, if $Y = \Pi_{<n}$ and $\Lambda : f \mapsto (f(t_i) : i = 1,\ldots,n)$, with $t_1 < \cdots < t_n$, then, by (4.3)Example, for arbitrary $f : \mathbb{R} \to \mathbb{R}$, there is exactly one polynomial $p$ of degree $< n$ for which $p(t_i) = f(t_i)$, $i = 1,\ldots,n$.

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 $\Pi_{<n}$, one picks two points, $t_1 < t_2$, and two nonnegative integers $r$ and $s$ with $r+1 + s+1 = n$, and defines

$$\Lambda : f \mapsto (f(t_1), Df(t_1), \ldots, D^r f(t_1), f(t_2), Df(t_2), \ldots, D^s f(t_2)).$$

Now the requirement that $\Lambda p = \Lambda f$ amounts to looking for $p \in \Pi_{<n}$ which agrees with $f$ in the sense that $p$ and $f$ have the same derivative values of order $0,1,\ldots,r$ at $t_1$ and the same derivative values of order $0,1,\ldots,s$ at $t_2$. □

**Example.** Recall from Calculus the bivariate **Taylor series**

$$f(x,y) = f(0) + D_1 f(0)\,x + D_2 f(0)\,y + \big(D_1^2 f(0)x^2 + D_1D_2 f(0)xy + D_2D_1 f(0)yx + D_2^2 f(0)y^2\big)/2 + h.o.t.$$

In particular, for any smooth function $f$, the quadratic polynomial

$$p : (x,y) \mapsto f(0) + D_1 f(0)\,x + D_2 f(0)\,y + \big(D_1^2 f(0)x^2 + 2D_1D_2 f(0)xy + D_2^2 f(0)y^2\big)/2$$

is the unique quadratic polynomial which matches the information about $f$ given by the data map

$$\Lambda : f \mapsto (f(0), D_1 f(0), D_2 f(0), D_1^2 f(0), D_1D_2 f(0), D_2^2 f(0)).$$

□

**Example.** When dealing with Fourier series, **Fourier series** one uses the data map

$$\Lambda : f \mapsto (\int_{-\pi}^{\pi} f(t) \operatorname{cis}(jt)\, \mathrm{d}t : j = 0, 1, 2, \ldots, N),$$

with cis standing for 'sine and cosine'. One looks for a **trigonometric polynomial**

$$p = [\operatorname{cis}(j\cdot) : j = 0, 1, \ldots, N]a$$

which satisfies $\Lambda p = \Lambda f$, and finds it in the **truncated Fourier series** for $f$. □

We continue with our assumption that $\Lambda$ maps $Y = \operatorname{ran} V$ onto $\mathbb{F}^n$. Since $V$ is onto, this makes the *square* matrix $\Lambda V$ onto, hence invertible. This provides the following pretty formula

$$x = Px := V(\Lambda V)^{-1}\Lambda x \tag{4.4}$$

for the interpolant $y \in Y$ to given $x \in X$ with respect to the data map $\Lambda$. The linear map $P := V(\Lambda V)^{-1}\Lambda$ so defined on $X$ is very special:

**(4.5) Proposition.** Let $V$ be a basis for the $n$-dimensional linear subspace $Y$ of the vector space $X$, and let $\Lambda \in L(X, \mathbb{F}^n)$ be 1-1 on $Y$, i.e., assume that $\Lambda V$ is invertible. Then $P := V(\Lambda V)^{-1}\Lambda$ is a linear map on $X$ with the following properties:

(i) $P$ is the identity on $Y = \operatorname{ran} V$. Indeed

$$PV = V(\Lambda V)^{-1}\Lambda\, V = V \operatorname{id} = V,$$

hence $P(Va) = Va$ for all $a \in \mathbb{F}^n$.

(ii) $\operatorname{ran} P = \operatorname{ran} V = Y$. Indeed, since $P = VA$ for some $A$, we have that $\operatorname{ran} P \subseteq \operatorname{ran} V$, while $PV = V$ implies that $\operatorname{ran} P \supseteq \operatorname{ran} V$.

(iii) $P$ is a **projector** or **idempotent**, i.e., $PP = P$. Indeed, by (i) and (ii), $P$ is the identity on its range.

(iv) $\operatorname{null} P = \operatorname{null} \Lambda$. Indeed, the fact that $P = A\Lambda$ for some $A$ implies that $\operatorname{null} P \supseteq \operatorname{null} \Lambda$, while also

$$\Lambda P = \Lambda\, V(\Lambda V)^{-1}\Lambda = \operatorname{id}_n \Lambda = \Lambda,$$

hence also $\operatorname{null} P \subseteq \operatorname{null} \Lambda$.

Any linear projector $P$ on the vector space $X$ provides a *splitting* of an arbitrary $x \in X$ into a sum

$$x = Px + (\operatorname{id} - P)x,$$

with $Px \in \operatorname{ran} P$; this much is trivial. However,

$$P(\operatorname{id} - P) = P - PP = P - P = 0,$$

hence $(\operatorname{id} - P)x$ is an element of $\operatorname{null} P$. Finally, if also

$$x = y + z$$

for some $y \in \operatorname{ran} P$ and some $z \in \operatorname{null} P$, then

$$Px = Py + Pz = y + 0 = y,$$

since $y \in \operatorname{ran} P$, hence $Px = x$, and $z \in \operatorname{null} P$ by construction. Therefore, necessarily $z = (\operatorname{id} - P)x$.

We have proved the following.

**(4.6) Theorem.** Every linear projector $P$ on a vector space $X$ provides a splitting of $X$ into the **direct sum**

$$X = (\operatorname{ran} P) \dot{+} (\operatorname{null} P)$$

in the sense that every $x \in X$ can be written in exactly one way as a sum

$$x = y + z,$$

with $y \in \operatorname{ran} P$ (hence $Py = y = Px$), and $z \in \operatorname{null} P$ (hence $(\operatorname{id} - P)z = z = (\operatorname{id} - P)x$).

Moreover, any decomposition $X = Y \dot{+} Z$ of a vector space $X$ into the direct sum of two subspaces $Y$ and $Z$ arises in this way.

**(4.7) Example.** This notion of 'direct sum' is a generalization of the familiar notion of a coordinate system, as the following example shows.

Take $X = \mathbb{R}^2$, and let $v, w \in \mathbb{R}^2$ be arbitrary vectors with $w'v \neq 0$. Consider the linear map

$$P := \frac{vw'}{w'v} : x \mapsto v\frac{w'x}{w'v}.$$

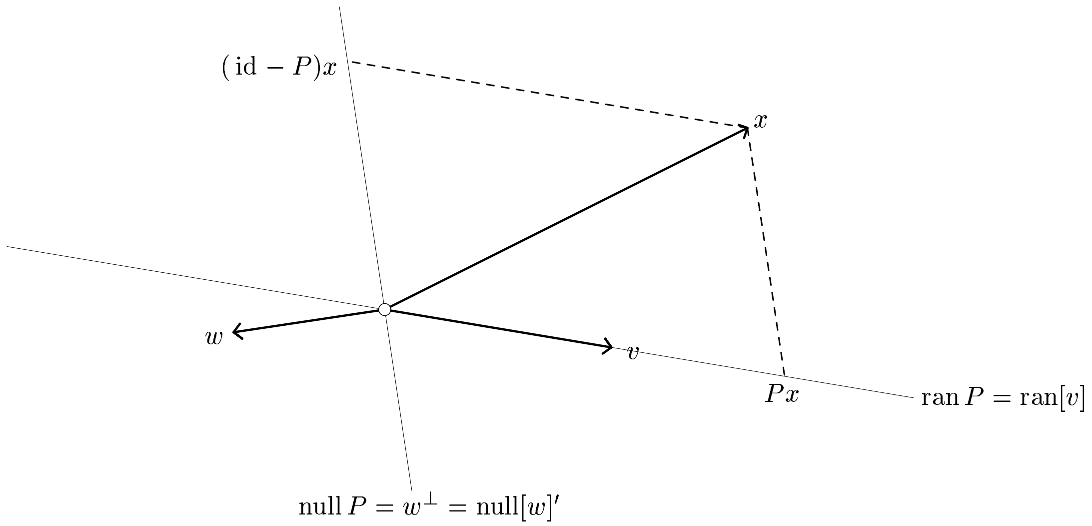
We readily compute that

$$PP = \frac{vw'}{w'v}\frac{vw'}{w'v} = \frac{v\,w'v\,w'}{(w'v)\,(w'v)} = \frac{vw'}{w'v} = P,$$

hence $P$ is a linear projector. Its range equals $\operatorname{ran}[v]$, and this is one-dimensional (since $v \neq 0$). Since, in this example, $\operatorname{dom} P = X$ is two-dimensional, it follows (from (3.15)Dimension Formula) that also $\operatorname{null} P$ is one-dimensional. Both $\operatorname{ran} P$ and $\operatorname{null} P$ are straight lines through the origin. The first one is given to us explicitly, as $\operatorname{ran}[v]$. The second one is given to us implicitly, as

$$w^\perp := \{x \in X : w'x = 0\}.$$

The two lines have only the origin in common since $x \in \operatorname{ran} P \cap \operatorname{null} P$ implies that $x = v\alpha$ for some scalar $\alpha$, therefore $0 = w'x = w'v\alpha$ and this implies that $\alpha = 0$ since $w'v \neq 0$ by assumption.



(4.8)Figure. A direct sum decomposition provides a kind of coordinate system.

We can locate the two summands in the split

$$x = Px + (\,\mathrm{id} - P)x$$

graphically: To find $Px$, draw the line through $x$ parallel to $\operatorname{null} P$; its unique intersection with $\operatorname{ran} P = \operatorname{ran}[v]$ is $Py$. The process of locating $(\,\mathrm{id} - P)x$ is the same, with the roles of $\operatorname{ran} P$ and $\operatorname{null} P$ reversed: Now draw the line through $x$ parallel to $\operatorname{ran} P$; its unique intersection with $\operatorname{null} P$ is the element $(\,\mathrm{id} - P)x$.

Finally, if you take $V := [v]$, $\Lambda := [w]' = w'$, then you recognize the condition $w'v \neq 0$ as saying that $\Lambda V$ is 1-1, hence you can infer from (4.5) that $P$ is the linear projector which associates with $y \in \mathbb{R}^2$ the unique element $y = Px \in \operatorname{ran}[v]$ for which $w'y = w'x$. □

**(4.9) Corollary.** If $V \in \mathbb{F}^{m\times n}$ is 1-1 and $W \in \mathbb{F}^{m\times n}$ is such that $\operatorname{ran} V \cap \operatorname{null} W' = \{0\}$, then $P := V(W'V)^{-1}W'$ is well-defined; it is the *unique* linear projector $P$ with

(4.10) $$\operatorname{ran} P = \operatorname{ran} V, \qquad \operatorname{null} P = \operatorname{null} W'.$$

In particular, then $W'$ is onto, and

(4.11) $$\mathbb{F}^m = \operatorname{ran} V \dotplus \operatorname{null} W'.$$

**Proof:** The assumption that $\operatorname{ran} V \cap \operatorname{null} W' = \{0\}$ implies that $\Lambda := W'$ is 1-1 on $\operatorname{ran} V$. By (4.5)Proposition, this implies that $P := V(W'V)^{-1}W'$ is a well-defined linear projector, and that it satisfies (4.10), hence, by (4.6)Theorem, we get the direct-sum decomposition (4.11).

Assume that also $Q$ is a linear projector on $\mathbb{F}^m$, with $\operatorname{ran} Q = \operatorname{ran} V = \operatorname{ran} P$ and $\operatorname{null} Q = \operatorname{null} W' = \operatorname{null} P = \operatorname{ran}(\,\mathrm{id}_m - P)$. Then, for any $x \in \mathbb{F}^m$,

$$Qx = Q(Px + (x - Px)) = QPx + Q(\,\mathrm{id}_m - P)x = Px,$$

since $Px \in \operatorname{ran} Q$ and $(\,\mathrm{id}_m - P)x \in \operatorname{null} P = \operatorname{null} Q$. In other words, $Q = P$.

Finally, since $W'$ is 1-1 on $\operatorname{ran} V$ by assumption, it follows from (3.15)Dimension Formula that $n = \dim W'(\operatorname{ran} V)$. However, $\dim W'(\operatorname{ran} V) \leq \dim \operatorname{ran} W' \leq \dim \operatorname{tar} W' = \dim \mathbb{F}^n = n$. Therefore, $W'$ is necessarily onto. □

## Orthogonal projectors and least-squares approximation

The formula in (4.2) for the coordinate map $V^{-1}$ for a given basis $V$ for the vector space $Y$ requires some data map, i.e., some linear map $\Lambda$ into $\mathbb{F}^n$, which is 1-1 on $Y$. If $Y \subset \mathbb{F}^m$ for some $m$, hence $V \in \mathbb{F}^{m\times n}$ is a matrix, then there are many $\Lambda$ available, namely $\Lambda = W'$, with $W$ any basis for $Y$ (including $V$ itself), as the following lemma readily shows.

**(4.12) Lemma.** If the range of the 1-1 matrix $V$ is contained in the range of some matrix $W$, then $W'V$ is 1-1, i.e., $W'$ is 1-1 on $\operatorname{ran} V$.

**Proof:** Assume that $W'Va = 0$ and let $b := Va$. Then $b \in \operatorname{ran} V \subseteq \operatorname{ran} W$, hence we must have $b = Wc$ for some vector $c$. Therefore,

$$0 = c'0 = c'\,W'Va = (Wc)'Va = b'b.$$

This implies that $Va = b = 0$, therefore that $a = 0$, since $V$ is assumed to be 1-1. □

This slick proof has two important ingredients. It uses the special case

$$c'W' = (Wc)'$$

of the fact that, for any two matrices $A$ and $B$ for which $AB$ makes sense,

$$(AB)' = B'A'. \tag{4.13}$$

It also uses the fact that the scalar product

$$\langle x, y\rangle := y'x = x{*}\overline{y} = \sum_j x_j\overline{y_j}$$

is **definite**, i.e., $b'b = \langle b, b\rangle = 0$ if and only if $b = 0$. This definiteness is due to the fact that we are using $x{*}\overline{y}$ instead of $x{*}y$; e.g., $(1, \mathrm{i}){*}(1, \mathrm{i}) = 1 - 1 = 0$. This is ultimately the reason that we follow `MATLAB` here and concentrate on the *conjugate* transpose, $A'$, rather than on the transpose. Of course, if $\mathbb{F} = \mathbb{R}$, then the two notions agree. However, work on eigenvalues requires consideration of *complex* scalars and, 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.

It follows that, for any basis $V$ of the linear subspace $Y$ of $\mathbb{F}^m$, the matrix $(V'V)^{-1}V'$ is well-defined, hence provides a formula for $V^{-1}$. It is, in effect, the formula used in `MATLAB` when calculating

```
V\
```

for a matrix `V` that is merely 1-1.

We conclude that, with $V$ a basis for the linear subspace $Y$ of $\mathbb{F}^m$, the linear projector

$$P_V := V(V'V)^{-1}V'$$

is well-defined. Moreover, by (4.5), $\operatorname{null} P_V = \operatorname{null} V' = \{x \in \mathbb{F}^m : V'x = 0\}$. Since $y \in \operatorname{ran} P_V = \operatorname{ran} V$ is necessarily of the form $y = Va$, it follows that, for any $y \in \operatorname{ran} P_V$ and any $z \in \operatorname{null} P_V$,

$$y'z = (Va)'z = a'V'z = 0.$$

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.

**Definition:** $u \perp v \ :=\ \ \langle u, v\rangle = 0 \ \ =: $ $u$ and $v$ are **orthogonal** or **perpendicular** to each other.
More generally, for any $F, G \subseteq \mathbb{F}^n$, we write $F \perp G$ to mean that, $\forall f \in F, g \in G,\ f \perp g$.
Note that $u \perp v$ iff $v \perp u$ since $\langle u, v\rangle = \overline{\langle v, u\rangle}$.

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 $\operatorname{ran} V$. This orthogonality, as we show in a moment, has the wonderful consequence that, for any $x$, $P_Vx$ is the unique element of $\operatorname{ran} P_V = \operatorname{ran} V$ which is closest to $x$ in the sense of the **Euclidean norm** or **2-norm**

$$\|\cdot\| : \mathbb{F}^n \to \mathbb{R} : x \mapsto \sqrt{x'x}.$$

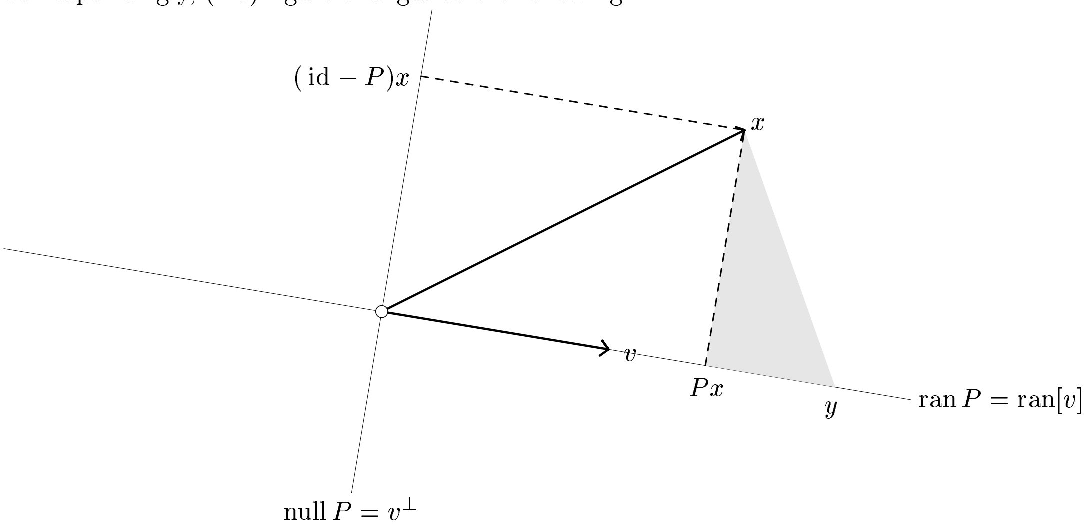
Thus, whether or not $b \in \mathbb{F}^m$ is in $\operatorname{ran} V$, the coordinate vector $a = (V'V)^{-1}V'b$ supplied by our formula gives the coordinates of the point in $\operatorname{ran} V$ closest to $b$. If $b \in \operatorname{ran} V$, then this is of course $b$ itself.

**(4.14) Example (cont.).** In (4.7)Example, the choice $\Lambda = V'$ amounts to choosing $w = v$. Now $P$ becomes $P = vv'/v'v$, and, correspondingly,

$$Px = v\frac{v'x}{v'v},$$

which we recognize as the standard formula for the orthogonal projection of the vector $x$ onto the line spanned by the vector $v$.

Correspondingly, (4.8)Figure changes to the following.




(4.15)Figure. If $x - Px$ is perpendicular to $\operatorname{ran} P$, then $Px$ is the closest point to $x$ from $\operatorname{ran} P$ since then, for any $y \in \operatorname{ran} P$, $\|x-y\|^2 = \|x - Px\|^2 + \|y - Px\|^2$.

□

The *proof* that, for any $x \in \mathbb{F}^m$, $P_V x$ is the unique element of $\operatorname{ran} V$ closest to $x$ in the Euclidean sense is based on nothing more than the following little calculation.

$$\|u+v\|^2 = \langle u+v, u+v\rangle = \|u\|^2 + \langle u, v\rangle + \langle v, u\rangle + \|v\|^2.$$

Since $\langle v, u\rangle = \overline{\langle u, v\rangle}$, this proves

**(4.16) Pythagoras:** $\|u+v\|^2 = \|u\|^2 + \|v\|^2$ if and only if $u \perp v$.

Since, for any $y \in Y$, $x - y = (x - P_V x) + (P_V x - y)$, while $(x - P_V x) \in \operatorname{null} P_V \perp \operatorname{ran} P_V = Y \ni (P_V x - y)$ we conclude that

(4.17) $$\|x-y\|^2 = \|x - P_V x\|^2 + \|P_V x - y\|^2.$$

Here, the first term on the right is *independent of* $y$. This shows that $\|x - y\|$ is uniquely minimized over $y \in Y$ by the choice $y = P_V x$, as we claimed.

Here is the formal statement.

**(4.18) Theorem:** For any basis $V \in \mathbb{F}^{m\times n}$ for the linear subspace $Y$ of $\mathbb{F}^m$, the linear map

$$P_V = V(V'V)^{-1}V'$$

is the orthogonal projector onto $Y$ in the sense that

$$Y = \operatorname{ran} P_V \perp \operatorname{null} P_V.$$

In other words, for any $x \in \mathbb{F}^m$, $P_V x \in Y$ and $x - P_V x \perp Y$. Therefore,

$$\forall\{x \in \mathbb{F}^m,\ y \in Y\} \quad \|x-y\| \ge \|x - P_V x\|,$$

with equality if and only if $y = P_V x$.

In MATLAB, the vector $P_V b$ is computed as `V*(V\b)`, in line with the fact mentioned earlier that the action of the matrix $(V'V)^{-1}V'$ is provided by the operator `V\`, i.e., (up to roundoff and for any vector `b`) the three vectors

```
c1 = V\b, c2 = inv(V'*V)*V'*b, c3 = (V'*V)\(V'*b)
```

are all the same. However, the first way is preferable since it avoids actually forming the matrix `V'*V` and, therefore, is less prone to roundoff effects.

□

Incidentally, by choosing $y = 0$ in (4.17), – legitimate since $\operatorname{ran} V$ is a linear subspace, – we find the following very useful fact.

**(4.19) Proposition:** For any 1-1 column map $V$ into $\mathbb{F}^m$ and any $x \in \mathbb{F}^m$,

$$\|x\| \ge \|P_V x\|,$$

with equality if and only if $x = P_V x$, i.e., if and only if $x \in \operatorname{ran} V$.

This says that $P_V$ *strictly reduces norms*, except for those elements which it doesn't change at all.

Since, for any nonzero $y \in \mathbb{F}^n$, $V := [y/\|y\|]$ is an o.n. basis for $\operatorname{ran}[y]$, we conclude that, for any $x \in \mathbb{F}^n$, $P_V x = y\langle x, y\rangle/\|y\|^2$ has norm smaller than $\|x\|$ unless $x = P_V x$. In other words

$$|\langle x, y\rangle| \le \|x\|\|y\| \tag{4.20}$$

with equality if and only if $[x, y]$ is not 1-1. This is the famous **Cauchy(-Bunyakovski-Schwarz) Inequality**, and it holds trivially in case $y = 0$. Be sure to remember not only the *in*equality, but also exactly when it is an *equal*ity.

### $\operatorname{ran} A$ and $\operatorname{null} A'$ form an orthogonal direct sum for $\operatorname{tar} A$.

The two basic linear subspaces associated with a matrix $A \in \mathbb{F}^{m\times n}$ are its range $\operatorname{ran} A$ and its kernel or nullspace $\operatorname{null} A$. However, it turns out to be more natural to consider the range of $A$ and the nullspace of the (conjugate) transpose $A'$ of $A$ together.

The connection between these two subspaces turns on the following basic observation, used already earlier.

**(4.21) Observation.** For any $A \in \mathbb{F}^{m\times n}$ and any $x \in \mathbb{F}^n, y \in \mathbb{F}^m$, $\langle Ax, y\rangle = \langle x, A'y\rangle$.

**Proof:** $\langle Ax, y\rangle = [y]'[Ax] = [y]'A[x] = (A'[y])'[x] = [A'y]'[x] = \langle x, A'y\rangle$. □

In particular, $x \perp \operatorname{ran} A$ if and only if $y \perp A'x$ for all $y$. However, one of the candidate $y$ here is the vector $A'x$. Since only the zero vector can be perpendicular to itself, it follows that

$$x \perp \operatorname{ran} A \iff x \in \operatorname{null} A'.$$

With the standard notation

$$M^\perp := \{x : x \perp M\}$$

for the **orthogonal complement of** $M$, we get the following.

**(4.22) Proposition.** For any matrix $A$, $(\operatorname{ran} A)^\perp = \operatorname{null} A'$.

**(4.23) Corollary.** For any matrix $A \in \mathbb{F}^{m\times n}$, $\mathbb{F}^m$ is the **orthogonal** sum $\mathbb{F}^m = \operatorname{ran} A \oplus \operatorname{null} A'$, i.e., the two summands are also orthogonal to each other.

**Proof:** Let $V$ be any basis for $\operatorname{ran} A$. By (4.6)Theorem and (4.9)Corollary,

$$\mathbb{F}^m = \operatorname{ran} V \dotplus \operatorname{null} V',$$

while, by choice of $V$, $\operatorname{ran} V = \operatorname{ran} A$, and so, by (4.22), $\operatorname{null} V' = (\operatorname{ran} V)^\perp = (\operatorname{ran} A)^\perp$. □

### Orthonormal matrices

The formula

$$P_V = V(V'V)^{-1}V'$$

for the orthogonal projector onto the range of the 1-1 matrix $V$ becomes particularly simple in case

$$V'V = \mathrm{id}; \tag{4.24}$$

it reduces to

$$P_V = VV'.$$

In this section, we show that every linear subspace of $\mathbb{F}^m$ has bases $V$ satisfying (4.24).

We begin with the following basic observation. If $Q = [q_1, q_2, \ldots, q_p] \in \mathbb{F}^{m\times k}$, $W = [w_1, w_2, \ldots, w_n] \in \mathbb{F}^{m\times n}$, then $W'Q$ is the matrix of inner products $\langle q_j, w_i\rangle$:

(4.25) $$(W'Q)(i,j) = [w_i]'[q_j] = \langle q_j, w_i\rangle, \quad i = 1,\ldots,n,\ j = 1,\ldots,k.$$

This implies that the condition $V'V = \text{id}$ is equivalent to having

$$\langle v_i, v_j\rangle = \begin{cases} 1 & i = j; \\ 0 & i \ne j. \end{cases}$$

Traditionally, this is described by saying that the sequence $v_1, v_2, \ldots, v_n$ is **orthonormal**, i.e., different entries are *ortho*gonal to each other and each entry is *normal*ized, meaning that it has Euclidean length 1. This motivates the following definition.

**Definition:** The matrix $Q \in \mathbb{F}^{m\times n}$ is called **orthonormal** (or **o.n.** for short) if $Q'Q = \text{id}$, i.e., if the sequence of its columns is orthonormal.

An o.n. matrix $Q$ has many special properties, all of which derive from the defining property, $Q'Q = \text{id}$, by the observation that therefore, for any $x, y \in \mathbb{F}^n$,

$$\langle Qx, Qy\rangle \ = \ \langle Q'Qx, y\rangle = \langle x, y\rangle.$$

For example, any o.n. $Q \in \mathbb{F}^{m\times n}$ is an **isometry** in the sense that

(4.26) $$\forall x \in \mathbb{F}^n \quad \|Qx\| = \|x\|.$$

More than that, any o.n. $Q \in \mathbb{F}^{m\times n}$ is **inner-product preserving** in the sense that $\langle Qx, Qy\rangle = \langle x, y\rangle$ for any $x, y \in \mathbb{F}^n$. This is also called **angle-preserving** since a standard definition of the **angle** $\varphi$ between two real nonzero $n$-vectors $x$ and $y$ is the following implicit one:

$$\cos(\varphi) := \frac{\langle x, y\rangle}{\|x\|\|y\|}.$$

Note that this definition makes good sense since, by (4.20)Cauchy's Inequality, the righthand side lies in the interval $[-1\,..\,1]$.

Finally, any o.n. $Q$ is necessarily 1-1, hence a basis for its range, an *orthonormal* basis. Remarkably, any linear subspace of $\mathbb{F}^m$ has orthonormal bases, as we now show.

The *existence* of an o.n. basis for a given linear subspace $Y$ of $\mathbb{F}^m$ can, in fact, be seen in many ways. The traditional argument uses a bootstrap procedure known as **Gram-Schmidt**: One starts with any basis $V \in \mathbb{F}^{m\times n}$ for $Y$. Assume already available an o.n. basis $Q_j =: [q_1, q_2, \ldots, q_j]$ for $\operatorname{ran} V_j$ with $V_j := [v_1, v_2, \ldots, v_j]$. This will certainly be so initially for $j = 0$ (in which case $\operatorname{ran} V_0 = \operatorname{ran}[\,] = \{0\}$ and its basis, namely, $[\,]$, has trivially orthonormal columns since it has no columns). Then one can compute the orthogonal projection $P_j y := Q_j Q_j' y$ onto $\operatorname{ran} V_j$ for any $y$. In particular, $w := v_{j+1} - Q_j Q_j' v_{j+1}$ is available as an element of $\operatorname{ran}[v_{j+1}] + \operatorname{ran} V_j = \operatorname{ran} V_{j+1}$ which is perpendicular to $\operatorname{ran} V_j$. Moreover, $w \ne 0$, since, in the contrary case, $v_{j+1} = Q_j Q_j' v_{j+1} \in \operatorname{ran} Q_j = \operatorname{ran} V_j$, hence $V$ would fail to be 1-1, contrary to assumption. This implies that

$$q_{j+1} := w/\|w\|$$

is of norm 1 and is perpendicular to $\operatorname{ran} V_j = \operatorname{ran} Q_j$, hence $Q_{j+1} := [Q_j, q_{j+1}]$ is also o.n., hence 1-1, therefore necessarily a basis for $\operatorname{ran} V_{j+1}$ since $\operatorname{ran} Q_{j+1} \subseteq \operatorname{ran} V_{j+1}$. This proves by induction on $j$:

**(4.27) Proposition:** For every 1-1 $V \in \mathbb{F}^{m\times n}$, there exists an o.n. $Q \in \mathbb{F}^{m\times n}$ 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 $V$. In particular, every linear subspace of $\mathbb{F}^m$ has an o.n. basis, hence is the range of an orthogonal projector.

**Proof:** Since $\operatorname{ran}[q_1, q_2, \ldots, q_j] = \operatorname{ran}[v_1, v_2, \ldots, v_j]$, then, in particular, for every $i > j$,

$$q_i \perp \operatorname{ran}[q_1, q_2, \ldots, q_j] = \operatorname{ran}[v_1, v_2, \ldots, v_j],$$

hence $(Q'V)(i,j) = q_i' v_j = 0$. This proves that $V = P_V V = QQ'V = QR$, with $R := Q'V$ upper triangular. □

Note that we have also proved the following.

**(4.28) Corollary.** Every o.n. $Q$ with $\operatorname{ran} Q \subseteq \operatorname{ran} V$ can be extended to an o.n. basis for $\operatorname{ran} V$.

Given any 1-1 matrix `V`, the `MATLAB` command `[q,r] = qr(V,0)` provides an o.n. basis for ran `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 ran`V`. Finally, the simplest possible statement, `p = qr(V)`, gives the most complicated result, namely a matrix `p` of the same size as `V` which 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 $j = 1, \ldots, \#_c\mathtt{V}$, $\operatorname{ran}\mathtt{V}(:,[1{:}j]) = \operatorname{ran}\mathtt{Q}(:,[1{:}j])$, 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 reflection**s $\mathrm{id} - 2ww'/\langle w, w\rangle$.

□

It is customary to call a *real* unitary matrix **orthogonal**. Unfortunately, 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.

### Map norms and the condition of a basis

One judges a basis $V$ by its 'condition' or 'condition number'. This number makes its appearance when one tries to bound the **relative error** $\|Va - V\widehat{a}\|/\|Va\|$ in a faulty representation $V\widehat{a}$ of the vector $y = Va$ in terms of the relative error $\|a - \widehat{a}\|/\|a\|$ in the (faulty) coordinate vector $\widehat{a}$. Its definition requires an understanding of map norms.

Assume that $V$ is a basis for the vector space $Y$. The coordinate vector $a$ for $y$ is the unique solution of the equation

$$V? = y.$$

We may not be able to compute the solution exactly. Even if we 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 $\widehat{a}$ on the accuracy of $V\widehat{a}$ as a representation for $y = Va$.

Consider first the **absolute error**, i.e., the numbers $\|a - \widehat{a}\|$ and $\|Va - V\widehat{a}\|$. With the tailor-made definition

$$\|V\| := \max_{x \neq 0} \frac{\|Vx\|}{\|x\|}, \tag{4.29}$$

we have at once the **error bound**

$$\|Va - V\widehat{a}\| \le \|V\| \|a - \widehat{a}\|. \tag{4.30}$$

Indeed, if $b := a - \widehat{a} = 0$, then also $Va - V\widehat{a} = 0$, and (4.30) holds trivially since both sides are zero. In the contrary case, $Va - V\widehat{a} = V(a - \widehat{a}) = Vb$, therefore

$$\frac{\|Va - V\widehat{a}\|}{\|a - \widehat{a}\|} = \frac{\|Vb\|}{\|b\|} \le \max_{x \neq 0} \frac{\|Vx\|}{\|x\|} = \|V\|,$$

and (4.30) follows.

Moreover, the error bound (4.30) is **sharp**, meaning, in this case, that, for any $a$, there is some choice of $\widehat{a}$ (in fact, there is one of any prescribed size, large or small) so that $\|Va - V\widehat{a}\| = \|V\|\|a - \widehat{a}\|$. The reason for this lies in the very definition (4.29) of the **map norm** $\|V\|$, as the maximum of the function

$$F : x \mapsto \|Vx\|/\|x\|$$

as $x$ varies over $\mathbb{F}^n\backslash 0$.

Of course, there is some argument needed here to establish that this maximum is actually taken on. Here it is, for the record, even though it may make no sense to you at this stage of your mathematical career. Notice that the norm is **absolutely homogeneous**, i.e., that

$$\|x\alpha\| = \|x\||\alpha|, \quad \forall x \in \mathbb{F}^n, \alpha \in \mathbb{F}.$$

Therefore, $F$ is **homogeneous of degree 0**, i.e.,

$$F(x\alpha) = F(x), \quad \forall x \in \mathbb{F}^n\backslash 0, \alpha \in \mathbb{F}\backslash 0.$$

In particular, $F(x) = F(x/\|x\|)$ for any $x \neq 0$. Therefore,

$$\|V\| = \max_{\|x\|=1} \|Vx\|.$$

Since $\{x \in \mathbb{F}^n : \|x\| = 1\} \to \mathbb{R} : x \mapsto \|Vx\|$ is a *continuous* function on a *compact* set, it achieves its maximum.

We conclude that *determination* of the map norm is a two-part process, as formalized in the following.

**(4.31) 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 $x \neq 0$, $\|Ax\| \ge c\|x\|$.

The first says that $\|A\| \le c$, while second says that $\|A\| \ge c$, hence, together they say that $\|A\| = c$.

**(4.32) Example.** We compute $\|A\|$ in case $A \in \mathbb{F}^{m\times n}$ is of the simple form

$$A = [v][w]' = vw'$$

for some $v \in \mathbb{F}^m$ and some $w \in \mathbb{F}^n$. Since

$$Ax = (vw')x = v(w'x),$$

we have

$$\|(vw')x\| = \|v\||w'x| \le \|v\|\|w\|\|x\|,$$

the equality by the absolute homogeneity of the norm, and the inequality by (4.20)Cauchy's Inequality. This shows that $\|vw'\| \le \|v\|\|w\|$. On the other hand, for the specific choice $x = w$, we get $(vw')w = v(w'w) = v\|w\|^2$, hence $\|(vw')w\| = \|v\|\|w\|\|w\|$. Assuming that $w \neq 0$, this shows that $\|vw'\| \ge \|v\|\|w\|$. However, this inequality is trivially true in case $w = 0$ since then $vw' = 0$. So, altogether, we have that

$$\|vw'\| = \|v\|\|w\|.$$

Note that we have, incidentally, proved that, for any $v \in \mathbb{F}^n$,

$$\|[v]\| = \|v\| = \|[v]'\|. \tag{4.33}$$

□

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 also gives the Euclidean norm when its argument is a 'vector'. Specifically, `norm(v)` and `norm(v,2)` both give $\|v\| = \sqrt{v'v}$. 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 (4.33), the norm of the *vector* $v$ equals the norm of the *matrices* $v$ and $v'$.

However, we will make use of other vector norms later, e.g., the max-norm

$$\|x\|_\infty := \max_j |x_j|,$$

because its associated map norm turns out to be easily computable. It is

$$\|A\|_\infty := \max_{x\neq 0} \|Ax\|_\infty / \|x\|_\infty = \max_i \sum_j |A(i,j)| = \max_i \|A(i,:)\|_1, \tag{4.34}$$

with

$$\|v\|_1 := \sum_j |v_j| \tag{4.35}$$

yet another vector norm, the socalled **1-norm**, and supplied in MATLAB by `norm(v,1)`. 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 $[v]$ or $[v]'$ for some vector $v$, then `norm(A,inf)` returns the max-norm of that vector, i.e., the number $\|v\|_\infty$. By (4.34), this is ok if $\mathtt{A} = [v]$, but gives, in general, the wrong result if $\mathtt{A} = v'$. This is an additional reason for sticking with the rule of using only $(n,1)$-matrices for representing $n$-vectors in MATLAB.

□

The best explicit expression available for $\|A\|$ for an arbitrary $A \in \mathbb{F}^{m\times n}$ is the following:

$$\|A\| := \max_{x\neq 0} \|Ax\|/\|x\| = \sqrt{\rho(A'A)} = \sigma_1(A). \tag{4.36}$$

This formula cannot be evaluated in finitely many steps since the number $\rho(A'A)$ is, by definition, the 'spectral radius' of $A'A$, i.e., the smallest possible radius of a disk centered at the origin which contains all the eigenvalues of $A'A$. The 2-norm of $A$ also equals $\sigma_1(A)$ which is, by definition, the largest 'singular value' of $A$. In general, one can only compute approximations to this number.

With this preparation, we now bound the *relative error* $\|Va - V\widehat{a}\|/\|Va\|$ in a faulty representation $V\widehat{a}$ of the vector $y = Va$ in terms of the relative error $\|a - \widehat{a}\|/\|a\|$ in the (faulty) coordinate vector $\widehat{a}$. In effect, we now want to know how large the ratio

$$\frac{\|Va - V\widehat{a}\|/\|Va\|}{\|a-\widehat{a}\|/\|a\|} = \frac{\|Va - V\widehat{a}\|}{\|a-\widehat{a}\|}\frac{\|a\|}{\|Va\|}$$

can become as both $a$ and $\widehat{a}$ vary freely. The first fraction on the righthand side we have already handled; we know it to be bounded by $\|V\|$. As to the second fraction, we use the following trick: since $V$ is a basis, hence invertible, the range of $V^{-1}$ is all of $\mathbb{F}^n$, hence

$$\max_{Va\neq 0} \frac{\|a\|}{\|Va\|} = \max_{b\neq 0} \frac{\|V^{-1}b\|}{\|VV^{-1}b\|} = \max_{b\neq 0} \frac{\|V^{-1}b\|}{\|b\|} = \|V^{-1}\|.$$

With this, we have obtained the standard relative error bound

$$\|Va - V\widehat{a}\|/\|Va\| \le \kappa(V)\ \|a - \widehat{a}\|/\|a\|, \tag{4.37}$$

in which the number

$$\kappa(V) := \|V\|\|V^{-1}\|$$

is called the **condition number** of the basis $V$.

The larger the condition number, the more tenuous is the relationship between the relative error in the coordinate vector and the relative error in the vector represented, i.e., the more **illconditioned** is the basis. How small can we make the condition number, i.e., what are the **bestconditioned** bases?

Since $\|x\| = \|VV^{-1}x\| \le \|V\|\|V^{-1}x\| \le \|V\|\|V^{-1}\|\|x\|$ for any $x$, we must have

$$1 \le \|V\|\|V^{-1}\| = \kappa(V),$$

i.e., the condition can never be smaller than 1 (except when $Y = \{0\}$, in which the condition of its sole basis is 0, but let's not worry about that). On the other hand, we already know that every linear subspace of $\mathbb{F}^n$ has o.n. bases $U$, and for any o.n. $U$, we even have $\|Ux\| = \|x\|$ for all $x$, hence any o.n. $U$ has $\|U\| = 1 = \|U^{-1}\|$, hence $\kappa(U) = 1$. In other words, o.n. bases are best-conditioned, and this is a great plus for them compared to some other bases.

`MATLAB` provides the condition for a *square* matrix `A` by the command `cond(A)`.

□

## 5. 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

$$A = V\Lambda, \qquad \text{with } \ V \in L(\mathbb{F}^r, Y), \ (\text{hence } \Lambda \in L(X, \mathbb{F}^r) \ ).$$

The following picture might be helpful:

$$\begin{tikzcd} X \arrow[rr, "A"] \arrow[dr, "\Lambda"'] & & Y \\ & \mathbb{F}^r \arrow[ur, "V"'] & \end{tikzcd}$$

For example, recall how you apply the linear map $D$ of differentiation to a polynomial $p \in \Pi_k$: 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 $\le k$ which matches given ordinates at given $k+1$ distinct abscissae, i.e., which satisfies

$$p(x_i) = y_i, \qquad i = 0, \ldots, k$$

for given data $(x_i, y_i), i = 0, \ldots, k$. Now, try, e.g., to compute the first derivative of the polynomial $p$ of degree $\le 3$ which satisfies $p(j) = (-1)^j$, $j = 1, 2, 3, 4$. Can you do it without factoring the linear map $D : \Pi_3 \to \Pi_3$ 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* $\Lambda : \mathbb{F}^m \to \mathbb{F}^n$ which is 1-1 on $X$, thereby obtaining the *factorization*

$$V^{-1} = (\Lambda V)^{-1}\Lambda$$

in which $\Lambda V$ is a square matrix, hence $(\Lambda V)^{-1}$ is also a matrix.

The smaller one can make $\#V$ in a factorization $A = V\Lambda$ of $A \in L(X, Y)$, the cheaper is the calculation of $A$.

**Definition:** The smallest $r$ for which $A \in L(X, Y)$ can be factored as $A = V\Lambda$ with $V \in L(\mathbb{F}^r, Y)$ (hence $\Lambda \in L(X, \mathbb{F}^r)$) is called the **rank** of $A$. This is written

$$r = \operatorname{rank} A.$$

Any factorization $A = V\Lambda$ 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 $A = V\Lambda$ with $\operatorname{dom} V = \mathbb{F}^1$, but we couldn't do it with $\operatorname{dom} V = \mathbb{F}^0$). To calculate $Ax$, we merely need to calculate the number $\alpha := (1,2,3,4,5)*x$, and then obtain $Ax$ as the particular scalar multiple $y\alpha$ of the vector $y := (1,1,1,1)$. That is much cheaper than computing the matrix product of the $4\times 5$-matrix $A$ with the 1-column matrix $[x]$.

As the example illustrates, any matrix

$$A := [v][w]' = vw'$$

with $v \in \mathbb{F}^m$ and $w \in \mathbb{F}^n$ has rank 1 unless it is trivial, i.e., unless either $v$ or $w$ is the zero vector. This explains why an *elementary* matrix is also called a **rank-one perturbation of the identity**.

The only linear map of rank 0 is the 0 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 $A = V\Lambda$ 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 $\mathrm{id}_X = V\Lambda$ for some $V \in L(\mathbb{F}^r, X)$, $V$ must necessarily be onto, hence $\dim X \le r$, by (3.8)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 $\mathrm{id}_X$.

When $A$ is an $m\times n$-matrix, then, trivially, $A = A\,\mathrm{id}_n = \mathrm{id}_m A$, hence $\operatorname{rank} A \le \min\{m,n\}$.

When $A$ is a matrix, then it is often more convenient to write the factorization in the form

$$A = VW' = \sum_{j=1}^{r} v_j w_j', \tag{5.1}$$

with $V =: [v_1, v_2, \ldots, v_r]$ and $W =: [w_1, w_2, \ldots, w_r]$. Since each of the maps $v_j w_j'$ has rank $\le 1$, this shows that *the rank of $A$ gives the smallest number of terms necessary to write $A$ as a sum of rank-one maps.*

**(5.2) Proposition.** $A = V\Lambda$ 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 $A = V\Lambda$. Then $\operatorname{ran} A \subseteq \operatorname{ran} V$, 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 (3.11)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 $\operatorname{ran} A$. □

One can prove in a similar way that $A = V\Lambda$ is minimal if and only if $\Lambda$ is onto and $\operatorname{null} A = \operatorname{null} \Lambda$.

**(5.3) Corollary:** The factorization $A = A(:,\texttt{bound})\operatorname{rrref}(A)$ provided by elimination (see (2.8)) is minimal.

**(5.4) Corollary:** If $A = V\Lambda$ is minimal and $A$ is invertible, then also $V$ and $\Lambda$ are invertible.

**Proof:** By (5.2)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 $\Lambda = V^{-1}A$ is invertible. □

But note that the matrix $[\,1\,] = [\,1 \quad 1\,]\begin{bmatrix}1\\1\end{bmatrix}$ is invertible, even though neither of its two factors is.

**The rank of a matrix and its (conjugate) transpose**

If $A$ is a matrix, then $A = VW'$ 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 iff $A^T = \overline{W}V^T$ is minimal. Therefore:

**Proposition.** $\operatorname{rank} A = \operatorname{rank} A' = \operatorname{rank} A^T$.

**(5.5) Corollary:** If $A$ is a matrix, then $\dim \operatorname{ran} A = \dim \operatorname{ran} A' = \dim \operatorname{ran} A^T$.

**(5.6) Corollary:** For any matrix $A$, $A'$ and/or $A^T$ 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 $(A')' = A$. □

**SVD**

Let $A = VW'$ be a minimal factorization for the $m \times n$-matrix $A$ of rank $r$. Then $A' = WV'$ is a minimal factorization for $A'$. By (5.2), this implies that $V$ is a basis for $\operatorname{ran} A$ and $W$ is a basis for $\operatorname{ran} A'$.

Moreover, for any basis $U$ of $\operatorname{ran} A$, we have

$$A = \operatorname{id}_m A = UU^{-1}A,$$

with $U^{-1}A$ a linear map from $\mathbb{F}^n$ to $\mathbb{F}^r$, hence an $r \times n$ matrix. Thus, setting $T := (U^{-1}A)'$, we get the factorization

$$A = UT'$$

which is also minimal, since $U$ is a basis for $\operatorname{ran} A$. In other words, for every basis $U$ of $\operatorname{ran} A$, there is some basis $T$ of $\operatorname{ran} A'$ so that $A = UT'$ is a minimal factorization.

By applying this last result to $A'$, we conclude that also, for *every* basis $W$ of $\operatorname{ran} A'$, there is some basis $U$ of $\operatorname{ran} A$ so that $A = UW'$ is minimal.

So, assume without loss that $W$ is an *o.n.* basis for $\operatorname{ran} A'$. Then $A = AWW'$, i.e., $V = AW$ is the corresponding basis for $\operatorname{ran} A$ for which $A = VW'$. Can we so choose $W$ that also $V = AW$ is o.n.?

Well, if both $V$ and $W$ are o.n., then, for any $x$, $\|Ax\| = \|VW'x\| = \|W'x\|$, while, for $x \in \operatorname{ran} A'$, $x = WW'x$, hence $\|x\| = \|W'x\|$. Therefore, altogether, in such a case, $A$ is an isometry on $\operatorname{ran} A'$, a very special situation.

Perhaps surprisingly, nevertheless there is an o.n. basis $W$ for $\operatorname{ran} A'$ for which $AW$ is *orthogonal*, i.e., of the form $V\Sigma$ with $V$ o.n. and $\Sigma$ diagonal.

**(5.7) Theorem.** For every $A \in \mathbb{F}^{m\times n}$, there exist o.n. bases $V$ and $W$ for $\operatorname{ran} A$ and $\operatorname{ran} A'$, respectively, and a diagonal matrix $\Sigma$ with positive diagonal entries, so that

$$A = V\Sigma W'. \tag{5.8}$$

**Proof:** For efficiency, the proof given here uses results, concerning the eigenstructure of hermitian positive definite matrices, which are established only later in these notes. This may help to motivate the study of the eigenstructure of matrices.

For motivation of the proof, assume for the moment that $A = V\Sigma W'$ is a factorization of the kind we claim to exist. Then, with $\Sigma =: \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$, it follows that

$$A'A = W\Sigma' V'\, V\Sigma W' = W\Sigma'\Sigma W',$$

hence

$$A'AW = W \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_r), \tag{5.9}$$

with $W$ o.n., and the $\lambda_j = \overline{\sigma_j}\sigma_j = |\sigma_j|^2$ all positive.

Just such an o.n. $W \in \mathbb{F}^{m\times r}$ and positive scalars $\lambda_j$ do exist by (6.38)Corollary, since the matrix $A'A$ is hermitian and positive semidefinite and has rank $r$.

With $W$ and the $\lambda_j$ so chosen, it follows that $W$ is an o.n. basis for $\operatorname{ran} A'$, since (5.9) implies that $\operatorname{ran} W \subseteq \operatorname{ran} A'$, and $W$ is a 1-1 column map of order $r = \dim \operatorname{ran} A'$. This implies, as we saw earlier, that $A = UW'$ for some $U$. Further, we compute with (5.9) that

$$(AW)'(AW) = W'A'AW = W'W \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_r) = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_r).$$

Consequently, with

$$\Sigma := \operatorname{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_r}),$$

the matrix

$$V := AW\Sigma^{-1} = U\Sigma^{-1}$$

is well-defined and o.n., and $A = UW' = V\Sigma W'$. □

It is customary to order the numbers

$$\sigma_j := \sqrt{\lambda_j}, \quad j = 1, \ldots, r.$$

Specifically, one assumes that

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r.$$

These numbers $\sigma_j$ are called the (nonzero) **singular value**s of $A$, and with this ordering, the factorization

$$A = \sum_{j=1}^{\operatorname{rank} A} v_j \sigma_j w_j'$$

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 $V \operatorname{id}_n V'$ for $\operatorname{id}_n$.

Some prefer to have a factorization $A = \tilde{V}\tilde{\Sigma}\tilde{W}'$ 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 $\operatorname{ran} A'$, respectively). This can always be achieved by extending $V$ and $W$ from (5.8) in any manner whatsoever to o.n. bases $\tilde V := [V, V_1]$ and $\tilde W := [W, W_1]$ and, correspondingly, extending $\Sigma$ to

$$\tilde\Sigma := \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \in \mathbb{F}^{m\times n}$$

by the adjunction of blocks of 0 of appropriate size. With this, we have

$$A = \tilde V \tilde\Sigma \tilde W' = \sum_{j=1}^{\min\{m,n\}} v_j \sigma_j w_j', \tag{5.10}$$

and the diagonal entries

$$\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r > 0 = \sigma_{r+1} = \cdots = \sigma_{\min\{m,n\}}$$

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 (5.10) 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

$$A? = b$$

in case $A$ is not invertible (for whatever reason). In the previous chapter, we looked in this case for a solution of the 'projected' problem **projected problem**

$$A? = P_{\operatorname{ran} A} b =: \widehat b \tag{5.11}$$

for the simple reason that any solution $x$ of this equation makes the residual **residual** $\|Ax - b\|_2$ as small as it can be made by any $x$. For this reason, any solution of (5.11) is called a **least-squares solution** for $A? = b$.

If now $A$ is 1-1, then (5.11) 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 solution of minimal norm. The svd for $A$ makes it easy to find this particular solution.

If $A = V\Sigma W'$ is a svd for $A$, then $V$ is an o.n. basis for $\operatorname{ran} A$, hence

$$\widehat b = P_{\operatorname{ran} A} b = VV'b.$$

Therefore, (5.11) is equivalent to the equation

$$V\Sigma W'? = VV'b.$$

Since $V$ is o.n., hence 1-1, and $\Sigma$ is invertible, this equation is, in turn, equivalent to

$$W'? = \Sigma^{-1} V' b,$$

hence to

$$WW'? = W\Sigma^{-1}V'b. \tag{5.12}$$

Since $W$ is also o.n., $WW' = P_W$ is an o.n. projector, hence, by (4.19)Proposition, strictly reduces norms unless it is applied to something in its range. Since the right-hand side of (5.12) is in $\operatorname{ran} W$, it follows that the solution of smallest norm of (5.12), i.e., the best least-squares solution of $A? = b$, is that right-hand side, i.e., the vector

$$\widehat{x} := A^+ b,$$

with the matrix

$$A^+ := W\Sigma^{-1}V'$$

the **Moore-Penrose pseudo-inverse** of $A$.

Note that

$$A^+A = W\Sigma^{-1}V'V\Sigma W' = WW',$$

hence $A^+$ is a left inverse for $A$ in case $W$ is square, i.e., in case $\operatorname{rank} A = \#_c A$. Similarly,

$$AA^+ = V\Sigma W'W\Sigma^{-1}V' = VV',$$

hence $A^+$ is a right inverse for $A$ in case $V$ is square, i.e., in case $\operatorname{rank} A = \#_r A$. In any case,

$$A^+A = P_{\operatorname{ran} A'}, \qquad AA^+ = P_{\operatorname{ran} A},$$

therefore, in particular,

$$AA^+A = A.$$

## 2-norm and 2-condition of a matrix

Recall from (4.26) that o.n. matrices are 2-norm-preserving, i.e.,

$$\|x\|_2 = \|Ux\|_2, \qquad \forall x \in \mathbb{F}^n, \text{o.n. } U \in \mathbb{F}^{m\times n}.$$

This implies that

$$\|VB\|_2 = \|B\|_2 = \|BW'\|_2, \qquad \forall \text{ o.n. } V \in \mathbb{F}^{r\times m}, B \in \mathbb{F}^{m\times n}, \text{o.n. } W \in \mathbb{F}^{r\times n}.$$

Indeed,

$$\|VB\|_2 = \max_{x\neq 0} \frac{\|VBx\|_2}{\|x\|_2} = \max_{x\neq 0} \frac{\|Bx\|_2}{\|x\|_2} = \|B\|_2.$$

Also

$$\|BW'\|_2 = \max_{x\neq 0} \frac{\|BW'x\|_2}{\|x\|_2} = \max_{y\neq 0} \frac{\|BW'Wy\|_2}{\|Wy\|_2} = \max_{y\neq 0} \frac{\|By\|_2}{\|y\|_2} = \|B\|_2.$$

It follows that, with $A = V\Sigma W' \in \mathbb{F}^{m\times n}$ a svd for $A$,

$$\|A\|_2 = \|\Sigma\|_2 = \sigma_1, \tag{5.13}$$

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 $A^{-1} = A^+ = V\Sigma^{-1}W'$. 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.

**(5.14) Proposition:** If $A = V\Sigma W'$ 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

$$A_k := \sum_{j=1}^{k} v_j \sigma_j w_j'.$$

**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

$$W_{k+1} := [w_1, w_2, \ldots, w_{k+1}].$$

Therefore, by (3.14)Corollary, the intersection $\operatorname{null}(B) \cap \operatorname{ran} W_{k+1}$ contains a vector $z$ of norm 1. Then $Bz = 0$, and $W'z = W_{k+1}'z$, and $\|W_{k+1}'z\|_2 = \|z\|_2 = 1$. Therefore, $Az = V\Sigma W'z = V_{k+1}\Sigma_{k+1}W_{k+1}'z$, hence

$$\|A - B\|_2 \ge \|Az - Bz\|_2 = \|Az\|_2 = \|\Sigma_{k+1}W_{k+1}z\|_2 \ge \sigma_{k+1}\|W_{k+1}z\|_2 = \sigma_{k+1}.$$

On the other hand, for the specific choice $B = A_k$, we get $\|A - A_k\|_2 = \sigma_{k+1}$ by (5.13), since $A - A_k = \sum_{j>k} v_j\sigma_j w_j'$ is a svd for it, hence its largest singular value is $\sigma_{k+1}$. □

In particular, if we have in hand a svd

$$A + E = V \operatorname{diag}(\hat\sigma_1, \hat\sigma_2, \ldots, \hat\sigma_{\hat r})W'$$

for the *perturbed* matrix $A + E$, and 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 $\mathrm{id}_3$ 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 only rank 2, – which happens to be the case in this particular example.

### Equivalence

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 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:

$$\begin{array}{ccc} & A & \\ X & \longrightarrow & Y \\ W \uparrow & & \uparrow V \\ \mathbb{F}^n & \longrightarrow & \mathbb{F}^m \\ & \widehat{A} & \end{array}$$

In this situation, $\widehat{A} = V^{-1}AW$ is called a **matrix representation for** $A$.

For example, we noted earlier that the matrix

$$\widehat{D}_k := \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 2 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & k \end{bmatrix}$$

is the standard matrix representation used in Calculus for the linear map $D : \Pi_k \to \Pi_{k-1}$ of differentiation of polynomials of degree $\le k$.

In practice, one looks, for given $A \in L(X,Y)$, for matrix representations $\widehat{A}$ which 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

$$\widehat{A} = \begin{bmatrix} \mathrm{id}_{\mathrm{rank}\,A} & 0 \\ 0 & 0 \end{bmatrix},$$

with 0 indicating zero matrices of the appropriate size to make $\widehat{A}$ of size $\dim \mathrm{tar}\, A \times \dim \mathrm{dom}\, A$. It can be obtained from any minimal factorization $A = \tilde{V}\Lambda$ by extending $\tilde{V}$ to a basis $V = [\tilde{V}, V_1]$ of $\mathrm{tar}\, A$ and extending $\Lambda$ to the inverse $W^{-1}$ of a basis $W$ for $\mathrm{dom}\, A$.

The situation becomes much more interesting and challenging when $\mathrm{dom}\, A = \mathrm{tar}\, A$ 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 these notes.

## 6. The powers of a linear map

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 which show the importance of understanding the powers of a linear map.

### Examples

**Fixed-point iteration:** A standard method for solving a large linear system $A? = b$ (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 $x_0, x_1, x_2, \ldots$ of approximate solutions by the **iteration**

$$x_k := M^{-1}(Nx_{k-1} + b), \quad k = 1, 2, \ldots. \tag{6.1}$$

Assuming this iteration to converge, with $x := \lim_{k\to\infty} x_k$ its limit, it follows that

$$x = M^{-1}(Nx + b), \tag{6.2}$$

hence that $Mx = Nx + b$, therefore finally that $Ax = (M - N)x = b$, i.e., the limit solves our original problem $A? = b$.

Let $\varepsilon_k := x - x_k$ be the **error** in our $k$th approximate solution. Then on subtracting the iteration equation (6.1) from the exact equation (6.2), we find that

$$\varepsilon_k = x - x_k = M^{-1}(Nx + b\ - (Nx_{k-1} + b)) = M^{-1}N\varepsilon_{k-1}.$$

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 $x$, we have no way of choosing the **initial guess** $x_0$ in any special way. For convergence, we must therefore demand that

$$\lim_{k\to\infty} B^k z = 0 \quad \text{for all } z \in \mathbb{F}^n.$$

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.,

$$M \ge 0, \qquad Me = e,$$

with $e$ the vector with all entries equal to 1,

$$e := (1, 1, 1, \ldots, 1).$$

The entries of $M$ are interpreted as probabilities: $M(i,j)$ 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

$$\sum_k M(i,k)M(k,j) = M^2(i,j).$$

More generally, the probability that we have gone after $m$ steps from vertex $i$ to vertex $j$ is the number $M^m(i,j)$, i.e., the $(i,j)$-entry of the $m$th 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 $i$,

$$\lim_{m\to\infty} M^m(i,:) = x_\infty$$

for a certain ($i$-independent) vector $x_\infty$ with nonnegative entries which 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 $x_\infty$ as a properly scaled eigenvector of the transpose $M'$ 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 these notes.

**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**:

$$\exp(A) := \mathrm{id} + A + A^2/2 + A^3/6 + \cdots + A^k/k! + \cdots. \tag{6.3}$$

The matrix exponential is used in solving the first-order system

$$Dy(t) = Ay(t) \text{ for } t > 0, \quad y(0) = c \tag{6.4}$$

of constant-coefficient ordinary differential equations. Here $A$ is a square matrix, of order $n$ say, and $y(t)$ is an $n$-vector that depends on $t$. Further,

$$Dy(t) := \lim_{h\to 0}(y(t+h) - y(t))/h$$

is the first derivative at $t$ of the vector-valued function $y$. One verifies that the particular function

$$y(t) := \exp(tA)c, \quad t \ge 0,$$

solves the differential equation (6.4). 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 $\mu$. For, in that case, $A^2x = A(Ax) = A(x\mu) = Ax\mu = x\,(\mu)^2$ and, more generally,

$$Ax = x\mu \quad\Longrightarrow\quad A^k x = x\,(\mu)^k, \quad k = 0, 1, 2, \ldots. \tag{6.5}$$

If $x = 0$, this will be so for any scalar $\mu$. If $x \neq 0$, then this will be true for at most one scalar $\mu$. That scalar is called an *eigenvalue for* $A$ with associated *eigenvector* $x$.

> **(6.6) Definition:** Let $A \in L(X)$. Any scalar $\mu$ for which there is a *nontrivial* vector $x \in X$ so that $Ax = x\mu$ is called an **eigenvalue** of $A$. The collection of all eigenvalues of $A$ is called the **spectrum** of $A$ and is denoted $\operatorname{spec}(A)$.
> Thus
> $$\operatorname{spec}(A) = \{\mu \in \mathbb{F} : \ A - \mu\,\mathrm{id} \ \text{ is not invertible}\}.$$
> If $\mu \in \operatorname{spec}(A)$, then the elements of $\operatorname{null}(A - \mu\,\mathrm{id})$ are called the **eigenvectors** of $A$ associated with $\mu$. The number
> $$\rho(A) := \max|\operatorname{spec}(A)| = \max\{|\mu| : \mu \in \operatorname{spec}(A)\}$$
> is called the **spectral radius of** $A$.

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, $Av_j = v_j\mu_j$, $j = 1, \ldots, n$, hence

$$AV = [Av_1, \ldots, Av_n] = [v_1\mu_1, \ldots, v_n\mu_n] = V\mathrm{M},$$

with M the *diagonal* matrix

$$\mathrm{M} := \operatorname{diag}(\mu_1, \mu_2, \ldots, \mu_n).$$

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 $x = Va$ for some $n$-vector $a$ and thus

$$A^k x = A^k Va = V\mathrm{M}^k a = v_1\mu_1^k a_1 + v_2\mu_2^k a_2 + \cdots + v_n\mu_n^k a_n$$

for any $k$. For such a matrix and for any $t$,

$$\exp(tA) = V\exp(t\mathrm{M})V^{-1} = V\operatorname{diag}(\ldots, \exp(t\mu_j), \ldots)V^{-1}.$$

**(6.7) Example:** The matrix $A := \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ maps the 2-vector $x := (1, 1)$ to $3x$ and the 2-vector $y := (1, -1)$ to itself. Hence, $A[x, y] = [3x, y] = [x, y]\operatorname{diag}(3, 1)$ or

$$A = V\operatorname{diag}(3, 1)V^{-1}, \quad \text{with} \ \ V := [x, y] = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

Elimination gives

$$[V, \mathrm{id}] = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & -1 & 0 & 1 \end{bmatrix} \to \begin{bmatrix} \mathbf{1} & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \end{bmatrix} \to \begin{bmatrix} 1 & 0 & 1/2 & 1/2 \\ 0 & \mathbf{-2} & -1 & 1 \end{bmatrix} \to \begin{bmatrix} 1 & 0 & 1/2 & 1/2 \\ 0 & 1 & 1/2 & -1/2 \end{bmatrix},$$

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,

$$\exp(tA) = V \operatorname{diag}(e^{3t}, e^t)V^{-1} = \begin{bmatrix} e^{3t}+e^t & e^{3t}-e^t \\ e^{3t}-e^t & e^{3t}+e^t \end{bmatrix}.$$

□

## Similarity

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

$$A = VBV^{-1}.$$

Thus, in the best of circumstances, $A$ is similar to a diagonal matrix, i.e.,

$$A = V\mathrm{M}V^{-1}$$

for some *diagonal* matrix M. Any $A$ similar to a diagonal matrix is called *diagona(liza)ble*.

A linear map $A \in L(X)$ is called **diagona(liza)ble** if it is similar to a diagonal matrix, i.e., if there is a basis for its domain $X$ consisting entirely of eigenvectors for $A$.

In trying to decide whether or not a given linear map $A$ is diagonable, it is sufficient to decide this question for any convenient linear map $B$ similar to $A$. For, if such a $B$ is diagonable, 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:

**(6.8) 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:** Certainly, $A \sim A$, since $A = \mathrm{id} A\, \mathrm{id}$. 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!) 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 diagonable, it is sufficient to know how to do this for *matrices*.

### Not all linear maps are diagonable

Now, depending on whether you like your life simple or interesting, here is some bad or good news: *Not all square matrices are diagonable.*

The simplest example is provided by the matrix

$$N := \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} = [0, e_1].$$

$N$ maps $e_2$ to $e_1$ and maps $e_1$ to 0. This implies that $N^2$ maps everything to 0, i.e., $N$ is **nilpotent**. Suppose there were a basis $V$ for $X = \mathbb{F}^2$ so that $V^{-1}NV =: \mathrm{M} = \mathrm{diag}(\mu_1, \mu_2)$. Then

$$0 = V^{-1}0V = V^{-1}N^2V = \mathrm{M}^2 = \mathrm{diag}((\mu_1)^2, (\mu_2)^2),$$

therefore $\mu_1 = \mu_2 = 0$, therefore $N = V\mathrm{M}V^{-1} = V0V^{-1} = 0$, and this would be nonsense since $N \neq 0$. But we will see shortly that almost all linear maps are diagonable.

### Does every square matrix have an eigenvalue?

Since an eigenvalue for $A$ is any *scalar* $\mu$ for which $\mathrm{null}(A - \mu\,\mathrm{id})$ 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

$$A := \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} = [e_2, -e_1].$$

This linear map rotates every $x \in \mathbb{R}^2$ 90 degrees counter-clockwise, hence the only vector $x$ mapped by it to a scalar multiple of itself is the zero vector. In other words, this linear map has no nontrivial 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 these notes. Now every (square) matrix has eigenvalues, as follows from the following simple argument.

**(6.9) 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

$$V := [x, Ax, A^2x, \ldots, A^nx].$$

Since $V$ has more columns than its target, $X$, has dimensions, $V$ cannot be 1-1. This implies that $Va = 0$ for some $a \neq 0$. In fact, assume that $A^dx$ is the first free column, i.e., the first column which is in the range of the columns preceding it. Then $\mathrm{null}\,V$ contains exactly one vector of the form

$$a = (a_0, a_1, \ldots, a_{d-1}, 1, 0, \cdots, 0),$$

and this is the vector we choose. Then, writing the equation $Va = 0$ out in full, we get

$$a_0 x + a_1 Ax + \cdots + a_{d-1} A^{d-1} x + A^d x \;=\; 0. \tag{6.10}$$

Now here comes the trick: Consider the *polynomial*

$$p : t \mapsto a_0 + a_1 t + \cdots + a_{d-1} t^{d-1} + t^d. \tag{6.11}$$

Then, substituting for $t$ our map $A$, we get the linear map

$$p(A) := a_0 \,\mathrm{id} + a_1 A + \cdots + a_{d-1} A^{d-1} + A^d.$$

With this, (6.10) can be written, very consisely,

$$p(A)x = 0.$$

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 $z_1, \ldots, z_d$. This implies (see (6.40)Lemma below) that

$$p(A) = (A - z_1 \,\mathrm{id})(A - z_2 \,\mathrm{id}) \cdots (A - z_d \,\mathrm{id}).$$

Now, $p(A)$ is not 1-1 since it maps the nonzero vector $x$ to zero. Therefore, *not all the maps* $(A - z_j \,\mathrm{id})$, $j = 1, \ldots, d$, *can be 1-1.* In other words, for some $j$, $(A - z_j \,\mathrm{id})$ fails to be 1-1, i.e., has a nontrivial nullspace, and that makes $z_j$ an eigenvalue for $A$. □

**(6.12) Example.** Let's try this out on our earlier example, the rotation matrix

$$A := [e_2, -e_1].$$

Choosing $x = e_1$, we have

$$[x, Ax, A^2 x] = [e_1, e_2, -e_1],$$

hence the first free column is $A^2 x = -e_1$, and, by inspection,

$$x + A^2 x = 0.$$

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 the Backgrounder on complex numbers). In fact, we conclude that, with $y := (A + \mathrm{i}\,\mathrm{id})x$, $(A - \mathrm{i}\,\mathrm{id})y = p(A)x = 0$, while $y = Ae_1 + \mathrm{i}e_1 = e_2 + \mathrm{i}e_1 \neq 0$, showing that i is an eigenvalue for this $A$.

## Polynomials in a linear map

The proof of (6.9)Theorem uses in an essential way the following fact.

**(6.13) 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 $r, p, q$, then 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

$$p(t)q(t) = (t^2 + t + 1)(t - 1) = t^2(t - 1) + t(t - 1) + (t - 1) = t^3 - t^2 + t^2 - t + t - 1 = t^3 - 1,$$

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 (6.9) 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 (6.11) with $d$ the *smallest* integer for which $A^d x \in \operatorname{ran}[x, Ax, \ldots, A^{d-1}x]$. Thus, with $\mu$ any zero of $p$, we can write

(6.14) $$p(t) = (t - \mu)q(t)$$

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 $[x, Ax, A^2, \ldots]$. Since

$$0 = p(A)x = (A - \mu\,\mathrm{id})q(A)x,$$

it follows that $\mu$ is an eigenvalue for $A$ with associated nontrivial eigenvector $q(A)x$.

This is exactly how we got a nontrivial eigenvector for the eigenvalue i in (6.12)Example.

**(6.7)Example continued** As another example, consider again the matrix $A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$. We choose $x = e_1$ and consider

$$[x, Ax, \ldots, A^n x] = [e_1, Ae_1, A(Ae_1)] = \begin{bmatrix} 1 & 2 & 5 \\ 0 & 1 & 4 \end{bmatrix}.$$

Since $[e_1, Ae_1, A^2e_1]$ is in row echelon form, we conclude that the first two unknowns are bound. Elimination gives the rref

$$\begin{bmatrix} 1 & 0 & -3 \\ 0 & 1 & 4 \end{bmatrix},$$

hence $(3, -4, 1) \in \operatorname{null}[e_1, Ae_1, A^2e_1]$. Correspondingly, $p(A)e_1 = 0$, with

$$p(t) = 3 - 4t + t^2 = (t - 3)(t - 1).$$

Consequently, $\mu = 3$ is an eigenvalue for $A$, with corresponding (nontrivial) eigenvector

$$(A - \mathrm{id})e_1 = (1, 1);$$

also, $\mu = 1$ is an eigenvalue for $A$, with corresponding (nontrivial)

$$(A - 3\,\mathrm{id})e_1 = (-1, 1).$$

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 (6.7)Example. After all, if $v$ is an eigenvector, then so is $\alpha v$ for any scalar $\alpha$. □

Here is some standard language concerning the items in our discussion so far. One calls any nontrivial polynomial $r$ for which $r(A)x = 0$ 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 $x$. If such a polynomial is of exact degree $k$, say, then it has the form

$$r(t) = b_0 + b_1 t + \cdots + b_{k-1}t^{k-1} + t^k.$$

Since $r(A)x = 0$, we conclude that

$$b_0 x + b_1 Ax + \cdots + b_{k-1}A^{k-1}x + A^k x = 0.$$

In particular, $A^k x$ is in $\operatorname{ran}[x, Ax, \ldots, A^{k-1}x]$, i.e., the column $A^k x$ of $[x, Ax, A^2x, \ldots]$ is free. This implies that $k \geq d$, with $d$ the degree of the polynomial $p$ constructed in the proof of (6.9)Theorem. For, there we chose $d$ as the *smallest* index for which $A^d x$ is a free column of $[x, Ax, A^2, \ldots]$. In particular, all prior columns must be bound. This makes $p$ the unique monic polynomial of smallest degree for which $p(A)x = 0$.

Here, for the record, is a formal account of what we have proved.

**(6.15) 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 $d$ the smallest index for which $A^d x$ is a free column of $[x, Ax, A^2x, \ldots]$. Moreover, $(a_0, a_1, \ldots, 1)$ is the unique vector in $\operatorname{null}[x, Ax, \ldots, A^d x]$ with its last entry equal to 1.

Assuming that $X$ is a vector space over $\mathbb{F} = \mathbb{C}$, every zero $\mu$ of $p_{A,x}$ is an eigenvalue of $A$, with associated nontrivial eigenvector $q(A)x$, where $p_{A,x}(t) =: (t - \mu)q(t)$. (See the Backgrounder on Horner's method for the standard way to compute $q$ from $p_{A,x}$ and $\mu$.)

For *example*, consider the **permutation matrix** $P = [e_2, e_3, e_1]$ and take $x = e_1$. Then

$$[x, Px, P^2x, P^3x] = [e_1, e_2, e_3, e_1].$$

Hence, $P^x$ is the first column which is in the span of the columns to the left of it, and the element in the nullspace corresponding to it is the vector $(-1, 0, 0, 1)$. Hence, the minimal polynomial for $P$ at $x = e_1$ 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 nontrivial eigenvector is obtainable in the form $q(P)e_1$ with $q(t) := p(t)/(t-1) = t^2 + t + 1$, hence the eigenvector is $e_1 + e_2 + e_3$.

### 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 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 the spectrum of an upper triangular matrix is the set of its diagonal entries and similar matrices have the same spectrum. Here are the various facts.

**(6.16) Proposition:** If $A$ and $\widehat{A}$ are similar, then $\operatorname{spec}(A) = \operatorname{spec}(\widehat{A})$.

**Proof:** If $\widehat{A} = V^{-1}AV$ for some invertible $V$, then, for any scalar $\mu$, $\widehat{A} - \mu\,\mathrm{id} = V^{-1}(A - \mu\,\mathrm{id})V$. In particular, $\widehat{A} - \mu\,\mathrm{id}$ is not invertible (i.e., $\mu \in \operatorname{spec}(\widehat{A})$) if and only if $A - \mu\,\mathrm{id}$ is not invertible (i.e., $\mu \in \operatorname{spec}(A)$).

□

**(6.17) Proposition:** A triangular matrix is invertible if and only if all its diagonal entries are nonzero.

**Proof:** Assume first that $A$ is an upper triangular matrix, i.e., $A$ is square and $A(i,j) = 0$ for all $i > j$. If $A(j,j) \neq 0$ for all $j$, then every unknown in $A? = 0$ is bound, hence $A$ is 1-1, hence (being square) it is invertible. Conversely, if some $A(j,j) = 0$, then there is a smallest $j$ for which $A(j,j) = 0$. For this $j$, the $j$th unknown is free (since all rows possibly involving this unknown are rows 1 through $j-1$, and these are all taken as pivot rows for prior unknowns), hence $A$ fails to be 1-1, hence fails to be invertible.

If now $A$ is *lower* triangular, then $A^T$ is upper triangular, hence $A^T$ is invertible iff all its diagonal entries are nonzero. However, $A^T$, being the transpose of $A$, is (by (5.6)) invertible iff $A$ is invertible, while the diagonal entries of $A^T$ are the same as the diagonal entries of $A$. This proves that a lower triangular matrix is invertible if and only if all its diagonal entries are nonzero. □

**(6.18) Corollary:** If $A \in L(X)$ is similar to a triangular matrix $\widehat{A}$, then $\mu$ is an eigenvalue for $A$ if and only if $\mu = \widehat{A}(j,j)$ for some $j$. In formulae,

$$\operatorname{spec}(A) = \{\widehat{A}(j,j) : \text{ all } j\}.$$

More precisely, if $\widehat{A} = V^{-1}AV$ is upper triangular and $j$ is the smallest index for which $\mu = \widehat{A}(j,j)$, then there is a nontrivial eigenvector for $A$ belonging to $\mu$ available in the form $w = Va$, with $a$ the element in the standard basis for $\operatorname{null}(\widehat{A} - \mu\,\mathrm{id})$ associated with the (free) $j$th unknown, i.e., $a \in \operatorname{null}(\widehat{A} - \mu\,\mathrm{id})$, $a_j = 1$, and all other entries corresponding to free unknowns are 0.

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 $A$ which 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 reliable in the presence of round-off as well as fast.

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 (if available), in the sense that $A$`V`$(:,j)$ = `V`$(:,j)$`M`$(j,j)$, all $j$.

□

**(6.19) 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 $1 \times 1$-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 a nontrivial eigenvector, $u_1$, say, with corresponding eigenvalue, $\mu_1$ say. Extend $u_1$ to a basis $U = [u_1, u_2, \ldots, u_n]$ for $\mathbb{C}^n$. Then

$$AU = [Au_1, \cdots, Au_n] = [u_1\mu_1, Au_2, \cdots, Au_n].$$

We want to compute $U^{-1}AU$. For this, observe that $U^{-1}u_1 = U^{-1}Ue_1 = e_1$. Therefore,

$$U^{-1}AU = [e_1\mu_1, U^{-1}Au_2, \cdots, U^{-1}Au_n].$$

Writing this out in detail, we have

$$U^{-1}AU = \widehat{A} := \begin{bmatrix} \mu_1 & \times & \cdots & \times \\ 0 & \times & \cdots & \times \\ \vdots & \vdots & \cdots & \vdots \\ 0 & \times & \cdots & \times \end{bmatrix} =: \begin{bmatrix} \mu_1 & C \\ 0 & A_1 \end{bmatrix}.$$

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

$$\mathrm{diag}(1, W^{-1})\widehat{A}\ \mathrm{diag}(1, W) = \begin{bmatrix} 1 & 0 \\ 0 & W^{-1} \end{bmatrix} \begin{bmatrix} \mu_1 & C \\ 0 & A_1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & W \end{bmatrix} = \begin{bmatrix} \mu_1 & CW \\ 0 & W^{-1}A_1W \end{bmatrix}.$$

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. Since $\mathrm{diag}(1, W^{-1})\,\mathrm{diag}(1, W) = \mathrm{diag}(1, \mathrm{id}_{n-1}) = \mathrm{id}_n$, 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), 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:

**(6.20) 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 $B_\varepsilon$ similar to $A$ whose off-diagonal entries are all $< \varepsilon$ in absolute value.

**Proof:** By (6.19)Theorem, we know that any such $A$ is similar to an upper triangular matrix. Since similarity is transitive (see (6.8)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

$$W := \mathrm{diag}(\delta^1, \delta^2, \ldots, \delta^n),$$

and the *scalar* $\delta$ 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,

$$B(i,j) = (W^{-1}AW)(i,j) = A(i,j)\delta^{j-i}, \quad \text{all } i,j.$$

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 $B(i,j)$ with $j > i$, hence are the corresponding entries of $A$ multiplied with some *positive* power of the scalar $\delta$. Thus, if

$$c := \max_{i<j} |A(i,j)|$$

and we choose $\delta := \varepsilon / c$, then we can be certain that

$$|B(i,j)| \leq \varepsilon, \qquad \text{all } i \neq j,$$

regardless of how small we choose that positive $\varepsilon$. □

### Vector norms, matrix norms, and convergence of sequences

Our discussion of the power sequence $A^0, A^1, A^2, \cdots$ 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.

We have already met the Euclidean norm

$$\|x\|_2 := \sqrt{x'x}$$

of an $n$-vector $x$. It turns out, for a reason to be made clear in a moment, to be more convenient for us to use instead the socalled **max-norm**

$$\|x\|_\infty := \max_j |x_j|.$$

Both the Euclidean or 2-norm $\|\ \|_2$ and the max-norm $\|\ \|_2$ satisfy the following definition of a vector norm:

**(6.21) 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., $\forall \{\alpha \in \mathbb{F}, x \in X\}\ \|\alpha x\| = |\alpha| \|x\|$;

(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 norm plays exactly the role in $\mathbb{F}^n$ that the absolute value plays in $\mathbb{R}$ or $\mathbb{C}$. It provides a convenient means of measuring the distance between elements. In particular, it is used for describing when a sequence of $n$-vectors converges.

Suppose $z_1, z_2, z_3, \cdots$ is an infinite sequence of $n$-vectors. Recall our agreement to refer to the $j$th entry of the $k$th term $z_k$ in such a vector sequence by $z_k(j)$. We say that this sequence **converges to the $n$-vector** $z_\infty$ and write

$$z_\infty = \lim_{k \to \infty} z_k,$$

in case

$$\lim_{k\to\infty} \|z_\infty - z_k\| = 0.$$

It is not hard to see that

$$z_\infty = \lim_{k\to\infty} z_k \quad\Longleftrightarrow\quad \forall\{i\}\ z_\infty(i) = \lim_{k\to\infty} z_k(i).$$

Note that $z_\infty = \lim_{k\to\infty} z_k$ if and only if, for every $\varepsilon > 0$, there is some $k_0$ so that, for all $k > k_0$, $\|z_\infty - z_k\|_\infty < \varepsilon$. 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"

$$\{y \in \mathbb{F}^n : \|y - z_\infty\|_\infty < \varepsilon\}$$

whose center is $z_\infty$ and whose radius is $\varepsilon$.

**(6.22) Lemma:** A convergent sequence is necessarily bounded. More explicitly, if the sequence $(x_k)$ of $n$-vectors converges, then $\sup_k \|x_k\|_\infty < \infty$, i.e., there is some $c$ so that, for all $k$, $\|x_k\|_\infty \le c$.

The proof is a verbatim repeat of the proof of this assertion for scalar sequences, as given in the Backgrounder on scalar sequences.

Next we consider norms for matrices. Suppose $A \in \mathbb{F}^{m\times n}$. Then a natural definition for $\|A\|_\infty$ would be the number $\max_{ij} |A(i,j)|$. However, the standard definition is the **map norm associated with the vector norm** $\|\ \|_\infty$, i.e., the number

$$\|A\|_\infty := \max_{x\ne 0} \|Ax\|_\infty / \|x\|_\infty \quad = \quad \max_i \sum_j |A(i,j)|$$

mentioned earlier, in (4.34), since it is the smallest number $c$ for which the inequality $\|Ax\|_\infty \le c\|x\|_\infty$ holds for all $x \in \mathbb{F}^n$.

For example, if $A = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$ and $y = (1,1)$, then $Ay = (1,1) + (-1,1) = (0,2)$, hence $\|y\|_\infty = 1$ and $\|Ay\|_\infty = 2$. Therefore, if we are to have

$$\|Ax\|_\infty \le c\|x\|_\infty$$

for all 2-vectors $x$, then $c$ must be at least 2. Now $\|A\|_\infty$, as we defined it, is indeed equal to 2, while $\max_{i,j} |A(i,j)|$ is only 1.

Now we can say why, in the present discussions, we prefer the max-norm $\|\cdot\|_\infty$ to the Euclidean norm $\|x\|_2 := \sqrt{\sum_i |x_i|^2}$: It is usually impossible to compute exactly the associated map norm

$$\|A\|_2 := \max_{x\ne 0} \|Ax\|_2 / \|x\|_2,$$

since it equals $\sqrt{\rho(A'A)}$, i.e., requires the calculation of eigenvalues, a non-finite process.

We observe that $\mathbb{F}^{m\times n} \to \mathbb{R} : A \mapsto \|A\|_\infty$ is (just like any other map norm) a vector norm for the vector space $\mathbb{F}^{m\times n} = L(\mathbb{F}^n, \mathbb{F}^m)$, in the sense of (6.21). In addition, the max-norm (just like any other map norm) also satisfies the inequality

$$\|BA\|_\infty \le \|B\|_\infty \|A\|_\infty \tag{6.23}$$

whenever the product $BA$ makes sense. Indeed, if $A \in \mathbb{F}^{m\times n}$ and $B \in \mathbb{F}^{k\times m}$ and $x$ is any $n$-vector, then

$$\|BAx\|_\infty \le \|B\|_\infty \|Ax\|_\infty \le \|B\|_\infty \|A\|_\infty \|x\|_\infty.$$

This proves (6.23) since (see (4.31)) $\|BA\|_\infty$ is the smallest $c$ so that $\|ABx\|_\infty \le c\|x\|_\infty$ for all $x \in \mathbb{F}^n$. In particular, for any square matrix $A$ and any $k$, $\|A^k\|_\infty \le \|A\|_\infty^k$, by induction on $k$.

We say that the sequence $A_1, A_2, A_3, \cdots$ 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.

## 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

$$A^0, A^1, A^2, \cdots \tag{6.24}$$

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 (6.24) for any square *matrix* $A$, then we understand (6.24) 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 Backgrounder on the convergence of scalar sequences and, in particular, on the scalar sequence $(\zeta^0, \zeta^1, \zeta^2, \cdots)$.

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 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 stabilityy is equivalent to the power boundedness of $A$.

The third property is fundamental in the study of iterative processes, as discussed earlier.

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 a nontrivial eigenvector for $A$, with corresponding eigenvalue $\mu$. Then $Ax = \mu x$, hence $A^k x = \mu^k x$ for $k = 1, 2, 3, \ldots$. Suppose $A$ is powerbounded. 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 nontrivial eigenvector, we conclude that

$$A \text{ powerbounded} \quad \Longrightarrow \quad \rho(A) \le 1. \tag{6.25}$$

Actually, more is true. Suppose that $\mu$ is a *defective* eigenvalue for $A$. This means that

$$\operatorname{null}(A - \mu\,\mathrm{id}) \cap \operatorname{ran}(A - \mu\,\mathrm{id}) \ne \{0\}.$$

In other words, there exists a nontrivial eigenvector for $A$ belonging to $\mu$ of the form $x = (A - \mu\,\mathrm{id})y$. This implies that

$$Ay = x + \mu y.$$

Therefore

$$A^2y = Ax + \mu Ay = \mu x + \mu(x + \mu y) = 2\mu x + \mu^2 y.$$

Therefore

$$A^3y = 2\mu Ax + \mu^2 Ay = 2\mu^2 x + \mu^2(x + \mu y) = 3\mu^2 x + \mu^3 y.$$

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 powerbounded.

We conclude:

**(6.26) Proposition:** If the matrix $A$ is powerbounded, then, for all $\mu \in \mathrm{spec}(A)$, $|\mu| \le 1$, with equality only if $\mu$ is a nondefective eigenvalue for $A$.

Now we consider the case that $A$ is convergent (hence, in particular, powerbounded). If $A$ is convergent, then, for any nontrivial eigenvector $x$ with associated eigenvalue $\mu$, the sequence $(\mu^k x)$ must converge. Since $x$ stays fixed, this implies that the scalar sequence $(\mu^k)$ 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 nontrivial eigenvector $x$ with associated eigenvalue $\mu$, the sequence $(\mu^k x)$ must converge to 0. Since $x$ stays fixed (and is nonzero), this implies that the scalar sequence $(\mu^k)$ must converge to 0. This, to recall, implies that $|\mu| < 1$.

Remarkably, these simple necessary conditions just derived, for powerboundedness, convergence, and convergence to 0, are also sufficient.

For the proof, we need one more piece of information, namely a better understanding of the distinction between defective and nondefective eigenvalues.

### Defective and nondefective eigenvalues

Here is the formal definition of the term 'defective eigenvalue'. A better term might be **interesting**, since it is the existence of defective eigenvalues that makes eigenvalue discussions nontrivial. For example, as we shall see, a matrix (over the complex scalars) is diagonable if and only if none of its eigenvalues is defective.

**Definition:** One calls the scalar $\mu$ a **defective** eigenvalue for $A \in L(X)$ in case

$$\mathrm{null}(A - \mu\,\mathrm{id}) \cap \mathrm{ran}(A - \mu\,\mathrm{id}) \ne \{0\}.$$

This says that $\mu$ is an eigenvalue for $A$ since it says that $\mathrm{null}(A - \mu\,\mathrm{id})$ is not trivial. But it says much more. It says that we can find a nontrivial eigenvector for $A$ belonging to $\mu$ of the specific form

$Ay - \mu y$ (i.e., in $\mathrm{ran}(A - \mu\,\mathrm{id})$). The nilpotent matrix $[0, e_1]$ mentioned earlier is a simple example for this, as $\mathrm{null}[0, e_1] = \mathrm{ran}[e_1] = \mathrm{ran}[0, e_1]$, hence its sole eigenvalue, 0, is defective.

Suppose now that $\mu$ is a *non*defective eigenvalue for $A \in L(X)$. Then $\mathrm{null}(A-\mu\,\mathrm{id}) \cap \mathrm{ran}(A-\mu\,\mathrm{id}) = \{0\}$. Let $V$ be any basis for $\mathrm{null}(A - \mu\,\mathrm{id})$ and $W$ any basis for $\mathrm{ran}(A - \mu\,\mathrm{id})$. I claim that $[V, W]$ *is a basis for* $X$. Indeed, by (3.10)Proposition, $[V, W]$ is 1-1, while, by the Dimension Formula,

$$\#[V, W] = \#V + \#W = \dim \mathrm{null}(A - \mu\,\mathrm{id}) + \dim \mathrm{ran}(A - \mu\,\mathrm{id}) = \dim \mathrm{dom}\, A = \dim X,$$

showing that $[V, W]$ is maximally 1-1, hence a basis for $X$.

This proves

**(6.27) Proposition:** Let $\mu$ be a nondefective eigenvalue of the linear map $A \in L(X)$, with $X$ a finite-dimensional vector space. Then, for any basis $V$ for $\mathrm{null}(A - \mu\,\mathrm{id})$ and any basis $W$ for $\mathrm{ran}(A - \mu\,\mathrm{id})$, $[V, W]$ is a basis for $X$. Note that any such $V$ consists entirely of (nontrivial) eigenvectors for $A$ (all belonging to the eigenvalue $\mu$).

In particular, if $\mu$ is a nondefective eigenvalue for $A$, then

$$X = \mathrm{null}(A - \mu\,\mathrm{id}) \dotplus \mathrm{ran}(A - \mu\,\mathrm{id}),$$

i.e., any $x \in X$ can be written in a unique way as $x = y + z$ with $y \in Y := \mathrm{null}(A - \mu\,\mathrm{id})$ and $z \in Z := \mathrm{ran}(A - \mu\,\mathrm{id})$. Further, $Z = \mathrm{ran}(A - \mu\,\mathrm{id})$ is $A$**-invariant** in the sense that $A(Z) \subset Z$. This implies that the rule

$$B : Z \to Z : y \mapsto Ay$$

defines a linear map on $Z$ (which we will also denote by $B = A_{|Z}$). This linear map $B = A_{|Z}$ is important to us because

$$\mathrm{spec}(B) = \mathrm{spec}(A)\backslash\{\mu\}. \tag{6.28}$$

Indeed, if $Bz = \nu z$ for some scalar $\nu$ and some $z \in Z\backslash 0$, then, in as much as $Bz = Az$ for all $z \in \mathrm{ran}\, A$ (by the very definition of $B$), $Az = \nu z$, i.e., $\nu \in \mathrm{spec}(A)$. But since $\mathrm{null}(A - \mu\,\mathrm{id}) \cap Z = \{0\}$, $\nu \neq \mu$, i.e., $\nu \in \mathrm{spec}(A)\backslash\{\mu\}$. Conversely, if $\nu \in \mathrm{spec}(A)$, then there must be some $x \in X\backslash 0$ so that $Ax = \nu x$. Writing this as $x = y + z$ with $y \in Y$ and $z \in Z$, we find

$$\nu y + \nu z = \nu(y + z) = A(y + z) = Ay + Az = \mu y + Az = \mu y + Bz,$$

or, $Y \ni (\nu - \mu)y = (Bz - \nu z) \in Z$. Since $Y \cap Z = \{0\}$, this implies that $(\nu - \mu)y = 0 = Bz - \nu z$. Hence, if $\nu \in \mathrm{spec}(A)\backslash\{\mu\}$, then necessarily $y = 0$, therefore $z = x \neq 0$, thus $\nu \in \mathrm{spec}(B)$ (and $x \in Z = \mathrm{ran}(A - \mu\,\mathrm{id})$). We have proved the following

**(6.29) Corollary:** If $\mu$ is a nondefective eigenvalue of $A \in L(X)$ for some finite-dimensional vector space $X$, then $X$ has a basis $U = [V, W]$, with $V$ a(ny) basis for $\mathrm{null}(A-\mu\,\mathrm{id})$ (hence consisting entirely of eigenvectors of $A$ belonging to $\mu$), and $W$ a(ny) basis for $Z := \mathrm{ran}(A - \mu\,\mathrm{id})$. Further,

$$\mathrm{spec}(A_{|Z}) = \mathrm{spec}(A)\backslash\{\mu\},$$

and any eigenvector for $A$ belonging to some eigenvalue other than $\mu$ is contained in $Z$. In particular, $\nu \in \mathrm{spec}(B)$ is a defective eigenvalue for $B := A_{|Z}$ if and only if it is a defective eigenvalue for $A$.

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, $\mu := 3 \in \operatorname{spec}(A)$. 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 elimination applied to $(A - \mu\,\mathrm{id})? = 0$ will designate the first unknown as bound and the second unknown as free. This also implies that $\operatorname{null}(A - \mu\,\mathrm{id})$ is one-dimensional, with $V := \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ a basis for it.

It follows, by inspection, that $\operatorname{null}(A - \mu\,\mathrm{id}) \cap \operatorname{ran}(A - \mu\,\mathrm{id}) = \{0\}$ 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 $B := A_{|Z}$, with $Z = \operatorname{ran}(A - \mu\,\mathrm{id})$? In this case, things are very simple since $Z$ is one-dimensional. Since $A(Z) \subseteq Z$, $A$ must map any $z \in Z$ to a scalar multiple of itself! In particular, since $z = (-1,1) \in \operatorname{ran}(A - \mu\,\mathrm{id})$, $A$ must map it 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 a nontrivial eigenvector for $A$ belonging to the eigenvalue 1.

Altogether therefore,

$$AU = [Ax, Az] = [3x, z] = U \operatorname{diag}(3,1),$$

showing that $A$ is actually diagonable.

The 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 $e_1$ in common. Hence, 2 is a defective eigenvalue for this matrix $A$. □

The Corollary invites repetition: If $A$ has a further nondefective eigenvalue $\nu \neq \mu$, then, by the Corollary, it is a nondefective eigenvalue for $B = A_{|Z}$, hence, by the Corollary, we can find a basis $W = W_1 := [V_\nu, W_2]$ for $Z$ with $V_\nu$ a basis for $\operatorname{null}(B - \nu\,\mathrm{id})$ and $W_2$ a basis for $Z_2 := \operatorname{ran}(B - \nu\,\mathrm{id})$. Further, $V_\nu$ consists entirely of eigenvectors for $B$, *hence for* $A$, belonging to the eigenvalue $\nu$, and $Z_2$ is invariant under $B$, hence under $A$. Finally, any eigenvalue $\kappa$ of $A$ different from $\mu$ and $\nu$ is nessarily an eigenvalue for $C := B_{|Z_2} = A_{|Z_2}$. Hence, if also $\kappa$ is nondefective, we can continue further, obtaining a split of $Z_2$ into $\operatorname{ran} V_\kappa = \operatorname{null}(C - \kappa\,\mathrm{id}) = \operatorname{null}(A - \kappa\,\mathrm{id})$ and $Z_3 := \operatorname{ran}(C - \kappa\,\mathrm{id})$, with $A_{|Z_3}$ having the same spectrum as $A$, except for $\mu$, $\nu$, $\kappa$.

A formal proof, by induction on the order of $A$, along these lines establishes the following useful

**(6.30)Proposition:** If $\mu_1, \ldots, \mu_r$ are 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 $Z := \operatorname{ran} W$ $A$-invariant, and

$$\operatorname{spec}(A_{|Z}) = \operatorname{spec}(A) \backslash \{\mu_1, \ldots, \mu_r\}.$$

In particular, *if $A$ has no defective eigenvalues, then $A$ is diagonable*, since then we can achieve $W = []$, hence $V$ is already a basis for $X$.

**Three interesting properties of the power sequence of a linear map: The sequel**

**(6.31) Theorem:** Let $A \in \mathbb{C}^{n\times n}$. Then:

(i) $A$ is powerbounded iff , for all $\mu \in \mathrm{spec}(A)$, $|\mu| \le 1$, with $|\mu| = 1$ only if $\mu$ is not defective.

(ii) $A$ is convergent iff , for all $\mu \in \mathrm{spec}(A)$, $|\mu| \le 1$, with $|\mu| = 1$ only if $\mu$ 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.

We begin with (iii). Since $A$ is a matrix over the complex scalars, we know from (6.20)Corollary that, for any $\varepsilon > 0$, we can find an upper triangular matrix $B_\varepsilon$ similar to $A$ and with all off-diagonal entries less than $\varepsilon$ 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

$$\|B_\varepsilon\|_\infty = \max_i \sum_j |B_\varepsilon(i,j)| \le \max_i |B_\varepsilon(i,i)| + (n-1)\varepsilon,$$

since each of those sums involves $n-1$ off-diagonal entries and each such entry is less than $\varepsilon$ in absolute value. Further, $B_\varepsilon$ is upper triangular and similar to $A$, hence

$$\max_i |B_\varepsilon(i,i)| = \max\{|\mu| : \mu \in \mathrm{spec}(A)\} = \rho(A).$$

By assumption, $\rho(A) < 1$. This makes it possible to choose $\varepsilon$ 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 (6.30)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 $Z := \mathrm{ran}\, W$ is $A$-invariant and $\mathrm{spec}(A_{|Z})$ 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 $\mathrm{spec}(B) = \mathrm{spec}(A_{|Z})$, hence $\rho(B) < 1$. Consequently, for any $k$,

$$A^k U = A^k[V, W] = [A^k V, A^k W] = [V\mathrm{M}^k, WB^k] = U\,\mathrm{diag}(\mathrm{M}^k, B^k).$$

In other words,

$$A^k = U\,\mathrm{diag}(\mathrm{M}^k, B^k)U^{-1}.$$

Therefore, $\|A^k\|_\infty \le \|U\|_\infty \max\{\|\mathrm{M}\|_\infty^k, \|B^k\|_\infty\}\|U^{-1}\|_\infty$, and this last expression converges 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 actual equal to 1. Then $\mathrm{M} = \mathrm{id}$, and so, $\lim_{k\to\infty} A^k = C := U\,\mathrm{diag}(\mathrm{M}, 0)U^{-1}$ since $A^k - C = U\,\mathrm{diag}(0, B^k)U^{-1}$, hence

$$\|A^k - C\|_\infty \le \|U\|_\infty \|B^k\|_\infty \|U^{-1}\|_\infty \le \mathrm{const}\|B^k\|_\infty \to 0$$

as $k \to \infty$. □

## The power method

The simple background for the success of the **power method** is the following

**(6.32) Proposition:** If $A$ has just one eigenvalue $\mu$ of absolute value $\rho(A)$ and $\mu$ is nondefective, then, for almost any $x$ and almost any $y$, the sequence

$$A^k x/(y{*}A^k x), \qquad k = 1, 2, \ldots$$

converges to an eigenvector of $A$ belonging to that absolutely maximal eigenvalue $\mu$. In particular, for almost any vector $y$, the ratio

$$y{*}A^{k+1}x/y{*}A^k x$$

converges to $\mu$.

**Proof:** By assumption, there is (by (6.30)Proposition) a basis $U := [V, W]$, with $V$ a basis for the space $\operatorname{null}(A - \mu\,\mathrm{id})$ of all eigenvectors of $A$ belonging to that absolutely largest eigenvalue $\mu$ 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 $x =: [V, W](a, b)$,

$$A^k x = \mu^k V a + B^k W b = \mu^k \left(V a + (B/\mu)^k W b\right)$$

and $(B/\mu)^k W b \to 0$ as $k \to \infty$. Consequently, for any $y$,

$$\frac{y{*}A^{k+1}x}{y{*}A^k x} = \frac{\mu^{k+1}(y{*}Va + y{*}(B/\mu)^{k+1}Wb)}{\mu^k(y{*}Va + y{*}(B/\mu)^k Wb)} = \mu\frac{y{*}Va + y{*}(B/\mu)^{k+1}Wb}{y{*}Va + y{*}(B/\mu)^k Wb} \to \mu$$

provided $y{*}Va \neq 0$. □

Note that the speed with which $y{*}A^{k+1}x/y{*}A^k x$ converges to $\mu$ depends on the speed with which $(B/\mu)^k W b \to 0$ as $k \to \infty$, hence, ultimately, on $\rho(B/\mu)$.

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.

## Nondefective eigenvalues; the sequel

We proved earlier that a matrix without any defective eigenvalues is diagonable. It is worthwhile to point out that the converse also holds, i.e., that all the eigenvalues of a diagonable matrix are nondefective.

**(6.33) Theorem:** Let $A \in L(X)$, with $X$ a finite-dimensional vector space over the complex scalars. Then, $A$ is diagonable if and only if none of its eigenvalues is defective.

**Proof:** We already observed '$\Longleftarrow$' as an extreme case of (6.30)Proposition.

For the proof of '$\Longrightarrow$', assume that $A$ is diagonable. This means that $A = V\mathrm{M}V^{-1}$ for some basis $V = [v_1, v_2, \ldots, v_n]$ for $X$ and some diagonal matrix $\mathrm{M} = \operatorname{diag}(\mu_1, \mu_2, \ldots, \mu_n)$. Let $\mu$ be an eigenvalue of $A$. We need to show that $\operatorname{null}(A - \mu\,\mathrm{id}) \cap \operatorname{ran}(A - \mu\,\mathrm{id}) = \{0\}$. For this, let $z \in \operatorname{null}(A - \mu\,\mathrm{id}) \cap \operatorname{ran}(A - \mu\,\mathrm{id})$. Then $z = (A - \mu\,\mathrm{id})Va$ for some coordinate vector $a$ and

$$0 = (A - \mu\,\mathrm{id})z = (A - \mu\,\mathrm{id})^2 Va = V(\mathrm{M} - \mu\,\mathrm{id})^2 a = \sum_j v_j\,(\mu_j - \mu)^2 a_j,$$

therefore $a_j = 0$, for all $j$ with $\mu_j \neq \mu$. This implies that

$$z = (A - \mu\,\mathrm{id})Va = (A - \mu\,\mathrm{id})\sum_{\mu_j = \mu} v_j a_j = \sum_{\mu_j = \mu}(A - \mu\,\mathrm{id})v_j a_j = 0,$$

since the sum now involves explicitly only $v_j$ for which $\mu_j = \mu$, hence for which $(A - \mu\,\mathrm{id})v_j = 0$. □

**(6.34) Example** Here is a concrete example, chosen for its simplicity.

Let $A = \begin{bmatrix} 1 & 1 \\ 0 & \alpha \end{bmatrix}$. Then $\mathrm{spec}(A) = \{1, \alpha\}$. In particular, $A$ is diagonable if $\alpha \neq 1$ (by (6.36)Proposition below) since then $A$ has two eigenvalues. On the other hand, if $\alpha = 1$, then $A$ is not diagonable since it then looks like $\mathrm{id}_2 + N$, with $N := \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ the simplest example of a non-diagonable matrix. Also, in the latter case, the sole eigenvalue, 1, is certainly defective since $e_1$ is both in $\mathrm{null}(A - \mathrm{id})$ and in $\mathrm{ran}(A - \mathrm{id})$.

Also,

$$A^k = \begin{bmatrix} 1 & 1 + \alpha + \cdots + \alpha^{k-1} \\ 0 & \alpha^k \end{bmatrix} = \begin{cases} \begin{bmatrix} 1 & \frac{1-\alpha^k}{1-\alpha} \\ 0 & \alpha^k \end{bmatrix}, & \alpha \neq 1; \\ \begin{bmatrix} 1 & k \\ 0 & 1 \end{bmatrix}, & \alpha = 1. \end{cases}$$

We see that $A$ is powerbounded whenever $|\alpha| \leq 1$ *except* when $\alpha = 1$, i.e., except when there is a defective absolutely largest eigenvalue. □

**(6.35) Example** Let $A := [x][y]'$ 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 a nontrivial eigenvector for $A$. Since $Az = x\ (y'z)$, we have, in particular,

$$Ax = x\ (y'x),$$

hence $x$ is a nontrivial eigenvector for $A$ belonging to the eigenvalue $\mu := y'x$.

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 $U := [x, V]$ is 1-1 (hence a basis for $\mathbb{R}^n$) if and only if $x \notin \mathrm{ran}\, V$, i.e., if and only if $x \notin \mathrm{null}\, A$.

case $x \notin \mathrm{ran}\, V$: Then $U = [x, V]$ is a basis for $\mathbb{R}^n$. Consider the representation $\widehat{A} = U^{-1}AU$ for $A$ with respect to this basis: We have $Ax = x\ (y'x)$, therefore

$$\widehat{A}e_1 = U^{-1}AUe_1 = U^{-1}Ax = (y'x)e_1,$$

(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., $U(U^{-1}z) = z$). Further, if $V =: [v_1, v_2, \ldots, v_{n-1}]$, then $Au_j = Av_{j-1} = 0$ for $j = 2, \ldots, n$, therefore

$$\widehat{A}e_j = 0, \qquad j = 2, \ldots, n.$$

Hence, altogether,

$$\widehat{A} = [\mu e_1, 0, \ldots, 0].$$

In particular, $A$ is diagonable, with eigenvalues $y'x$ and 0.

case $x \in \mathrm{ran}\, V$: Then $U = [x, V]$ is not a basis for $\mathbb{R}^n$. Worse than that, $A$ is now not diagonable. 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

$$\{0\} \neq \mathrm{ran}(A - 0\,\mathrm{id}) = \mathrm{ran}\, A = \mathrm{ran}[x] \subseteq \mathrm{null}\, A = \mathrm{null}(A - 0\,\mathrm{id}).$$

Therefore $\mathrm{null}(A - 0\,\mathrm{id}) \cap \mathrm{ran}(A - 0\,\mathrm{id}) \neq \{0\}$. □

Whether or not an eigenvalue for $A$ is defective has to be decided case by case. But, if $A \in L(X)$ has as many eigenvalues as $X$ has dimensions, then they are all necessarily nondefective. This is due to the following proposition.

**(6.36) Proposition:** If $V = [v_1, v_2, \ldots, v_r]$ consists of nonzero eigenvectors of $A$ belonging to different eigenvalues, then $V$ is 1-1. In particular, $A$ can have at most $\dim X$ different eigenvalues. Further, if $A$ has as many different eigenvalues as $X$ has dimensions, i.e., if $\#\mathrm{spec}(A) = \dim X$, then $A$ is diagonable.

**Proof:** By induction on $r$, it being true for $r = 1$ (since each $v_j$ is nonzero, by assumption). Assuming that $V := [v_1, v_2, \ldots, v_{r-1}]$ is 1-1, we need, by (3.4)Lemma, only show that $v_r \notin \operatorname{ran} V$ in order to conclude that also $[V, v_r]$ is 1-1. So, assume to the contrary that $v_r = Va$ and let $Av_j =: v_j\mu_j$, all $j$. Then

$$0 = (A - \mu_r \operatorname{id})v_r = (A - \mu_r \operatorname{id})Va = V(\mathrm{M} - \mu_r \operatorname{id})a = V\ (\ldots, (\mu_j - \mu_r)a_j, \ldots),$$

and, since $V$ is 1-1 by induction hypothesis, it would follow that $(\mu_j - \mu_r)a_j = 0$ for all $j < r$. But since all the eigenvalues are different, this would imply that $a_j = 0$ for all $j < r$, therefore $a = 0$, hence $v_r = Va = 0$, and this contradicts the assumption that $v_r$ is nonzero. □

It is hard to tell just by looking at a matrix whether or not it is diagonable. There is one exception: If $A$ is hermitian, i.e., equal to its conjugate transpose, then it is not only diagonable, but has an orthonormal basis of eigenvectors, as we will show next.

## 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

$$V'V = \operatorname{id}.$$

**(6.37) 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 $\widehat{A} := U^{-1}AU = U'AU$ is upper triangular.

**Proof:** Simply repeat the proof of (6.19)Theorem, with the following modifications: Normalize the nontrivial eigenvector $u_1$, 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 $[u_1, \ldots]$ 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 $\operatorname{diag}(1, W)$. □

Here is one of the many consequences of Schur's theorem. It concerns **hermitian** matrices, i.e., matrices $A$ for which $A' = A$. 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 $U$ , $\widehat{A} := U'AU$ is upper triangular. On the other hand, for any matrix $A$ and any unitary matrix $U$,

$$(U'AU)' = U'(A')U.$$

In other words: if $\widehat{A}$ is the matrix representation for $A$ with respect to a *unitary* basis, then $\widehat{A}'$ is the matrix representation for $A'$ with respect to the very same basis. For our hermitian matrix $A$ with its upper triangular matrix representation $\widehat{A} = U'AU$ with respect to the unitary basis $U$, this means that also $\widehat{A}' = \widehat{A}$, 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

**(6.38) Corollary:** A matrix $A \in \mathbb{C}^n$ is hermitian if and only it is unitarily similar to a real diagonal matrix.

**Proof:** We still have to prove that if $\widehat{A} := U'AU$ is real and diagonal for some unitary $U$, then $A$ is necessarily hermitian. But that follows at once from the fact that then $\widehat{A}' = \widehat{A}$, therefore $A' = (U\widehat{A}U')' = U\widehat{A}'U' = U\widehat{A}U' = A$. □

A slightly more involved argument makes it possible to characterize all those matrices which 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 $\widehat{A} := U'AU$ is diagonal, then

$$A'A = (U\widehat{A}'U')(U\widehat{A}U') = U\widehat{A}'\widehat{A}U' = U\widehat{A}\widehat{A}'U' = (U\widehat{A}U')(U\widehat{A}'U') = AA',$$

hence having $A'A = AA'$ 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 diagona(liza)bility of a matrix available which 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 which 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 nontrivial eigenvector in common.

Assuming this refinement of Schur's theorem (cf. (6.41)Theorem below), one would obtain, for a given normal matrix $A$, a unitary basis $U$ so that both $U'AU$ and $U'A'U$ are upper triangular. Since one of these is the conjugate transpose of the other, they must both be diagonal. This finishes the proof of

**(6.39) Theorem:** A matrix $A \in \mathbb{C}^n$ is unitarily similar to a diagonal matrix if and only if $AA' = A'A$.

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

**(6.40) Lemma:** If $A, B \in \mathbb{C}^n$ commute, then there exists a nontrivial vector which is eigenvector for both of them.

**Proof:** Let $x \neq 0$ be an eigenvector for $A$, $Ax = x\mu$ say, and consider the $B$-invariant subspace generated by $x$, i.e., the **Krylov space**

$$K = K_{B,x} = \{p(B)x : p \in \Pi\} = \operatorname{ran}[x, Bx, B^2x, \ldots].$$

This is a nontrivial $B$-invariant linear subspace, hence $B_{|K}$ has a nontrivial eigenvector, i.e., $Bv = v\nu$ for some $v \in K\backslash 0$ (and some scalar $\nu$). Since $v \in K$, we have $v = p(B)x$ for some $p \in \Pi$, and since $AB = BA$, we also have $Ap(B) = p(B)A$, therefore

$$Av = Ap(B)x = p(B)Ax = p(B)x\mu = v\mu,$$

showing that our nontrivial eigenvector $v$ for $B$ is also an eigenvector for $A$. □

**(6.41) Schur's refined theorem:** For every $A, B \in L(\mathbb{C}^n)$ which commute, there exists a unitary basis $U$ for $\mathbb{C}^n$ so that both $U'AU$ and $U'BU$ are upper triangular.

**Proof:** This is a further refinement of the proof of (6.19)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 (6.40)Lemma, we may choose $u_1$ 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 $\nu$, we have

$$U'AU = \begin{bmatrix} \mu & C \\ 0 & A_1 \end{bmatrix} \qquad U'BU = \begin{bmatrix} \nu & D \\ 0 & B_1 \end{bmatrix}.$$

Therefore

$$\begin{aligned} \begin{bmatrix} \mu\nu & \mu D + CB_1 \\ 0 & A_1B_1 \end{bmatrix} &= \begin{bmatrix} \mu & C \\ 0 & A_1 \end{bmatrix} \begin{bmatrix} \nu & D \\ 0 & B_1 \end{bmatrix} \\ &= U'AUU'BU = U'ABU = U'BAU = U'BU\,U'AU = \begin{bmatrix} \nu\mu & \nu C + DA_1 \\ 0 & B_1A_1 \end{bmatrix}, \end{aligned}$$

hence also $A_1$ and $B_1$ commute. □

## 7. 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 which must contain parts or all of the spectrum of $A$. To give a simple example, we proved (in (6.38)Corollary) that all the eigenvalues of a hermitian matrix must be real, i.e., that $\mathrm{spec}(A) \subset \mathbb{R}$ in case $A' = A$.

Since $\mu \in \mathrm{spec}(A)$ iff $(A - \mu\,\mathrm{id})$ is not invertible, it is not surprising that many localization theorems derive from tests for invertibility.

### Gershgorin's circles

Let $\mu$ be an eigenvalue for $A$ with corresponding eigenvector $x$. Without loss of generality, we may assume that $\|x\| = 1$ in whatever vector norm on $X = \mathrm{dom}\,A$ 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 ball $B_{\|A\|}$ of radius $\|A\|$ centered at the origin. In other words,

$$\rho(A) \le \|A\| \tag{7.1}$$

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 surely 1-1 since then

$$\|Ax\| = \|x - Ex\| \ge \|x\| - \|Ex\| \ge \|x\| - \|E\|\|x\| \ge \|x\|(1 - \|E\|)$$

with the factor $(1 - \|E\|)$ *positive*, hence $Ax = 0$ implies that $\|x\| = 0$.

Now consider a **diagonally dominant** $A$, i.e., a matrix $A$ with the property that

$$\forall i \quad |A(i,i)| > \sum_{j\ne i} |A(i,j)|. \tag{7.2}$$

For example, of the three matrices

$$\begin{bmatrix} 2 & -1 \\ 2 & 3 \end{bmatrix}, \quad \begin{bmatrix} -2 & -1 \\ 3 & 3 \end{bmatrix}, \quad \begin{bmatrix} -2 & -1 \\ 4 & 3 \end{bmatrix}, \tag{7.3}$$

only the first is diagonally dominant. Setting

$$D := \mathrm{diag}\,A = \mathrm{diag}(\ldots, A(i,i), \ldots),$$

we notice that (i) $D$ is invertible (since all its diagonal entries are nonzero); and (ii) the matrix

$$E := -D^{-1}(L+U) : (i,j) \mapsto \begin{cases} -A(i,j)/A(i,i) & i \ne j; \\ 0 & \text{otherwise} \end{cases}$$

has norm

$$\|E\|_\infty = \max_i \sum_{j\ne i} |A(i,j)/A(i,i)| < 1,$$

by the assumed diagonal dominance of $A$. This implies that the matrix $\mathrm{id} - E$ is invertible, therefore, since

$$\mathrm{id}_n - E = \mathrm{id}_n + D^{-1}(L+U) = D^{-1}(D + L + U) = D^{-1}A,$$

also $A$ is invertible. This proves

**(7.4) Proposition:** Any diagonally dominant matrix is invertible.

In particular, the first of the three matrices in (7.3) 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.,

$$\exists i \quad |A(i,i) - \mu| \le \sum_{j\ne i} |A(i,j)|. \tag{7.5}$$

This gives the famous

(7.6) **Gershgorin Circle Theorem**: The spectrum of $A \in \mathbb{C}^{n\times n}$ is contained in the union of circles

$$B_{r_i}(A(i,i)) := \{z \in \mathbb{C} : |A(i,i) - z| \le r_i := \sum_{j\ne i} |A(i,j)|\}, \quad i = 1,\ldots,n.$$

For the three matrices in (7.3), this says that

$$\operatorname{spec}(\begin{bmatrix} 2 & -1 \\ 2 & 3 \end{bmatrix}) \subseteq B_1(2)\cup B_2(3) \quad \operatorname{spec}(\begin{bmatrix} -2 & -1 \\ 3 & 3 \end{bmatrix}) \subseteq B_1(-2)\cup B_3(3) \quad \operatorname{spec}(\begin{bmatrix} -2 & -1 \\ 4 & 3 \end{bmatrix}) \subseteq B_1(-2)\cup B_4(3).$$

More than that, according to a *refinement of the Gershgorin Circle Theorem*, the second matrix must have one eigenvalue in the ball $B_1(-2)$ and another one in the ball $B_3(3)$, since these two balls 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 $-1$ common to the two balls $B_1(-2)$ and $B_4(3)$.

## determinants

The **determinant** is, by definition, the unique multilinear alternating form

$$\det : [a_1,\ldots,a_n] \to \mathbb{F}$$

for which

$$\det(\mathrm{id}_n) = 1. \tag{7.7}$$

Here, **multilinear** means that det is *linear* in each of its $n$ arguments, i.e.,

$$\det[\ldots, a + \alpha b, \ldots] = \det[\ldots, a, \ldots] + \alpha \det[\ldots, b, \ldots]. \tag{7.8}$$

(Here and below, the various ellipses ... indicate the other arguments, the ones that are kept fixed.) Further, **alternating** means that the interchange of two arguments reverses the sign, i.e.,

$$\det[\ldots, a, \ldots, b, \ldots] = -\det[\ldots, b, \ldots, a, \ldots].$$

In particular, $\det A = 0$ in case two columns of $A$ are the same, i.e.,

$$\det[\ldots, b, \ldots, b, \ldots] = 0.$$

Combining this last with (7.8), we find that

$$\det[\ldots, a + \alpha b, \ldots, b, \ldots] = \det[\ldots, a, \ldots, b, \ldots],$$

i.e., *addition of a scalar multiple of one argument to a different argument does not change the determinant.*

In particular, *if* $A = [a_1, a_2, \ldots, a_n]$ *is not invertible, then* $\det A = 0$ since then there must be some column $a_j$ of $A$ writable as a linear combination of other columns, i.e.,

$$\det A = \det[\ldots, a_j, \ldots] = \det[\ldots, 0, \ldots] = 0,$$

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

$$\det(AB) = \det(A)\det(B) \tag{7.9}$$

since, for an invertible $A$,

$$1 = \det(\,\mathrm{id}_n) = \det(AA^{-1}) = \det(A)\det(A^{-1}),$$

the first equality by (7.7).

**(7.10) Theorem:** For all $A \in \mathbb{C}^{n\times n}$, $\mathrm{spec}(A) = \{\mu \in \mathbb{C} : \det(A - \mu\,\mathrm{id}) = 0\}$.

Of course, this theorem is quite useless unless we have in hand an explicit formula for the determinant. Here is the standard formula:

$$\det[a_1, a_2, \ldots, a_n] = \sum_{\mathbf{i}\in \$_\mathbf{n}} (-)^{\mathbf{i}} \prod_j a_j(\mathbf{i}(j)) \tag{7.11}$$

in which the sum is over all permutations of order $n$, i.e., all 1-1 (hence invertible) maps $\mathbf{i} : \{1, \ldots, n\} \to \{1, \ldots, n\}$, and the number $(-)^{\mathbf{i}}$ is 1 or $-1$ depending on the **parity** of the number of interchanges it takes to bring the sequence $\mathbf{i}$ back into increasing order.

For $n = 1$, we get the trivial fact that, for any scalar $\alpha$ $\mathrm{spec}([a]) = \{a\}$.

For $n = 2$, (7.10) implies that

$$\mathrm{spec}(\begin{bmatrix} a & b \\ c & d \end{bmatrix}) = \{\mu \in \mathbb{C} : (a - \mu)(d - \mu) = bc\}.$$

For $n = 3$, we get

$$\mathrm{spec}(\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}) = \{\mu \in \mathbb{C} : p(\mu) = 0\},$$

with

$$p(\mu) := (a - \mu)(e - \mu)(i - \mu) + bfg + chd - c(e - \mu)g - (a - \mu)fh - bd(i - \mu).$$

For $n = 4$, (7.11) 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 (7.9) and on the following

**(7.12) Lemma:** The determinant of any triangular matrix is just the product of its diagonal entries.

**Proof:** This observation follows at once from (7.11) since any permutation $\mathbf{i}$ other than the identity $(1, 2, \ldots, n)$ must have $\mathbf{i}(k) < k$ for some $k$, hence the corresponding product $\prod_j a_j(\mathbf{i}(j))$ in (7.11) will be zero for any lower triangular matrix. Since any such $\mathbf{i}$ must also have $\mathbf{i}(h) > h$ for some $h$, the corresponding product will also vanish for any upper triangular matrix. Thus, in either case, only the product $\prod_j a_j(j)$ is possibly nonzero. $\square$

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

$$\det A = (-)^P \prod_j U(j,j),$$

with the number $(-)^P$ 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 (7.11) is often taken as the definition of $\det A$. It is a simple consequence of the fundamental identity (7.9), and the latter follows readily from the multilinearity and alternation property of the determinant. See the chapter 'More on determinants'.

### Annihilating polynomials

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**:

**(7.13) Theorem:** For any polynomial $p$ and any linear map $A \in L(X)$ with $\mathbb{F} = \mathbb{C}$,

$$\operatorname{spec}(p(A)) = p(\operatorname{spec}(A)) := \{p(\mu) : \mu \in \operatorname{spec}(A)\}.$$

**Proof:** If $\mu \in \operatorname{spec}(A)$, then, for some nonzero $x$, $Ax = \mu x$, therefore also $p(A)x = p(\mu)x$, hence $p(\mu) \in \operatorname{spec}(p(A))$.

Conversely, if $\nu \in \operatorname{spec}(p(A))$, 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 $\mu_j \in \operatorname{spec}(A)$, while $p(\mu_j) - \nu = 0$. In other words, $\operatorname{spec}(p(A)) \subset p(\operatorname{spec}(A))$. $\square$

In particular, if $p$ annihilates $A$, then $p(A) = 0$, hence $\operatorname{spec}(p(A)) = \{0\}$, therefore $\operatorname{spec}(A) \subset \{\mu \in \mathbb{C} : p(\mu) = 0\}$.

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

$$\chi_A : t \mapsto \det(t\,\mathrm{id}_n - A).$$

To be sure, by (6.19), we can write any such $A$ as the product $A = V\widehat{A}V^{-1}$ with $\widehat{A}$ upper triangular. Correspondingly,

$$\chi_A(t) = \det V \det(t\,\mathrm{id}_n - \widehat{A})(\det V)^{-1} = \det(t\,\mathrm{id}_n - \widehat{A}) = \chi_{\widehat{A}}(t) = \prod_j (t - \widehat{A}(j,j)),$$

the last equation by (7.12)Lemma. Consequently, $\chi_A(A) = V\chi_A(\widehat{A})V^{-1}$, with

$$\chi_A(\widehat{A}) = (\widehat{A} - \mu_1\,\mathrm{id})\cdots(\widehat{A} - \mu_n\,\mathrm{id}), \qquad \mu_j := \widehat{A}(j,j), \quad j = 1,\ldots,n,$$

and this, we claim, is necessarily the zero map, for the following reason: The factor $(\widehat{A} - \mu_j\,\mathrm{id})$ is upper triangular, with the $j$th diagonal entry equal to zero. This implies that, for each $i$, $(\widehat{A} - \mu_j\,\mathrm{id})$ maps

$$T_i := \operatorname{ran}[e_1,\ldots,e_i]$$

into itself, but maps $T_j$ into $T_{j-1}$. Therefore

$$\begin{aligned}\operatorname{ran}\chi_A(\widehat{A}) = \chi_A(\widehat{A})T_n &= (\widehat{A} - \mu_1\,\mathrm{id})\cdots(\widehat{A} - \mu_n\,\mathrm{id})T_n\\ &\subseteq (\widehat{A} - \mu_1\,\mathrm{id})\cdots(\widehat{A} - \mu_{n-1}\,\mathrm{id})T_{n-1}\\ &\subseteq (\widehat{A} - \mu_1\,\mathrm{id})\cdots(\widehat{A} - \mu_{n-2}\,\mathrm{id})T_{n-2}\\ &\cdots\\ &\subseteq (\widehat{A} - \mu_1\,\mathrm{id})T_1 \subseteq T_0 = \{0\},\end{aligned}$$

or, $\chi_A(\widehat{A}) = 0$, therefore also $\chi_A(A) = 0$. This is known as the **Cayley-Hamilton Theorem**.

### The multiplicity of an eigenvalue

Since $\chi_A$ is of exact degree $n$ in case $A \in \mathbb{C}^n$, $\chi_A$ has exactly $n$ zeros *counting multiplicities.* This means that

$$\chi_A(t) = (t - \mu_1)\cdots(t - \mu_n)$$

for a certain $n$-sequence $\mu$. Further,

$$\operatorname{spec}(A) = \{\mu_j : j = 1,\ldots,n\},$$

and this set may well contain only one number, as it does when $A = 0$ or $A = \mathrm{id}$. However, it is customary to associate with each eigenvalue, $\mu$, its **algebraic multiplicity**, $\#_a\mu$, which, by definition, is its multiplicity as a zero of the characteristic polynomial. For example, the matrix $\mathrm{id}_n$ as only the eigenvalue 1, but it has algebraic multiplicity $n$. In this way, each $A \in \mathbb{C}^{n\times n}$ has $n$ eigenvalues *counting algebraic multiplicity.*

I have been saying '*algebraic* multiplicity' rather than just 'multiplicity', since there is a second way of counting eigenvalues, and that is by geometric multiplicity. The **geometric multiplicity**, $\#_g\mu$, of the eigenvalue $\mu$ for $A$ is, by definition, the dimension of the space of corresponding eigenvectors, i.e.,

$$\#_g\mu := \dim\operatorname{null}(A - \mu\,\mathrm{id}).$$

For the sole eigenvalue, 1, of $\mathrm{id}_n$, algebraic and geometric multiplicity coincide. 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.

An eigenvalue is (algebraically or geometrically) **simple** if it as (algebraic or geometric) multiplicity 1.

> **7.14 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.

### Perron-Frobenius

We call the matrix $A$ **positive** (**nonnegative**) and write $A > 0$ ($A \ge 0$) in case all its entries are positive (nonnegative). A positive (nonnegative) matrix $A$ of order $n$ maps the **positive orthant**

$$\mathbb{R}^n_+ := \{y \in \mathbb{R}^n \backslash 0 : y \ge 0\}$$

into its interior (into itself). Thus the (scaled) power method started with a nonnegative vector would converge to a nonnegative vector if it converges. This suggests that the absolutely largest eigenvalue for such a matrix is nonnegative, with a corresponding nonnegative eigenvector. The Perron-Frobenius theorem makes this intuition precise.

Since $A$ maps $\mathbb{R}^n_+$ into itself, it makes sense to consider, for given $y \in \mathbb{R}^n_+$, scalars $\alpha$ for which $Ay \ge \alpha y$ (in the sense that $(Ay)_j \ge \alpha y_j$, all $j$), i.e., for which $Ay - \alpha y \ge 0$. The largest such scalar is the nonnegative number

$$r(y) := \min\{(Ay)_j / y_j : y_j > 0\}.$$

The basic observation is that

$$Ay - \alpha y > 0 \quad \Longrightarrow \quad r(y) > \alpha. \tag{7.15}$$

The function $r$ so defined is continuous, hence takes on its supremum on the compact set $S_+ := \{y \ge 0 : \|y\| = 1\}$. I.e., there exists $x \in S_+$ for which $\mu := r(x) = \sup r(S_+)$.

Assume now, in addition to $A \ge 0$, that also $p(A) > 0$ for some polynomial $p$.

Claim: $Ax = \mu x$.

**Proof:** Assume that $Ax \ne \mu x$. Since $\mu = r(x)$, we have $Ax - \mu x \ge 0$, therefore $A(p(A)x) - \mu p(A)x = p(A)(Ax - \mu x) > 0$, hence $r(p(A)x) > \mu = \sup r(S_+)$, a contradiction by (7.15). □

Claim: $x > 0$.

Proof. Since $0 \ne x \ge 0$ and $p(A) > 0$, we have $p(\mu)x = p(A)x > 0$, hence $x > 0$. □

Consequence: $x$ is the unique maximizer for $r$.

**Proof:** If also $r(y) = \mu$ for some $y \in S_+$, then by the same argument $Ay = \mu y$, therefore $Az = \mu z$ for all $z = x + \alpha(y - x)$, and each of these $z$ must be positive if it is nonnegative, and this is possible only if $y - x = 0$. □

Consequence: For any eigenvalue $\nu$ of any matrix $B$ with nontrivial eigenvector $y$, if $|B| \le A$, then $|\nu| \le \mu$, with equality only if $|\ y/\|y\|\ | = x$ and $|B| = A$. (More precisely, equality implies that $B = \exp(\mathrm{i}\varphi) DAD^{-1}$, with $D := \mathrm{diag}(\ldots, y_j/|y_j|, \ldots)$ and $\exp(\mathrm{i}\varphi) := \nu/|\nu|$.)

**Proof:** Observe that

$$|\nu||y| = |By| \le |B|\,|y| \le A|y|, \tag{7.16}$$

hence $|\nu| \le r(|y|) \le \mu$. If now there is equality, then, by the uniqueness of the minimizer $x$ (and assuming without loss that $\|y\| = 1$), we must have $|y| = x$ and equality throughout (7.16), and this implies $|B| = A$. More precisely, $D := \mathrm{diag}(\ldots, y_j/|y_j|, \ldots)$ is then welldefined and satisfies $y = D|y|$, hence $C|y| = \mu|y| = A|y|$, with $C := \exp(-\mathrm{i}\varphi) D^{-1} B D \le A$ and $\nu =: \mu \exp(\mathrm{i}\varphi)$, therefore $C = A$. □

Consequences. By choosing $B = A$, we get that $\mu = \rho(A) := \max\{|\nu| : \nu \in \sigma(A)\}$, and that $\mu$ has geometric multiplicity 1 (as an eigenvalue of $A$).

We also get that $\rho(A)$ is strictly monotone in the entries of $A$, i.e., that $\rho(\widehat{A}) > \rho(A)$ in case $\widehat{A} \ge A \ne A$ (using the fact that $p(A) > 0$ and $\widehat{A} \ge A$ implies that also $q(\widehat{A}) > 0$ for some polynomial $q$; see below).

As a consequence, we find *computable* upper and lower bounds for the spectral radius of $A$:

Claim:

$$\forall \{y > 0\} \;\; r(y) \le \rho(A) \le R(y) := \max_j (Ay)_j / y_j,$$

with equality if and only if $y = \alpha x$ (for some positive $\alpha$). In particular, $\rho(A)$ is the only eigenvalue of $A$ with positive eigenvector.

**Proof:** Assume without loss that $\|y\| = 1$. We already know that for any such $y > 0$, $r(y) \le \rho(A)$ with equality if and only if $y = x$. For the other inequality, observe that $R(y) = \|D^{-1}ADe\|_\infty$ with $D := \mathrm{diag}(\ldots, y_j/|y_j|, \ldots)$ and $e := (1, \ldots, 1)$. Since $D^{-1}AD \ge 0$, it takes on its max-norm at $e$, hence

$$R(y) = \|D^{-1}AD\|_\infty \ge \rho(D^{-1}AD) = \rho(A).$$

Now assume that $r(y) = R(y)$. Then $Ay = r(y)y$, hence $r(y) \le r(x) = \rho(A) \le R(y) = r(y)$, therefore equality must hold throughout and, in particular, $y = x$.

If, on the other hand, $r(y) < R(y)$, then we can find $\widehat{A} \ne A \le \widehat{A}$ so that $\widehat{A}y = R(y)y$, therefore $r_{\widehat{A}}(y) = R(y) = R_{\widehat{A}}(y)$, hence $R(y) = \rho(\widehat{A}) > \rho(A)$. □

Claim: $\mu$ has simple *algebraic* multiplicity.

**Proof:** Since we already know that $\mu$ has simple geometric multiplicity, it suffices to show that $\mu$ is not a defective eigenvalue, i.e., that $\mathrm{null}(A - \mu) \cap \mathrm{ran}(A - \mu) = \{0\}$. So assume to the contrary that $Ay - \mu y$ is a nontrivial eigenvector of $A$ belonging to $\mu$. Then, by the simple geometric multiplicity of $\mu$, we may assume without loss that $Ay - \mu y = x$, or $Ay = \mu y + x$, therefore, for all $k$, $A^k y = \mu^k y + k\mu^{k-1}x$, hence, finally,

$$(A/\mu)^k y = y + k(x/\mu).$$

Hence, for large enough $k$, $z := (A/\mu)^k y$ has all its entries positive, and $Az = Ay + kx = \mu y + (k+1)x = \mu(z + x/\mu) > \mu z$, therefore $r(z) > \mu$, 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., $p(t) = t$). Frobenius extended it to all $A \ge 0$ which 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: 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 $A^k(i, j) > 0$.

**(iii)** No proper $A$-invariant subspace is spanned by unit-vectors.

**(iv)** For no permutation matrix $P$ is

$$PAP^{-1} = \begin{bmatrix} B & C \\ 0 & D \end{bmatrix} \qquad (7.17)$$

with $B, D$ square matrices of positive order.

**(v)** The directed graph for $A$ is strongly connected.

Indeed, (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 $A^k(i, j) = 0$ for all $k$. But then also $p(A)(i, j) = 0$ for all polynomials $p$; in other words, (i)$\Longrightarrow$(ii). Further, it says that the set $J := J(j) := \{r : \exists\{k\}\; A^k(r, j) \ne 0\}$

is a proper subset of $\{1,\ldots,n\}$ (since it doesn't contain $i$), but neither is it empty ( since it contains $j$, as $A^0(j,j)\neq 0$). Since $A^{k+\ell}(r,j)=\sum_m A^k(r,m)A^\ell(m,j)$, it follows that $J(m)\subseteq J(j)$ for all $m\in J(j)$. This implies, in particular, that $A(r,m)=0$ for all $r\notin J(j), m\in J(j)$, hence that $\operatorname{span}(e_m)_{m\in J}$ is a proper $A$-invariant subspace, thus implying not(iii). It also implies not(iv), since it shows that the columns $A(:,m), m\in J$, have zero entries in rows $r, r\notin J$, i.e., that (7.17) holds for the permutation $P=[(e_m)_{m\in J},(e_r)_{r\notin J}]$, with both $B$ and $D$ of order $<n$.

Conversely, if e.g., (iii) does not hold, and $\operatorname{span}(e_m)_{m\in J}$ is that proper $A$-invariant subspace, then it is also invariant under any $p(A)$, hence also $p(A)(r,m)=0$ for every $r\notin J$, $m\in J$, i.e., (i) does not hold.

The final characterization is explicitly that given by Frobenius. He observed that

$$A^k(i,j)=\sum_{j_1}\cdots\sum_{j_{k-1}}A(i,j_1)\cdots A(j_{k-1},j),$$

hence that, for $i\neq j$, $A^k(i,j)\neq 0$ 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 $r$. He introduced the **directed graph** of $A$ as the graph with $n$ vertices in which the directed edge $(i,j)$ is present iff $A(i,j)\neq 0$. 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 $A^k(i,j)\neq 0$ for some $k>0$.

## 8. Some applications

### Markov Chains

Recall 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 square matrix $M$ all nonnegative and all entries in any particular row adding up to 1. In other words, $M \ge 0$ and $Me = e$, with

$$e := (1, 1, \ldots, 1).$$

In particular, $1 \in \operatorname{spec}(M)$. Further, since $\|M\|_\infty = 1$, we conclude from (7.1) that $\rho(M) \le 1$. 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 (6.31)Theorem, this implies that $M$ is convergent. In fact, since 1 is a nondefective simple eigenvalue of $M$ with corresponding eigenvector $e$, there is a basis $V = [e, W]$, with $W$ a basis for $\operatorname{ran}(M - \mathrm{id})$, hence

$$MV = [e, MW] = V \operatorname{diag}(1, B)$$

for some $B$ with $\rho(B) < 1$. Therefore,

$$M^k V = V \operatorname{diag}(1, B^k) \xrightarrow[k \to \infty]{} V \operatorname{diag}(1, 0).$$

In other words,

$$\lim_{k \to \infty} M^k = eu',$$

with $M'u = u$, i.e., $u$ is a nontrivial eigenvector of $M'$ belonging to the eigenvalue 1. In particular, all rows of $M^k$ converge to this particular nonnegative vector whose entries sum to 1.

### 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 $x_1, x_2, \ldots, x_n, x_{n+1} := x_1 \in \mathbb{R}^2$ of a closed polygon.

**for** $j = 1 : n$**, do:** $y_{2j-1} \leftarrow (3x_j + x_{j+1})/4$; $y_{2j} \leftarrow (x_j + 3x_{j+1})/4$; **enddo**

**output:** the vertices $y_1, y_2, \ldots, y_{2n}, y_{2n+1} := y_1 \in \mathbb{R}^2$ of a closed polygon that is inscribed into the input polygon.

In other words,

$$[y_1, \ldots, y_{2n}] = [x_1, \ldots, x_n] C_n,$$

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

$$[x_1^{(k)}, \ldots, x_{2^k n}^{(k)}] := [x_1, \ldots, x_n] C_n C_{2n} \cdots C_{2^k n}$$

converges to a smooth curve, namely the curve

$$t \mapsto \sum_j x_j B_2(t-j),$$

with $B_2$ a certain smooth piecewise quadratic function, a so-called quadratic B-spline (whatever that may be).

Here, we consider the following much simpler and more radical corner-cutting:

$$[y_1, \dots, y_n] = [x_1, \dots, x_n]A,$$

with

$$A := \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 1 \\ 1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 1 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 1 \end{bmatrix} /2. \tag{8.1}$$

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

$$[x_1^{(k)}, \dots, x_n^{(k)}] := [x_1, \dots, x_n]A^k$$

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, $\sum_j x_j/n$, of the original vertices.

(i) The matrix $A$, defined in (8.1), 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 current row. All such matrices have eigenvectors of the form

$$u_\lambda := (\lambda^1, \lambda^2, \dots, \lambda^n),$$

with the scalar $\lambda$ chosen so that $\lambda^n = 1$, hence $\lambda^{n+1} = \lambda$. For our $A$, we compute

$$Au_\lambda = (\lambda^n + \lambda^1, \lambda^1 + \lambda^2, \dots, \lambda^{n-1} + \lambda^n)/2.$$

Hence, if $\lambda^n = 1$, then

$$Au_\lambda = \frac{1+\lambda}{2\lambda} u_\lambda.$$

(ii) The equation $\lambda^n = 1$ has exactly $n$ distinct solutions, namely the $n$ **roots of unity**

$$\lambda_j := \exp(2\pi \mathrm{i} j/n) = \omega^j, \quad j = 1, \dots, n.$$

Here,

$$\omega := \omega_n := \exp(2\pi \mathrm{i}/n)$$

is a **primitive $n$th root of unity**. Note that

$$\overline{\omega} = 1/\omega.$$

Let

$$V = [v_1, \dots, v_n] := [u_{\lambda_1}, \dots, u_{\lambda_n}]$$

be the column map whose $j$th column is the eigenvector

$$v_j := (\omega^j, \omega^{2j}, \ldots, \omega^{nj})$$

of $A$, with corresponding eigenvalue

$$\mu_j := \frac{1+\lambda_j}{2\lambda_j} = (\omega^{-j}+1)/2, \quad j = 1, \ldots, n.$$

Since these eigenvalues are distinct, $V$ is 1-1 (by (6.34)Proposition), 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

$$A^k = V \operatorname{diag}(\ldots, \mu_j^k, \ldots) V^{-1} \xrightarrow[k \to \infty]{} V \operatorname{diag}(0, \ldots, 0, 1) V^{-1}$$

since $|\mu_j| < 1$ for $j < n$, while $\mu_n = 1$. Hence

$$\lim_{k\to\infty} A^k = v_n V^{-1}(n, :).$$

(iv) In order to compute $V^{-1}(n,:)$, we compute $V'V$ (recalling that $\overline{\omega^r} = \omega^{-1}$):

$$(V'V)(j,k) = v_j' v_k = \sum_{r=1}^n \omega^{-rj}\, \omega^{rk} = \sum_{r=1}^n \omega^{(k-j)r}.$$

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

$$V'V = n \operatorname{id}_n,$$

i.e., $V/\sqrt{n}$ is *unitary*, i.e., an o.n. basis for $\mathbb{C}^n$. In particular, $V^{-1} = V'/n$, therefore

$$V^{-1}(n,:) = v_n'/n.$$

(v) It follows that

$$\lim_{k\to\infty} A^k = (1/n) v_n v_n',$$

with

$$v_n = (1, 1, \ldots, 1).$$

Consequently,

$$\lim_{k\to\infty} [\ldots, x_j^{(k)}, \ldots] \;=\; \sum_j x_j/n\, v_n' \;=\; [\ldots, \sum_j x_j/n, \ldots],$$

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.

### Tridiagonal Toeplitz matrix

Circulants are a special case of *Toeplitz* matrices, i.e., of matrices which are constant along diagonals. Precisely, the matrix $A$ is **Toeplitz** if

$$A(i,j) = a_{i-j}, \quad \forall i, j,$$

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 $a$ 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 still works, i.e., we can fashion some eigenvectors For such a matrix, we can still **to be continued**

## 9. Optimization and quadratic forms

### Minimization

We are interested in *minimizing* a given function

$$f : \mathrm{dom}\, f \subseteq \mathbb{R}^n \to \mathbb{R},$$

i.e., we are looking for $x \in \mathrm{dom}\, f$ so that

$$\forall y \in \mathrm{dom}\, f \quad f(x) \le f(y).$$

Any such $x$ is called a **minimizer for** $f$; in symbols:

$$x \in \mathrm{argmin}\, f.$$

The discussion applies, of course, also to finding some $x \in \mathrm{argmax}\, f$, i.e., finding a **maximizer** for $f$, since $x \in \mathrm{argmax}\, f$ iff $x \in \mathrm{argmin}(-f)$.

Finding minimizers is, in general, an impossible problem since one cannot tell whether or not $x \in \mathrm{argmin}\, f$ except by checking *every* $y \in \mathrm{dom}\, f$ to make certain that, indeed, $f(x) \le f(y)$. However, if $f$ is a 'smooth' function, then one can in principle check whether, at least, $x$ is a **local minimizer**, i.e., whether $f(x) \le f(y)$ for all 'nearby' $y$, by checking whether the **gradient**

$$Df(x) = (D_i f(x) : i = 1, \ldots, n)$$

of $f$ at $x$ is zero. Here, $D_i f = \partial f / \partial x_i$ is the derivative of $f$ with respect to its $i$th argument.

To be sure, the vanishing of the gradient of $f$ at $x$ is only a *necessary* condition for $x$ 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 which are neither local minima nor local maxima but are, perhaps, only saddle points. By definition, any point $x$ for which $Df(x) = 0$ is a **critical point** for $f$.

At a critical point, $f$ is locally flat. This means that, in the Taylor expansion

$$f(x+h) = f(x) + (Df(x))'h + h'(D^2 f(x)/2)h + \text{h.o.t.}(h)$$

for $f$ at $x$, the linear term, $(Df(x))'h$, is zero. Thus, if the matrix

$$H := D^2 f(x) = (D_i D_j f(x) : i, j = 1, \ldots, n)$$

of second derivatives of $f$ is not just the zero matrix, then $x$ is a local minimizer (maximizer) for $f$ if and only if 0 is a minimizer (maximizer) for the **quadratic form**

$$\mathbb{R}^n \to \mathbb{R} : h \mapsto h'Hh$$

associated with the **Hessian** $H = D^2 f(x)$ **for** $f$ **at** $x$.

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

$$H' = H.$$

However, in the contrary case, one simply defines $H$ to be

$$H := (D^2 f(x) + (D^2 f(x))')/2,$$

thus making it real symmetric while, still,

$$h'Hh = h'D^2 f(x)h, \quad \forall h \in \mathbb{R}^n.$$

In any case, it follows that *quadratic forms model the behavior of a smooth function 'near' a critical 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, via

$$q_A : \mathbb{R}^n \to \mathbb{R} : x \mapsto x'Ax.$$

However, as we already observed, the quadratic form 'sees' only the **symmetric part**

$$(A + A')/2$$

of $A$, i.e.,

$$\forall x \in \mathbb{R}^n \quad x'Ax = x'(A + A')/2\; x.$$

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

$$q_A(x+h) = (x+h)'A(x+h) = x'Ax + x'Ah + h'Ax + h'Ah = q_A(x) + 2(Ax)'h + h'Ah,$$

using the fact that $A' = A$, thus $h'Ax = x'Ah = (Ax)'h$, hence

$$Dq_A(x) = 2Ax, \quad D^2 q_A(x) = 2A.$$

It follows that, for any 1-1 $A$, 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:

$$A \text{ is } \begin{array}{c} \textbf{positive} \\ \textbf{positive semi-} \\ \textbf{negative semi-} \\ \textbf{negative} \end{array} \textbf{definite} \; := \; \begin{array}{ll} \forall x \ne 0 & x'Ax > 0 \\ \forall x & x'Ax \ge 0 \\ \forall x & x'Ax \le 0 \\ \forall x \ne 0 & x'Ax < 0 \end{array} \iff 0 \text{ is } \begin{array}{l} \textbf{the unique minimizer} \\ \textbf{a minimizer} \\ \textbf{a maximizer} \\ \textbf{the unique maximizer} \end{array} \text{ for } q_A.$$

If none of these conditions obtain, i.e., if there exist $x$ and $y$ so that $x'Ax < 0 < y'Ay$, then $q_A$ is called **indefinite** and, in this case, 0 is a **saddle point** for $q_A$.

(xxx)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 (xxx)Figure, in which the level lines are parabolas 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 $x$, a smooth $f$ behaves like its quadratic term $h \mapsto h'(D^2 f(x)/2)h$, 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.

These two patterns turn out to be the only two possible ones for definite quadratic forms on $\mathbb{R}^2$. On $\mathbb{R}^n$, there are only $\lceil (n+1)/2 \rceil$ possible distinct patterns, as follows from the fact that, *for every quadratic form $q_A$, there are o.n. coordinate systems $U$ for which*

$$q_A(x) = \sum_{i=1}^{n} d_i \; (Ux)_i^2.$$

### 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 $f(x) = (Vx)'AVx = x'(V'AV)x$, hence

$$q_A \circ V = q_{V'AV}.$$

This makes it interesting to look for bases $V$ for which $V'AV$ is as simple as possible. Matrices $A$ and $B$ for which $B = V'AV$ 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 (9.8) 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.

**(9.1) Proposition:** Every quadratic form $q_A$ on $\mathbb{R}^n$ can be written in the form

$$q_A(x) = \sum_{j=1}^{n} d_j (u_j' x)^2,$$

for some suitable o.n. basis $U = [u_1, u_2, \ldots, u_n]$ for which $U'AU = \operatorname{diag}(d_1, \ldots, d_n) \in \mathbb{R}^n$.

**Proof:** Since $A$ is hermitian, there exists, by (6.38)Corollary, some o.n. basis $U = [u_1, u_2, \ldots, u_n]$ for $\mathbb{F}^n$ for which $U'AU = \operatorname{diag}(d_1, d_2, \ldots, d_n) \in \mathbb{R}^n$. Now use the facts that $U'U = \mathrm{id}_n$ and therefore $q_A(x) = q_{U'AU}(U'x)$ to obtain for $q_A(x)$ the displayed expression. □

What about the classification introduced earlier, into positive or negative (semidefinite)? The proposition permits us to visualize $q_A(x)$ as a weighted sum of squares (with real weights $d_1, d_2, \ldots, d_n$) and $U'x$ 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

$$c_r := \{x \in \mathbb{R}^2 : q_A(x) = r\} = \{x \in \mathbb{R}^2 : d_1(u_1'x)^2 + d_2(u_2'x)^2 = r\}$$

for $r > 0$ is an ellipse, with axes parallel to $u_1$ and $u_2$. If $0 = d_1 < d_2$, then these ellipses turn into parabolas Similarly, if $d_2 < 0$, then the contour line

$$c_r := \{x \in \mathbb{R}^2 : q_A(x) = r\} = \{x \in \mathbb{R}^2 : d_1(u_1'x)^2 + d_2(u_2'x)^2 = r\}$$

for $r < 0$ is an ellipse, with axes parallel to $u_1$ and $u_2$. Finally, if $d_1 < 0 < d_2$, then, for any $r$, the contour line

$$c_r := \{x \in \mathbb{R}^2 : q_A(x) = r\} = \{x \in \mathbb{R}^2 : d_1(u_1'x)^2 + d_2(u_2'x)^2 = r\}$$

is a hyperbola, with axes parallel to $u_1$ and $u_2$.

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**
> $$R_A(x) := \langle Ax, x\rangle / \langle x, x\rangle,$$
> and $R_A(x) = \mu$ in case $Ax = \mu x$.

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 $Ax \notin \mathrm{ran}[x]$. Then the error $h := Ax - R_A(x)x$ in the least-suares approximation to $Ax$ from $\mathrm{ran}[x]$ is not zero, and is perpendicular to $\mathrm{ran}[x]$. Consequently, $\langle Ax, h\rangle = \langle h, h\rangle > 0$, and therefore the value

$$\langle Ax + th, x + th\rangle = \langle Ax, x\rangle + 2t\langle Ax, h\rangle + t^2\langle Ah, h\rangle$$

of the numerator of $R_A(x + th)$ grows linearly for positive $t$, while its denominator

$$\langle x + th, x + th\rangle = \langle x, x\rangle + t^2\langle h, h\rangle$$

grows only quadratically, i.e., much less fast for $t$ near zero. It follows that, in this situation, $R_A(x + th) > R_A(x)$ for all 'small' positive $t$, hence $x$ cannot be a critical point for $R_A$. – To put it differently, for any critical point $x$ for $R_A$, we necessarily have $Ax \in \mathrm{ran}[x]$, therefore $Ax = R_A(x)x$. Of course, that makes any such $x$ an eigenvector with corresponding eigenvalue $R_A(x)$. □

Next, recall from (6.38) that a hermitian matrix is unitarily similar to a real diagonal matrix. This means that we may assume, after some reordering if necessary, that

$$A = UDU'$$

with $U$ unitary and with $\mathrm{M} = \mathrm{diag}(\mu_1, \ldots, \mu_n)$ where

$$\mu_1 \le \mu_2 \le \cdots \le \mu_n.$$

At times, we will write, more explicitly,

$$\mu_j(A)$$

to denote the $j$th eigenvalue of the hermitian matrix $A$ in this ordering. Note that there may be coincidences here, i.e., $\mu_j(A)$ is the $j$th 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 $\langle Ux, Uy\rangle = \langle x, y\rangle$ for any orthonormal $U$. Thus

$$\langle Ax, x\rangle = \langle U\mathrm{M}U'x, x\rangle = \langle \mathrm{M}(U'x), U'x\rangle,$$

and $\langle x, x\rangle = \langle U'x, U'x\rangle$. Therefore

$$R_A(x) = \langle Ax, x\rangle / \langle x, x\rangle = \langle \mathrm{M}(U'x), U'x\rangle / \langle U'x, U'x\rangle = R_{\mathrm{M}}(U'x).$$

This implies that

$$\begin{matrix}\max_x \\ \min_x\end{matrix} R_A(x) = \begin{matrix}\max_y \\ \min_y\end{matrix} R_{\mathrm{M}}(y).$$

On the other hand, since M is diagonal, $\langle \mathrm{M}y, y\rangle = \sum_j \mu_j |y_j|^2$, therefore

$$R_{\mathrm{M}}(y) = \sum_j \mu_j |y_j|^2 / \sum_j |y_j|^2,$$

and this shows that

$$\min_x R_A(x) = \min_y R_{\mathrm{M}}(y) = \mu_1, \qquad \max_x R_A(x) = \max_y R_{\mathrm{M}}(y) = \mu_n.$$

This is **Rayleigh's Principle**. It characterizes the extreme eigenvalues of a hermitian matrix. The intermediate eigenvalues are the solution of more subtle extremum problems. This is the content of the **Courant-Fischer minimax Theorem** and the **N.N. maximin Theorem**. It seems most efficient to combine both in the following

**(9.2) MMM** (or, **maximinimaxi**) **Theorem:** Let $A$ be a hermitian matrix of order $n$, hence $A = UMU'$ for some unitary $U$ and some real diagonal matrix $\mathrm{M} = \mathrm{diag}(\cdots, \mu_j, \ldots)$ with $\mu_1 \le \cdots \le \mu_n$. Then, for $j = 1, \ldots, n$,

$$\max_{\dim G < j} \min_{x \perp G} R_A(x) = \mu_j = \min_{j \le \dim H} \max_{x \in H} R_A(x),$$

with $G$ and $H$ otherwise arbitrary linear subspaces.

**Proof:** If $\dim G < j \le \dim H$, then one can find $y \in H \backslash 0$ with $y \perp G$ (since, with $V$ a basis for $G$ and $W$ a basis for $H$, this amounts to finding a nontrivial solution to the equation $V'W? = 0$, and this system is homogeneous with more unknowns than equations). Therefore

$$\min_{x \perp G} R_A(x) \ \le R_A(y) \le \ \max_{x \in H} R_A(x).$$

Hence,

$$\max_{\dim G < j} \min_{x \perp G} R_A(x) \ \le \ \min_{j \le \dim H} \max_{x \in H} R_A(x).$$

On the other hand, for $G = \mathrm{ran}[u_1, \ldots, u_{j-1}]$ and $H = \mathrm{ran}[u_1, \ldots, u_j]$,

$$\min_{x \perp G} R_A(x) \ = \ \mu_j(A) \ = \ \max_{x \in H} R_A(x).$$

□

The MMM theorem has various useful (and immediate) corollaries.

**(9.3) Interlacing Theorem:** If the matrix $B$ is obtained from the hermitian matrix $A$ by crossing out the $k$th row and column (i.e., $B = A(I, I)$ with $I := (1, \ldots, k-1, k+1, \ldots, n)$), 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 $\mathrm{spec}(A)$. Let $J : \mathbb{F}^{n-1} \to \mathbb{F}^n : x \mapsto (x, 0)$. Then $R_B(x) = R_A(Jx)$ and $\mathrm{ran}\, J = \mathrm{ran}[e_n]\perp$, therefore $\{JG + \mathrm{ran}[e_n] : \dim G < j, G \subset \mathbb{F}^{n-1}\} \subseteq \{G : \dim G < j+1, G \subseteq \mathbb{F}^n\}$. Hence

$$\mu_j(B) = \max_{\dim G < j} \min_{x \perp G} R_A(Jx) = \max_{\dim G < j} \min_{y \perp JG + \mathrm{ran}[e_n]} R_A(y) \le \max_{\dim G < j+1} \min_{y \perp G} R_A(y) = \mu_{j+1}(A).$$

Also, since $\{JH : j \le \dim H, H \subseteq \mathbb{F}^{n-1}\} \ \subseteq \ \{H : j \le \dim H, H \subseteq \mathbb{F}^n\}$,

$$\mu_j(B) \ = \ \min_{j \le \dim H} \max_{x \in H} R_A(Jx) = \min_{j \le \dim H} \max_{y \in JH} R_A(y) \ge \min_{j \le \dim H} \max_{y \in H} R_A(y) = \mu_j(A).$$

□

**(9.4) 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 $\le \max \operatorname{spec}(B)$ and at least $r$ eigenvalues of $A$ must be $\ge \min \operatorname{spec}(B)$.

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

$$f(t) \le g(t) \quad \forall t,$$

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 miminum 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,

**(9.5) Corollary:** If $A$, $B$ are hermitian, and $R_A(x) \le R_B(x) + c$ for some constant $c$ and all $x$, then

$$\mu_j(A) \le \mu_j(B) + c, \quad \forall j.$$

This gives

**(9.6) Weyl's inequalities:** If $A = B + C$, with $A, B, C$ hermitian, then

$$\mu_j(B) + \mu_1(C) \le \mu_j(A) \le \mu_j(B) + \mu_n(C), \quad \forall j.$$

**Proof:** Since $\mu_1(C) \le R_C(x) \le \mu_n(C)$ (by Rayleigh's principle), while $R_B(x) + R_C(x) = R_A(x)$, the preceding corollary provides the proof. □

A typical *application of Weyl's Inequalities* is the observation that, for $A = BB' + C \in \mathbb{F}^{n\times n}$ 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 $\operatorname{rank} BB' \le \operatorname{rank} B \le k$, hence $\mu_j(BB')$ 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

**(9.7) 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

> **(9.8) 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 $B = V'AV$ 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 $H$,

$$\mu_j(B) \;=\; \max_{x\in H} R_B(x) = \max_{x\in H} R_A(Vx) R_{V'V}(x),$$

and this is necessarily positive, since $\dim VH = j$ and $R_{V'V}(x) = (\|Vx\|^2/\|x\|^2)^2$ is positive for any $x \neq 0$. □

It follows that we don't have to diagonalize the real symmetrix matrix $A$ (as we did in the proof of (9.1)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 $A = LDL'$, 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 $D_\mu$ obtained in the factorization $L_\mu D_\mu L'_\mu$ for $(A - \mu\,\mathrm{id})$ tells us the number of eigenvalues of $A$ to the right (left) of $\mu$, hence makes it easy to locate and refine intervals which contain just one eigenvalue of $A$.

## 10. More on determinants

### Definition and basic properties

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. I have found that, of the many, many determinant identities available, in the end I have always managed with just one nontrivial one, viz. *Sylvester's determinant identity*, and this is nothing but Gauss elimination.

The determinant (map) associates with each square matrix $A$ a scalar $\det(A)$ in such a way that

$$\det(AB) = \det(A)\det(B) \tag{10.1}$$

$$\det(\mathrm{id}) = 1. \tag{10.2}$$

Since for an invertible $A$, $\mathrm{id} = AA^{-1}$, this implies that $\det(A) \neq 0$ for any invertible $A$. In fact, the determinant is defined in such a way that the invertible matrices are the only **nonsingular** matrices, i.e., the only matrices with a nonzero determinant (see two paragraphs below).

This fact is routinely used in the form: *the linear system $A? = b$ has exactly one solution if and only if* $\det(A) \neq 0$. On the other hand, the calculation of $\det(A)$ is more expensive than simply solving the linear system by Gauss elimination, during which one is bound to find out whether or not there is a solution. Hence it is not clear what the use of the determinant is in the context of solving $A? = b$ *numerically*.

Another popular use of the determinant is **Cramer's rule**. Here, one makes use of the fact that, as a function of the $n$ columns $a_1, a_2, \ldots, a_n$ of the matrix $A = [a_1, a_2, \ldots, a_n] \in \mathbb{F}^{n\times n}$, $\det(A) = \det(a_1, a_2, \ldots, a_n)$ is linear in each of its $n$ arguments and is zero if two of the arguments coincide. Thus, if $b = Ax = \sum_j a^j x_j$, then, for any $i$,

$$\det(\cdots, a^{i-1}, b, a^{i+1}, \ldots) = \det(\cdots, a^{i-1}, \sum_j a^j x_j, a^{i+1}, \ldots) = \sum_j \det(\cdots, a^{i-1}, a^j, a^{i+1}, \ldots)x_j = \det(A)x_i.$$

Therefore, for a nonsingular $A$,

$$\det(\cdots, a^{i-1}, Ax, a^{i+1}, \ldots)/\det(A) = x_i, \qquad \text{all } i. \tag{10.3}$$

In particular, a nonsingular $A$ must be 1-1, hence invertible.

But the suggestion that one actually construct the solution to $A? = b$ 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-manipulation 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.

Cramer's rule and other determinant identities are also useful in the *analysis* of linear algebra problems, e.g., in determining the continuous dependence of the solution on the given data. Cramer's rule may also have its place in the numerical solution of $A? = b$ on parallel computers.

We continue to use `MATLAB`'s notation for submatrices:

$$A(\mathbf{i};\mathbf{j})(i,j) := A(\mathbf{i}(i),\mathbf{j}(j)), \quad \forall i,j.$$

Here, finally, is the detailed discussion of determinants.

For $[a_1, a_2, \ldots, a_n] = A \in \mathbb{F}^{n\times n}$,

$$\det[a_1, a_2, \ldots, a_n] = \det(A)$$

is, by definition, a **multilinear alternating form** in its $n$ arguments. At a minimum, this means that

$$\mathbb{F}^n :\to \mathbb{F} : a \mapsto \det[a, \ldots]$$

is linear, and that

$$\det[\ldots, a, \ldots, b, \ldots] = -\det[\ldots, b, \ldots, a, \ldots]$$

(with the various ellipses '$\ldots$' indicating the other arguments, held fixed). Note that this **alternation** property implies that the determinant is zero if two of its arguments coincide, which, together with the linearity, implies that

$$\det[\cdots, a + b\alpha, \ldots, b, \ldots] = \det[\cdots, a, \ldots, b, \ldots].$$

It follows that

$$\det(BA) = \det[Ba_1, Ba_2, \ldots, Ba_n] = \det[\ldots, \sum_i b_i a_j(i), \ldots] = \sum_{\mathbf{i}\in\{1,\ldots,n\}^n} \det[b_{\mathbf{i}(1)}, \ldots, b_{\mathbf{i}(n)}] \prod_j a_j(\mathbf{i}(j)),$$

just using linearity. By the alternation property, most of these summands are zero. Only those determinants $\det[b_{\mathbf{i}(1)}, \ldots, b_{\mathbf{i}(n)}]$ for which $\mathbf{i} \in \$_{\mathbf{n}} :=$ symmetric group of order $n$, i.e., for which $\mathbf{i}$ is a permutation of the sequence $\underline{n} := (1, \ldots, n)$, are not automatically zero. Further, for such $\mathbf{i}$,

$$\det[b_{\mathbf{i}(1)}, \ldots, b_{\mathbf{i}(n)}] = (-)^{\mathbf{i}} \det(B)$$

by the alternation property, with $(-)^{\mathbf{i}} = \pm 1$ depending on whether it takes an even or an odd number of interchanges to change $\mathbf{i}$ into a strictly increasing sequence. (This requires the proof that it is always possible to convert any such $\mathbf{i}$ into an increasing sequence by interchanges (easy) and that, while this might be done in many ways, the parity of such number of interchanges is independent of how we got from the sequence $\mathbf{i}$ to the sequence $\underline{n} = (1, 2, \ldots, n)$.) Thus

$$\det(BA) = \det(B) \sum_{\mathbf{i}\in\$_{\mathbf{n}}} (-)^{\mathbf{i}} \prod_j a_j(\mathbf{i}(j)). \tag{10.4}$$

In particular, det is entirely pinned down once we know the number $\det(B)$ for some invertible $B$. It is customary to choose the *normalization* (10.2), i.e., $\det(\mathrm{id}) := 1$. With this choice, (10.4) implies the formula (10.5), hence also that $\det(BA) = \det(B)\det(A)$ for arbitrary $B$ and $A$. On the other hand, starting with the resulting formula

$$\det[a_1, a_2, \ldots, a_n] = \sum_{\mathbf{i}\in\$_{\mathbf{n}}} (-)^{\mathbf{i}} \prod_j a_j(\mathbf{i}(j)) = \sum_{\mathbf{i}\in\$_{\mathbf{n}}} (-)^{\mathbf{i}} \prod_j A(\mathbf{i}(j), j) \tag{10.5}$$

as a definition, one readily verifies that det so defined satisfies the three properties (multilinear, alternating, with $\det(\mathrm{id}) = 1$) claimed for it.

One verifies directly from the formula (10.5) that $\det A^T = \det A$ (since, for every $\mathbf{i} \in \$_{\mathbf{n}}$, $\mathbf{i}^{-1}$ exists and is again in $\$_{\mathbf{n}}$). Also from that formula, one obtains the very useful **Laplace's expansion by the last column**:

$$\begin{aligned}\det A &= \sum_{\mathbf{i}\in\$_{\mathbf{n}}} a_n(\mathbf{i}(n))(-)^{\mathbf{i}} \prod_{j<n} a_j(\mathbf{i}(j)) \\ &= \sum_{i=1}^n a_n(i) \sum_{\mathbf{i}\in\$_{\underline{n}\backslash i}} a_n(\mathbf{i}(n))(-)^{\mathbf{i},i} \prod_{j<n} a_j(\mathbf{i}(j)) \\ &= \sum_{i=1}^n a_n(i)(-)^{n-i} \det A(\underline{n}\backslash i, 1 : n-1)\end{aligned}$$

using the fact that $(-)^{\mathbf{i},i} = (-)^{\mathbf{i}}(-)^{n-i}$. Finally, also from (10.5),

$$\det T = \prod_j T(j,j)$$

for any *triangular* matrix $T$, since for any $\mathbf{i}$ other than $\underline{n}$, there is a $j$ with $\mathbf{i}(j) < j$ and a $j$ with $\mathbf{i}(j) > j$, hence triangularity implies that $\prod_j T(j, \mathbf{i}(j)) = 0$ for all $\mathbf{i} \neq \underline{n}$. Thus, for general $A$, the fastest way to compute $\det A$ is via an LU (or PLU) factorization.

## Sylvester

Now comes the promised Sylvester's Identity which is a direct consequence of the following neat description of Gauss elimination.

With $\mathbf{k} := \{1, \ldots, k\}$, consider the matrix

$$B(i,j) := \det A((\mathbf{k}i), (\mathbf{k}j)).$$

Then, on expanding by entries of the last row,

$$B(i,j) = A(i,j) \det A(\mathbf{k},\mathbf{k}) - \sum_{r \le k} A(i,r)(-)^{k-r} \det A(\mathbf{k}, ((\mathbf{k} \backslash r)j)).$$

This shows that

$$B(:,j) \in A(:,j) \det A(\mathbf{k},\mathbf{k}) + \operatorname{ran} A(:,\mathbf{k}),$$

while, directly, $B(i,j) = 0$ for $i \in \mathbf{k}$ since then $\det A((\mathbf{k}i), (\mathbf{k}j))$ has two rows the same.

In the same way,

$$B(i,:) \in A(i,:) \det A(\mathbf{k},\mathbf{k}) + \operatorname{ran} A(\mathbf{k},:),$$

while,directly, $B(i,j) = 0$ for $j \in \mathbf{k}$. Thus, if $\det A(\mathbf{k},\mathbf{k}) \neq 0$, then, for $i > k$,

$$B(i,:)/ \det A(\mathbf{k},\mathbf{k})$$

provides the $i$th row of the matrix obtained from $A$ after $k$ steps of Gauss elimination. In other words, the matrix $S := B/ \det A(\mathbf{k},\mathbf{k})$ provides the **Schur complement** $S((k+1 \ldots n), (k+1 \ldots n))$ in $A$ of the **pivot block** $A(\mathbf{k},\mathbf{k})$.

Since such row elimination is done by elementary matrices with determinant equal to 1, it follows that

$$\det A = \det A(\mathbf{k},\mathbf{k}) \det S((k+1 \ldots n), (k+1 \ldots n)).$$

Since, for any $\#\mathbf{i} = \#\mathbf{j}$, $B(\mathbf{i},\mathbf{j})$ depends only on the square matrix $\det A((\mathbf{k}\mathbf{i}), (\mathbf{k}\mathbf{j}))$, this implies

**(10.6) Sylvester's determinant identity:** If

$$S(i,j) := \det A((\mathbf{k}i), (\mathbf{k}j))/ \det A(\mathbf{k},\mathbf{k}), \quad \forall i,j,$$

then

$$\det S(\mathbf{i},\mathbf{j}) = \det A((\mathbf{k}\mathbf{i}), (\mathbf{k}\mathbf{j}))/ \det A(\mathbf{k},\mathbf{k}).$$

## Cauchy-Binet

## 11. Background

### 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 $-\mathrm{i}$. 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) sqareroot 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 $\zeta$,

$$|z\zeta| = |z||\zeta|.$$

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. 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*, and $\varphi \in \mathbb{R}$ is called its **argument**. Indeed, for any real $\varphi$, $\exp(\mathrm{i}\varphi) = \cos(\varphi) + \mathrm{i}\sin(\varphi)$ has absolute value 1, and $\varphi$ is the angle (in radians) which the vector $(a, b)$ makes with the positive real axis. Note that, for $z \neq 0$, the argument, $\varphi$, is only defined up to a multiple of $2\pi$, 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 $\varphi$, and therefore $z^k = \exp(\mathrm{i}(k\varphi))$. Hence, the sequence $z^0, z^1, z^2, \cdots$ appears as a sequence of points on the unit circle, equally spaced around that circle, never accumulating anywhere unless $\varphi = 0$, i.e., unless $z = 1$.

**(11.1) 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.

> **(11.2) Lemma:** The sequence $(\zeta^1, \zeta^2, \zeta^3, \cdots)$ is bounded if and only if $|\zeta| \le 1$. Here, $\zeta^k$ denotes the $k$th power of the scalar $z$.

**Proof:** Assume that $|\zeta| > 1$. I claim that, for all $m$,

$$|\zeta^m| - 1 > (|\zeta| - 1)m. \tag{11.3}$$

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 (11.3) 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 $(\zeta^1, \zeta^2, \zeta^3, \cdots)$ 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 $(\zeta^1, \zeta^2, \zeta^3, \cdots)$ is not only bounded, it lies entirely in the **unit disk**

$$B_1^- := \{z \in \mathbb{C} : |z| \le 1\}.$$

A sequence $(\zeta_1, \zeta_2, \zeta_3, \cdots)$ of (real or complex) scalars is said to **converge to the scalar** $\zeta$, in symbols:

$$\lim_{m\to\infty} \zeta_m = \zeta,$$

if, for all $\varepsilon > 0$, there is some $m_\varepsilon$ 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 $\varepsilon$ we draw around the point $\zeta$, *all* the terms of the sequence except the first few are inside that circle.

> **(11.4) 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 $m$,

$$|\zeta_m| \le r := |\zeta| + 1 + \max\{|\zeta_k| : k = 1, \ldots, m_0\}.$$

Note that $r$ is indeed a well-defined nonnegative number, since a *finite* set of real numbers always has a largest element. □

**(11.5) Lemma:** The sequence $(\zeta^1, \zeta^2, \zeta^3, \cdots)$ 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, $\lim_{m\to\infty} \zeta^m = 0$. □

### Horner, or: How to divide a polynomial by a linear factor

Recall that, given the polynomial $p$ and one of its roots, $\mu$, 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 $\le 3$.

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

$$(11.6) \qquad b_j := a_j + z b_{j+1},$$

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

$$\begin{array}{rcccccccc}
b_0 + (t-z)(b_1 + b_2 t + b_3 t^2) & = & b_0 & + & b_1 t & + & b_2 t^2 & + & b_3 t^3 \\
 & & -z b_1 & - & z b_2 t & - & z b_3 t^2 & & \\
 & = & b_0 - z b_1 & + & (b_1 - z b_2) t & + & (b_2 - z b_3) t^2 & + & b_3 t^3 \\
 & = & a_0 & + & a_1 t & + & a_2 t^2 & + & a_3 t^3
\end{array}$$

The last equality holds since, by (11.6),

$$b_j - z b_{j+1} = a_j$$

for $j < 3$ while $b_3 = a_3$ by definition.

**(11.7) Nested Multiplication (aka 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

$$b_j := \begin{cases} a_j, & j = k; \\ a_j + z b_{j+1}, & j < k. \end{cases}$$

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 (4.3)) for the *easy* part of the *Fundamental Theorem of Algebra*, namely that a polynomial of degree $k$ has at most $k$ roots.

**`MATLAB` grammar**