Chapter 6

Direct Sums

In this chapter all vector spaces are defined over an arbitrary field K. For the sake of concreteness, the reader may safely assume that $K = \mathbb{R}$.

6.1 Sums, Direct Sums, Direct Products

There are some useful ways of forming new vector spaces from older ones, in particular, direct products and direct sums. Regarding direct sums, there is a subtle point, which is that if we attempt to define the direct sum $E \coprod F$ of two vector spaces using the cartesian product $E \times F$, we don't quite get the right notion because elements of $E \times F$ are ordered pairs, but we want $E \coprod F = F \coprod E$. Thus, we want to think of the elements of $E \coprod F$ as unordrered pairs of elements. It is possible to do so by considering the direct sum of a *family* $(E_i)_{i\in\{1,2\}}$, and more generally of a family $(E_i)_{i\in I}$. For simplicity, we begin by considering the case where $I = \{1, 2\}$.

Definition 6.1. Given a family $(E_i)_{i\in\{1,2\}}$ of two vector spaces, we define the *(external) direct sum* $E_1 \coprod E_2$ (or *coproduct*) of the family $(E_i)_{i\in\{1,2\}}$ as the set

$$E_1 \coprod E_2 = \{\{\langle 1, u\rangle, \langle 2, v\rangle\} \mid u \in E_1,\ v \in E_2\},$$

with addition

$$\{\langle 1, u_1\rangle, \langle 2, v_1\rangle\} + \{\langle 1, u_2\rangle, \langle 2, v_2\rangle\} = \{\langle 1, u_1 + u_2\rangle, \langle 2, v_1 + v_2\rangle\},$$

and scalar multiplication

$$\lambda\{\langle 1, u\rangle, \langle 2, v\rangle\} = \{\langle 1, \lambda u\rangle, \langle 2, \lambda v\rangle\}.$$

We define the *injections* $in_1\colon E_1 \to E_1 \coprod E_2$ and $in_2\colon E_2 \to E_1 \coprod E_2$ as the linear maps defined such that,

$$in_1(u) = \{\langle 1, u\rangle, \langle 2, 0\rangle\},$$

and

$$in_2(v) = \{\langle 1, 0\rangle, \langle 2, v\rangle\}.$$

Note that

$$E_2 \coprod E_1 = \{\{\langle 2, v\rangle, \langle 1, u\rangle\} \mid v \in E_2,\ u \in E_1\} = E_1 \coprod E_2.$$

Thus, every member $\{\langle 1, u\rangle, \langle 2, v\rangle\}$ of $E_1 \coprod E_2$ can be viewed as an *unordered pair* consisting of the two vectors u and v, tagged with the index 1 and 2, respectively.

Remark: In fact, $E_1 \coprod E_2$ is just the product $\prod_{i\in\{1,2\}} E_i$ of the family $(E_i)_{i\in\{1,2\}}$.

This is not to be confused with the cartesian product $E_1 \times E_2$. The vector space $E_1 \times E_2$ is the set of all ordered pairs $\langle u, v\rangle$, where $u \in E_1$, and $v \in E_2$, with addition and multiplication by a scalar defined such that

$$\begin{aligned}
\langle u_1, v_1\rangle + \langle u_2, v_2\rangle &= \langle u_1 + u_2, v_1 + v_2\rangle, \\
\lambda\langle u, v\rangle &= \langle \lambda u, \lambda v\rangle.
\end{aligned}$$

There is a bijection between $\prod_{i\in\{1,2\}} E_i$ and $E_1 \times E_2$, but as we just saw, elements of $\prod_{i\in\{1,2\}} E_i$ are certain sets. The product $E_1 \times \cdots \times E_n$ of any number of vector spaces can also be defined. We will do this shortly.

The following property holds.

Proposition 6.1. *Given any two vector spaces, E_1 and E_2, the set $E_1 \coprod E_2$ is a vector space. For every pair of linear maps, $f\colon E_1 \to G$ and $g\colon E_2 \to G$, there is a unique linear map, $f + g\colon E_1 \coprod E_2 \to G$, such that $(f + g) \circ in_1 = f$ and $(f + g) \circ in_2 = g$, as in the following diagram:*

$$\begin{tikzcd}
E_1 \arrow[d, "in_1"'] \arrow[dr, "f"] & \\
E_1 \coprod E_2 \arrow[r, "f+g"] & G \\
E_2 \arrow[u, "in_2"] \arrow[ur, "g"'] &
\end{tikzcd}$$

Proof. Define

$$(f + g)(\{\langle 1, u\rangle, \langle 2, v\rangle\}) = f(u) + g(v),$$

for every $u \in E_1$ and $v \in E_2$. It is immediately verified that $f + g$ is the unique linear map with the required properties. $\square$

We already noted that $E_1 \coprod E_2$ is in bijection with $E_1 \times E_2$. If we define the *projections* $\pi_1\colon E_1 \coprod E_2 \to E_1$ and $\pi_2\colon E_1 \coprod E_2 \to E_2$, such that

$$\pi_1(\{\langle 1, u\rangle, \langle 2, v\rangle\}) = u,$$

and

$$\pi_2(\{\langle 1, u\rangle, \langle 2, v\rangle\}) = v,$$

we have the following proposition.

Proposition 6.2. *Given any two vector spaces, E_1 and E_2, for every pair of linear maps, $f\colon D \to E_1$ and $g\colon D \to E_2$, there is a unique linear map, $f \times g\colon D \to E_1 \coprod E_2$, such that $\pi_1 \circ (f \times g) = f$ and $\pi_2 \circ (f \times g) = g$, as in the following diagram:*

$$
\begin{tikzcd}
 & E_1 \\
D \arrow[ur, "f"] \arrow[r, "f\times g"] \arrow[dr, "g"'] & E_1 \coprod E_2 \arrow[u, "\pi_1"'] \arrow[d, "\pi_2"] \\
 & E_2
\end{tikzcd}
$$

Proof. Define
$$(f \times g)(w) = \{\langle 1, f(w)\rangle, \langle 2, g(w)\rangle\},$$
for every $w \in D$. It is immediately verified that $f \times g$ is the unique linear map with the required properties. ∎

Remark: It is a peculiarity of linear algebra that direct sums and products of finite families are isomorphic. However, this is no longer true for products and sums of infinite families.

When U, V are subspaces of a vector space E, letting $i_1\colon U \to E$ and $i_2\colon V \to E$ be the inclusion maps, if $U \coprod V$ is isomomorphic to E under the map $i_1 + i_2$ given by Proposition 6.1, we say that E *is a direct sum* of U and V, and we write $E = U \coprod V$ (with a slight abuse of notation, since E and $U \coprod V$ are only isomorphic). It is also convenient to define the sum $U_1 + \cdots + U_p$ and the internal direct sum $U_1 \oplus \cdots \oplus U_p$ of any number of subspaces of E.

Definition 6.2. Given $p \geq 2$ vector spaces $E_1, \ldots, E_p$, the product $F = E_1 \times \cdots \times E_p$ can be made into a vector space by defining addition and scalar multiplication as follows:

$$
\begin{aligned}
(u_1, \ldots, u_p) + (v_1, \ldots, v_p) &= (u_1 + v_1, \ldots, u_p + v_p) \\
\lambda(u_1, \ldots, u_p) &= (\lambda u_1, \ldots, \lambda u_p),
\end{aligned}
$$

for all $u_i, v_i \in E_i$ and all $\lambda \in \mathbb{R}$. The zero vector of $E_1 \times \cdots \times E_p$ is the p-tuple

$$(\underbrace{0, \ldots, 0}_{p}),$$

where the ith zero is the zero vector of E_i.

With the above addition and multiplication, the vector space $F = E_1 \times \cdots \times E_p$ is called the *direct product* of the vector spaces $E_1, \ldots, E_p$.

As a special case, when $E_1 = \cdots = E_p = \mathbb{R}$, we find again the vector space $F = \mathbb{R}^p$. The *projection maps* $pr_i\colon E_1 \times \cdots \times E_p \to E_i$ given by

$$pr_i(u_1, \ldots, u_p) = u_i$$

are clearly linear. Similarly, the maps $\mathrm{in}_i\colon E_i \to E_1 \times \cdots \times E_p$ given by

$$\mathrm{in}_i(u_i) = (0, \ldots, 0, u_i, 0, \ldots, 0)$$

are injective and linear. If $\dim(E_i) = n_i$ and if $(e_1^i, \ldots, e_{n_i}^i)$ is a basis of E_i for $i = 1, \ldots, p$, then it is easy to see that the $n_1 + \cdots + n_p$ vectors

$$\begin{array}{ccc}
(e_1^1, 0, \ldots, 0), & \ldots, & (e_{n_1}^1, 0, \ldots, 0), \\
\vdots & \vdots & \vdots \\
(0, \ldots, 0, e_1^i, 0, \ldots, 0), & \ldots, & (0, \ldots, 0, e_{n_i}^i, 0, \ldots, 0), \\
\vdots & \vdots & \vdots \\
(0, \ldots, 0, e_1^p), & \ldots, & (0, \ldots, 0, e_{n_p}^p)
\end{array}$$

form a basis of $E_1 \times \cdots \times E_p$, and so

$$\dim(E_1 \times \cdots \times E_p) = \dim(E_1) + \cdots + \dim(E_p).$$

Let us now consider a vector space E and p subspaces $U_1, \ldots, U_p$ of E. We have a map

$$a\colon U_1 \times \cdots \times U_p \to E$$

given by

$$a(u_1, \ldots, u_p) = u_1 + \cdots + u_p,$$

with $u_i \in U_i$ for $i = 1, \ldots, p$. It is clear that this map is linear, and so its image is a subspace of E denoted by

$$U_1 + \cdots + U_p$$

and called the *sum* of the subspaces $U_1, \ldots, U_p$. By definition,

$$U_1 + \cdots + U_p = \{u_1 + \cdots + u_p \mid u_i \in U_i,\ 1 \leq i \leq p\},$$

and it is immediately verified that $U_1 + \cdots + U_p$ is the smallest subspace of E containing $U_1, \ldots, U_p$. This also implies that $U_1 + \cdots + U_p$ does not depend on the order of the factors U_i; in particular,

$$U_1 + U_2 = U_2 + U_1.$$

Definition 6.3. For any vector space E and any $p \geq 2$ subspaces $U_1, \ldots, U_p$ of E, if the map a defined above is injective, then the sum $U_1 + \cdots + U_p$ is called a *direct sum* and it is denoted by

$$U_1 \oplus \cdots \oplus U_p.$$

The space E is the *direct sum* of the subspaces U_i if

$$E = U_1 \oplus \cdots \oplus U_p.$$

As in the case of a sum, $U_1 \oplus U_2 = U_2 \oplus U_1$.

If the map a is injective, then by Proposition 3.17 we have $\operatorname{Ker} a = \{(\underbrace{0, \ldots, 0}_{p})\}$ where each 0 is the zero vector of E, which means that if $u_i \in U_i$ for $i = 1, \ldots, p$ and if

$$u_1 + \cdots + u_p = 0,$$

then $(u_1, \ldots, u_p) = (0, \ldots, 0)$, that is, $u_1 = 0, \ldots, u_p = 0$.

Proposition 6.3. *If the map $a \colon U_1 \times \cdots \times U_p \to E$ is injective, then every $u \in U_1 + \cdots + U_p$ has a unique expression as a sum*

$$u = u_1 + \cdots + u_p,$$

with $u_i \in U_i$, for $i = 1, \ldots, p$.

Proof. If

$$u = v_1 + \cdots + v_p = w_1 + \cdots + w_p,$$

with $v_i, w_i \in U_i$, for $i = 1, \ldots, p$, then we have

$$w_1 - v_1 + \cdots + w_p - v_p = 0,$$

and since $v_i, w_i \in U_i$ and each U_i is a subspace, $w_i - v_i \in U_i$. The injectivity of a implies that $w_i - v_i = 0$, that is, $w_i = v_i$ for $i = 1, \ldots, p$, which shows the uniqueness of the decomposition of u. $\square$

Proposition 6.4. *If the map $a \colon U_1 \times \cdots \times U_p \to E$ is injective, then any p nonzero vectors $u_1, \ldots, u_p$ with $u_i \in U_i$ are linearly independent.*

Proof. To see this, assume that

$$\lambda_1 u_1 + \cdots + \lambda_p u_p = 0$$

for some $\lambda_i \in \mathbb{R}$. Since $u_i \in U_i$ and U_i is a subspace, $\lambda_i u_i \in U_i$, and the injectivity of a implies that $\lambda_i u_i = 0$, for $i = 1, \ldots, p$. Since $u_i \neq 0$, we must have $\lambda_i = 0$ for $i = 1, \ldots, p$; that is, $u_1, \ldots, u_p$ with $u_i \in U_i$ and $u_i \neq 0$ are linearly independent. $\square$

Observe that if a is injective, then we must have $U_i \cap U_j = (0)$ whenever $i \neq j$. However, this condition is generally not sufficient if $p \geq 3$. For example, if $E = \mathbb{R}^2$ and U_1 the line spanned by $e_1 = (1, 0)$, U_2 is the line spanned by $d = (1, 1)$, and U_3 is the line spanned by $e_2 = (0, 1)$, then $U_1 \cap U_2 = U_1 \cap U_3 = U_2 \cap U_3 = \{(0, 0)\}$, but $U_1 + U_2 = U_1 + U_3 = U_2 + U_3 = \mathbb{R}^2$, so $U_1 + U_2 + U_3$ is not a direct sum. For example, d is expressed in two different ways as

$$d = (1, 1) = (1, 0) + (0, 1) = e_1 + e_2.$$

See Figure 6.1.

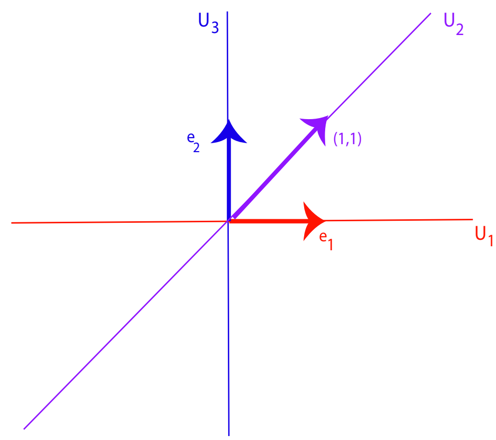

Figure 6.1: The linear subspaces U_1, U_2, and U_3 illustrated as lines in $\mathbb{R}^2$.

As in the case of a sum, $U_1 \oplus U_2 = U_2 \oplus U_1$. *Observe that when the map a is injective, then it is a linear isomorphism between $U_1 \times \cdots \times U_p$ and $U_1 \oplus \cdots \oplus U_p$.* The difference is that $U_1 \times \cdots \times U_p$ is defined even if the spaces U_i are not assumed to be subspaces of some common space.

If E is a direct sum $E = U_1 \oplus \cdots \oplus U_p$, since any p nonzero vectors $u_1, \ldots, u_p$ with $u_i \in U_i$ are linearly independent, if we pick a basis $(u_k)_{k \in I_j}$ in U_j for $j = 1, \ldots, p$, then $(u_i)_{i \in I}$ with $I = I_1 \cup \cdots \cup I_p$ is a basis of E. Intuitively, E is split into p independent subspaces.

Conversely, given a basis $(u_i)_{i \in I}$ of E, if we partition the index set I as $I = I_1 \cup \cdots \cup I_p$, then each subfamily $(u_k)_{k \in I_j}$ spans some subspace U_j of E, and it is immediately verified that we have a direct sum

$$E = U_1 \oplus \cdots \oplus U_p.$$

Definition 6.4. Let $f \colon E \to E$ be a linear map. For any subspace U of E, if $f(U) \subseteq U$ we say that U is *invariant under f*.

Assume that E is finite-dimensional, a direct sum $E = U_1 \oplus \cdots \oplus U_p$, and that each U_j is invariant under f. If we pick a basis $(u_i)_{i \in I}$ as above with $I = I_1 \cup \cdots \cup I_p$ and with each $(u_k)_{k \in I_j}$ a basis of U_j, since each U_j is invariant under f, the image $f(u_k)$ of every basis vector u_k with $k \in I_j$ belongs to U_j, so the matrix A representing f over the basis $(u_i)_{i \in I}$ is a *block diagonal* matrix of the form

$$A = \begin{pmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_p \end{pmatrix},$$

with each block A_j a $d_j \times d_j$-matrix with $d_j = \dim(U_j)$ and all other entries equal to 0. If $d_j = 1$ for $j = 1, \ldots, p$, the matrix A is a diagonal matrix.

There are natural injections from each U_i to E denoted by $\mathrm{in}_i \colon U_i \to E$.

Now, if $p = 2$, it is easy to determine the kernel of the map $a \colon U_1 \times U_2 \to E$. We have

$$a(u_1, u_2) = u_1 + u_2 = 0 \quad \text{iff} \quad u_1 = -u_2,\ u_1 \in U_1, u_2 \in U_2,$$

which implies that

$$\mathrm{Ker}\, a = \{(u, -u) \mid u \in U_1 \cap U_2\}.$$

Now, $U_1 \cap U_2$ is a subspace of E and the linear map $u \mapsto (u, -u)$ is clearly an isomorphism between $U_1 \cap U_2$ and $\mathrm{Ker}\, a$, so $\mathrm{Ker}\, a$ is isomorphic to $U_1 \cap U_2$. As a consequence, we get the following result:

Proposition 6.5. *Given any vector space E and any two subspaces U_1 and U_2, the sum $U_1 + U_2$ is a direct sum iff $U_1 \cap U_2 = (0)$.*

An interesting illustration of the notion of direct sum is the decomposition of a square matrix into its symmetric part and its skew-symmetric part. Recall that an $n \times n$ matrix $A \in \mathrm{M}_n$ is *symmetric* if $A^\top = A$, *skew -symmetric* if $A^\top = -A$. It is clear that s

$$\mathbf{S}(n) = \{A \in \mathrm{M}_n \mid A^\top = A\} \quad \text{and} \quad \mathbf{Skew}(n) = \{A \in \mathrm{M}_n \mid A^\top = -A\}$$

are subspaces of M_n, and that $\mathbf{S}(n) \cap \mathbf{Skew}(n) = (0)$. Observe that for any matrix $A \in \mathrm{M}_n$, the matrix $H(A) = (A + A^\top)/2$ is symmetric and the matrix $S(A) = (A - A^\top)/2$ is skew-symmetric. Since

$$A = H(A) + S(A) = \frac{A + A^\top}{2} + \frac{A - A^\top}{2},$$

we see that $\mathrm{M}_n = \mathbf{S}(n) + \mathbf{Skew}(n)$, and since $\mathbf{S}(n) \cap \mathbf{Skew}(n) = (0)$, we have the direct sum

$$\mathrm{M}_n = \mathbf{S}(n) \oplus \mathbf{Skew}(n).$$

Remark: The vector space $\mathbf{Skew}(n)$ of skew-symmetric matrices is also denoted by $\mathfrak{so}(n)$. It is the *Lie algebra* of the group $\mathbf{SO}(n)$.

Proposition 6.5 can be generalized to any $p \geq 2$ subspaces at the expense of notation. The proof of the following proposition is left as an exercise.

Proposition 6.6. *Given any vector space E and any $p \geq 2$ subspaces $U_1, \ldots, U_p$, the following properties are equivalent:*

(1) The sum $U_1 + \cdots + U_p$ is a direct sum.

(2) We have

$$U_i \cap \left(\sum_{j=1, j\neq i}^{p} U_j \right) = (0), \quad i = 1, \ldots, p.$$

(3) We have

$$U_i \cap \left(\sum_{j=1}^{i-1} U_j \right) = (0), \quad i = 2, \ldots, p.$$

Because of the isomorphism

$$U_1 \times \cdots \times U_p \approx U_1 \oplus \cdots \oplus U_p,$$

we have

Proposition 6.7. *If E is any vector space, for any (finite-dimensional) subspaces $U_1, \ldots,$ U_p of E, we have*

$$\dim(U_1 \oplus \cdots \oplus U_p) = \dim(U_1) + \cdots + \dim(U_p).$$

If E is a direct sum

$$E = U_1 \oplus \cdots \oplus U_p,$$

since every $u \in E$ can be written in a unique way as

$$u = u_1 + \cdots + u_p$$

with $u_i \in U_i$ for $i = 1 \ldots, p$, we can define the maps $\pi_i \colon E \to U_i$, called *projections*, by

$$\pi_i(u) = \pi_i(u_1 + \cdots + u_p) = u_i.$$

It is easy to check that these maps are linear and satisfy the following properties:

$$\pi_j \circ \pi_i = \begin{cases} \pi_i & \text{if } i = j \\ 0 & \text{if } i \neq j, \end{cases}$$
$$\pi_1 + \cdots + \pi_p = \mathrm{id}_E.$$

For example, in the case of the direct sum

$$\mathrm{M}_n = \mathbf{S}(n) \oplus \mathbf{Skew}(n),$$

the projection onto $\mathbf{S}(n)$ is given by

$$\pi_1(A) = H(A) = \frac{A + A^\top}{2},$$

and the projection onto $\mathbf{Skew}(n)$ is given by

$$\pi_2(A) = S(A) = \frac{A - A^\top}{2}.$$

Clearly, $H(A)+S(A) = A$, $H(H(A)) = H(A)$, $S(S(A)) = S(A)$, and $H(S(A)) = S(H(A)) = 0$.

A function f such that $f \circ f = f$ is said to be *idempotent*. Thus, the projections π_i are idempotent. Conversely, the following proposition can be shown:

Proposition 6.8. *Let E be a vector space. For any $p \geq 2$ linear maps $f_i \colon E \to E$, if*

$$f_j \circ f_i = \begin{cases} f_i & \text{if } i = j \\ 0 & \text{if } i \neq j, \end{cases}$$
$$f_1 + \cdots + f_p = \mathrm{id}_E,$$

then if we let $U_i = f_i(E)$, we have a direct sum

$$E = U_1 \oplus \cdots \oplus U_p.$$

We also have the following proposition characterizing idempotent linear maps whose proof is also left as an exercise.

Proposition 6.9. *For every vector space E, if $f \colon E \to E$ is an idempotent linear map, i.e., $f \circ f = f$, then we have a direct sum*

$$E = \mathrm{Ker}\, f \oplus \mathrm{Im}\, f,$$

so that f is the projection onto its image $\mathrm{Im}\, f$.

We now give the definition of a direct sum for any arbitrary nonempty index set I. First, let us recall the notion of the product of a family $(E_i)_{i\in I}$. Given a family of sets $(E_i)_{i\in I}$, its product $\prod_{i\in I} E_i$, is the set of all functions $f \colon I \to \bigcup_{i\in I} E_i$, such that, $f(i) \in E_i$, for every $i \in I$. It is one of the many versions of the axiom of choice, that, if $E_i \neq \emptyset$ for every $i \in I$, then $\prod_{i\in I} E_i \neq \emptyset$. A member $f \in \prod_{i\in I} E_i$, is often denoted as $(f_i)_{i\in I}$. For every $i \in I$, we have the *projection* $\pi_i \colon \prod_{i\in I} E_i \to E_i$, defined such that, $\pi_i((f_i)_{i\in I}) = f_i$. We now define direct sums.

Definition 6.5. Let I be any nonempty set, and let $(E_i)_{i\in I}$ be a family of vector spaces. The *(external) direct sum* $\coprod_{i\in I} E_i$ (or *coproduct*) of the family $(E_i)_{i\in I}$ is defined as follows:

$\coprod_{i\in I} E_i$ consists of all $f \in \prod_{i\in I} E_i$, which have finite support, and addition and multiplication by a scalar are defined as follows:

$$(f_i)_{i\in I} + (g_i)_{i\in I} = (f_i + g_i)_{i\in I},$$
$$\lambda(f_i)_{i\in I} = (\lambda f_i)_{i\in I}.$$

We also have *injection maps* $in_i \colon E_i \to \coprod_{i\in I} E_i$, defined such that, $in_i(x) = (f_i)_{i\in I}$, where $f_i = x$, and $f_j = 0$, for all $j \in (I - \{i\})$.

The following proposition is an obvious generalization of Proposition 6.1.

Proposition 6.10. *Let I be any nonempty set, let $(E_i)_{i\in I}$ be a family of vector spaces, and let G be any vector space. The direct sum $\coprod_{i\in I} E_i$ is a vector space, and for every family $(h_i)_{i\in I}$ of linear maps $h_i\colon E_i \to G$, there is a unique linear map*

$$\Big(\sum_{i\in I} h_i\Big)\colon \coprod_{i\in I} E_i \to G,$$

such that, $(\sum_{i\in I} h_i)\circ in_i = h_i$, for every $i\in I$.

Remarks:

(1) One might wonder why the direct sum $\coprod_{i\in I} E_i$ consists of familes of finite support instead of arbitrary families; in other words, why didn't we define the direct sum of the family $(E_i)_{i\in I}$ as $\prod_{i\in I} E_i$? The product space $\prod_{i\in I} E_i$ with addition and scalar multiplication defined as above is also a vector space but the problem is that any linear map $\widehat{h}\colon \prod_{i\in I} E_i \to G$ such that $\widehat{h}\circ in_i = h_i$ for all $\in I$ must be given by

$$\widehat{h}((u_i)_{\in I}) = \sum_{i\in I} h_i(u_i),$$

and if I is infinite, the sum on the right-hand side is infinite, and thus undefined! If I is finite then $\prod_{i\in I} E_i$ and $\coprod_{i\in I} E_i$ are isomorphic.

(2) When $E_i = E$, for all $i\in I$, we denote $\coprod_{i\in I} E_i$ by $E^{(I)}$. In particular, when $E_i = K$, for all $i\in I$, we find the vector space $K^{(I)}$ of Definition 3.11.

We also have the following basic proposition about injective or surjective linear maps.

Proposition 6.11. *Let E and F be vector spaces, and let $f\colon E\to F$ be a linear map. If $f\colon E\to F$ is injective, then there is a surjective linear map $r\colon F\to E$ called a retraction, such that $r\circ f = \mathrm{id}_E$. See Figure 6.2. If $f\colon E\to F$ is surjective, then there is an injective linear map $s\colon F\to E$ called a section, such that $f\circ s = \mathrm{id}_F$. See Figure 6.3.*

Proof. Let $(u_i)_{i\in I}$ be a basis of E. Since $f\colon E\to F$ is an injective linear map, by Proposition 3.18, $(f(u_i))_{i\in I}$ is linearly independent in F. By Theorem 3.7, there is a basis $(v_j)_{j\in J}$ of F, where $I\subseteq J$, and where $v_i = f(u_i)$, for all $i\in I$. By Proposition 3.18, a linear map $r\colon F\to E$ can be defined such that $r(v_i) = u_i$, for all $i\in I$, and $r(v_j) = w$ for all $j\in (J-I)$, where w is any given vector in E, say $w = 0$. Since $r(f(u_i)) = u_i$ for all $i\in I$, by Proposition 3.18, we have $r\circ f = \mathrm{id}_E$.

Now, assume that $f\colon E\to F$ is surjective. Let $(v_j)_{j\in J}$ be a basis of F. Since $f\colon E\to F$ is surjective, for every $v_j\in F$, there is some $u_j\in E$ such that $f(u_j) = v_j$. Since $(v_j)_{j\in J}$ is a basis of F, by Proposition 3.18, there is a unique linear map $s\colon F\to E$ such that $s(v_j) = u_j$. Also, since $f(s(v_j)) = v_j$, by Proposition 3.18 (again), we must have $f\circ s = \mathrm{id}_F$. $\square$

The converse of Proposition 6.11 is obvious.

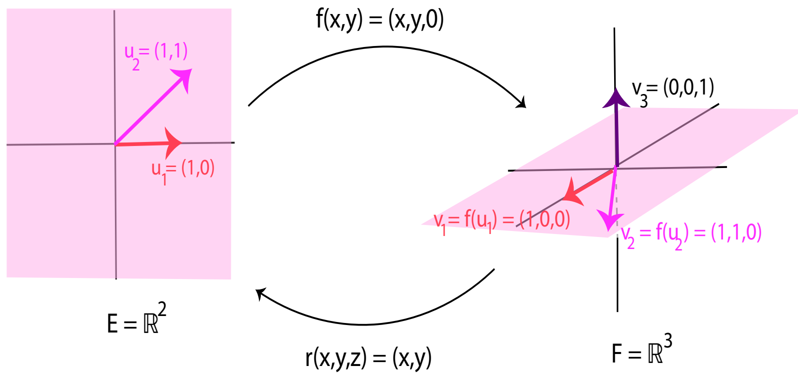

Figure 6.2: Let $f\colon E \to F$ be the injective linear map from $\mathbb{R}^2$ to $\mathbb{R}^3$ given by $f(x, y) = (x, y, 0)$. Then a surjective retraction is given by $r\colon \mathbb{R}^3 \to \mathbb{R}^2$ is given by $r(x, y, z) = (x, y)$. Observe that $r(v_1) = u_1$, $r(v_2) = u_2$, and $r(v_3) = 0$.

6.2 Matrices of Linear Maps and Multiplication by Blocks

Direct sums yield a fairly easy justification of matrix block multiplication. The key idea is that the representation of a linear map $f\colon E \to F$ over a basis $(u_1, \ldots, u_n)$ of E and a basis $(v_1, \ldots, v_m)$ of F by a matrix $A = (a_{ij})$ of scalars (in K) can be generalized to the representation of f over a direct sum decomposition $E = E_1 \oplus \cdots \oplus E_n$ of E and a direct sum decomposition $F = F_1 \oplus \cdots \oplus F_m$ of F in terms of a matrix (f_{ij}) of linear maps $f_{ij}\colon E_j \to F_i$. Futhermore, matrix multiplication of scalar matrices extends naturally to matrix multiplication of matrices of linear maps. We simply have to replace multiplication of scalars in K by the composition of linear maps.

Let E and F be two vector spaces and assume that they are expressed as direct sums

$$E = \bigoplus_{j=1}^{n} E_j, \quad F = \bigoplus_{i=1}^{m} F_i.$$

Definition 6.6. Given any linear map $f\colon E \to F$, we define the linear maps $f_{ij}\colon E_j \to F_i$ as follows. Let $pr_i^F\colon F \to F_i$ be the projection of $F = F_1 \oplus \cdots \oplus F_m$ onto F_i. If $f_j\colon E_j \to F$ is the restriction of f to E_j, which means that for every vector $x_j \in E_j$,

$$f_j(x_j) = f(x_j),$$

then we define the map $f_{ij}\colon E_j \to F_i$ by

$$f_{ij} = pr_i^F \circ f_j,$$

so that if $x_j \in E_j$, then

$$f(x_j) = f_j(x_j) = \sum_{i=1}^{m} f_{ij}(x_j), \quad \text{with } f_{ij}(x_j) \in F_i. \tag{$\dagger_1$}$$

Observe that we are summing over the index i, which eventually corresponds to summing the entries in the jth column of the matrix representing f; see Definition 6.7.

We see that for every vector $x = x_1 + \cdots + x_n \in E$, with $x_j \in E_j$, we have

$$f(x) = \sum_{j=1}^{n} f_j(x_j) = \sum_{j=1}^{n}\sum_{i=1}^{m} f_{ij}(x_j) = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}(x_j).$$

Observe that conversely, given a family $(f_{ij})_{1\leq i\leq m, 1\leq j\leq n}$ of linear maps $f_{ij}\colon E_j \to F_i$, we obtain the linear map $f\colon E \to F$ defined such that for every $x = x_1 + \cdots + x_n \in E$, with $x_j \in E_j$, we have

$$f(x) = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}(x_j).$$

As a consequence, it is easy to check that there is an isomorphism between the vector spaces

$$\mathrm{Hom}(E, F) \quad \text{and} \quad \prod_{\substack{1\leq i\leq m\\ 1\leq j\leq n}} \mathrm{Hom}(E_j, F_i).$$

Example 6.1. Suppose that $E = E_1 \oplus E_2$ and $F = F_1 \oplus F_2 \oplus F_3$, and that we have six maps $f_{ij}\colon E_j \to F_i$ for $i = 1, 2, 3$ and $j = 1, 2$. For any $x = x_1 + x_2$, with $x_1 \in E_1$ and $x_2 \in E_2$, we have

$$\begin{aligned} y_1 &= f_{11}(x_1) + f_{12}(x_2) \in F_1 \\ y_2 &= f_{21}(x_1) + f_{22}(x_2) \in F_2 \\ y_3 &= f_{31}(x_1) + f_{32}(x_2) \in F_3, \end{aligned}$$

$$\begin{aligned} f_1(x_1) &= f_{11}(x_1) + f_{21}(x_1) + f_{31}(x_1) \\ f_2(x_2) &= f_{12}(x_2) + f_{22}(x_2) + f_{32}(x_2), \end{aligned}$$

and

$$\begin{aligned} f(x) &= f_1(x_1) + f_2(x_2) = y_1 + y_2 + y_3 \\ &= f_{11}(x_1) + f_{12}(x_2) + f_{21}(x_1) + f_{22}(x_2) + f_{31}(x_1) + f_{32}(x_2). \end{aligned}$$

These equations suggest the matrix notation

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \\ f_{31} & f_{32} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

In general we proceed as follows. For any $x = x_1 + \cdots + x_n$ with $x_j \in E_j$, if $y = f(x)$, since $F = F_1 \oplus \cdots \oplus F_m$, the vector $y \in F$ has a unique decomposition $y = y_1 + \cdots + y_m$ with $y_i \in F_i$, and since $f_{ij} \colon E_j \to F_i$, we have $\sum_{j=1}^n f_{ij}(x_j) \in F_i$, so $\sum_{j=1}^n f_{ij}(x_j) \in F_i$ is the ith component of $f(x)$ over the direct sum $F = F_1 \oplus \cdots \oplus F_m$; equivalently

$$pr_i^F(f(x)) = \sum_{j=1}^n f_{ij}(x_j), \quad 1 \leq i \leq m.$$

Consequently, we have

$$y_i = \sum_{j=1}^n f_{ij}(x_j), \quad 1 \leq i \leq m. \tag{$\dagger_2$}$$

This time we are summing over the index j, which eventually corresponds to multiplying the ith row of the matrix representing f by the n-tuple $(x_1, \ldots, x_n)$; see Definition 6.7.

All this suggests a generalization of the matrix notation Ax, where A is a matrix of scalars and x is a column vector of scalars, namely to write

$$\begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} f_{1\,1} & \ldots & f_{1\,n} \\ \vdots & \ddots & \vdots \\ f_{m\,1} & \ldots & f_{m\,n} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \tag{$\dagger_3$}$$

which is an abbreviation for the m equations

$$y_i = \sum_{j=1}^n f_{ij}(x_j), \quad i = 1, \ldots, m.$$

The interpretation of the multiplication of an $m \times n$ matrix of linear maps f_{ij} by a column vector of n vectors $x_j \in E_j$ is to apply each f_{ij} to x_j to obtain $f_{ij}(x_j)$ and to sum over the index j to obtain the ith output vector. This is the generalization of multiplying the scalar a_{ij} by the scalar x_j. Also note that the jth column of the matrix (f_{ij}) consists of the maps $(f_{1j}, \ldots, f_{mj})$ such that $(f_{1j}(x_j), \ldots, f_{mj}(x_j))$ are the components of $f(x_j) = f_j(x_j)$ over the direct sum $F = F_1 \oplus \cdots \oplus F_m$.

In the special case in which each E_j and each F_i is one-dimensional, this is equivalent to choosing a basis $(u_1, \ldots, u_n)$ in E so that E_j is the one-dimensional subspace $E_j = Ku_j$, and a basis $(v_1, \ldots, v_m)$ in F_j so that F_i is the one-dimensional subspace $F_i = Kv_i$. In this case every vector $x \in E$ is expressed as $x = x_1u_1 + \cdots + x_nu_n$, where the $x_i \in K$ are scalars and similarly every vector $y \in F$ is expressed as $y = y_1v_1 + \cdots + y_mv_m$, where the $y_i \in K$ are scalars. Each linear map $f_{ij} \colon E_j \to F_i$ is a map between the one-dimensional spaces Ku_j and Kv_i, so it is of the form $f_{ij}(x_ju_j) = a_{ij}x_jv_i$, with $x_j \in K$, and so the matrix (f_{ij}) of linear maps f_{ij} is in one-to-one correspondence with the matrix (a_{ij}) of scalars in K, and Equation ($\dagger_3$) where the x_j and y_i are vectors is equivalent to the same familar equation where the x_j and y_i are the scalar coordinates of x and y over the respective bases.

Definition 6.7. Let E and F be two vector spaces and assume that they are expressed as direct sums

$$E = \bigoplus_{j=1}^{n} E_j, \quad F = \bigoplus_{i=1}^{m} F_i.$$

Given any linear map $f\colon E \to F$, if $(f_{ij})_{1\leq i\leq m, 1\leq j\leq n}$ is the familiy of linear maps $f_{ij}\colon E_j \to F_i$ defined in Definition 6.6, the $m \times n$ matrix of linear maps

$$M(f) = \begin{pmatrix} f_{1\,1} & \dots & f_{1\,n} \\ \vdots & \ddots & \vdots \\ f_{m\,1} & \dots & f_{m\,n} \end{pmatrix}$$

is called the *matrix of f with respect to the decompositions $\bigoplus_{j=1}^{n} E_j$, and $\bigoplus_{i=1}^{m} F_i$ of E and F as direct sums.*

For any $x = x_1 + \cdots + x_n \in E$ with $x_j \in E_j$ and any $y = y_1 + \cdots + y_m \in F$ with $y_i \in F_i$, we have $y = f(x)$ iff

$$\begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} f_{1\,1} & \dots & f_{1\,n} \\ \vdots & \ddots & \vdots \\ f_{m\,1} & \dots & f_{m\,n} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix},$$

where the matrix equation above means that the system of m equations

$$y_i = \sum_{j=1}^{n} f_{ij}(x_j), \quad i = 1\dots, m, \tag{$\dagger$}$$

holds.

But now we can also promote matrix multiplication. Suppose we have a third space G written as a direct sum. It is more convenient to write

$$E = \bigoplus_{k=1}^{p} E_k, \quad F = \bigoplus_{j=1}^{n} F_j, \quad G = \bigoplus_{i=1}^{m} G_i.$$

Assume we also have two linear maps $f\colon E \to F$ and $g\colon F \to G$. Now we have the $n \times p$ matrix of linear maps $B = (f_{jk})$ and the $m \times n$ matrix of linear maps $A = (g_{ij})$. We would like to find the $m \times p$ matrix associated with $g \circ f$.

By definition of $f_k\colon E_k \to F$ and $f_{jk}\colon E_k \to F_j$, if $x_k \in E_k$, then

$$f_k(x_k) = f(x_k) = \sum_{j=1}^{n} f_{jk}(x_k), \quad \text{with } f_{jk}(x_k) \in F_j, \tag{$*_1$}$$

and similarly, by definition of $g_j\colon F_j \to G$ and $g_{ij}\colon F_j \to G_i$, if $y_j \in F_j$, then

$$g_j(y_j) = g(y_j) = \sum_{i=1}^{m} g_{ij}(y_j), \quad \text{with } g_{ij}(y_j) \in G_i. \tag{$*_2$}$$

If we write $h = g \circ f$, we need to find the maps $h_k \colon E_k \to G$, with

$$h_k(x_k) = h(x_k) = (g \circ f)(x_k),$$

and $h_{ik} \colon E_k \to G_i$ given by

$$h_{ik} = pr_i^G \circ h_k.$$

We have

$$\begin{aligned} h_k(x_k) &= (g \circ f)(x_k) = g(f(x_k)) = g(f_k(x_k)) \\ &= g\Bigg(\sum_{j=1}^n f_{jk}(x_k)\Bigg) && \text{by } (*_1) \\ &= \sum_{j=1}^n g(f_{jk}(x_k)) = \sum_{j=1}^n g_j(f_{jk}(x_k)) && \text{since } g \text{ is linear} \\ &= \sum_{j=1}^n \sum_{i=1}^m g_{ij}(f_{jk}(x_k)) = \sum_{i=1}^m \sum_{j=1}^n g_{ij}(f_{jk}(x_k)), && \text{by } (*_2) \end{aligned}$$

and since $\sum_{j=1}^n g_{ij}(f_{jk}(x_k)) \in G_i$, we conclude that

$$h_{ik}(x_k) = \sum_{j=1}^n g_{ij}(f_{jk}(x_k)) = \sum_{j=1}^n (g_{ij} \circ f_{jk})(x_k), \qquad (*_3)$$

which can also be expressed as

$$h_{ik} = \sum_{j=1}^n g_{ij} \circ f_{jk}. \qquad (*_4)$$

Equation $(*_4)$ is exactly the analog of the formula for the multiplication of matrices of scalars! We just have to replace multiplication by composition. The $m \times p$ matrix of linear maps (h_{ik}) is the "product" AB of the matrices of linear maps $A = (g_{ij})$ and $B = (f_{jk})$, except that multiplication is replaced by composition.

In summary we just proved the following result.

Proposition 6.12. *Let E, F, G be three vector spaces expressed as direct sums*

$$E = \bigoplus_{k=1}^p E_k, \quad F = \bigoplus_{j=1}^n F_j, \quad G = \bigoplus_{i=1}^m G_i.$$

For any two linear maps $f \colon E \to F$ and $g \colon F \to G$, let $B = (f_{jk})$ be the $n \times p$ matrix of linear maps associated with f (with respect to the decomposition of E and F as direct sums) and let $A = (g_{ij})$ be the $m \times n$ matrix of linear maps associated with g (with respect to the decomposition of F and G as direct sums). Then the $m \times p$ matrix $C = (h_{ik})$ of linear maps *associated with $h = g \circ f$ (with respect to the decomposition of E and G as direct sums) is given by*

$$C = AB,$$

with

$$h_{ik} = \sum_{j=1}^{n} g_{ij} \circ f_{jk}, \quad 1 \leq i \leq m,\ 1 \leq k \leq p.$$

We will use Proposition 6.12 to justify the rule for the block multiplication of matrices. The difficulty is mostly notational. Again suppose that E and F are expressed as direct sums

$$E = \bigoplus_{j=1}^{n} E_j, \quad F = \bigoplus_{i=1}^{m} F_i,$$

and let $f\colon E \to F$ be a linear map. Furthermore, suppose that E has a finite basis $(u_t)_{t\in T}$, where T is the disjoint union $T = T_1 \cup \cdots \cup T_n$ of nonempty subsets T_j so that $(u_t)_{t\in T_j}$ is a basis of E_j, and similarly F has a finite basis $(v_s)_{s\in S}$, where S is the disjoint union $S = S_1 \cup \cdots \cup S_m$ of nonempty subsets S_i so that $(v_s)_{s\in S_i}$ is a basis of F_i. Let $M = |S|$, $N = |T|$, $s_i = |S_i|$, and let $t_j = |T_j|$. Since s_i is the number of elements in the basis $(v_s)_{s\in S_i}$ of F_i and $F = F_1 \oplus \cdots \oplus F_m$, we have $M = \dim(F) = s_1 + \cdots + s_m$. Similarly, since t_j is the number of elements in the basis $(u_t)_{t\in T_j}$ of E_j and $E = E_1 \oplus \cdots \oplus E_n$, we have $N = \dim(E) = t_1 + \cdots + t_n$.

Let $A = (a_{st})_{(s,t)\in S\times T}$ be the (ordinary) $M \times N$ matrix of scalars (in K) representing f with respect to the basis $(u_t)_{t\in T}$ of E and the basis $(v_s)_{s\in S}$ of F with $M = r_1 + \cdots + r_m$ and $N = s_1 + \cdots + s_n$, which means that for any $t \in T$, the tth column of A consists of the components a_{st} of $f(u_t)$ over the basis $(v_s)_{s\in S}$, as in the beginning of Section 4.1.

For any i and any j such that $1 \leq i \leq m$ and $1 \leq j \leq n$, we can form the $s_i \times t_j$ matrix A_{S_i,T_j} obtained by deleting all rows in A of index $s \notin S_i$ and all columns in A of index $t \notin T_j$. The matrix A_{S_i,T_j} is the indexed family $(a_{st})_{(s,t)\in S_i\times T_j}$, as explained at the beginning of Section 4.1.

Observe that the matrix A_{S_i,T_j} is actually the matrix representing the linear map $f_{ij}\colon E_j \to F_i$ of Definition 6.7 with respect to the basis $(u_t)_{t\in T_j}$ of E_j and the basis $(v_s)_{s\in S_i}$ of F_i, in the sense that for any $t \in T_j$, the tth column of A_{S_i,T_j} consists of the components a_{st} of $f_{ij}(u_t)$ over the basis $(v_s)_{s\in S_i}$.

Definition 6.8. Given an $M \times N$ matrix A (with entries in K), we define the $m \times n$ matrix $(A_{ij})_{1\leq i\leq m, 1\leq j\leq n}$ whose entry A_{ij} is the matrix $A_{ij} = A_{S_i,T_j}$, $1 \leq i \leq m$, $1 \leq j \leq n$, and we call it the *block matrix of A associated with the partitions $S = S_1 \cup \cdots \cup S_m$ and $T = T_1 \cup \cdots \cup T_n$*. The matrix A_{S_i,T_j} is an $s_i \times t_j$ matrix called the (i,j)th *block* of this block matrix.

Here we run into a notational dilemma which does not seem to be addressed in the literature. Horn and Johnson [95] (Section 0.7) define partitioned matrices as we do, but they do not propose a notation for block matrices.

The problem is that the block matrix $(A_{ij})_{1\leq i\leq m, 1\leq j\leq n}$ is *not* equal to the original matrix A. First of all, the block matrix is $m \times n$ and its entries are matrices, but the matrix A is $M \times N$ and its entries are scalars. But even if we think of the block matrix as an $M \times N$ matrix of scalars, some rows and some columns of the original matrix A may have been *permuted* due to the choice of the partitions $S = S_1 \cup \cdots \cup S_m$ and $T = T_1 \cup \cdots \cup T_n$; see Example 6.3.

We propose to denote the block matrix $(A_{ij})_{1\leq i\leq m, 1\leq j\leq n}$ by $[A]$. This is not entirely satisfactory since all information about the partitions of S and T are lost, but at least this allows us to distinguish between A and a block matrix arising from A.

To be completely rigorous we may proceed as follows. Let $[m] = \{1, \ldots, m\}$ and $[n] = \{1, \ldots, n\}$.

Definition 6.9. For any two finite sets of indices S and T, an $S \times T$ *matrix* A is an $S \times T$-indexed family with values in K, that is, a function

$$A\colon S \times T \to K.$$

Denote the space of $S \times T$ matrices with entries in K by $\mathrm{M}_{S,T}(K)$.

An $S \times T$ matrix A is an $S \times T$-indexed family $(a_{st})_{(s,t)\in S\times T}$, but the standard representation of a matrix by a rectangular array of scalars is not quite correct because it assumes that the rows are indexed by indices in the "canonical index set" $[m]$ and that the columns are indexed by indices in the "canonical index set" $[n]$. Also the index sets need not be ordered, but in practice they are, so if $S = \{s_1, \ldots, s_m\}$ and $T = \{t_1, \ldots, t_n\}$, we denote an $S \times T$ matrix A by the rectangular array

$$A = \begin{pmatrix} a_{s_1 t_1} & \cdots & a_{s_1 t_n} \\ \vdots & \ddots & \vdots \\ a_{s_m t_1} & \cdots & a_{s_m t_n} \end{pmatrix}.$$

Even if the index sets are not ordered, the product of an $R \times S$ matrix A and of an $S \times T$ matrix B is well defined and $C = AB$ is an $R \times T$ matrix (where R, S, T are finite index sets); see Proposition 6.13.

Then an $m \times n$ block matrix X induced by two partitions $S = S_1 \cup \cdots \cup S_m$ and $T = T_1 \cup \cdots \cup T_n$ is an $[m] \times [n]$-indexed family

$$X\colon [m] \times [n] \to \prod_{(i,j)\in [m]\times[n]} \mathrm{M}_{S_i,T_j}(K),$$

such that $X(i,j) \in \mathrm{M}_{S_i,T_j}(K)$, which means that $X(i,j)$ is an $S_i \times T_j$ matrix with entries in K. The map X also defines the $M \times N$ matrix $A = (a_{st})_{s\in S, t\in T}$, with

$$a_{st} = X(i,j)_{st},$$

for any $s \in S_i$ and any $j \in T_j$, so in fact $X = [A]$ and $X(i,j) = A_{S_i,T_j}$. But remember that we abbreviate $X(i,j)$ as X_{ij}, so the (i,j)th entry in the block matrix $[A]$ of A associated with the partitions $S = S_1 \cup \cdots \cup S_m$ and $T = T_1 \cup \cdots \cup T_n$ should be denoted by $[A]_{ij}$. To minimize notation we will use the simpler notation A_{ij}. Schematically we represent the block matrix $[A]$ as

$$[A] = \begin{pmatrix} A_{S_1,T_1} & \cdots & A_{S_1,T_n} \\ \vdots & \ddots & \vdots \\ A_{S_m,T_1} & \cdots & A_{S_m,T_n} \end{pmatrix} \quad \text{or simply as} \quad [A] = \begin{pmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{m1} & \cdots & A_{mn} \end{pmatrix}.$$

In the simplified notation we lose the information about the index sets of the blocks.

Remark: It is easy to check that the set of $m \times n$ block matrices induced by two partitions $S = S_1 \cup \cdots \cup S_m$ and $T = T_1 \cup \cdots \cup T_n$ is a vector space. In fact, it is isomorphic to the direct sum

$$\bigoplus_{(i,j)\in[m]\times[n]} \mathrm{M}_{S_i,T_j}(K).$$

Addition and rescaling are performed blockwise.

Example 6.2. Let $S = \{1,2,3,4,5,6\}$, with $S_1 = \{1,2\}$, $S_2 = \{3\}$, $S_3 = \{4,5,6\}$, and $T = \{1,2,3,4,5\}$, with $T_1 = \{1,2\}$, $T_2 = \{3,4\}$, $T_3 = \{5\}$, and Then $s_1 = 2$, $s_2 = 1$, $s_3 = 3$ and $t_1 = 2, t_2 = 2, t_3 = 1$. The original matrix is a 6×5 matrix $A = (a_{ij})$. Schematically we obtain a 3×3 matrix of nine blocks. where A_{11}, A_{12}, A_{13} are respectively 2×2, 2×2 and 2×1, A_{21}, A_{22}, A_{23} are respectively 1×2, 1×2 and 1×1, and A_{31}, A_{32}, A_{33} are respectively 3×2, 3×2 and 3×1, as illustrated below.

$$[A] = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix} = \begin{pmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} & \begin{bmatrix} a_{13} & a_{14} \\ a_{23} & a_{24} \end{bmatrix} & \begin{bmatrix} a_{15} \\ a_{25} \end{bmatrix} \\ \begin{bmatrix} a_{31} & a_{32} \end{bmatrix} & \begin{bmatrix} a_{33} & a_{34} \end{bmatrix} & \begin{bmatrix} a_{35} \end{bmatrix} \\ \begin{bmatrix} a_{41} & a_{42} \\ a_{51} & a_{52} \\ a_{61} & a_{62} \end{bmatrix} & \begin{bmatrix} a_{43} & a_{44} \\ a_{53} & a_{54} \\ a_{63} & a_{64} \end{bmatrix} & \begin{bmatrix} a_{45} \\ a_{55} \\ a_{65} \end{bmatrix} \end{pmatrix}.$$

Technically, the blocks are obtained from A in terms of the subsets S_i, T_j. For example,

$$A_{12} = A_{\{1,2\},\{3,4\}} = \begin{bmatrix} a_{13} & a_{14} \\ a_{23} & a_{24} \end{bmatrix}.$$

Example 6.3. Let $S = \{1,2,3\}$, with $S_1 = \{1,3\}$, $S_2 = \{2\}$, and $T = \{1,2,3\}$, with $T_1 = \{1,3\}$, $T_2 = \{2\}$. Then $s_1 = 2$, $s_2 = 1$, and $t_1 = 2, t_2 = 1$. The block 2×2 matrix $[A]$ associated with above partitions is

$$[A] = \begin{pmatrix} A_{\{1,3\},\{1,3\}} & A_{\{1,3\},\{2\}} \\ A_{\{2\},\{1,3\}} & A_{\{2\},\{2\}} \end{pmatrix} = \begin{pmatrix} \begin{bmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{bmatrix} & \begin{bmatrix} a_{12} \\ a_{32} \end{bmatrix} \\ \begin{bmatrix} a_{21} & a_{23} \end{bmatrix} & \begin{bmatrix} a_{22} \end{bmatrix} \end{pmatrix}.$$

Observe that $[A]$ viewed as a 3×3 scalar matrix is definitely different from

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

In practice, $S = \{1, \ldots, M\}$ and $T = \{1, \ldots, N\}$, so there are bijections $\alpha_i \colon \{1, \ldots, s_i\} \to S_i$ and $\beta_j \colon \{1, \ldots, t_j\} \to T_j$, $1 \leq i \leq m$, $1 \leq j \leq n$. Each $s_i \times t_j$ matrix A_{S_i,T_j} is considered as a submatrix of A, this is why the rows are indexed by S_i and the columns are indexed by T_j, but this matrix can also be viewed as an $s_i \times t_j$ matrix $A'_{ij} = \big((a'_{ij})_{st}\big)$ by itself, with the rows indexed by $\{1, \ldots, s_i\}$ and the columns indexed by $\{1, \ldots, t_j\}$, with

$$(a'_{ij})_{st} = a_{\alpha(s)\beta(t)}, \quad 1 \leq s \leq s_i,\ 1 \leq t \leq t_j.$$

Symbolic systems like `Matlab` have commands to construct such matrices. But be careful that to put a matrix such as A'_{ij} back into A at the correct row and column locations requires viewing this matrix as A_{S_i,T_j}. Symbolic systems like `Matlab` also have commands to assign row vectors and column vectors to specific rows or columns of a matrix. Technically, to be completely rigorous, the matrices A_{S_i,T_j} and A'_{ij} are different, even though they contain the same entries. The reason they are different is that in A_{S_i,T_j} the entries are indexed by the index sets S_i and T_j, but in A'_{ij} they are indexed by the index sets $\{1, \ldots, s_i\}$ and $\{1, \ldots, t_j\}$. This depends whether we view A_{S_i,T_j} as a submatrix of A or as a matrix on its own.

In most cases, the partitions $S = S_1 \cup \cdots \cup S_m$ and $T = T_1 \cup \cdots \cup T_n$ are chosen so that

$$\begin{aligned} S_i &= \{s \mid s_1 + \cdots + s_{i-1} + 1 \leq s \leq s_1 + \cdots + s_i\} \\ T_j &= \{t \mid t_1 + \cdots + t_{i-1} + 1 \leq t \leq t_1 + \cdots + t_j\}, \end{aligned}$$

with $s_i = |S_i| \geq 1$, $t_j = |T_j| \geq 1$, $1 \leq i \leq m, 1 \leq j \leq n$. For $i = 1$, we have $S_1 = \{1, \ldots, s_1\}$ and $T_1 = \{1, \ldots, t_1\}$. *This means that we partition into consecutive subsets of consecutive integers and we preserve the order of the bases.* In this case, $[A]$ can be viewed as A. But the results about block multiplication hold in the general case.

Finally we tackle block multiplication. But first we observe that the computation made in Section 4.2 can be immediately adapted to matrices indexed by arbitrary finite index sets I, J, K, not necessary of the form $\{1, \ldots, p\}$, $\{1, \ldots, n\}$, $\{1, \ldots, m\}$. We need this to deal with products of matrices occurring as blocks in other matrices, since such matrices are of the form A_{S_i,T_j}, *etc.*

We can prove immediately the following result generalizing Equation (4) proven in Section 4.2 (also see the fourth equation of Proposition 4.2).

Proposition 6.13. *Let I, J, K be any nonempty finite index sets. If the $I \times J$ matrix $A = (a_{ij})_{(i,j) \in I \times J}$ represents the linear map $g \colon F \to G$ with respect to the basis $(v_j)_{j \in J}$ of F and the basis $(w_i)_{i \in I}$ of G and if the $J \times K$ matrix $B = (b_{jk})_{(j,k) \in J \times K}$ represents the linear map $f \colon E \to F$ with respect to the basis $(u_k)_{k \in K}$ of E and the basis $(v_j)_{j \in J}$ of F, then the*

$I \times K$ matrix $C = (c_{ik})_{(i,k)\in I\times K}$ representing the linear map $g \circ f\colon E \to G$ with respect to the basis $(u_k)_{k\in K}$ of E and the basis $(w_i)_{i\in I}$ of G is given by

$$C = AB,$$

where for all $i \in I$ and all $k \in K$,

$$c_{ik} = \sum_{j\in J} a_{ij} b_{jk}.$$

Let E, F, G be three vector spaces expressed as direct sums

$$E = \bigoplus_{k=1}^{p} E_k, \quad F = \bigoplus_{j=1}^{n} F_j, \quad G = \bigoplus_{i=1}^{m} G_i,$$

and let $f\colon E \to F$ and $g\colon F \to G$ be two linear maps. Furthermore, assume that E has a finite basis $(u_t)_{t\in T}$, where T is the disjoint union $T = T_1 \cup \cdots \cup T_p$ of nonempty subsets T_k so that $(u_t)_{t\in T_k}$ is a basis of E_k, F has a finite basis $(v_s)_{s\in S}$, where S is the disjoint union $S = S_1 \cup \cdots \cup S_n$ of nonempty subsets S_j so that $(v_s)_{s\in S_j}$ is a basis of F_j, and G has a finite basis $(w_r)_{r\in R}$, where R is the disjoint union $R = R_1 \cup \cdots \cup R_m$ of nonempty subsets R_i so that $(w_r)_{r\in R_i}$ is a basis of G_i. Also let $M = |R|$, $N = |S|$, $P = |T|$, $r_i = |R_i|$, $s_j = |S_j|$, $t_k = |T_k|$, so that $M = \dim(G) = r_1 + \cdots + r_m$, $N = \dim(F) = s_1 + \cdots + s_n$, and $P = \dim(E) = t_1 + \cdots + t_p$.

Let B be the $N \times P$ matrix representing f with respect to the basis $(u_t)_{t\in T}$ of E and the basis $(v_s)_{s\in S}$ of F, let A be the $M \times N$ matrix representing g with respect to the basis $(v_s)_{s\in S}$ of F and the basis $(w_r)_{r\in R}$ of G, and let C be the $M \times P$ matrix representing $h = g \circ f$ with respect to the basis basis $(u_t)_{t\in T}$ of E and the basis $(w_r)_{r\in R}$ of G.

The matrix $[A]$ is an $m \times n$ block matrix of $r_i \times s_j$ matrices A_{ij} $(1 \le i \le m, 1 \le j \le n)$, the matrix $[B]$ is an $n \times p$ block matrix of $s_j \times t_k$ matrices B_{jk} $(1 \le j \le n, 1 \le k \le p)$, and the matrix $[C]$ is an $m \times p$ block matrix of $r_i \times t_k$ matrices C_{ik} $(1 \le i \le m, 1 \le k \le p)$. Furthermore, to be precise, $A_{ij} = A_{R_i,S_j}$, $B_{jk} = B_{S_j,T_k}$, and $C_{ik} = C_{R_i,T_k}$.

Now recall that the matrix A_{R_i,S_j} represents the linear map $g_{ij}\colon F_j \to G_i$ with respect to the basis $(v_s)_{s\in S_j}$ of F_j and the basis $(w_r)_{r\in R_i}$ of G_i, the matrix B_{S_j,T_k} represents the linear map $f_{jk}\colon E_k \to F_j$ with respect to the basis $(u_t)_{t\in T_k}$ of E_k and the basis $(v_s)_{s\in S_j}$ of F_j, and the matrix C_{R_i,T_k} represents the linear map $h_{ik}\colon E_k \to G_i$ with respect to the basis $(u_t)_{t\in T_k}$ of E_k and the basis $(w_r)_{r\in R_i}$ of G_i.

By Proposition 6.12, h_{ik} is given by the formula

$$h_{ik} = \sum_{j=1}^{n} g_{ij} \circ f_{jk}, \quad 1 \le i \le m,\ 1 \le k \le p, \tag{$*_5$}$$

and since the matrix A_{R_i,S_j} represents $g_{ij}\colon F_j \to G_i$, the matrix B_{S_j,T_k} represents $f_{jk}\colon E_k \to F_j$, and the matrix C_{R_i,T_k} represents $h_{ik}\colon E_k \to G_i$, so $(*_5)$ implies the matrix equation

$$C_{ik} = \sum_{j=1}^{n} A_{ij} B_{jk}, \quad 1 \leq i \leq m,\ 1 \leq k \leq p, \tag{$*_6$}$$

establishing (when combined with Proposition 6.13) the fact that $[C] = [A][B]$, namely the product $C = AB$ of the matrices A and B can be performed by blocks, using the same product formula on matrices that is used on scalars.

We record the above fact in the following proposition.

Proposition 6.14. *Let M, N, P be any positive integers, and let $\{1, \ldots, M\} = R_1 \cup \cdots \cup R_m$, $\{1, \ldots, N\} = S_1 \cup \cdots \cup S_n$, and $\{1, \ldots, P\} = T_1 \cup \cdots \cup T_p$ be any partitions into nonempty subsets R_i, S_j, T_k, and write $r_i = |R_i|$, $s_j = |S_j|$ and $t_k = |T_k|$ $(1 \leq i \leq m, 1 \leq j \leq n, 1 \leq k \leq p)$. Let A be an $M \times N$ matrix, let $[A]$ be the corresponding $m \times n$ block matrix of $r_i \times s_j$ matrices A_{ij} $(1 \leq i \leq m, 1 \leq j \leq n)$, and let B be an $N \times P$ matrix and $[B]$ be the corresponding $n \times p$ block matrix of $s_j \times t_k$ matrices B_{jk} $(1 \leq j \leq n, 1 \leq k \leq p)$. Then the $M \times P$ matrix $C = AB$ corresponds to an $m \times p$ block matrix $[C]$ of $r_i \times t_k$ matrices C_{ik} $(1 \leq i \leq m, 1 \leq k \leq p)$, and we have*

$$[C] = [A][B],$$

which means that

$$C_{ik} = \sum_{j=1}^{n} A_{ij} B_{jk}, \quad 1 \leq i \leq m,\ 1 \leq k \leq p.$$

Remark: The product $A_{ij}B_{jk}$ of the blocks A_{ij} and B_{jk}, which are really the matrices A_{R_i,S_j} and B_{S_j,T_k}, can be computed using the matrices A'_{ij} and B'_{jk} (discussed after Example 6.3) that are indexed by the "canonical" index sets $\{1, \ldots, r_i\}$, $\{1, \ldots, s_j\}$ and $\{1, \ldots, t_k\}$. But after computing $A'_{ij}B'_{jk}$, we have to remember to insert it as a block in $[C]$ using the correct index sets R_i and T_k. This is easily achieved in `Matlab`.

Example 6.4. Consider the partition of the index set $R = \{1, 2, 3, 4, 5, 6\}$ given by $R_1 = \{1, 2\}$, $R_2 = \{3\}$, $R_3 = \{4, 5, 6\}$; of the index set $S = \{1, 2, 3\}$ given by $S_1 = \{1, 2\}$, $S_2 = \{3\}$; and of the index set $T = \{1, 2, 3, 4, 5, 6\}$ given by $T_1 = \{1\}$, $T_2 = \{2, 3\}$, $T_3 = \{4, 5, 6\}$. Let $[A]$ be the 3×2 block matrix

$$[A] = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \\ A_{31} & A_{32} \end{pmatrix} = \begin{pmatrix} \begin{bmatrix} \quad & \quad \\ \quad & \quad \end{bmatrix} & \begin{bmatrix} \quad \\ \quad \end{bmatrix} \\ \begin{bmatrix} \quad & \quad \end{bmatrix} & \begin{bmatrix} \quad \end{bmatrix} \\ \begin{bmatrix} \quad & \quad \\ \quad & \quad \\ \quad & \quad \end{bmatrix} & \begin{bmatrix} \quad \\ \quad \\ \quad \end{bmatrix} \end{pmatrix}$$

where A_{11}, A_{12} are 2×2, 2×1; A_{21}, A_{22} are 1×2, 1×1; and A_{31}, A_{32} are 3×2, 3×1, and $[B]$ be the 2×3 block matrix

$$[B] = \begin{pmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \end{pmatrix} = \begin{pmatrix} \begin{bmatrix} \, \\ \, \end{bmatrix} & \begin{bmatrix} \, & \, \\ \, & \, \end{bmatrix} & \begin{bmatrix} \, & \, & \, \\ \, & \, & \, \end{bmatrix} \\ \begin{bmatrix} \, \end{bmatrix} & \begin{bmatrix} \, & \, \end{bmatrix} & \begin{bmatrix} \, & \, & \, \end{bmatrix} \end{pmatrix},$$

where B_{11}, B_{12}, B_{13} are 2×1, 2×2, 2×3; and B_{21}, B_{22}, B_{23} are 1×1, 1×2, 1×3. Then $[C] = [A][B]$ is the 3×3 block matrix

$$[C] = \begin{pmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{pmatrix} = \begin{pmatrix} \begin{bmatrix} \, \\ \, \end{bmatrix} & \begin{bmatrix} \, & \, \\ \, & \, \end{bmatrix} & \begin{bmatrix} \, & \, & \, \\ \, & \, & \, \end{bmatrix} \\ \begin{bmatrix} \, \end{bmatrix} & \begin{bmatrix} \, & \, \end{bmatrix} & \begin{bmatrix} \, & \, & \, \end{bmatrix} \\ \begin{bmatrix} \, \\ \, \\ \, \end{bmatrix} & \begin{bmatrix} \, & \, \\ \, & \, \\ \, & \, \end{bmatrix} & \begin{bmatrix} \, & \, & \, \\ \, & \, & \, \\ \, & \, & \, \end{bmatrix} \end{pmatrix},$$

where C_{11}, C_{12}, C_{13} are 2×1, 2×2, 2×3; C_{21}, C_{22}, C_{23} are 1×1, 1×2, 1×3; and C_{31}, C_{32}, C_{33} are 3×1, 3×2, 3×3. For example,

$$C_{32} = A_{31}B_{12} + A_{32}B_{22}.$$

Example 6.5. This example illustrates some of the subtleties having to do with the partitioning of the index sets. Consider the 1×3 matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \end{pmatrix}$$

and the 3×2 matrix

$$B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix}.$$

Consider the partition of the index set $R = \{1\}$ given by $R_1 = \{1\}$; of the index set $S = \{1, 2, 3\}$ given by $S_1 = \{1, 3\}$, $S_2 = \{2\}$; and of the index set $T = \{1, 2\}$ given by $T_1 = \{2\}$, $T_2 = \{1\}$. The corresponding block matrices are the 1×2 block matrix

$$[A] = \begin{pmatrix} A_{\{1\},\{1,3\}} & A_{\{1\},\{2\}} \end{pmatrix} = \left(\begin{bmatrix} a_{11} & a_{13} \end{bmatrix} \quad \begin{bmatrix} a_{12} \end{bmatrix} \right),$$

and the 2×2 block matrix

$$[B] = \begin{pmatrix} B_{\{1,3\},\{2\}} & B_{\{1,3\},\{1\}} \\ B_{\{2\},\{2\}} & B_{\{2\},\{1\}} \end{pmatrix} = \begin{pmatrix} \begin{bmatrix} b_{12} \\ b_{32} \end{bmatrix} & \begin{bmatrix} b_{11} \\ b_{31} \end{bmatrix} \\ \begin{bmatrix} b_{22} \end{bmatrix} & \begin{bmatrix} b_{21} \end{bmatrix} \end{pmatrix}.$$

The product of the 1×2 block matrix $[A]$ and the 2×2 block matrix $[B]$ is the 1×2 block matrix $[C]$ given by

$$\begin{aligned}
[C] = [A][B] &= \begin{pmatrix} \begin{bmatrix} a_{11} & a_{13} \end{bmatrix} & \begin{bmatrix} a_{12} \end{bmatrix} \end{pmatrix} \begin{pmatrix} \begin{bmatrix} b_{12} \\ b_{32} \end{bmatrix} & \begin{bmatrix} b_{11} \\ b_{31} \end{bmatrix} \\ \begin{bmatrix} b_{22} \end{bmatrix} & \begin{bmatrix} b_{21} \end{bmatrix} \end{pmatrix} \\
&= \left(\begin{bmatrix} a_{11} & a_{13} \end{bmatrix} \begin{bmatrix} b_{12} \\ b_{32} \end{bmatrix} + \begin{bmatrix} a_{12} \end{bmatrix} \begin{bmatrix} b_{22} \end{bmatrix} \quad \begin{bmatrix} a_{11} & a_{13} \end{bmatrix} \begin{bmatrix} b_{11} \\ b_{31} \end{bmatrix} + \begin{bmatrix} a_{12} \end{bmatrix} \begin{bmatrix} b_{21} \end{bmatrix} \right) \\
&= \begin{pmatrix} \begin{bmatrix} a_{11}b_{12} + a_{13}b_{32} + a_{12}b_{22} \end{bmatrix} & \begin{bmatrix} a_{11}b_{11} + a_{13}b_{31} + a_{12}b_{21} \end{bmatrix} \end{pmatrix} \\
&= \begin{pmatrix} \begin{bmatrix} a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \end{bmatrix} & \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} \end{bmatrix} \end{pmatrix}.
\end{aligned}$$

The block matrix $[C]$ is obtained from the 1×2 matrix $C = AB$ using the partitions of $R = \{1\}$ given by $R_1 = \{1\}$ and of $T = \{1, 2\}$ given by $T_1 = \{2\}$, $T_2 = \{1\}$, so

$$[C] = \begin{pmatrix} C_{\{1\},\{2\}} & C_{\{1\},\{1\}} \end{pmatrix},$$

which means that $[C]$ is obtained from C by permuting its two columns. Since

$$\begin{aligned}
C = AB &= \begin{pmatrix} a_{11} & a_{12} & a_{13} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix} \\
&= \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \end{pmatrix},
\end{aligned}$$

we have confirmed that $[C]$ is correct.

Example 6.6. Matrix block multiplication is a very effective method to prove that if an upper-triangular matrix A is invertible, then its inverse is also upper-triangular. We proceed by induction on the dimensiopn n of A. If $n = 1$, then $A = (a)$, where a is a scalar, so A is invertible iff $a \neq 0$, and $A^{-1} = (a^{-1})$, which is trivially upper-triangular. For the induction step we can write an $(n+1) \times (n+1)$ upper triangular matrix A in block form as

$$A = \begin{pmatrix} T & U \\ 0_{1,n} & \alpha \end{pmatrix},$$

where T is an $n \times n$ upper triangular matrix, U is an $n \times 1$ matrix and $\alpha \in \mathbb{R}$. Assume that A is invertible and let B be its inverse, written in block form as

$$B = \begin{pmatrix} C & V \\ W & \beta \end{pmatrix},$$

where C is an $n \times n$ matrix, V is an $n \times 1$ matrix, W is a $1 \times n$ matrix, and $\beta \in \mathbb{R}$. Since B is the inverse of A, we have $AB = I_{n+1}$, which yields

$$\begin{pmatrix} T & U \\ 0_{1,n-1} & \alpha \end{pmatrix} \begin{pmatrix} C & V \\ W & \beta \end{pmatrix} = \begin{pmatrix} I_n & 0_{n,1} \\ 0_{1,n} & 1 \end{pmatrix}.$$

By block multiplication we get

$$\begin{aligned} TC + UW &= I_n \\ TV + \beta U &= 0_{n,1} \\ \alpha W &= 0_{1,n} \\ \alpha\beta &= 1. \end{aligned}$$

From the above equations we deduce that $\alpha, \beta \neq 0$ and $\beta = \alpha^{-1}$. Since $\alpha \neq 0$, the equation $\alpha W = 0_{1,n}$ yields $W = 0_{1,n}$, and so

$$TC = I_n, \quad TV + \beta U = 0_{n,1}.$$

It follows that T is invertible and C is its inverse, and since T is upper triangular, by the induction hypothesis, C is also upper triangular.

The above argument can be easily modified to prove that if A is invertible, then its diagonal entries are nonzero.

We are now ready to prove a very crucial result relating the rank and the dimension of the kernel of a linear map.

6.3 The Rank-Nullity Theorem; Grassmann's Relation

We begin with the following fundamental proposition.

Proposition 6.15. *Let E, F and G, be three vector spaces, $f \colon E \to F$ an injective linear map, $g \colon F \to G$ a surjective linear map, and assume that $\mathrm{Im}\, f = \mathrm{Ker}\, g$. Then, the following properties hold. (a) For any section $s \colon G \to F$ of g, we have $F = \mathrm{Ker}\, g \oplus \mathrm{Im}\, s$, and the linear map $f + s \colon E \oplus G \to F$ is an isomorphism.*[1]

(b) For any retraction $r \colon F \to E$ of f, we have $F = \mathrm{Im}\, f \oplus \mathrm{Ker}\, r$.[2]

$$\begin{tikzcd} E \arrow[r, shift left, "f"] & F \arrow[l, shift left, "r"] \arrow[r, shift left, "g"] & G \arrow[l, shift left, "s"] \end{tikzcd}$$

Proof. (a) Since $s \colon G \to F$ is a section of g, we have $g \circ s = \mathrm{id}_G$, and for every $u \in F$,

$$g(u - s(g(u))) = g(u) - g(s(g(u))) = g(u) - g(u) = 0.$$

Thus, $u - s(g(u)) \in \mathrm{Ker}\, g$, and we have $F = \mathrm{Ker}\, g + \mathrm{Im}\, s$. On the other hand, if $u \in \mathrm{Ker}\, g \cap \mathrm{Im}\, s$, then $u = s(v)$ for some $v \in G$ because $u \in \mathrm{Im}\, s$, $g(u) = 0$ because $u \in \mathrm{Ker}\, g$, and so,

$$g(u) = g(s(v)) = v = 0,$$

[1] The existence of a section $s \colon G \to F$ of g follows from Proposition 6.11.

[2] The existence of a retraction $r \colon F \to E$ of f follows from Proposition 6.11.

because $g \circ s = \mathrm{id}_G$, which shows that $u = s(v) = 0$. Thus, $F = \operatorname{Ker} g \oplus \operatorname{Im} s$, and since by assumption, $\operatorname{Im} f = \operatorname{Ker} g$, we have $F = \operatorname{Im} f \oplus \operatorname{Im} s$. But then, since f and s are injective, $f + s\colon E \oplus G \to F$ is an isomorphism. The proof of (b) is very similar. $\square$

Note that we can choose a retraction $r\colon F \to E$ so that $\operatorname{Ker} r = \operatorname{Im} s$, since
$F = \operatorname{Ker} g \oplus \operatorname{Im} s = \operatorname{Im} f \oplus \operatorname{Im} s$ and f is injective so we can set $r \equiv 0$ on $\operatorname{Im} s$.

Given a sequence of linear maps $E \xrightarrow{f} F \xrightarrow{g} G$, when $\operatorname{Im} f = \operatorname{Ker} g$, we say that the sequence $E \xrightarrow{f} F \xrightarrow{g} G$ is *exact at F*. If in addition to being exact at F, f is injective and g is surjective, we say that we have a *short exact sequence*, and this is denoted as

$$0 \longrightarrow E \xrightarrow{f} F \xrightarrow{g} G \longrightarrow 0.$$

The property of a short exact sequence given by Proposition 6.15 is often described by saying that $0 \longrightarrow E \xrightarrow{f} F \xrightarrow{g} G \longrightarrow 0$ is a (short) *split exact sequence*.

As a corollary of Proposition 6.15, we have the following result which shows that given a linear map $f\colon E \to F$, its domain E is the direct sum of its kernel $\operatorname{Ker} f$ with some isomorphic copy of its image $\operatorname{Im} f$.

Theorem 6.16. *(Rank-nullity theorem) Let E and F be vector spaces, and let $f\colon E \to F$ be a linear map. Then, E is isomorphic to* $\operatorname{Ker} f \oplus \operatorname{Im} f$*, and thus,*

$$\dim(E) = \dim(\operatorname{Ker} f) + \dim(\operatorname{Im} f) = \dim(\operatorname{Ker} f) + \operatorname{rk}(f).$$

See Figure 6.3.

Proof. Consider

$$\operatorname{Ker} f \xrightarrow{i} E \xrightarrow{f'} \operatorname{Im} f,$$

where $\operatorname{Ker} f \xrightarrow{i} E$ is the inclusion map, and $E \xrightarrow{f'} \operatorname{Im} f$ is the surjection associated with $E \xrightarrow{f} F$. Then, we apply Proposition 6.15 to any section $\operatorname{Im} f \xrightarrow{s} E$ of f' to get an isomorphism between E and $\operatorname{Ker} f \oplus \operatorname{Im} f$, and Proposition 6.7, to get $\dim(E) = \dim(\operatorname{Ker} f) + \dim(\operatorname{Im} f)$. $\square$

Definition 6.10. The dimension $\dim(\operatorname{Ker} f)$ of the kernel of a linear map f is called the *nullity* of f.

We now derive some important results using Theorem 6.16.

Proposition 6.17. *Given a vector space E, if U and V are any two subspaces of E, then*

$$\dim(U) + \dim(V) = \dim(U + V) + \dim(U \cap V),$$

an equation known as Grassmann's relation.

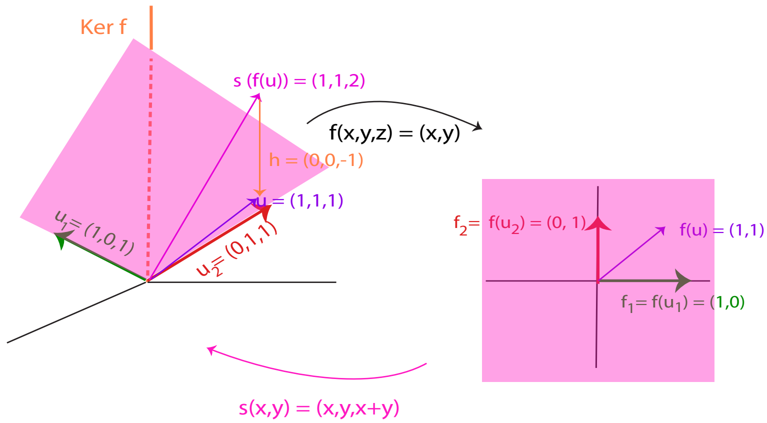

Figure 6.3: Let $f\colon E \to F$ be the linear map from $\mathbb{R}^3$ to $\mathbb{R}^2$ given by $f(x, y, z) = (x, y)$. Then $s\colon \mathbb{R}^2 \to \mathbb{R}^3$ is given by $s(x, y) = (x, y, x + y)$ and maps the pink $\mathbb{R}^2$ isomorphically onto the slanted pink plane of $\mathbb{R}^3$ whose equation is $-x - y + z = 0$. Theorem 6.16 shows that $\mathbb{R}^3$ is the direct sum of the plane $-x - y + z = 0$ and the kernel of f which the orange z-axis.

Proof. Recall that $U + V$ is the image of the linear map

$$a\colon U \times V \to E$$

given by

$$a(u, v) = u + v,$$

and that we proved earlier that the kernel $\operatorname{Ker} a$ of a is isomorphic to $U \cap V$. By Theorem 6.16,

$$\dim(U \times V) = \dim(\operatorname{Ker} a) + \dim(\operatorname{Im} a),$$

but $\dim(U \times V) = \dim(U) + \dim(V)$, $\dim(\operatorname{Ker} a) = \dim(U \cap V)$, and $\operatorname{Im} a = U + V$, so the Grassmann relation holds. □

The Grassmann relation can be very useful to figure out whether two subspace have a nontrivial intersection in spaces of dimension > 3. For example, it is easy to see that in $\mathbb{R}^5$, there are subspaces U and V with $\dim(U) = 3$ and $\dim(V) = 2$ such that $U \cap V = (0)$; for example, let U be generated by the vectors $(1, 0, 0, 0, 0), (0, 1, 0, 0, 0)$, $(0, 0, 1, 0, 0)$, and V be generated by the vectors $(0, 0, 0, 1, 0)$ and $(0, 0, 0, 0, 1)$. However, we claim that if $\dim(U) = 3$ and $\dim(V) = 3$, then $\dim(U \cap V) \geq 1$. Indeed, by the Grassmann relation, we have

$$\dim(U) + \dim(V) = \dim(U + V) + \dim(U \cap V),$$

namely

$$3 + 3 = 6 = \dim(U + V) + \dim(U \cap V),$$

and since $U + V$ is a subspace of $\mathbb{R}^5$, $\dim(U + V) \leq 5$, which implies

$$6 \leq 5 + \dim(U \cap V),$$

that is $1 \leq \dim(U \cap V)$.

As another consequence of Proposition 6.17, if U and V are two hyperplanes in a vector space of dimension n, so that $\dim(U) = n - 1$ and $\dim(V) = n - 1$, the reader should show that

$$\dim(U \cap V) \geq n - 2,$$

and so, if $U \neq V$, then

$$\dim(U \cap V) = n - 2.$$

Here is a characterization of direct sums that follows directly from Theorem 6.16.

Proposition 6.18. *If $U_1, \ldots, U_p$ are any subspaces of a finite dimensional vector space E, then*

$$\dim(U_1 + \cdots + U_p) \leq \dim(U_1) + \cdots + \dim(U_p),$$

and

$$\dim(U_1 + \cdots + U_p) = \dim(U_1) + \cdots + \dim(U_p)$$

iff the U_is form a direct sum $U_1 \oplus \cdots \oplus U_p$.

Proof. If we apply Theorem 6.16 to the linear map

$$a \colon U_1 \times \cdots \times U_p \to U_1 + \cdots + U_p$$

given by $a(u_1, \ldots, u_p) = u_1 + \cdots + u_p$, we get

$$\begin{aligned}\dim(U_1 + \cdots + U_p) &= \dim(U_1 \times \cdots \times U_p) - \dim(\mathrm{Ker}\, a) \\ &= \dim(U_1) + \cdots + \dim(U_p) - \dim(\mathrm{Ker}\, a),\end{aligned}$$

so the inequality follows. Since a is injective iff $\mathrm{Ker}\, a = (0)$, the U_is form a direct sum iff the second equation holds. ☐

Another important corollary of Theorem 6.16 is the following result:

Proposition 6.19. *Let E and F be two vector spaces with the same finite dimension $\dim(E) = \dim(F) = n$. For every linear map $f \colon E \to F$, the following properties are equivalent:*

(a) f is bijective.

(b) f is surjective.

(c) f is injective.

(d) $\mathrm{Ker}\, f = (0)$.

Proof. Obviously, (a) implies (b).

If f is surjective, then $\mathrm{Im}\, f = F$, and so $\dim(\mathrm{Im}\, f) = n$. By Theorem 6.16,

$$\dim(E) = \dim(\mathrm{Ker}\, f) + \dim(\mathrm{Im}\, f),$$

and since $\dim(E) = n$ and $\dim(\mathrm{Im}\, f) = n$, we get $\dim(\mathrm{Ker}\, f) = 0$, which means that $\mathrm{Ker}\, f = (0)$, and so f is injective (see Proposition 3.17). This proves that (b) implies (c).

If f is injective, then by Proposition 3.17, $\mathrm{Ker}\, f = (0)$, so (c) implies (d).

Finally, assume that $\mathrm{Ker}\, f = (0)$, so that $\dim(\mathrm{Ker}\, f) = 0$ and f is injective (by Proposition 3.17). By Theorem 6.16,

$$\dim(E) = \dim(\mathrm{Ker}\, f) + \dim(\mathrm{Im}\, f),$$

and since $\dim(\mathrm{Ker}\, f) = 0$, we get

$$\dim(\mathrm{Im}\, f) = \dim(E) = \dim(F),$$

which proves that f is also surjective, and thus bijective. This proves that (d) implies (a) and concludes the proof. $\square$

One should be warned that Proposition 6.19 fails in infinite dimension.

The following Proposition will also be useful.

Proposition 6.20. *Let E be a vector space. If $E = U \oplus V$ and $E = U \oplus W$, then there is an isomorphism $f \colon V \to W$ between V and W.*

Proof. Let R be the relation between V and W, defined such that

$$\langle v, w \rangle \in R \quad \text{iff} \quad w - v \in U.$$

We claim that R is a functional relation that defines a linear isomorphism $f \colon V \to W$ between V and W, where $f(v) = w$ iff $\langle v, w \rangle \in R$ (R is the graph of f). If $w - v \in U$ and $w' - v \in U$, then $w' - w \in U$, and since $U \oplus W$ is a direct sum, $U \cap W = (0)$, and thus $w' - w = 0$, that is $w' = w$. Thus, R is functional. Similarly, if $w - v \in U$ and $w - v' \in U$, then $v' - v \in U$, and since $U \oplus V$ is a direct sum, $U \cap V = (0)$, and $v' = v$. Thus, f is injective. Since $E = U \oplus V$, for every $w \in W$, there exists a unique pair $\langle u, v \rangle \in U \times V$, such that $w = u + v$. Then, $w - v \in U$, and f is surjective. We also need to verify that f is linear. If

$$w - v = u$$

and

$$w' - v' = u',$$

where $u, u' \in U$, then, we have

$$(w + w') - (v + v') = (u + u'),$$

where $u + u' \in U$. Similarly, if

$$w - v = u$$

where $u \in U$, then we have

$$\lambda w - \lambda v = \lambda u,$$

where $\lambda u \in U$. Thus, f is linear. $\square$

Given a vector space E and any subspace U of E, Proposition 6.20 shows that the dimension of any subspace V such that $E = U \oplus V$ depends only on U. We call $\dim(V)$ the *codimension* of U, and we denote it by $\mathrm{codim}(U)$. A subspace U of codimension 1 is called a *hyperplane*.

The notion of rank of a linear map or of a matrix is an important one, both theoretically and practically, since it is the key to the solvability of linear equations. Recall from Definition 3.19 that the *rank* $\mathrm{rk}(f)$ of a linear map $f\colon E \to F$ is the dimension $\dim(\mathrm{Im}\, f)$ of the image subspace $\mathrm{Im}\, f$ of F.

We have the following simple proposition.

Proposition 6.21. *Given a linear map $f\colon E \to F$, the following properties hold:*

(i) $\mathrm{rk}(f) = \mathrm{codim}(\mathrm{Ker}\, f)$.

(ii) $\mathrm{rk}(f) + \dim(\mathrm{Ker}\, f) = \dim(E)$.

(iii) $\mathrm{rk}(f) \leq \min(\dim(E), \dim(F))$.

Proof. Since by Proposition 6.16, $\dim(E) = \dim(\mathrm{Ker}\, f) + \dim(\mathrm{Im}\, f)$, and by definition, $\mathrm{rk}(f) = \dim(\mathrm{Im}\, f)$, we have $\mathrm{rk}(f) = \mathrm{codim}(\mathrm{Ker}\, f)$. Since $\mathrm{rk}(f) = \dim(\mathrm{Im}\, f)$, (ii) follows from $\dim(E) = \dim(\mathrm{Ker}\, f) + \dim(\mathrm{Im}\, f)$. As for (iii), since $\mathrm{Im}\, f$ is a subspace of F, we have $\mathrm{rk}(f) \leq \dim(F)$, and since $\mathrm{rk}(f) + \dim(\mathrm{Ker}\, f) = \dim(E)$, we have $\mathrm{rk}(f) \leq \dim(E)$. $\square$

The rank of a matrix is defined as follows.

Definition 6.11. Given a $m \times n$-matrix $A = (a_{i\,j})$ over the field K, the *rank* $\mathrm{rk}(A)$ of the matrix A is the maximum number of linearly independent columns of A (viewed as vectors in K^m).

In view of Proposition 3.8, the rank of a matrix A is the dimension of the subspace of K^m generated by the columns of A. Let E and F be two vector spaces, and let $(u_1, \ldots, u_n)$ be a basis of E, and $(v_1, \ldots, v_m)$ a basis of F. Let $f\colon E \to F$ be a linear map, and let $M(f)$ be its matrix w.r.t. the bases $(u_1, \ldots, u_n)$ and $(v_1, \ldots, v_m)$. Since the rank $\mathrm{rk}(f)$ of f is the dimension of $\operatorname{Im} f$, which is generated by $(f(u_1), \ldots, f(u_n))$, the rank of f is the maximum number of linearly independent vectors in $(f(u_1), \ldots, f(u_n))$, which is equal to the number of linearly independent columns of $M(f)$, since F and K^m are isomorphic. Thus, we have $\operatorname{rk}(f) = \operatorname{rk}(M(f))$, for every matrix representing f.

We will see later, using duality, that the rank of a matrix A is also equal to the maximal number of linearly independent rows of A.

If U is a hyperplane, then $E = U \oplus V$ for some subspace V of dimension 1. However, a subspace V of dimension 1 is generated by any nonzero vector $v \in V$, and thus we denote V by Kv, and we write $E = U \oplus Kv$. Clearly, $v \notin U$. Conversely, let $x \in E$ be a vector such that $x \notin U$ (and thus, $x \neq 0$). We claim that $E = U \oplus Kx$. Indeed, since U is a hyperplane, we have $E = U \oplus Kv$ for some $v \notin U$ (with $v \neq 0$). Then, $x \in E$ can be written in a unique way as $x = u + \lambda v$, where $u \in U$, and since $x \notin U$, we must have $\lambda \neq 0$, and thus, $v = -\lambda^{-1}u + \lambda^{-1}x$. Since $E = U \oplus Kv$, this shows that $E = U + Kx$. Since $x \notin U$, we have $U \cap Kx = 0$, and thus $E = U \oplus Kx$. This argument shows that a hyperplane is a maximal proper subspace H of E.

In Chapter 11, we shall see that hyperplanes are precisely the Kernels of nonnull linear maps $f\colon E \to K$, called linear forms.

6.4 Summary

The main concepts and results of this chapter are listed below:

- *Direct products, sums, direct sums*.
- *Projections*.
- The fundamental equation

  $$\dim(E) = \dim(\operatorname{Ker} f) + \dim(\operatorname{Im} f) = \dim(\operatorname{Ker} f) + \operatorname{rk}(f)$$

  (Proposition 6.16).
- *Grassmann's relation*

  $$\dim(U) + \dim(V) = \dim(U + V) + \dim(U \cap V).$$

- Characterizations of a bijective linear map $f\colon E \to F$.
- *Rank* of a matrix.

6.5 Problems

Problem 6.1. Let V and W be two subspaces of a vector space E. Prove that if $V \cup W$ is a subspace of E, then either $V \subseteq W$ or $W \subseteq V$.

Problem 6.2. Prove that for every vector space E, if $f\colon E \to E$ is an idempotent linear map, i.e., $f \circ f = f$, then we have a direct sum

$$E = \operatorname{Ker} f \oplus \operatorname{Im} f,$$

so that f is the projection onto its image $\operatorname{Im} f$.

Problem 6.3. Let $U_1, \ldots, U_p$ be any $p \geq 2$ subspaces of some vector space E and recall that the linear map

$$a\colon U_1 \times \cdots \times U_p \to E$$

is given by

$$a(u_1, \ldots, u_p) = u_1 + \cdots + u_p,$$

with $u_i \in U_i$ for $i = 1, \ldots, p$.

(1) If we let $Z_i \subseteq U_1 \times \cdots \times U_p$ be given by

$$Z_i = \left\{ \left(u_1, \ldots, u_{i-1}, -\sum_{j=1, j\neq i}^{p} u_j, u_{i+1}, \ldots, u_p \right) \;\middle|\; \sum_{j=1, j\neq i}^{p} u_j \in U_i \cap \left(\sum_{j=1, j\neq i}^{p} U_j \right) \right\},$$

for $i = 1, \ldots, p$, then prove that

$$\operatorname{Ker} a = Z_1 = \cdots = Z_p.$$

In general, for any given i, the condition $U_i \cap \left(\sum_{j=1, j\neq i}^{p} U_j \right) = (0)$ does not necessarily imply that $Z_i = (0)$. Thus, let

$$Z = \left\{ \left(u_1, \ldots, u_{i-1}, u_i, u_{i+1}, \ldots, u_p \right) \;\middle|\; u_i = -\sum_{j=1, j\neq i}^{p} u_j,\ u_i \in U_i \cap \left(\sum_{j=1, j\neq i}^{p} U_j \right),\ 1 \leq i \leq p \right\}.$$

Since $\operatorname{Ker} a = Z_1 = \cdots = Z_p$, we have $Z = \operatorname{Ker} a$. Prove that if

$$U_i \cap \left(\sum_{j=1, j\neq i}^{p} U_j \right) = (0) \quad 1 \leq i \leq p,$$

then $Z = \operatorname{Ker} a = (0)$.

(2) Prove that $U_1 + \cdots + U_p$ is a direct sum iff

$$U_i \cap \left(\sum_{j=1, j\neq i}^{p} U_j \right) = (0) \quad 1 \leq i \leq p.$$