Chapter 3
Algebraic Aspects of Saddle Point Problems

The examples of Chap. 1 clearly showed that several formulations typically lead to linear systems of the general form

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix}, \tag{3.0.1}$$

where A and B are linear differential operators from some functional space to another (which often is its dual space). The general abstract theory for systems of the type (3.0.1) in Hilbert spaces will be given in Chap. 4. As we shall see, it involves from time to time non-trivial results in functional analysis that can be difficult to understand for readers with a weaker mathematical background.

The purpose of this chapter is to present first the basic results of the general abstract theory in the much simpler context of *finite dimensional spaces*, where we can avoid all the subtleties of functional analysis. We shall therefore study systems of the form (3.0.1) where A and B are respectively an $n \times n$ matrix and an $m \times n$ matrix, while $\mathbf{x}$ and $\mathbf{f}$ are $n \times 1$ vectors and $\mathbf{y}$ and $\mathbf{g}$ are $m \times 1$ vectors.

It is clear that the present finite dimensional case will usually be reached after the discretisation of more general systems in abstract Hilbert spaces, so that we cannot be afraid of wasting our time in analysing it in detail. Moreover, many results that will be proved in the next chapter can be seen, formally, as simple extensions of the present algebraic version (although the proofs in the infinite dimensional case are often more tricky).

Hence, in a sense, the present chapter is dedicated to the readers that have a weaker background in mathematics, and in particular in functional analysis. We hope that, for them, a good grasp of the finite dimensional cases will be sufficient to understand *the results* (if not the proofs) that will be discussed in the next chapter.

In the study of linear systems of the type (3.0.1), our first need will be to express in proper form the conditions for their *solvability* in terms of the properties of the matrices A and B. By solvability we mean that, for every right-hand side $\mathbf{f}$ and $\mathbf{g}$,

D. Boffi et al., *Mixed Finite Element Methods and Applications*, Springer Series in Computational Mathematics 44, DOI 10.1007/978-3-642-36519-5_3,

the system (3.0.1) has a unique solution. It is well known that this property holds *if and only if* the $(n+m)\times(n+m)$ matrix

$$M = \begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \tag{3.0.2}$$

is *non-singular*, i.e. if and only if its determinant is different from zero. We shall therefore give necessary and sufficient conditions on the sub-matrices A and B for producing a non-singular M.

In order to have a good numerical method, however, solvability is not enough. An additional property that we also require is *stability*. Let us see in more detail what we mean by that. For a solvable finite-dimensional linear system, we always have continuous dependence of the solution upon the data. This means that there exists a constant c such that for every set of vectors $\mathbf{x}, \mathbf{y}, \mathbf{f}, \mathbf{g}$ satisfying (3.0.1) we have

$$\|\mathbf{x}\| + \|\mathbf{y}\| \le c(\|\mathbf{f}\| + \|\mathbf{g}\|). \tag{3.0.3}$$

In turn, this property implies solvability. Indeed, if we assume that (3.0.3) holds for every set of vectors $\mathbf{x}, \mathbf{y}, \mathbf{f}, \mathbf{g}$ satisfying (3.0.1), then, whenever $\mathbf{f}$ and $\mathbf{g}$ are both zero, $\mathbf{x}$ and $\mathbf{y}$ must also be equal to zero. This is another way of saying that the homogeneous system has only the trivial solution, which implies that the determinant of the matrix (3.0.2) is different from zero, and hence the system is solvable.

However, formula (3.0.3) deserves another very important comment. Actually, we did not specify the norms adopted for $\mathbf{x}, \mathbf{y}, \mathbf{f}, \mathbf{g}$. We had the right to do so since, in finite dimension, all norms are equivalent. Hence, the change of one norm with another would only result in a change of the numerical value of the constant c, but it would not change the basic fact that such a constant exists. However, in dealing with linear systems resulting from the discretisation of a partial differential equation, we face a slightly different situation. In fact, if we want to analyse the behaviour of a given *method* when the mesh-size becomes smaller and smaller, we must ideally consider *a sequence* of linear systems whose dimension increases and approaches infinity when the mesh-size tends to zero. As it is well known (and it can also be easily verified), the constants involved in the equivalence of different norms depend on the dimension of the space. For instance, in $\mathbb{R}^n$, the two norms

$$\|\mathbf{x}\|_1 := \sum_{i=1}^{n} |x_i| \qquad \text{and} \qquad \|\mathbf{x}\|_2 := \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2} \tag{3.0.4}$$

are indeed equivalent, in the sense that there exist two positive constants c_1 and c_2 such that

$$c_1 \|\mathbf{x}\|_2 \le \|\mathbf{x}\|_1 \le c_2 \|\mathbf{x}\|_2 \tag{3.0.5}$$

for all $\mathbf{x}$ in $\mathbb{R}^n$. However, it can be rather easily checked that the *best* constants one can choose in (3.0.5) are

$$\|\mathbf{x}\|_2 \le \|\mathbf{x}\|_1 \le \sqrt{n}\,\|\mathbf{x}\|_2; \tag{3.0.6}$$

in particular, the first inequality becomes an equality, for instance, when x_1 is equal to 1 and all the other x_i's are zero, while the second inequality becomes an equality, for instance, when all the x_i are equal to 1.

When considering a discretisation method for a boundary value problem, which gives rise to a sequence of algebraic problems with increasing dimension, we have to take into account that n becomes unbounded. It is then most natural to ask the following question. *Is it possible, for a given choice of the sequence of matrices A and B and norms $\|\mathbf{x}\|$, $\|\mathbf{y}\|$, $\|\mathbf{f}\|$, and $\|\mathbf{g}\|$, to find a constant c independent of the mesh-size that makes (3.0.3) hold true for all mesh-sizes?* If this is true (with some additional relations between the matrices and the norms that will be made precise later on, in Sect. 3.4), we consider *the method* to be *stable*. We point out that, in this context, stability is a property of methods and not a property of linear systems.

However, in this preliminary chapter, we will not deal directly with boundary value problems and related methods. We will consider generic sequences of matrices A and B with the corresponding sequences of norms; then we will require A and B to satisfy suitable properties expressed in terms of constants (say, α and β) that will be assumed to be *the same constants for all the sequence*; finally, we will show that this gives rise to a constant c in (3.0.3) that depends only on α and β, and is therefore valid for all the linear systems of the sequence.

To read the present chapter, only a rudimentary background in linear algebra will be needed, but we hope that the basic ideas will still come out clear enough. The chapter is therefore mostly recommended for readers with a weak mathematical background. Some proofs, in particular in the last two sections, although simple, are somewhat lengthy. The readers with less mathematical inclination might skip them. On the other hand, the chapter could be considered as useless for people with a stronger mathematical formation. Indeed, essentially everything will be repeated, in the more general context of Hilbert spaces, in the next chapter. However, the examples and the counterexamples of the last two Sections might still have some interest, and at least a glance at them is recommended for everybody.

We summarise the outline of the chapter: we first (in Sect. 3.1) recall some elementary facts in linear algebra. The main goal for that is to fix the notation, and to refresh the memory for people with a low mathematical background. Then, in Sect. 3.2 we consider the unique solvability of problems of the type (3.0.1), and we describe necessary and sufficient conditions in terms of properties of matrices A and B. At this level, all norms are considered to be equivalent. Next, in Sect. 3.3 we extend part of the theory to matrices of the type

$$M = \begin{pmatrix} A & B^T \\ B & C \end{pmatrix}, \tag{3.0.7}$$

which is indeed *very generic*. However, we shall play the game that (3.0.7) is, in some sense, *a perturbation of* (3.0.2). Roughly speaking, we shall assume that A and B are such that, for $C = 0$, the matrix (3.0.7) is non-singular, and we look for conditions on C that would preserve this non-singularity. In that section as well, all

norms will be considered as equivalent. In the following Sect. 3.4, we start dealing with *big matrices*, and for this we introduce different norms, together with the problem of *stability* of a sequence of problems for a given choice of the sequences of norms. As announced, our conditions will involve stability constants (to be precise: M_a, M_b, α, and β, that will be defined later on), depending on properties of matrices A and B, respectively. The dependence of the global stability constants upon M_a, M_b, α, and β (and in particular upon α and β) will be tracked down with care, and some simple examples will show the optimality of our results. Some additional results are presented in Sect. 3.5. Finally, the stability conditions for the perturbed problems of the type (3.0.7) will be considered in Sect. 3.6.

3.1 Notation, and Basic Results in Linear Algebra

3.1.1 Basic Definitions

Let r and s be positive integers, and $M : \mathbb{R}^r \to \mathbb{R}^s$ an $s \times r$ real matrix. We denote by M^T the **transposed matrix** of M, given by

$$M^T_{i,j} = M_{j,i} \qquad i = 1, \dots, r, \;\; j = 1, \dots, s. \tag{3.1.1}$$

It is clear that M^T is an $r \times s$ matrix, and therefore $M^T : \mathbb{R}^s \to \mathbb{R}^r$. It is also immediate to check that

$$(M^T)^T \equiv M. \tag{3.1.2}$$

If we have two matrices $M : \mathbb{R}^r \to \mathbb{R}^s$ and $N : \mathbb{R}^k \to \mathbb{R}^r$, the **product** MN of the two matrices will be the usual *rows times columns* one, namely

$$(M\,N)_{m,n} = \sum_{i=1}^{r} M_{m,i} N_{i,n} \qquad 1 \le m \le s, \;\; 1 \le n \le k. \tag{3.1.3}$$

Vectors in $\mathbb{R}^n$ will be considered as *columns*, that is as $n \times 1$ matrices. It is elementary to check that, in the above assumptions on N and M, we have

$$(M\,N)^T = N^T M^T \tag{3.1.4}$$

and (since the transposed of a 1×1 matrix is the matrix itself)

$$\mathbf{y}^T M \mathbf{x} \equiv \mathbf{x}^T M^T \mathbf{y} \qquad \forall \mathbf{x} \in \mathbb{R}^r, \; \forall \mathbf{y} \in \mathbb{R}^s. \tag{3.1.5}$$

Throughout this section, which is very elementary, we shall denote by $\mathbf{0}_r$ and $\mathbf{0}_s$ the zero vectors in $\mathbb{R}^r$ and in $\mathbb{R}^s$ respectively. This notation will be abandoned in the sequel, with only a few exceptions. Throughout the first three sections of this

chapter, unless it is otherwise explicitly specified, the **norm** in $\mathbb{R}^r$, for every integer $r \geq 1$, will be the usual *Euclidean norm* defined by

$$\|\mathbf{x}\|^2 := \sum_{i=1}^{r} x_i^2 \equiv \mathbf{x}^T \mathbf{x}. \tag{3.1.6}$$

We define the **kernel** and the **Range** (or **image**) of M and M^T as follows:

$$\begin{aligned}
&(i)\ \mathrm{Ker}M := \{\mathbf{x} \in \mathbb{R}^r \text{ such that } M\mathbf{x} = \mathbf{0}_s\},\\
&(ii)\ \mathrm{Ker}M^T := \{\mathbf{y} \in \mathbb{R}^s \text{ such that } M^T\mathbf{y} = \mathbf{0}_r\},\\
&(iii)\ \mathrm{Im}M := \{\mathbf{y} \in \mathbb{R}^s \text{ such that } M\mathbf{x} = \mathbf{y} \text{ for some } \mathbf{x} \in \mathbb{R}^r\},\\
&(iv)\ \mathrm{Im}M^T := \{\mathbf{x} \in \mathbb{R}^r \text{ such that } M^T\mathbf{y} = \mathbf{x} \text{ for some } \mathbf{y} \in \mathbb{R}^s\}.
\end{aligned} \tag{3.1.7}$$

3.1.2 Subspaces

As usual, we shall say that Z is a subspace of $\mathbb{R}^r$ if $Z \subset \mathbb{R}^r$ and Z is itself a linear space.

Remark 3.1.1. We recall that a subset Z of a linear space $\mathbb{R}^r$ *is itself a linear space* (and hence is *a subspace*) if, for any two elements $\mathbf{z}_1$ and $\mathbf{z}_2$ in Z, their sum $\mathbf{z}_1 + \mathbf{z}_2$ also belongs to Z and moreover, for any $z \in Z$ and for any real number λ, the product $\lambda\mathbf{z}$ also belongs to Z. □

Remark 3.1.2. According to the previous definition, when, for instance, $r = 3$, any subspace Z of $\mathbb{R}^3$ has to be made of triplets. However, it is quite common to consider, say, $\mathbb{R}^2$ as a subspace of $\mathbb{R}^3$ by considering $(x_1, x_2)^T$ as identified with the triplet $(x_1, x_2, 0)^T$. This, strictly speaking, is not 100% correct. However, on some occasion, it might turn out to be convenient, as we are going to see immediately in the Example 3.1.1 here below. Therefore we will accept it sometimes, while being very careful with what we do. □

If Z is a linear subspace of $\mathbb{R}^r$, the image of the restriction of M to Z will be denoted by $M(Z)$. Hence,

$$M(Z) := \{\mathbf{y} \in \mathbb{R}^s \text{ such that } M\mathbf{z} = \mathbf{y} \text{ for some } \mathbf{z} \in Z\}. \tag{3.1.8}$$

It is clear that $M(\mathbb{R}^r) \equiv \mathrm{Im}M$.

Example 3.1.1. Assume that $r = 5, s = 2$, and consider the operator $M : \mathbb{R}^5 \to \mathbb{R}^2$ defined by

$$M = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}. \tag{3.1.9}$$

If Z is the subspace $Z := \{x_3 = x_4 = x_5 = 0\}$ (that is the space of quintuples of the type $(x_1, x_2, 0, 0, 0)^T$), the temptation to identify the restriction of M to Z with the matrix

$$M_Z = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{3.1.10}$$

is actually quite strong. If, instead of a 2×5 matrix, we had a 2×500 matrix, then the temptation would be much stronger (as well as the economy in using the form (3.1.10)). □

Definition 3.1.1. Let M be an $s \times r$ matrix. Let Z be a subspace of $\mathbb{R}^r$ and S a subspace of $\mathbb{R}^s$. We say that M **restricted to** Z is **injective** if

$$\forall \mathbf{z}^1 \in Z,\ \forall \mathbf{z}^2 \in Z \text{ we have: } \{M\mathbf{z}^1 = M\mathbf{z}^2\} \Rightarrow \{\mathbf{z}^1 = \mathbf{z}^2\}. \tag{3.1.11}$$

We say that M **from** Z **to** S is **surjective** if

$$\forall \mathbf{w} \in S\ \exists\, \mathbf{z} \in Z \text{ such that } M\mathbf{z} = \mathbf{w}. \tag{3.1.12}$$

It is easy to see that, if for instance $Z \equiv \mathbb{R}^r$, then M is injective if and only if $\mathrm{Ker}M = \mathbf{0}_r$. More generally, M restricted to Z is injective if and only if $\mathrm{Ker}M \cap Z = \mathbf{0}_r$. On the other hand, if $S \equiv \mathbb{R}^s$, then M is surjective if and only if $M(Z) = \mathbb{R}^s$. More generally, M is surjective from Z to S if and only if $M(Z) \supseteq S$.

From now on, if we say that an $s \times r$ matrix M is injective or surjective, without specifying the subspaces Z and S, we intend that $\mathrm{Ker}M = \mathbf{0}_r$ or $\mathrm{Im}M = \mathbb{R}^s$, respectively. In other words, by default we intend that $Z = \mathbb{R}^r$ and $S = \mathbb{R}^s$.

The **dimension** of a linear space will be denoted by *dim*. Hence, for instance, $dim(\mathbb{R}^r) = r$, and if Z is a subspace $\subseteq \mathbb{R}^r$, then $dim(Z) \leq r$. Moreover,

$$Z \text{ subspace of } \mathbb{R}^r \text{ and } dim(Z) = r \quad \Rightarrow \quad Z \equiv \mathbb{R}^r. \tag{3.1.13}$$

The **rank** of M is defined as the dimension of its range:

$$rank(M) := dim(\mathrm{Im}M). \tag{3.1.14}$$

Example 3.1.2. In order to become familiar with the notation, it will be convenient to consider an elementary example, made by the family of matrices

$$M_\alpha = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & \alpha \end{pmatrix}, \tag{3.1.15}$$

where α is a real parameter. We have clearly $r = 5$ and $s = 3$. For our present purposes, only the cases $\alpha = 0$ and $\alpha = 1$ will be relevant. The transposed matrix will be

$$M_\alpha^T = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \alpha \end{pmatrix}. \tag{3.1.16}$$

It is immediate to check that for $\alpha = 0$ we have:

$$\begin{aligned} &\mathrm{Ker}M_0 = \{\mathbf{x} \in \mathbb{R}^5 \text{ s. t. } x_3 = x_4 = 0\} \;\; dim(\mathrm{Ker}M_0) = 3, \\ &\mathrm{Ker}M_0^T = \{\mathbf{y} \in \mathbb{R}^3 \text{ s. t. } y_1 = y_2 = 0\} \;\; dim(\mathrm{Ker}M_0^T) = 1, \\ &\mathrm{Im}M_0 = \{\mathbf{y} \in \mathbb{R}^3 \text{ s. t. } y_3 = 0\} \;\; dim(\mathrm{Im}M_0) = 2, \\ &\mathrm{Im}M_0^T = \{\mathbf{x} \in \mathbb{R}^5 \text{ s. t. } x_1 = x_2 = x_5 = 0\} \;\; dim(\mathrm{Im}M_0^T) = 2, \end{aligned} \tag{3.1.17}$$

while for $\alpha = 1$, instead, we have

$$\begin{aligned} &\mathrm{Ker}M_1 = \{\mathbf{x} \in \mathbb{R}^5 \text{ s. t. } x_3 = x_4 = x_5 = 0\} \;\; dim(\mathrm{Ker}M_1) = 2, \\ &\mathrm{Ker}M_1^T = \mathbf{0_3} \;\; dim(\mathrm{Ker}M_1^T) = 0, \\ &\mathrm{Im}M_1 = \mathbb{R}^3 \;\; dim(\mathrm{Im}M_1) = 3, \\ &\mathrm{Im}M_1^T = \{\mathbf{x} \in \mathbb{R}^5 \text{ s. t. } x_1 = x_2 = 0\} \;\; dim(\mathrm{Im}M_1^T) = 3. \end{aligned} \tag{3.1.18}$$

In particular, M_1 is surjective from $\mathbb{R}^5$ to $\mathbb{R}^3$, and M_1^T is injective from $\mathbb{R}^3$ to $\mathbb{R}^5$. The same properties are not true for M_0 and M_0^T respectively. These simple cases might also be useful to check several other properties that will be discussed in the rest of the section. □

3.1.3 Orthogonal Subspaces

For a given linear subspace Z of $\mathbb{R}^r$, we define its **orthogonal subspace** $Z^\perp$ as follows

$$Z^\perp := \{\mathbf{x} \in \mathbb{R}^r \text{ such that } \mathbf{x}^T\mathbf{z} = 0 \; \forall \mathbf{z} \in Z\}. \tag{3.1.19}$$

It is not difficult (and quite intuitive) to check that

$$dim\,(Z^\perp) + dim\,(Z) = r, \tag{3.1.20}$$

and each $\mathbf{x}$ of $\mathbb{R}^r$ can be split *in a unique way* in its two components $\mathbf{x}_Z \in Z$ and $\mathbf{x}_\perp$

$$\mathbf{x} = \mathbf{x}_Z + \mathbf{x}_\perp. \tag{3.1.21}$$

We also have that

$$Z \cap Z^{\perp} = \mathbf{0}_r, \tag{3.1.22}$$

that

$$(Z^{\perp})^{\perp} \equiv Z \tag{3.1.23}$$

and that for two subspaces Z_1 and Z_2

$$Z_1 \subseteq Z_2 \;\; \Rightarrow Z_2^{\perp} \subseteq Z_1^{\perp}. \tag{3.1.24}$$

Example 3.1.3. For instance, with the notation of the previous example, if $Z = \mathrm{Ker}M_\alpha$, we have in $\mathbb{R}^5$: for $\alpha = 0$

$$\begin{aligned} &(\mathrm{Ker}M_0)^{\perp} = \{\mathbf{x} \in \mathbb{R}^5 \text{ such that } x_1 = x_2 = x_5 = 0\} \\ &dim((\mathrm{Ker}M_0)^{\perp}) = 2, \end{aligned} \tag{3.1.25}$$

and for $\alpha = 1$

$$\begin{aligned} &(\mathrm{Ker}M_1)^{\perp} = \{\mathbf{x} \in \mathbb{R}^5 \text{ such that } x_1 = x_2 = 0\} \\ &dim((\mathrm{Ker}M_1)^{\perp}) = 3. \end{aligned} \tag{3.1.26}$$

Always referring to the previous example, we have instead, in $\mathbb{R}^3$: for $\alpha = 0$

$$(\mathrm{Ker}M_0^T)^{\perp} = \{\mathbf{y} \in \mathbb{R}^3 \text{ such that } y_3 = 0\} \;\; dim((\mathrm{Ker}M_0^T)^{\perp}) = 2, \tag{3.1.27}$$

and for $\alpha = 1$

$$(\mathrm{Ker}M_1^T)^{\perp} = \{ \text{ the whole } \mathbb{R}^3\} \;\; dim((\mathrm{Ker}M_1^T)^{\perp}) = 3. \tag{3.1.28}$$

□

Remark 3.1.3. Note that the definition of the orthogonal subspace relies on the choice of the whole space. For instance, as we have already seen in Remark 3.1.2, it is quite common to accept that $\mathbb{R}^r \subset \mathbb{R}^{r+1}$ by identifying $(x_1, \ldots, x_r)$ with $(x_1, \ldots, x_r, 0)$. In this case, for $Z \subseteq \mathbb{R}^r$ we could consider Z both to be a subspace of $\mathbb{R}^r$ and a subspace of $\mathbb{R}^{r+1}$. Clearly, its orthogonal in $\mathbb{R}^r$ and its orthogonal in $\mathbb{R}^{r+1}$ will be different. We will try to be careful whenever this type of confusion can occur. □

3.1.4 Orthogonal Projections

The notion of orthogonal projection on a subspace will play an important role in the next Section. We recall it here, briefly.

For a given subspace Z, say, of $\mathbb{R}^r$, we introduce the **orthogonal projection** $\pi_Z\colon \mathbb{R}^r \to Z$ as follows. For a given $\mathbf{x} \in \mathbb{R}^r$, its orthogonal projection $\pi_Z\mathbf{x}$ is the minimiser in Z of the quantity $\|\mathbf{x} - \mathbf{z}\|$. Hence, we have

$$\pi_Z\mathbf{x} \in Z \quad \text{and} \quad \|\mathbf{x} - \pi_Z\mathbf{x}\| \leq \|\mathbf{x} - \mathbf{z}\|, \quad \forall \mathbf{z} \in Z. \tag{3.1.29}$$

An alternative and equivalent way of writing (3.1.29) is

$$\pi_Z\mathbf{x} := \arg\min_{\mathbf{z}\in Z} \|\mathbf{z} - \mathbf{x}\|. \tag{3.1.30}$$

It is easy to see that such a minimiser exists, is unique and is the unique solution of

$$\pi_Z\mathbf{x} \in Z \qquad \text{and} \qquad \mathbf{z}^T\pi_Z\mathbf{x} = \mathbf{z}^T\mathbf{x}, \qquad \forall \mathbf{z} \in Z. \tag{3.1.31}$$

An obvious consequence of (3.1.31) is

$$\{\mathbf{x} \in Z^{\perp}\} \quad \Leftrightarrow \quad \{\pi_Z\mathbf{x} = \mathbf{0}_r\}. \tag{3.1.32}$$

Example 3.1.4. Always referring to the cases of Example 3.1.2, if, for instance, $Z = \mathrm{Ker}M_0$ and $\mathbf{x} = (1, 2, 3, 4, 5)^T$, then $\pi_Z\mathbf{x} = (1, 2, 0, 0, 5)^T$. □

It will also be convenient to associate to a subspace $Z \subseteq \mathbb{R}^r$ the **extension** operator E_Z, defined as the linear operator that to every $z \in Z$ associates the same z, thought as a member of $\mathbb{R}^r$. At first sight, this appears to be **obnoxiously redundant**. However, as we have seen in Remark 3.1.2, it is quite common, for instance, to identify $Z = \mathbb{R}^2$ as the subspace of $\mathbb{R}^3$ made by the triplets $(x_1, x_2, 0)^T$. Note that, if we consider

$$Z := \{(x_1, x_2, 0)^T\}, \tag{3.1.33}$$

then E_Z is just the *identity matrix*. If however we consider

$$Z := \{(x_1, x_2)^T\}, \tag{3.1.34}$$

then the operator E_Z would correspond to the matrix

$$E_Z = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, \tag{3.1.35}$$

and its *transposed operator* would be

$$E_Z^T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \equiv \pi_Z. \tag{3.1.36}$$

Considering now the general case, we note that if we follow a notation of the type of (3.1.34), then the equality

$$E_Z^T \equiv \pi_Z, \tag{3.1.37}$$

in fact, holds for a general Z. Indeed, for every $\mathbf{z} \in Z$, we can consider the element $E_Z\mathbf{z}$ defined as $\mathbf{z} + \mathbf{0}_{Z^\perp}$ and for every $\mathbf{y} \in \mathbb{R}^r$, we can split it into its components on Z and on $Z^\perp$ and write $\mathbf{y} = \mathbf{y}_Z + \mathbf{y}_{Z^\perp}$, getting

$$\mathbf{y}^T E_Z \mathbf{z} = (\mathbf{y}_Z + \mathbf{y}_{Z^\perp})^T(\mathbf{z} + \mathbf{0}_{Z^\perp}) = (\mathbf{z} + \mathbf{0}_{Z^\perp})^T(\mathbf{y}_Z + \mathbf{y}_{Z^\perp}) = \mathbf{z}^T \pi_Z \mathbf{y}. \tag{3.1.38}$$

On the other hand, following a notation of the type (3.1.33), we would have (also in general)

$$E_Z \equiv E_Z^T \equiv \pi_Z \equiv \pi_Z^T. \tag{3.1.39}$$

3.1.5 Basic Results

We start by proving an easy but useful proposition.

Proposition 3.1.1. *Let M be an $s \times r$ matrix. Then, the restriction of M to $(\mathrm{Ker}M)^\perp$ is a one-to-one mapping between $(\mathrm{Ker}M)^\perp$ and $\mathrm{Im}M$.*

Proof. Let us see first that M, restricted to $(\mathrm{Ker}M)^\perp$, is injective: according to the definition (3.1.11), we have to prove that, if $\mathbf{z}^1$ and $\mathbf{z}^2$ belong to $(\mathrm{Ker}M)^\perp$, and $M\mathbf{z}^1 = M\mathbf{z}^2$, then we must have $\mathbf{z}^1 = \mathbf{z}^2$. Indeed, setting $\tilde{\mathbf{z}} := \mathbf{z}^1 - \mathbf{z}^2$ we have $M\tilde{\mathbf{z}} = 0$ and hence $\tilde{\mathbf{z}} \in \mathrm{Ker}M$. On the other hand, the vector $\tilde{\mathbf{z}}$, as the difference between two elements of $(\mathrm{Ker}M)^\perp$, must also be in $(\mathrm{Ker}M)^\perp$. Hence, $\tilde{\mathbf{z}}$ belongs, at the same time, to $\mathrm{Ker}M$ and to $(\mathrm{Ker}M)^\perp$. Due to (3.1.22), this implies $\tilde{\mathbf{z}} = \mathbf{0}_r$, that means $\mathbf{z}^1 = \mathbf{z}^2$, as we wanted.

Let us now see that M, as a mapping from $(\mathrm{Ker}M)^\perp$ to $\mathrm{Im}M$, is surjective. According to the definition (3.1.12), we have to prove that, for every element $\mathbf{w} \in \mathrm{Im}M$, there exists a $\mathbf{z} \in (\mathrm{Ker}M)^\perp$ such that $M\mathbf{z} = \mathbf{w}$. For this, let $\mathbf{w}$ be an element of $\mathrm{Im}M$. By definition, there exists an $\mathbf{x} \in \mathbb{R}^r$ such that $M\mathbf{x} = \mathbf{w}$. Split this $\mathbf{x}$ into its components along $\mathrm{Ker}M$ and $(\mathrm{Ker}M)^\perp$. Let $\mathbf{x} = \mathbf{x}_K + \mathbf{z}$ be the splitting, with $\mathbf{x}_K \in \mathrm{Ker}M$ and $\mathbf{z} \in (\mathrm{Ker}M)^\perp$. By definition of kernel, $M\mathbf{x}_K = 0$, so that $M\mathbf{z} = M\mathbf{x}_K + M\mathbf{z} = M\mathbf{x} = \mathbf{w}$, as we wanted. □

As immediate consequences, we have now the following properties.

Corollary 3.1.1. *Let M be an $s \times r$ matrix. Then, there exists a lifting L_M, linear from $\mathrm{Im}M$ to $(\mathrm{Ker}M)^\perp$, such that*

$$L_M M\mathbf{x} = \mathbf{x} \quad \forall \mathbf{x} \in (\mathrm{Ker}M)^\perp. \tag{3.1.40}$$

Moreover, there exists a $\mu > 0$ such that

$$\mu\|L_M\mathbf{y}\| \leq \|\mathbf{y}\| \quad \forall \mathbf{y} \in \mathrm{Im}M \text{ and } \quad \mu\|\mathbf{x}\| \leq \|M\mathbf{x}\| \quad \forall \mathbf{x} \in (\mathrm{Ker}M)^{\perp}. \quad (3.1.41)$$

Proof. The existence of L_M satisfying (3.1.40) is obvious. Since all linear operators are continuous in finite dimension, the two inequalities in (3.1.41) (that are actually, in this context, *the same* inequality) are also obvious. □

Remark 3.1.4. We point out that (3.1.40) easily implies (applying M to both sides) that

$$ML_M\mathbf{y} = \mathbf{y} \quad \forall\, \mathbf{y} \in \mathrm{Im}M. \quad (3.1.42)$$

We also point out that exchanging M with M^T in (3.1.41) we have that there exists a $\mu > 0$ such that

$$\mu\|\mathbf{y}\| \leq \|M^T\mathbf{y}\| \quad \forall \mathbf{y} \in (\mathrm{Ker}M^T)^{\perp}. \quad (3.1.43)$$

□

Remark 3.1.5. We used the same letter (μ) to denote the two constants that appear in (3.1.41) and (3.1.43). This was not by chance. Actually, as we shall see in a while (see e.g. Proposition 3.4.3), the two constants coincide, in the sense that if, for a certain value of μ, (3.1.41) is verified, then (3.1.43) is also verified, and vice-versa. □

Corollary 3.1.2. *Let M be an $s \times r$ matrix. Then,*

$$dim((\mathrm{Ker}M)^{\perp}) = dim(\mathrm{Im}M), \quad (3.1.44)$$

$$dim(\mathrm{Im}M) + dim(\mathrm{Ker}M) = r, \quad (3.1.45)$$

$$dim((\mathrm{Ker}M^T)^{\perp}) = dim(\mathrm{Im}M^T), \quad (3.1.46)$$

and:

$$dim(\mathrm{Im}M^T) + dim(\mathrm{Ker}M^T) = s. \quad (3.1.47)$$

Proof. Equation (3.1.44) is an obvious consequence of Proposition 3.1.1. Equation (3.1.45) follows from (3.1.44) using (3.1.20). Then (3.1.46) and (3.1.47) follow by exchanging M and M^T in (3.1.44) and in (3.1.45).

Remark 3.1.6. Note that (3.1.44) is well in agreement with the previous examples: for $\alpha = 0$, we have from (3.1.17) that $dim(\mathrm{Im}M_0) = 2$ and from (3.1.25) that $dim((\mathrm{Ker}M_0)^{\perp}) = 2$, while for $\alpha = 1$, we have from (3.1.18) that $dim(\mathrm{Im}M_1) = 3$ and from (3.1.26) that $dim((\mathrm{Ker}M_1)^{\perp}) = 3$. The agreement of (3.1.46), (3.1.45) and (3.1.47) with the previous examples can be checked in a similar way. We leave it as an exercise. □

Moreover, the following property is very commonly used.

Corollary 3.1.3. *A square $r \times r$ matrix M is injective if and only if it is surjective.*

Proof. The proof follows immediately from (3.1.45). Indeed,

$$\begin{aligned} &M \text{ is injective} \Leftrightarrow \mathrm{Ker}M = \{\mathbf{0}_r\} \Leftrightarrow dim(\mathrm{Ker}M) = 0 \\ &\Leftrightarrow dim(\mathrm{Im}M) = r \Leftrightarrow \mathrm{Im}M = \mathbb{R}^r \Leftrightarrow M \text{ is surjective.} \end{aligned} \tag{3.1.48}$$

Remark 3.1.7. In different words, Corollary 3.1.3 says that, for a square $r \times r$ matrix M, the system

$$M\mathbf{x} = \mathbf{f} \tag{3.1.49}$$

has a unique solution for every right-hand side $\mathbf{f} \in \mathbb{R}^r$ (= surjectivity) if and only if the homogeneous system $M\mathbf{x} = \mathbf{0}_r$ has $\mathbf{x} = \mathbf{0}_r$ as a unique solution, that is if and only if

$$\{M\mathbf{x} = \mathbf{0}_r\} \Rightarrow \{\mathbf{x} = \mathbf{0}_r\} \tag{3.1.50}$$

(= injectivity). It can also be proved (although we are not going to do it here) that both properties are equivalent to say that *the determinant of the matrix M is different from zero*. □

In particular, we recall the following definition.

Definition 3.1.2. A square $r \times r$ matrix M is said to be **non-singular** if it is injective (or, which is the same, if it is surjective, or, which is again the same, if its determinant is different from zero).

It is well known that if M is a non-singular $r \times r$ matrix, then it has an *inverse matrix*, denoted by M^{-1} such that

$$M^{-1}M = M\,M^{-1} = \mathbb{I}_r \tag{3.1.51}$$

where $\mathbb{I}_r$ is the identity matrix in $\mathbb{R}^r$. It is easy to check that whenever M is non-singular, then M^T is also non-singular, and its inverse is given by $(M^T)^{-1} = (M^{-1})^T$. With a (quite common) abuse of notation, we will indicate it simply by M^{-T}, that is

$$M^{-T} = (M^T)^{-1} = (M^{-1})^T. \tag{3.1.52}$$

An important property is given by the following proposition.

Proposition 3.1.2. *Let M be an $s \times r$ matrix. Then,*

$$\mathrm{Ker}M^T = (\mathrm{Im}M)^{\perp}. \tag{3.1.53}$$

Proof. We start by proving that $\mathrm{Ker}M^T \subseteq (\mathrm{Im}M)^{\perp}$. Let $\mathbf{y} \in \mathbb{R}^s$ be in $\mathrm{Ker}M^T$ (that is, $M^T\mathbf{y} = \mathbf{0}_r$). We want to prove that $\mathbf{y} \in (\mathrm{Im}M)^{\perp}$, that is

$$\mathbf{y}^T(M\mathbf{x}) = 0 \quad \forall \mathbf{x} \in \mathbb{R}^r. \tag{3.1.54}$$

This, however, is immediate since

$$\mathbf{y}^T(M\mathbf{x}) = \mathbf{x}^T M^T \mathbf{y} = 0. \tag{3.1.55}$$

Now, we prove that $(\mathrm{Im}M)^{\perp} \subseteq \mathrm{Ker}M^T$. Let therefore $\mathbf{z} \in \mathbb{R}^s$ be in $(\mathrm{Im}M)^{\perp}$ (that is $\mathbf{z}^T M\mathbf{x} = 0$ for all $\mathbf{x} \in \mathbb{R}^r$). Then,

$$\mathbf{x}^T(M^T\mathbf{z}) = 0 \quad \forall \mathbf{x} \in \mathbb{R}^r, \tag{3.1.56}$$

implying that $M^T\mathbf{z} = \mathbf{0}_r$, that is, $\mathbf{z} \in \mathrm{Ker}M^T$.

□

We then have the following theorem.

Theorem 3.1.1. *Let M be an $s \times r$ matrix. Then:*

$$\mathrm{Ker}M^T = (\mathrm{Im}M)^{\perp}, \tag{3.1.57}$$

$$\mathrm{Im}M = (\mathrm{Ker}M^T)^{\perp}, \tag{3.1.58}$$

$$\mathrm{Ker}M = (\mathrm{Im}M^T)^{\perp}, \tag{3.1.59}$$

$$\mathrm{Im}M^T = (\mathrm{Ker}M)^{\perp}. \tag{3.1.60}$$

Proof. Property (3.1.57) has already been seen in (3.1.53). Property (3.1.58) follows from (3.1.53) and (3.1.23). Properties (3.1.59) and (3.1.60) then follow exchanging M and M^T. □

We note that from Theorem 3.1.1 we can easily deduce some useful properties:

$$\{\mathrm{Im}M \equiv \mathbb{R}^s\} \Leftrightarrow \{\mathrm{Ker}M^T = \mathbf{0}_s\}, \quad \mathrm{Im}\{M^T \equiv \mathbb{R}^r\} \Leftrightarrow \{\mathrm{Ker}M = \mathbf{0}_r\}. \tag{3.1.61}$$

All the above properties can also be easily checked on the example of matrices M_α in (3.1.15) and their transposed.

Remark 3.1.8. In spite of its immediate proof, Theorem 3.1.1 can be considered as the finite dimensional version of a very important theorem of functional analysis (that we shall see in the next chapter) which goes under the name of the *Banach Closed Range Theorem*. □

Collecting the results of Proposition 3.1.1, of Corollary 3.1.40 and of Theorem 3.1.1, we now have immediately the following result.

Corollary 3.1.4. *Let M be an $s \times r$ matrix. Then, setting $K := KerM$ and $H := KerM^T$, we have:*

$$M \text{ is one-to-one from } K^{\perp} \text{ to } ImM \equiv H^{\perp}, \tag{3.1.62}$$

$$M^T \text{ is one-to-one from } H^{\perp} \text{ to } ImM^T \equiv K^{\perp}, \tag{3.1.63}$$

$$\exists L_M : H^{\perp} \to K^{\perp} \text{ such that } L_M(M\mathbf{x}) = \mathbf{x} \quad \forall\, \mathbf{x} \in K^{\perp}, \tag{3.1.64}$$

$$\exists L_{M^T} : K^{\perp} \to H^{\perp} \text{ such that } L_{M^T}(M^T\mathbf{y}) = \mathbf{y} \quad \forall\, \mathbf{y} \in H^{\perp}, \tag{3.1.65}$$

$$(L_M)^T = L_{M^T}. \tag{3.1.66}$$

Example 3.1.5. Assume that the matrix M has the following form

$$M = \begin{pmatrix} \mu_1 & 0 & \cdot & 0 & 0 & 0 & 0 & 0 \\ 0 & \mu_2 & \cdot & 0 & 0 & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \mu_k & 0 & 0 & 0 & 0 \\ 0 & 0 & \cdot & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \cdot & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \tag{3.1.67}$$

where k is the dimension of $K^{\perp} \equiv (\mathrm{Ker} M)^{\perp}$, which, due to (3.1.44) and (3.1.58) coincides with the dimension of $H^{\perp} \equiv (\mathrm{Ker} M^T)^{\perp}$. Here we have $r = k + 4$ and $s = k + 2$. We obviously have

$$M^T = \begin{pmatrix} \mu_1 & 0 & \cdot & 0 & 0 & 0 \\ 0 & \mu_2 & \cdot & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \mu_k & 0 & 0 \\ 0 & 0 & \cdot & 0 & 0 & 0 \\ 0 & 0 & \cdot & 0 & 0 & 0 \\ 0 & 0 & \cdot & 0 & 0 & 0 \\ 0 & 0 & \cdot & 0 & 0 & 0 \end{pmatrix}, \tag{3.1.68}$$

and

$$L_M = \begin{pmatrix} \mu_1^{-1} & 0 & \cdot & 0 & 0 & 0 \\ 0 & \mu_2^{-1} & \cdot & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \mu_k^{-1} & 0 & 0 \\ 0 & 0 & \cdot & 0 & 0 & 0 \\ 0 & 0 & \cdot & 0 & 0 & 0 \\ 0 & 0 & \cdot & 0 & 0 & 0 \\ 0 & 0 & \cdot & 0 & 0 & 0 \end{pmatrix} \qquad L_{M^T} = \begin{pmatrix} \mu_1^{-1} & 0 & \cdot & 0 & 0 & 0 & 0 & 0 \\ 0 & \mu_2^{-1} & \cdot & 0 & 0 & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \mu_k^{-1} & 0 & 0 & 0 & 0 \\ 0 & 0 & \cdot & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \cdot & 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{3.1.69}$$

Remark 3.1.9. Although the form of the matrix M in Example 3.1.5 might appear very special, using the so-called *singular-value decomposition* (see e.g. [228]) for every $s \times r$ matrix B, we can always choose an orthonormal basis in $\mathbb{R}^r$ and an orthonormal basis in $\mathbb{R}^s$ that will transform the matrix B in the form (3.1.67). We shall come back to this later on. □

3.1.6 Restrictions of Operators

Assume that we have a subspace $Z \subseteq \mathbb{R}^r$ and an $s \times r$ matrix M. To M we can associate its restriction M_Z to Z defined as

$$M_Z\mathbf{z} = M(E_Z(\mathbf{z})) \quad \forall \mathbf{z} \in Z \qquad \text{that is } M_Z = M\ E_Z, \tag{3.1.70}$$

where E_Z, here and in all this chapter, is the *extension operator* as defined in Sect. 3.1.4.

If now S is a subspace of $\mathbb{R}^s$, we can consider **the operator M_{ZS}, from Z to S**, defined as

$$M_{ZS} = \pi_S\ M\ E_Z. \tag{3.1.71}$$

Clearly, the transposed operator $(M_{ZS})^T$ will be

$$(M_{ZS})^T = \pi_Z\ M^T\ E_S = (M^T)_{SZ}. \tag{3.1.72}$$

Remark 3.1.10. We point out that all the results that we have seen in the previous subsections (and in particular Theorem 3.1.1) still hold for operators like M_{ZS}, but we have to be careful in the interpretation of the orthogonal complement. In particular, we have

$$\mathrm{Ker} M_{SZ}^T = (\mathrm{Im} M_{ZS})^{\perp_S}, \tag{3.1.73}$$

$$\mathrm{Im} M_{ZS} = (\mathrm{Ker} M_{SZ}^T)^{\perp_S}, \tag{3.1.74}$$

$$\mathrm{Ker} M_{ZS} = (\mathrm{Im} M_{SZ}^T)^{\perp_Z}, \tag{3.1.75}$$

$$\mathrm{Im} M_{SZ}^T = (\mathrm{Ker} M_{ZS})^{\perp_Z}, \tag{3.1.76}$$

where, for three spaces $U \subseteq V \subseteq W$, the notation $U^{\perp_V}$ stands (rather obviously) for *the elements of V that are orthogonal to all the elements of U*. □

Example 3.1.6. In the same spirit, considering once more the matrix (3.1.15) (which describes a linear operator from $\mathbb{R}^5$ to $\mathbb{R}^3$), if the subspace $Z \subseteq \mathbb{R}^5$ is defined by $\{x_1 = x_4 = 0\}$, we can indeed either follow the example of (3.1.33) and consider Z as the set of quintuplets $(0, x_2, x_3, 0, x_5)^T$ and describe *the restriction of M to Z* again with the matrix (3.1.15). Otherwise, we can follow the example of (3.1.34), and consider Z as a set of triples $(x_2, x_3, x_5)^T$, and describe it with the matrix

$$M_Z = M\ E_Z = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & \alpha \end{pmatrix}. \tag{3.1.77}$$

So far there is no big difference, and the first option seems actually much cleaner. □

Example 3.1.7. Coming back to the Example 3.1.6 above, if we consider now the space $S \subset \mathbb{R}^3$, defined by $\{y_2 = 0\}$, and if we want to analyse the behaviour of M as an operator from Z to S, the first option would lead us to consider the matrix

$$M_{ZS}^{*} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \alpha \end{pmatrix}, \tag{3.1.78}$$

while the second option would lead to the (simpler) matrix

$$M_{ZS} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & \alpha \end{pmatrix}. \tag{3.1.79}$$

Apparently, the advantage of M_{ZS} over M_{ZS}^{*} is just simplicity. However, if you want to apply the general results of the previous subsection (as e.g. (3.1.58)) to the operator "M from Z to S", you see that the use of the form (3.1.79) makes life much easier: for instance, for $\alpha \neq 0$, the image of M_{ZS} coincides with the whole S while the kernel of $(M_{ZS})^T$ is reduced to $\mathbf{0}$. On the other hand, for $\alpha = 0$, then the image of M_{ZS} will be made by the pairs $(y_1, 0)^T$ and the kernel of $(M_{ZS})^T$ is made by the pairs $(0, y_2)^T$ so that again $\mathrm{Im}M_{ZS}$ is orthogonal to $\mathrm{Ker}(M_{ZS})^T$, and so on. Looking carefully, you can also see everything using the form (3.1.78), but with a bigger effort. □

Remark 3.1.11. We must be careful when discussing the *kernel* and the *image* of operators restricted to subspaces. Indeed, in general, $\mathrm{Ker}M_{ZS}$ will not be a subspace of $\mathrm{Ker}M$, and $\mathrm{Im}M_{ZS}$ will not be a subspace of $\mathrm{Im}M$. Let us see some examples. Assume that we consider operators $\mathbb{R}^2 \to \mathbb{R}^2$. We start with

$$M = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}. \tag{3.1.80}$$

Clearly, the kernel of M is given by $\mathrm{Ker}M = \{(x_1, x_2)^T | \text{ with } x_1 = -x_2\}$ and the image by $\mathrm{Im}M = \{(y_1, y_2)^T | \text{ with } y_1 = y_2\}$. If we take

$$Z := \{(x_1, x_2)^T | \text{ with } x_1 = x_2\} \qquad S := \{(y_1, y_2)^T | \text{ with } y_2 = 0\},$$

then $\mathrm{Ker}M_{ZS} = \{(0,0)^T\}$ and $\mathrm{Im}M_{ZS} := \{(y_1, y_2)^T | \text{ with } y_2 = 0\}$ so that $\mathrm{Ker}M_{ZS} \subseteq \mathrm{Ker}M$ but $\mathrm{Im}M_{ZS} \not\subseteq \mathrm{Im}M$. If we take instead

$$M = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}, \tag{3.1.81}$$

then $\mathrm{Ker}M = \{(0,0)^T\}$ and $\mathrm{Im}M := \mathbb{R}^2$. Choosing Z and S as before, we have now $\mathrm{Ker}M_{ZS} = \{(x_1, x_2)^T | \text{ with } x_1 = x_2\}$ and $\mathrm{Im}M_{ZS} := \{(0,0)^T\}$ so that now $\mathrm{Im}M_{ZS} \subseteq \mathrm{Im}M$ but $\mathrm{Ker}M_{ZS} \not\subseteq \mathrm{Ker}M$. □

The following result deals with the possible inclusions of kernels and images of M_{ZS} and M and their transposed operators.

Proposition 3.1.3. *Let M be an $s \times r$ matrix, let Z be a subspace of $\mathbb{R}^r$, let S be a subspace of $\mathbb{R}^s$ and let finally $M_{ZS} \equiv \pi_S M E_Z$ be the restriction of M operating from Z to S. Finally, let M^T and M_{SZ}^T be the transposed operators of M and M_{ZS}, respectively. Then, the two following inclusions are equivalent*

$$\mathrm{Ker} M_{ZS} \subseteq \mathrm{Ker} M \tag{3.1.82}$$

$$\mathrm{Im}(\pi_Z M^T) \subseteq \mathrm{Im} M_{SZ}^T. \tag{3.1.83}$$

Moreover, exchanging the operators with their transposed, we obviously also have

$$\mathrm{Ker} M_{SZ}^T \subseteq \mathrm{Ker} M^T \quad \Leftrightarrow \quad \mathrm{Im}(\pi_S M) \subseteq \mathrm{Im} M_{ZS}. \tag{3.1.84}$$

Proof. We start by noting that (3.1.82) is equivalent to

$$\mathrm{Ker} M_{ZS} = Z \cap \mathrm{Ker} M. \tag{3.1.85}$$

On the other hand, from (3.1.74) we have that an element of Z belongs to $\mathrm{Im} M_{SZ}^T$ if and only if it is orthogonal to all $\mathbf{z} \in \mathrm{Ker} M_{ZS}$. Taking into account that the generic element of $\mathrm{Im}(\pi_Z M^T)$ is $\pi_Z M \mathbf{y}$ (with $\mathbf{y}$ generic in $\mathbb{R}^s$), and that obviously (by transposition) $\mathbf{z}^T \pi_Z M^T \mathbf{y} = \mathbf{y}^T M E_Z \mathbf{z}$, we deduce that (3.1.83) is equivalent to

$$\mathbf{y}^T M E_Z \mathbf{z} = 0 \quad \forall \mathbf{y} \in \mathbb{R}^s, \quad \forall \mathbf{z} \in \mathrm{Ker} M_{ZS}, \tag{3.1.86}$$

which in turn is clearly equivalent to (3.1.85). □

Remark 3.1.12. An equivalent way of looking at Proposition 3.1.3 is as follows. Using (3.1.24), we immediately have that (3.1.82) holds if and only if $(\mathrm{Ker} M)^\perp \subseteq (\mathrm{Ker} M_{ZS})^\perp$, where both the orthogonals are taken in $\mathbb{R}^r$. On the other hand, from (3.1.60) we have that $(\mathrm{Ker} M)^\perp = \mathrm{Im} M^T$ while an elementary argument using (3.1.76) gives that

$$(\mathrm{Ker} M_{ZS})^{\perp_{\mathbb{R}^r}} = (\mathrm{Ker} M_{ZS})^{\perp_Z} \cup Z^{\perp_{\mathbb{R}^r}} = \mathrm{Im} M_{SZ}^T \cup Z^\perp. \tag{3.1.87}$$

Hence, (3.1.82) is equivalent to

$$\mathrm{Im} M^T \subseteq \mathrm{Im} M_{SZ}^T \cup Z^\perp, \tag{3.1.88}$$

which is clearly equivalent to (3.1.83). □

Example 3.1.8. In the case of the matrix M of Example 3.1.5, we see that the matrix $M_{K^\perp H^\perp}$ would be

$$M_{K^\perp H^\perp} = \begin{pmatrix} \mu_1 & 0 & \cdot & 0 \\ 0 & \mu_2 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \mu_k \end{pmatrix}, \tag{3.1.89}$$

showing its nice nature as a $k \times k$ non singular matrix. With this notation, $(L_M)_{K^\perp H^\perp}$ would just be the *inverse matrix*

$$L_{M_{K^\perp H^\perp}} = \begin{pmatrix} \mu_1^{-1} & 0 & \cdot & 0 \\ 0 & \mu_2^{-1} & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \mu_k^{-1} \end{pmatrix}. \tag{3.1.90}$$

□

3.2 Existence and Uniqueness of Solutions: The Solvability Problem

We go back to our general form (3.0.1), which we repeat here for the convenience of the reader:

$$A\mathbf{x} + B^T\mathbf{y} = \mathbf{f}, \tag{3.2.1}$$

$$B\mathbf{x} = \mathbf{g}. \tag{3.2.2}$$

We assume that $\mathbf{f}$ and $\mathbf{g}$ are given in $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively (n and m being given integer numbers ≥ 1), and that $\mathbf{x}$ and $\mathbf{y}$ are also sought in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. This implies that A must be a square matrix $n \times n$ and B a rectangular matrix $m \times n$.

An important role will be played by the kernels of the operators B and B^T. Hence, we set

$$K := \mathrm{Ker}B \qquad H := \mathrm{Ker}B^T. \tag{3.2.3}$$

An easy consequence of Theorem 3.1.1 that will be used quite often in the sequel is: *for all* $\mathbf{x} \in \mathbb{R}^n$ *and for all* $\mathbf{y} \in \mathbb{R}^m$,

$$\mathbf{x} \in \mathrm{Ker}B \quad \Rightarrow \quad \mathbf{x}^T B^T \mathbf{y} \equiv \mathbf{y}^T B\mathbf{x} = 0, \tag{3.2.4}$$

or equivalently, for $K = \mathrm{Ker}B$,

$$\pi_K B^T \mathbf{y} = 0 \qquad \forall\, \mathbf{y} \in \mathbb{R}^m. \tag{3.2.5}$$

3.2.1 A Preliminary Discussion

Our present aim is to give conditions on A and B in order that (3.2.1) and (3.2.2) have a unique solution.

Let us discuss first some heuristic ideas: according to Remark 3.1.7, the global matrix

$$M = \begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \tag{3.2.6}$$

will be non-singular if and only if the corresponding homogeneous system

$$A\mathbf{x} + B^T\mathbf{y} = 0, \tag{3.2.7}$$

$$B\mathbf{x} = 0, \tag{3.2.8}$$

has the pair $\mathbf{x} = 0$ and $\mathbf{y} = 0$ as a unique solution. Hence, we start our discussion assuming that $\mathbf{f}$ and $\mathbf{g}$ are both zero. What do we know about $\mathbf{x}$? From the second equation (3.2.8), we see that

$$\mathbf{x} \in K = \mathrm{Ker} B. \tag{3.2.9}$$

Moreover, we can take the projection π_K of the first equation (3.2.7). We note that, using (3.2.5), we have $\pi_K B^T\mathbf{y} = 0$ so that the projection of the first equation onto the kernel K is

$$\pi_K A\mathbf{x} = 0. \tag{3.2.10}$$

We wonder whether

$$\{\mathbf{x} \in K \text{ and } \pi_K A\mathbf{x} = 0\} \Rightarrow \{\mathbf{x} = 0\}\ . \tag{3.2.11}$$

Actually, it depends on the matrix A and on K. Either it does or it doesn't. For the moment, we just set, with the notation of (3.1.71),

$$A_{KK} := \pi_K A E_K. \tag{3.2.12}$$

Coming back to the question (3.2.11), let us analyse the two cases.

- If the answer to (3.2.11) is *no*, then we *surely lose* (meaning that the matrix will indeed be singular). Why do we say that? This is subtle, but not really difficult. We claim that *if the answer is* no*, then there exists a non-zero solution of the homogeneous system*. Let us see why. If the answer to (3.2.11) is *no*, it means that there exists an $\mathbf{x}^* \neq 0$ such that both (3.2.9) and (3.2.10) hold. Now, using (3.1.32), we note that (3.2.10) implies

$$A\mathbf{x}^* \in K^\perp. \tag{3.2.13}$$

Moreover, we remember that $K = \mathrm{Ker}B$ and that, from (3.1.60), $(\mathrm{Ker}B)^{\perp} = \mathrm{Im}B^T$. Hence, from (3.2.13), we have $A\mathbf{x}^* \in \mathrm{Im}B^T$, and therefore there must exist a $\mathbf{y}^*$ such that

$$B^T\mathbf{y}^* = A\mathbf{x}^*. \tag{3.2.14}$$

This is why we lose: indeed, the pair $(\mathbf{x}^*, -\mathbf{y}^*)$ satisfies *both* equations $A\mathbf{x}^* + B^T(-\mathbf{y}^*) = 0$ and $B\mathbf{x}^* = 0$, and we have a *non-zero* solution of the homogeneous problem (3.2.7) and (3.2.8), since at least $\mathbf{x}^* \neq 0$.

- If instead the answer to (3.2.11) is *yes*, we can conclude that, for every pair $(\mathbf{x}, \mathbf{y})$ solving the homogeneous system (3.2.7) and (3.2.8), we must have $\mathbf{x} = 0$. However, we still have to work on $\mathbf{y}$. Once we know that $\mathbf{x} = 0$, the first equation (3.2.7) becomes

$$B^T\mathbf{y} = 0, \tag{3.2.15}$$

and we face a second *dilemma*: do we have

$$\{B^T\mathbf{y} = 0\} \;\Rightarrow\; \{\mathbf{y} = 0\}\ ? \tag{3.2.16}$$

Clearly, the answer depends on the matrix B^T. If it is injective, the answer to (3.2.16) will be *yes*, otherwise it will be *no*. Here, however, the situation is simpler: indeed, if the answer is *no*, it means that there exists a $\hat{\mathbf{y}} \neq 0$ such that $B^T\hat{\mathbf{y}} = 0$, and we lose again because the pair $(\mathbf{0}_n, \hat{\mathbf{y}})$ will clearly be a non-zero solution to the homogeneous system (3.2.7) and (3.2.8). If instead the answer to (3.2.16) is also *yes*, then we can conclude: every solution $(\mathbf{x}, \mathbf{y})$ of the homogeneous system (3.2.7) and (3.2.8) will necessarily be zero, and the matrix M will be non-singular.

In conclusion to our heuristic analysis, it seems that, in order to have a non-singular global matrix M, we need a "yes" for both questions (3.2.11) and (3.2.16). This indeed is what we are going to *prove*, in a more precise way, in the next subsection.

3.2.2 The Necessary and Sufficient Condition

We start with the basic result that provides necessary and sufficient conditions for solvability.

Theorem 3.2.1. *Let n and m be two integers ≥ 1. Let A and B be an $n \times n$ matrix and an $m \times n$ matrix, respectively. Let K be the kernel of B as in (3.2.3), and let A_{KK} be defined as in (3.2.12). Then, the matrix*

$$M = \begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \tag{3.2.17}$$

is non-singular if and only if the following two conditions are both satisfied:

$$A_{KK} : K \to K \text{ is surjective (or, equivalently, is injective),} \tag{3.2.18}$$

$$B : \mathbb{R}^n \to \mathbb{R}^m \text{ is surjective (or, equivalently, } B^T \text{ is injective).} \tag{3.2.19}$$

Proof. We start by noting that the equivalence claimed in (3.2.18) has been made clear in Proposition 3.1.2, while the equivalence claimed in (3.2.19) is an easy consequence of (3.1.61). We also note that, in some sense, the theorem has been proved already during the heuristic discussion above. However, here we re-start and give a more detailed proof.

To this aim, assume first that (3.2.17) is non-singular, that is to say that the system (3.2.1) has a unique solution for every right-hand side $(\mathbf{f}, \mathbf{g}) \in \mathbb{R}^n \times \mathbb{R}^m$. In particular, looking at (3.2.2) we see that it must have a solution for every $\mathbf{g} \in \mathbb{R}^m$, and hence $\mathrm{Im} B \equiv \mathbb{R}^m$ and (3.2.19) holds. Moreover, for every $\mathbf{f} = \mathbf{f}_K \in K$ the system

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{f}_K \\ \mathbf{0}_m \end{pmatrix} \tag{3.2.20}$$

must have a solution. For every such solution, we clearly have $B\mathbf{x} = 0$, that is $\mathbf{x} \in K$. We also note that for every $\mathbf{y} \in \mathbb{R}^m$, from (3.2.5) we have that $\pi_K B^T \mathbf{y} = 0$. Hence, taking the projection π_K of the first equation of (3.2.20) yields:

$$\pi_K A\mathbf{x} = \mathbf{f}_K. \tag{3.2.21}$$

In other words, solving (3.2.20), we have that: for every $\mathbf{f} = \mathbf{f}_K \in K$, there exists an $\mathbf{x} \in K$ such that (3.2.21) holds. Hence, A_{KK} is surjective from K to K, and (3.2.18) holds.

Assume, conversely, that (3.2.18) and (3.2.19) hold. We want to show that the matrix (3.2.17) is non-singular. This will follow if we show that the homogeneous system (3.2.7) and (3.2.8) has $\mathbf{x} = 0$, $\mathbf{y} = 0$ as a unique solution. Indeed, from $B\mathbf{x} = 0$, we first get that $\mathbf{x} \in K$. Taking the projection π_K of the first equation (and noting again that $\pi_K B^T \mathbf{y} = 0$), we have then $\pi_K A\mathbf{x} = 0$. This, together with $\mathbf{x} \in K$, implies $\mathbf{x} = 0$ thanks to the injectivity in (3.2.18). Finally, the first equation now becomes $B^T \mathbf{y} = 0$, and this gives $\mathbf{y} = 0$ thanks to the injectivity in (3.2.19). □

Remark 3.2.1. It follows easily from (3.2.19), using for instance (3.1.45), that *a necessary condition* for the solvability is $n \geq m$. This was pretty obvious from the very beginning, but it could be a valuable first simple check for users that are truly illiterate from the mathematical point of view. □

Remark 3.2.2. We point out that a necessary and sufficient condition is somehow a delicate mathematical item: all possible necessary and sufficient conditions for a matrix to be non-singular are *mathematically* equivalent to each other, and all equivalent to the obvious *it is non-singular if and only if the determinant is different from zero* or even to the tautology *it is non-singular if and only if it is non-singular*.

It is only the commodity of usage that, in each context, makes a necessary and sufficient condition a useful instrument or a sterile mathematical exercise. In this respect, we may say that *it is not true, in practice*, that all necessary and sufficient conditions are equivalent. Moreover, it often happens that conditions that are *only necessary* or *only sufficient* are more useful, in practice, than the necessary and sufficient ones. This is what we shall discuss in the next subsection. □

Remark 3.2.3. We note that the result of Theorem 3.2.1 could have been obtained in a different, more algebraic way. As the result is particularly important, we report this alternative way as well, in the hope that two different points of view could provide a deeper understanding of the whole result.

For this, together with the kernel K of B, we now consider its orthogonal complement $K^\perp$ in $\mathbb{R}^n$ that we call J. Let n_K be the dimension of K and n_J the dimension of J. From (3.1.20) we have

$$n_K + n_J = n. \tag{3.2.22}$$

We now take a basis $\{\mathbf{x}_1^J, \ldots, \mathbf{x}_{n_J}^J\}$ in J and a basis $\{\mathbf{x}_1^K, \ldots, \mathbf{x}_{n_K}^K\}$ in K. It is clear that

$$\{\mathbf{x}_1^J, \ldots, \mathbf{x}_{n_J}^J, \mathbf{x}_1^K, \ldots, \mathbf{x}_{n_K}^K\} \tag{3.2.23}$$

will be a basis for $\mathbb{R}^n$. *With respect to this basis*, we can re-write the matrices A, B, and B^T as follows:

$$A = \begin{pmatrix} A_{JJ} & A_{JK} \\ A_{KJ} & A_{KK} \end{pmatrix} \qquad B = \begin{pmatrix} B_J & B_K \end{pmatrix} \qquad B^T = \begin{pmatrix} B_J^T \\ B_K^T \end{pmatrix}. \tag{3.2.24}$$

Now, from the definition (3.1.7) of K, we immediately have that $B_K = 0$ (that is the zero $m \times n_K$ matrix) so that $B_K^T = 0$ as well. Splitting $\mathbf{x}$ and $\mathbf{f}$ in their orthogonal components $\mathbf{x}_J$ and $\mathbf{x}_K$, and $\mathbf{f}_J$ and $\mathbf{f}_K$, respectively, we can now write the original system (3.2.1) as follows

$$\begin{pmatrix} A_{JJ} & A_{JK} & B_J^T \\ A_{KJ} & A_{KK} & 0 \\ B_J & 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{x}_J \\ \mathbf{x}_K \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{f}_J \\ \mathbf{f}_K \\ \mathbf{g} \end{pmatrix}. \tag{3.2.25}$$

With a little additional work we can see that B_J is a non-singular square matrix if and only if B is surjective, and the result of Theorem (3.2.1) follows from the block-triangular structure of (3.2.25) since $A_{KK} \equiv \pi_K A$. □

3.2.3 *Sufficient Conditions*

The problem of checking whether (3.2.18) holds or not could be simplified or even avoided in some particular cases, as pointed out in the following corollaries to the

basic Theorem 3.2.1. We recall that, in general, a square $r \times r$ matrix M is said to be **positive semi-definite** if

$$\mathbf{x}^T M\mathbf{x} \geq 0 \qquad \forall \mathbf{x} \in \mathbb{R}^r \tag{3.2.26}$$

and it is said to be **positive definite** if

$$\mathbf{x}^T M\mathbf{x} > 0 \qquad \forall \mathbf{x} \in \mathbb{R}^r \text{ with } \mathbf{x} \neq 0. \tag{3.2.27}$$

More generally, if Z is a subspace of $\mathbb{R}^r$, we say that M is **positive semi-definite on** Z if M_{ZZ} is positive semi-definite, that is

$$\forall \mathbf{x} \in Z \qquad \mathbf{x}^T M_{ZZ}\mathbf{x} \equiv \mathbf{x}^T M\mathbf{x} \geq 0, \tag{3.2.28}$$

and we say that M is **positive definite on** Z if M_{ZZ} is positive definite, that is

$$\forall \mathbf{x} \in Z \text{ with } \mathbf{x} \neq 0 \qquad \mathbf{x}^T M_{ZZ}\mathbf{x} \equiv \mathbf{x}^T M\mathbf{x} > 0. \tag{3.2.29}$$

We observe that a positive definite matrix is always non-singular, since (3.2.27) easily implies (3.1.50). Hence, in particular, if M is positive definite on a subspace Z, then M_{ZZ} will be non-singular $Z \to Z$. It is also obvious that if a matrix M is positive definite (or positive semi definite), then its restriction to every subspace Z will also be positive definite (resp. semi-definite).

From the above discussion, we have the following useful result.

Corollary 3.2.1. *Let A be an $n \times n$ matrix, and B an $m \times n$ matrix. If $B : \mathbb{R}^n \to \mathbb{R}^m$ is surjective and A is positive definite on the kernel K of B, then the matrix M in (3.2.17) is non-singular.*

The proof follows immediately from Theorem 3.2.1. The following corollary has more restrictive assumptions, but its use is even simpler.

Corollary 3.2.2. *Let A be an $n \times n$ positive definite matrix, and B an $m \times n$ matrix. If $B : \mathbb{R}^n \to \mathbb{R}^m$ is surjective then the matrix (3.2.17) is non-singular.*

Again, the proof is immediate. The advantage of Corollary 3.2.2 (when we can use it!) is that there is no need to characterise the kernel K, which, in some cases, can be a non-trivial task.

Among the various sufficient conditions, we could point out that if A_{KK} is an isomorphism from K to K, then the condition $\mathbf{g} \in \mathrm{Im}B$ will be *sufficient* to guarantee the *existence* of a solution for the system (3.2.1). We have in particular the following result.

Proposition 3.2.1. *Let n and m be two integers ≥ 1. Let A and B be an $n \times n$ matrix and an $m \times n$ matrix, respectively. Let K be the kernel of B as in (3.2.3), and let A_{KK} be defined as in (3.2.12). Assume that A_{KK} is an isomorphism from K to K and that $\mathbf{g} \in Im\,B$. Then the system (3.2.1) has at least one solution. Moreover, if $(\mathbf{x}_1, \mathbf{y}_1)$ and $(\mathbf{x}_2, \mathbf{y}_2)$ are two solutions of (3.2.1), then $\mathbf{x}_1 = \mathbf{x}_2$ and $(\mathbf{y}_1 - \mathbf{y}_2) \in H = \mathrm{Ker}B^T$.*

Proof. Indeed, if $\mathbf{g} \in \mathrm{Im}B$ then, by definition, there exists an $\mathbf{x}_g \in \mathbb{R}^n$ such that $B\mathbf{x}_g = \mathbf{g}$. Looking for $\mathbf{x}_0 \in K$, solution of the problem $A_{KK}\mathbf{x}_0 = \pi_K(\mathbf{f} - A\mathbf{x}_g)$, we can set $\mathbf{x} := \mathbf{x}_0 + \mathbf{x}_g$ and note that, projecting the first equation on K, we have $\pi_K(\mathbf{f} - A\mathbf{x}) = \mathbf{0}$, because $\pi_K\mathbf{f} - A_{KK}\mathbf{x}_0 - \pi_K A\mathbf{x}_g = \mathbf{0}$. In other words, $\mathbf{f} - A\mathbf{x} \in K^\perp$ which, thanks to (3.1.60), implies $\mathbf{f} - A\mathbf{x} \in \mathrm{Im}B^T$. Hence, there exists a $\mathbf{y} \in \mathbb{R}^m$ such that $B^T\mathbf{y} = \mathbf{f} - A\mathbf{x}$. It is immediate to check that $(\mathbf{x}, \mathbf{y})$ is a solution of (3.2.1). Assume now that $(\mathbf{x}_1, \mathbf{y}_1)$ and $(\mathbf{x}_2, \mathbf{y}_2)$ are two solutions of (3.2.1), and set $\mathbf{x}^* := \mathbf{x}_1 - \mathbf{x}_2$ and $\mathbf{y}^* := \mathbf{y}_1 - \mathbf{y}_2$. Clearly, $(\mathbf{x}^*, \mathbf{y}^*)$ is a solution of the homogeneous system (that is, (3.2.1) with $\mathbf{f} = \mathbf{0}$ and $\mathbf{g} = \mathbf{0}$). In particular we have, from the second equation, that $\mathbf{x}^* \in K$, and from the projection on K of the first equation we have $A_{KK}\mathbf{x}^* = \mathbf{0}$ and since A_{KK} is an isomorphism we have $\mathbf{x}^* = \mathbf{0}$. This implies $A\mathbf{x}^* = \mathbf{0}$ and, using again the first equation: $B^T\mathbf{y}^* = \mathbf{0}$ (that is $\mathbf{y}^* \in H$). □

Remark 3.2.4. In the framework of Proposition 3.2.1, the solution will never be unique, unless we have $H = \mathbf{0}_m$ (that however brings us back to Theorem 3.2.1). On the other hand, we could change the problem and look for $\mathbf{y}$ in $H^\perp$. This actually is *the* way to recover a well posed problem when B is not surjective. However, it obviously works only when $\mathbf{g} \in \mathrm{Im}B$. A particular case in which this would work systematically is whenever $\mathbf{g} \equiv \mathbf{0}$ (as it is often the case when the second equation expresses some incompressibility condition, or some sort of conservation property). □

3.2.4 Examples

Let us see now some examples and exercises. We start by emphasising that the part of A that *must* be non-singular is actually A_{KK}, and **not** A itself. Take for instance, for $n = 2$ and $m = 1$, the matrices

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \qquad B = \begin{pmatrix} 1 & 0 \end{pmatrix} \qquad B^T = \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \tag{3.2.30}$$

Then, the rank of B is 1 ($= m$), and (3.2.19) holds true. On the other hand we have that $K = \mathrm{Ker}B = \{\mathbf{x} \in \mathbb{R}^2 \text{ such that } x_1 = 0\}$. Hence, in this case, the *new basis* (3.2.23) coincides with the original one, and the matrices are in the form (3.2.24) already. It is then easy to check that A itself is non-singular, but $A_{KK} = (0)$ and hence (3.2.18) does not hold. Indeed, the whole matrix is

$$M = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \tag{3.2.31}$$

which is clearly singular.

On the other hand, consider the choice

$$A = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \qquad B = \begin{pmatrix} 1 & 0 \end{pmatrix} \qquad B^T = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \tag{3.2.32}$$

where A is singular. Since K is the same as before, the new coordinates (3.2.23) coincide again with the old ones, and we have easily that $A_{KK} = (1)$. This is clearly non-singular, so that (3.2.18) is now satisfied. Indeed, the whole matrix is now non-singular:

$$M = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}. \tag{3.2.33}$$

Along the same lines, referring to Corollary 3.2.2 we notice that it would not be enough to require that A is positive *semi*-definite (that is $\mathbf{x}^T A\mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$). Indeed, for the choice

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \qquad B = \begin{pmatrix} 1 & 0 \end{pmatrix} \qquad B^T = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \tag{3.2.34}$$

we have that A is positive semi-definite, we have that (3.2.19) is verified, but the whole matrix

$$M = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \tag{3.2.35}$$

is clearly singular.

In many cases, however, it is not immediate to see, at first glance, what the matrix A_{KK} is. Consider for instance the case

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \qquad B = \begin{pmatrix} 1 & -1 \end{pmatrix} \qquad B^T = \begin{pmatrix} 1 \\ -1 \end{pmatrix}. \tag{3.2.36}$$

We have in this case

$$K := \{\mathbf{x} \in \mathbb{R}^2 \text{ such that } x_1 - x_2 = 0\}. \tag{3.2.37}$$

Hence, K can be presented as the one-dimensional subset of $\mathbb{R}^2$ made of vectors of the type $(\alpha, \alpha)^T$ with $\alpha \in \mathbb{R}$. In its turn, J can now be presented as the one-dimensional subset of $\mathbb{R}^2$ made of vectors of the type $(\beta, -\beta)^T$ with $\beta \in \mathbb{R}$. In order to reach the form (3.2.24), we now have to express the matrix A in the new basis $\{\mathbf{x}_1^J, \ldots, \mathbf{x}_{n_J}^J, \mathbf{x}_1^K, \ldots, \mathbf{x}_{n_K}^K\}$ that is now simply $\{\mathbf{x}_1^J, \mathbf{x}_1^K\}$ with $\mathbf{x}_1^J = (1, -1)^T$ and $\mathbf{x}_1^K = (1, 1)^T$ (and if we want an orthonormal basis, we can take $\mathbf{x}_1^J = (1/\sqrt{2}, -1/\sqrt{2})^T$ and $\mathbf{x}_1^K = (1/\sqrt{2}, 1/\sqrt{2})^T$). After some classical computations, we can see that, *in this new basis*, the matrix A takes the form

$$\tilde{A} = \frac{1}{2}\begin{pmatrix} a-b-c+d & a+b-c-d \\ a-b+c-d & a+b+c+d \end{pmatrix}. \tag{3.2.38}$$

From (3.2.38) we have that A_{KK} is the 1×1 matrix $(\frac{1}{2}(a+b+c+d))$, which is non-singular if and only if $a+b+c+d \neq 0$.

Indeed, one can check easily (for instance, by computing the determinant) that the condition $a+b+c+d \neq 0$ is necessary and sufficient for the matrix

$$\begin{pmatrix} a & b & 1 \\ c & d & -1 \\ 1 & -1 & 0 \end{pmatrix} \tag{3.2.39}$$

to be non-singular. In cases like this (which are the majority), it would possibly be simpler to deal directly with the restriction of $\pi_K A$ to K, which is A_{KK} in the original variables. This would require to apply the (original) matrix A

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \tag{3.2.40}$$

to the general vector (in the original coordinates) $\mathbf{x}_K = (\alpha, \alpha)^T$ in K, obtaining the vector

$$A\mathbf{x}_K = \begin{pmatrix} \alpha(a+b) \\ \alpha(c+d) \end{pmatrix}. \tag{3.2.41}$$

Then, we have to check whether the component of $A\mathbf{x}_K$ in K (that is $\pi_K A\mathbf{x}$) is different from zero. As K is one-dimensional, this amounts to take the scalar product

$$(\mathbf{x}_1^K)^T A\mathbf{x}_K = (1/\sqrt{2}, \quad 1/\sqrt{2})A\mathbf{x}_K = \frac{\alpha}{\sqrt{2}}(a+b+c+d), \tag{3.2.42}$$

and see if it is different from zero when α is different from zero. We clearly obtain again the condition $a+b+c+d \neq 0$.

We point out, however, that if, by chance, a and d are positive and $ad > bc$, then A will be positive definite on the whole $\mathbb{R}^2$, and we can have the solvability directly from Corollary 3.2.2 without any additional work.

3.2.5 *Composite Matrices*

Sometimes, the matrix A itself has a block structure of the type

$$\mathbb{A} = \begin{pmatrix} C & D^T \\ D & 0 \end{pmatrix}. \tag{3.2.43}$$

Then again, one has to be careful and require the non-singularity of $\mathbb{A}$ just on the kernel of B. In some cases, together with an A with the structure (3.2.43), we have a B with the structure $\mathbb{B} = (E \quad 0)$ or $\mathbb{B} = (0 \quad E)$, so that the whole matrix has the block structure

$$M = \begin{pmatrix} C & D^T & E^T \\ D & 0 & 0 \\ E & 0 & 0 \end{pmatrix} \quad \text{or} \quad M = \begin{pmatrix} C & D^T & 0 \\ D & 0 & E^T \\ 0 & E & 0 \end{pmatrix}, \tag{3.2.44}$$

respectively. In these cases, it can be a useful exercise to rewrite conditions (3.2.18) and (3.2.19) in terms of properties of the matrices C, D, and E.

To fix the ideas, let us assume that, in the *first case* of (3.2.44), C is an $r \times r$ matrix, D is an $s \times r$ matrix, and E a $k \times r$ matrix. We also assume that $r \geq s + k$, otherwise, according to Remark 3.2.1, the Matrix M will surely be singular. It is clear that we can directly use Theorem 3.2.1, with

$$\mathbb{A} := C \text{ with } n = r \quad \text{and } \mathbb{B} := \begin{pmatrix} D \\ E \end{pmatrix} \text{ with } m = s + k. \tag{3.2.45}$$

With a minor effort, one can recognise that

$$\mathbb{K} := \mathrm{Ker}\mathbb{B} = \mathrm{Ker}D \cap \mathrm{Ker}E \qquad \mathrm{Im}\mathbb{B} = \begin{pmatrix} \mathrm{Im}D \\ \mathrm{Im}E \end{pmatrix} \tag{3.2.46}$$

and that

$$\begin{aligned} &\left\{\mathrm{Ker}\mathbb{B}^T = \begin{pmatrix} \mathbf{0}_s \\ \mathbf{0}_k \end{pmatrix}\right\} \\ &\Leftrightarrow \left\{\{D^T\mathbf{y} + E^T\mathbf{z} = \mathbf{0}_r\} \Rightarrow \{\mathbf{y} = \mathbf{0}_s \text{ and } \mathbf{z} = \mathbf{0}_k\}\right\} \\ &\Leftrightarrow \left\{\mathrm{Im}D^T \cap \mathrm{Im}E^T = \mathbf{0}_r\right\}. \end{aligned} \tag{3.2.47}$$

Conditions (3.2.18) and (3.2.19), in terms of the matrices C, D, and E, are then

$$\begin{aligned} &\mathrm{Im}D^T \cap \mathrm{Im}E^T = \mathbf{0}_r, \\ &C_{\mathbb{K}\mathbb{K}} \text{ is non-singular } \mathbb{K} \to \mathbb{K} \quad \text{where } \mathbb{K} = \mathrm{Ker}D \cap \mathrm{Ker}E. \end{aligned} \tag{3.2.48}$$

It is not difficult to verify that conditions (3.2.48) are necessary and sufficient for the non-singularity of the whole matrix M.

To deal with the *second case* of (3.2.44), we assume instead that C is an $r \times r$ matrix, D is an $s \times r$ matrix, and E a $k \times s$ matrix. We also assume, this time, that $r + k \geq s \geq k$, otherwise, according to Remark 3.2.1, the Matrix M will surely

be singular. Possibly the easiest way to apply Theorem 3.2.1 consists in performing first an exchange of rows and columns to reach the form

$$\begin{pmatrix} C & 0 & D^T \\ 0 & 0 & E \\ D & E^T & 0 \end{pmatrix}. \tag{3.2.49}$$

Then, we can take $n = r + k$ and $m = s$ with

$$\mathbb{A} = \begin{pmatrix} C & 0 \\ 0 & 0 \end{pmatrix} \qquad \mathbb{B} = \begin{pmatrix} D & E^T \end{pmatrix} \qquad \mathbb{B}^T = \begin{pmatrix} D^T \\ E \end{pmatrix}. \tag{3.2.50}$$

It is now immediate to see that

$$\mathrm{Ker}\mathbb{B}^T = \mathrm{Ker}D^T \cap \mathrm{Ker}E$$

so that condition (3.2.19) (that now becomes: $\mathrm{Ker}\mathbb{B}^T = \mathbf{0}_s = \mathbf{0}_m$) requires in this case that $\mathrm{Ker}D^T \cap \mathrm{Ker}E = \mathbf{0}_s$. Then, we have to look at the kernel of $\mathbb{B}$ and require the non-singularity of $\mathbb{A}$ on it. It is clear that the kernel of $\mathbb{B}$, in this case, is given by

$$\mathbb{K} = \{(\mathbf{x}, \mathbf{z}) \in \mathbb{R}^r \times \mathbb{R}^k \text{ such that } D\mathbf{x} + E^T\mathbf{z} = \mathbf{0}_s\}. \tag{3.2.51}$$

This includes all pairs of the form $(\mathbf{0}_r, \tilde{\mathbf{z}})$, with $\tilde{\mathbf{z}} \in \mathrm{Ker}E^T$. When we apply the matrix $\mathbb{A}$ to one of these vectors, we obviously obtain the zero vector. Hence, if we want the restriction of $\mathbb{A}$ to $\mathbb{K}$ to be non-singular, we must first require that these pairs are reduced to $(\mathbf{0}_r, \mathbf{0}_k)$, that is, we must require first that $\mathrm{Ker}E^T = \mathbf{0}_k$. However, $\mathbb{K}$ might also contain pairs $(\mathbf{x}, \mathbf{z})$ with $\mathbf{x} \neq \mathbf{0}_r$, provided $D\mathbf{x} \in \mathrm{Im}E^T$. This subset of $\mathbb{R}^r$ can be characterised, using also (3.1.60), as

$$\begin{aligned} \tilde{\mathbb{K}} &= \{\mathbf{x} \in \mathbb{R}^r \text{ such that } D\mathbf{x} = E^T\mathbf{z} \text{ for some } \mathbf{z} \in \mathbb{R}^k\} \\ &\equiv \{\mathbf{x} \in \mathbb{R}^r \text{ such that } \tilde{\mathbf{z}}^T D\mathbf{x} = 0 \quad \forall \tilde{\mathbf{z}} \in \mathrm{Ker}E\}. \end{aligned} \tag{3.2.52}$$

Hence, the conditions for the *second case* can be summarised in terms of the matrices C, D and E as:

$$\begin{aligned} &\mathrm{Ker}D^T \cap \mathrm{Ker}E = \mathbf{0}_s, \\ &\mathrm{Ker}E^T = \mathbf{0}_k, \\ &C_{\tilde{\mathbb{K}}\tilde{\mathbb{K}}} \text{ is non-singular } \tilde{\mathbb{K}} \to \tilde{\mathbb{K}} \quad \text{where } \tilde{\mathbb{K}} \text{ is given in (3.2.52).} \end{aligned} \tag{3.2.53}$$

Again, it is not difficult to verify that conditions (3.2.53) are necessary and sufficient for the non-singularity of the whole matrix.

There are obviously other equivalent ways to apply Theorem 3.2.1. For instance, we can, in both cases, consider directly $n = r + s$, $m = k$ and

$$\mathbb{A} = \begin{pmatrix} C & D^T \\ D & 0 \end{pmatrix} \qquad \mathbb{B} = (E \quad 0) \text{ or } \mathbb{B} = (0 \quad E). \tag{3.2.54}$$

As we are dealing with necessary and sufficient conditions, we would find exactly the same conditions as before, possibly with a longer argument.

In a similar way, one could treat the case when the space $\mathbb{R}^m \times \mathbb{R}^n$ is split into a bigger number of subspaces (four, five, etc.). We do not insist too much on these exercises.

3.3 The Solvability Problem for Perturbed Matrices

A different, more interesting variant arises when we consider the case of systems of the type

$$\begin{pmatrix} A & B^T \\ B & -C \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix}, \tag{3.3.1}$$

where again A and B are $n \times n$ and $m \times n$ matrices, respectively, and C is an $m \times m$ matrix. The name of the game here is to see C as a perturbation of the original problem (3.2.1). We shall therefore assume that matrices A and B satisfy (3.2.18) and (3.2.19), plus, possibly, some minor additional requirement, and we look for conditions on C in order to have the unique solvability of (3.3.1).

The minus sign in front of the matrix C is due to the fact that, in what follows, we are going to assume the perturbation, in some sense, to be *negative* (and hence C to be positive), in order to have existence and uniqueness results.

3.3.1 Preliminary Results

A first sufficient condition for solvability is quite obvious.

Proposition 3.3.1. *Assume that A and C are positive definite. Then problem (3.3.1) is uniquely solvable.* □

Indeed, it is easy to check that in this case the matrix

$$\begin{pmatrix} A & B^T \\ -B & C \end{pmatrix} \tag{3.3.2}$$

is positive definite.

Another more or less obvious sufficient condition is given in the following proposition.

Proposition 3.3.2. *Assume that* (3.2.18) *and* (3.2.19) *are satisfied. Then there exists an* $\varepsilon > 0$ *such that, for every* $m \times m$ *matrix* C *satisfying*

$$\|C\mathbf{y}\| \leq \varepsilon \|\mathbf{y}\|, \quad \forall \mathbf{y} \in \mathbb{R}^m, \tag{3.3.3}$$

problem (3.3.1) *is uniquely solvable.* □

The proof is based on the following obvious fact: if the determinant of a matrix is different from zero, and if we perturb the matrix by a small enough quantity, the determinant will still be different from zero. We omit the mathematical details.

In the next subsection, we shall provide a theorem that is more interesting, and more relevant for the applications. In order to prove it, however, we are going to need the following elementary (and classical) lemma, that will also be useful in other occasions.

Lemma 3.3.1. *Assume that* A *is a symmetric* $n \times n$ *matrix satisfying*

$$\mathbf{x}^T A\mathbf{x} \geq 0, \quad \forall \mathbf{x} \in \mathbb{R}^n \tag{3.3.4}$$

(that is: A *is positive semi-definite). Then, for every* $\mathbf{x} \in \mathbb{R}^n$ *and for every* $\mathbf{z} \in \mathbb{R}^n$, *we have*

$$(\mathbf{z}^T A\mathbf{x})^2 \leq (\mathbf{x}^T A\mathbf{x})\,(\mathbf{z}^T A\mathbf{z}), \tag{3.3.5}$$

and consequently, always for every $\mathbf{x} \in \mathbb{R}^n$,

$$\mathbf{x}^T A\mathbf{x} = 0 \quad \Rightarrow \quad A\mathbf{x} = 0. \tag{3.3.6}$$

Proof. Using (3.3.4), we easily have that, for any $\mathbf{z} \in \mathbb{R}^n$ and for any real number λ,

$$(\mathbf{x} + \lambda\mathbf{z})^T A(\mathbf{x} + \lambda\mathbf{z}) \geq 0. \tag{3.3.7}$$

Expanding (3.3.7) in powers of λ and using the symmetry of A, we have

$$\mathbf{x}^T A\mathbf{x} + 2\lambda\mathbf{z}^T A\mathbf{x} + \lambda^2\mathbf{z}^T A\mathbf{z} \geq 0, \tag{3.3.8}$$

implying that the equation (in the unknown λ) $\mathbf{x}^T A\mathbf{x} + 2\lambda\mathbf{z}^T A\mathbf{x} + \lambda^2\mathbf{z}^T A\mathbf{z} = 0$ cannot have distinct real roots, and therefore

$$\Delta \equiv (2\mathbf{z}^T A\mathbf{x})^2 - 4(\mathbf{x}^T A\mathbf{x})\,(\mathbf{z}^T A\mathbf{z}) \leq 0, \tag{3.3.9}$$

which, divided by four, gives exactly (3.3.5). From this we see that $\mathbf{x}^T A\mathbf{x} = 0$ implies that $\mathbf{z}^T A\mathbf{x} = 0$ *for all* $\mathbf{z} \in \mathbb{R}^n$, and therefore $A\mathbf{x} = 0$. This is what is claimed in (3.3.6). □

3.3.2 Main Results

We are now ready to present the main theorem of this section.

Theorem 3.3.1. *Let A be an $n \times n$ matrix, B an $m \times n$ matrix and let C be an $m \times m$ matrix. Assume (as in the basic Theorem 3.2.1) that B^T is injective and A_{KK} is non-singular from K to K, where $K = \mathrm{Ker}B$. Assume further that A and C are positive semi-definite and that, moreover, A is symmetric. Then, problem (3.3.1) is uniquely solvable for every right-hand side* $\mathbf{f}$, $\mathbf{g}$.

Proof. The proof can be easily done by showing that the homogeneous version of (3.3.1) (that is when $\mathbf{f}$ and $\mathbf{g}$ are both equal to zero) has $\mathbf{x} = 0$, $\mathbf{y} = 0$ as the unique solution. For this, let $(\mathbf{x}, \mathbf{y})$ be the solution of the homogeneous system. Taking the scalar product of the first equation of (3.3.1) times $\mathbf{x}$, we get

$$\mathbf{x}^T A\mathbf{x} + \mathbf{x}^T B^T \mathbf{y} = 0, \tag{3.3.10}$$

while, taking the scalar product of the second equation of (3.3.1) times y, we obtain

$$\mathbf{y}^T B\mathbf{x} - \mathbf{y}^T C\mathbf{y} = 0. \tag{3.3.11}$$

Subtracting (3.3.11) from (3.3.10), and using (3.1.5), we therefore have

$$\mathbf{x}^T A\mathbf{x} + \mathbf{y}^T C\mathbf{y} = 0. \tag{3.3.12}$$

Using the fact that A and C are positive semi-definite in (3.3.12), we then have

$$\mathbf{x}^T A\mathbf{x} = \mathbf{y}^T C\mathbf{y} = 0. \tag{3.3.13}$$

We can now use (3.3.13) and Lemma 3.3.1 to deduce that $A\mathbf{x} = 0$. Using this in the first equation, we obtain now $B^T\mathbf{y} = 0$ which, as B^T is injective, implies $\mathbf{y} = 0$. This, in turn, gives $C\mathbf{y} = 0$, so that, from the second equation, $B\mathbf{x} = 0$. Hence, $\mathbf{x}$ belongs to $\mathrm{Ker}B$. Having already $A\mathbf{x} = 0$, we deduce $A_{KK}\mathbf{x} = 0$, and since A_{KK} is non-singular $K \to K$, we conclude that $\mathbf{x}$ is also equal to zero. □

Remark 3.3.1. Looking at the proof of Theorem 3.3.1, we also see that we can trade the *symmetry* assumption on A with the condition that A is *positive definite on the whole* $\mathbb{R}^n$. Indeed, the symmetry was only used in Lemma 3.3.1 to show that $\mathbf{x}^T A\mathbf{x} = 0$ implies $A\mathbf{x} = 0$. If A is supposed to be positive definite, from $\mathbf{x}^T A\mathbf{x} = 0$ we have immediately $\mathbf{x} = 0$ and then $\mathbf{y} = 0$ as before. □

Theorem 3.3.1 has a counterpart, in which the symmetry assumption is shifted from A to C.

Theorem 3.3.2. *Let A be an $n \times n$ matrix, B an $m \times n$ matrix and let C be an $m \times m$ matrix. Assume (as in the basic Theorem 3.2.1) that B^T is injective and A_{KK} is non-singular from K to K, where $K = \mathrm{Ker}B$. Assume further that A and C are*

positive semi-definite and moreover that C is symmetric. Then, problem (3.3.1) *is uniquely solvable for every right-hand side* $\mathbf{f}$, $\mathbf{g}$.

Proof. We proceed exactly as in the proof of Theorem 3.3.1. Let $(\mathbf{x}, \mathbf{y})$ be a solution of the homogeneous system. Taking the scalar products of the first equation times x, the scalar product of the second equation times y, and finally taking the difference, we reach again (3.3.12) and (3.3.13). This time, we apply Lemma 3.3.1 to the matrix C, obtaining $C\mathbf{y} = 0$. Then we can go back to Theorem 3.2.1 and, using (3.2.18) and (3.2.19), we obtain $\mathbf{x} = 0$ and $\mathbf{y} = 0$. □

Remark 3.3.2. The above results could be summarised as follows. Assume that A and B verify the assumptions of the basic Theorem 3.2.1, that is: B is surjective (or, equivalently, B^T is injective) and A_{KK} is non-singular from K to K, where K is the kernel of B. Then, problem (3.3.1) is uniquely solvable under the following assumptions:

- A and C are positive semi-definite and A is symmetric;
- A is positive definite and C is positive semi-definite;
- A and C are positive semi-definite and C is symmetric. □

3.3.3 Examples

In the following Examples, we shall discuss the *necessity* of the conditions that we have used so far. The form (3.3.1) is clearly too general to allow non-trivial necessary and sufficient conditions. We shall therefore discuss the possibility of finding more general, but still easy, sufficient conditions.

In the first example, we shall see that the symmetry assumptions in Theorem 3.3.1 or in Theorem 3.3.2 cannot be easily reduced. Indeed, if we consider the case

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 1 & 0 \end{pmatrix} \quad B = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad C = \begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix}, \tag{3.3.14}$$

we see that A and C are positive semi-definite, B is surjective and A is non-singular when restricted to the Ker$B$ which in this case is $\{\mathbf{x} \in \mathbb{R}^3 \text{ such that } x_2 = x_3 = 0\}$. Hence, all the assumptions of Theorem 3.3.1 are satisfied but the symmetry assumption (since *neither A nor C is symmetric*). It is easy to see that the whole matrix

$$M = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & -1 & -1 \end{pmatrix} \tag{3.3.15}$$

is singular, since the third and fourth rows are equal. Note that A is symmetric when restricted to $\mathrm{Ker}\, B$, but this is not enough.

On the other hand, it is obvious that we cannot give up the assumption that A and C have, in some weak sense, the same sign, because the elementary choice

$$A = (1) \qquad B = (1) \qquad C = (-1) \tag{3.3.16}$$

gives rise to the singular matrix

$$M = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}. \tag{3.3.17}$$

Similarly, we cannot even accept that one of the two matrices, A or C, is indefinite: for instance, the choice

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \qquad B = \begin{pmatrix} 1 & 0 \end{pmatrix} \qquad C = (1) \tag{3.3.18}$$

with C symmetric and positive definite and A symmetric but indefinite, produces the singular matrix

$$M = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & -1 \end{pmatrix}. \tag{3.3.19}$$

Hence, although the conditions discussed in Theorems 3.3.1 and 3.3.2 are clearly only sufficient and by no way necessary, it does not seem easy to write down more convenient ones.

3.4 Stability

We saw at the beginning of this Chapter that *solvability* will not be sufficient to provide a *good method* to discretise partial differential equations, and some *stability* (in a sense to be made precise) is actually needed.

Here, we suppose that we are actually given a *sequence* of problems with increasing dimensions. It is clear that this will be the case when we are going to consider discretisations of a given, say, partial differential equation, with a sequence of finer and finer meshes. Consider therefore for $k = 1, 2, \ldots$ the problems

$$\begin{pmatrix} A_k & B_k^T \\ B_k & 0 \end{pmatrix} \begin{pmatrix} \mathbf{x}_k \\ \mathbf{y}_k \end{pmatrix} = \begin{pmatrix} \mathbf{f}_k \\ \mathbf{g}_k \end{pmatrix}, \tag{3.4.1}$$

where A_k is an $n_k \times n_k$ matrix, B_k an $m_k \times n_k$ matrix, and the dimensions n_k and m_k tend to infinity when k goes to infinity. Roughly speaking, we can imagine that each value of k will correspond to a different decomposition, and when we will say that some constant is *independent of the decomposition*, we will actually mean that it does not depend on the index k in (3.4.1).

We are therefore interested in conditions that ensure not only the unique solvability of each problem (3.4.1), but also a stability estimate of the type (3.0.3):

$$\|\mathbf{x}_k\| + \|\mathbf{y}_k\| \leq c(\|\mathbf{f}_k\| + \|\mathbf{g}_k\|), \tag{3.4.2}$$

where the constant c *does not depend on* k. This requirement is obviously meaningless, unless we specify the norms that we intend to use. As anticipated at the beginning of this chapter, the choice of the norms, in this case, is not irrelevant: although they are all equivalent, the *constants* involved in the equivalence may (and, in general, do) depend on the dimensions, which we are assuming to be going to infinity.

On the other hand, if we want to use these abstract results in order to provide a priori error bounds for some realistic discretisation of a differential problem, we are not totally free in the choice of the norms.

In general, in the finite element context, the norms to be used will be the norms in some functional space, where the differential problem itself is set. Hence, in practice, we are going to have little choice.

For instance (anticipating some ideas from the following chapters), our unknown vector $\mathbf{x}$ could represent the *nodal values* of a piecewise linear continuous function defined on a domain Ω that has been decomposed into triangles T. This means that we have a one-to-one mapping from $\mathbb{R}^n$ to the space $\mathcal{L}_1^1$ of piecewise linear continuous functions on Ω, that associates to a vector $\mathbf{v}$ in $\mathbb{R}^n$ the function $\varphi_{\mathbf{v}}$ such that, at every node N_j of the decomposition ($j = 1, 2 \ldots n$), we have $\varphi_{\mathbf{v}}(N_j) = v_j$. In this case, a very natural choice of norm for $\mathbf{v}$ would be

$$\|\mathbf{v}\|_0 := \Big(\int_\Omega \varphi_{\mathbf{v}}^2 \, d\Omega \Big)^{1/2}, \tag{3.4.3}$$

or, alternatively,

$$\|\mathbf{v}\|_1 := \Big(\int_\Omega |\nabla \varphi_{\mathbf{v}}|^2 \, d\Omega \Big)^{1/2}, \tag{3.4.4}$$

representing, respectively, the L^2-norm and the H_0^1-norm of the corresponding function $\varphi_{\mathbf{v}}$ (if this function vanishes on boundary nodes). At the present level, however, we have no functional spaces yet (nor, for what matters, a differential problem). Hence, we are going to consider norms, or, rather, families of norms, that are defined independently of functional spaces and discretisation schemes. However, having that target in mind, we shall make assumptions that are somehow tailored for it. In the present section, we shall then reconsider several aspects that were discussed in Sect. 3.2 but, this time, introducing norms, and analysing the behaviour of the various constants in dependence of the chosen norms.

For the sake of simplicity, from now on we shall drop the index k unless it will really be necessary, and we will just **remember** that m, n, A, and B depend on k.

Remark 3.4.1. We point out that, as we have already seen, **stability** is not a concept that can be applied to a single discretised problem, but only to a **sequence** of discretised problems, or to a discretisation **method** (that in turn gives rise to sequences of discretised problems). □

Remark 3.4.2. **Important warning** In what follows, we will often consider the *infimum* or the *supremum* of quotients of the type

$$\frac{\ell(\boldsymbol{\xi})}{\|\boldsymbol{\xi}\|} \quad \text{or} \quad \frac{|\ell(\boldsymbol{\xi})|}{\|\boldsymbol{\xi}\|} \tag{3.4.5}$$

where $\ell(\boldsymbol{\xi})$ is a real number depending linearly on $\boldsymbol{\xi}$. It is clear that the quotients in (3.4.5) make no sense for $\boldsymbol{\xi} = \mathbf{0}$, so that the value $\boldsymbol{\xi} = \mathbf{0}$ should be discarded when taking the *infimum* or the *supremum*. On the other hand, due to the linearity of ℓ, it is clear that for every $\boldsymbol{\xi}_0 \neq \mathbf{0}$ the quotients in (3.4.5) take the same value over the ray $\boldsymbol{\xi} = \kappa \boldsymbol{\xi}_0$ when κ ranges over the positive real numbers. Hence, the limit of the quotients (3.4.5) for $\boldsymbol{\xi} \to \mathbf{0}$, in general, will not exist (we would have a different limit on every ray coming out of the origin), but the meaning of, say,

$$\sup_{\boldsymbol{\xi}} \frac{\ell(\boldsymbol{\xi})}{\|\boldsymbol{\xi}\|} \tag{3.4.6}$$

will not be "seriously ambiguous". Hence, for the sake of brevity, we shall write in these cases

$$\sup_{\boldsymbol{\xi}} \frac{\ell(\boldsymbol{\xi})}{\|\boldsymbol{\xi}\|} \quad \text{instead of} \quad \sup_{\boldsymbol{\xi} \neq \mathbf{0}} \frac{\ell(\boldsymbol{\xi})}{\|\boldsymbol{\xi}\|}. \tag{3.4.7}$$

□

3.4.1 Assumptions on the Norms

We denote by $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{F}$, $\mathbf{G}$, respectively, the spaces of vectors $\mathbf{x}$, $\mathbf{y}$, $\mathbf{f}$, $\mathbf{g}$. Hence, we have

$$\mathbf{X} = \mathbb{R}^n, \quad \mathbf{Y} = \mathbb{R}^m, \quad \mathbf{F} = \mathbb{R}^n, \quad \mathbf{G} = \mathbb{R}^m. \tag{3.4.8}$$

Then, we assume that:

1. The spaces $\mathbf{X}$ and $\mathbf{Y}$ are equipped with norms $\|\cdot\|_X$ and $\|\cdot\|_Y$. For the sake of simplicity, we will assume that there exist two symmetric and positive definite matrices S_X (an $n \times n$ matrix) and S_Y (an $m \times m$ matrix) such that

$$\begin{aligned} \|\mathbf{x}\|_X^2 &= (S_X \mathbf{x})^T (S_X \mathbf{x}) \equiv \mathbf{x}^T S_X^T S_X \mathbf{x} \quad \forall\, \mathbf{x} \in \mathbf{X}, \\ \|\mathbf{y}\|_Y^2 &= (S_Y \mathbf{y})^T (S_Y \mathbf{y}) \equiv \mathbf{y}^T S_Y^T S_Y \mathbf{y} \quad \forall\, \mathbf{y} \in \mathbf{Y}. \end{aligned} \tag{3.4.9}$$

2. the spaces $\mathbf{F}$ and $\mathbf{G}$ are equipped with norms $\|\cdot\|_F$ and $\|\cdot\|_G$ defined as the **dual norms** of $\|\cdot\|_X$ and $\|\cdot\|_Y$, i.e.

$$\|\mathbf{f}\|_F := \sup_{\mathbf{x}\in\mathbf{X}} \frac{\mathbf{x}^T\mathbf{f}}{\|\mathbf{x}\|_X} \qquad \text{and} \qquad \|\mathbf{g}\|_G := \sup_{\mathbf{y}\in\mathbf{Y}} \frac{\mathbf{y}^T\mathbf{g}}{\|\mathbf{y}\|_Y}. \tag{3.4.10}$$

It is not difficult to check that

$$\begin{aligned} \|\mathbf{f}\|_F^2 &= (S_X^{-1}\,\mathbf{f})^T(S_X^{-1}\,\mathbf{f}) \equiv \mathbf{f}^T S_X^{-T} S_X^{-1}\mathbf{f} \;\; \forall\, \mathbf{f}\in\mathbf{F}, \\ \|\mathbf{g}\|_G^2 &= (S_Y^{-1}\,\mathbf{g})^T(S_Y^{-1}\,\mathbf{g}) \equiv \mathbf{g}^T S_Y^{-T} S_Y^{-1}\mathbf{g} \;\; \forall\, \mathbf{g}\in\mathbf{G}. \end{aligned} \tag{3.4.11}$$

3. Given the norms in $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{F}$ and $\mathbf{G}$, we can define the **induced norms** of the matrices A and B as follows

$$\|A\| := \sup_{\mathbf{x}\in\mathbf{X}} \frac{\|A\mathbf{x}\|_{\mathbf{F}}}{\|\mathbf{x}\|_{\mathbf{X}}} \qquad \|B\| := \sup_{\mathbf{x}\in\mathbf{X}} \frac{\|B\mathbf{x}\|_{\mathbf{G}}}{\|\mathbf{x}\|_{\mathbf{X}}}. \tag{3.4.12}$$

4. The norms of the transposed matrices A^T and B^T are obviously defined in the same way as in (3.4.12). Moreover, we have the following immediate result.

Proposition 3.4.1. *In the above assumptions, we have*

$$\|A\| = \|A^T\| \equiv \sup_{\mathbf{x}\in\mathbf{X}} \sup_{\mathbf{z}\in\mathbf{X}} \frac{\mathbf{z}^T A\mathbf{x}}{\|\mathbf{z}\|_{\mathbf{X}}\,\|\mathbf{x}\|_{\mathbf{X}}} \tag{3.4.13}$$

and

$$\|B\| = \|B^T\| \equiv \sup_{\mathbf{x}\in\mathbf{X}} \sup_{\mathbf{y}\in\mathbf{Y}} \frac{\mathbf{y}^T B\mathbf{x}}{\|\mathbf{y}\|_{\mathbf{Y}}\,\|\mathbf{x}\|_{\mathbf{X}}}. \tag{3.4.14}$$

□

The proof follows immediately from (3.1.5), which implies that $\mathbf{z}^T A\mathbf{x} = \mathbf{x}^T A^T\mathbf{z}$ and $\mathbf{y}^T B\mathbf{x} = \mathbf{x}^T B^T\mathbf{y}$.

5. We will assume that *there exist two constants M_a and M_b, independent of the mesh-size, such that*

$$\|A\| = \|A^T\| \le M_a \qquad \|B\| = \|B^T\| \le M_b. \tag{3.4.15}$$

Sometimes, for $\mathbf{K}$ a subspace of $\mathbf{X}$, we will also use the norm

$$\|\mathbf{f}\|_{K'} := \sup_{\mathbf{x}\in\mathbf{K}} \frac{\mathbf{x}^T\mathbf{f}}{\|\mathbf{x}\|_X}. \tag{3.4.16}$$

The following very useful properties are immediate consequences of the above assumptions.

Proposition 3.4.2. *Assume that the properties* (3.4.8)–(3.4.15) *hold true. Then, for every* $\mathbf{x}$ *and* $\mathbf{f}$ *in* $\mathbb{R}^n$ *and for every* $\mathbf{y}$ *and* $\mathbf{g}$ *in* $\mathbb{R}^m$*, we have*

$$\mathbf{x}^T\mathbf{f} \le \|\mathbf{x}\|_X \, \|\mathbf{f}\|_F, \qquad \mathbf{y}^T\mathbf{g} \le \|\mathbf{y}\|_Y \, \|\mathbf{g}\|_G, \tag{3.4.17}$$

$$\|A\mathbf{x}\|_F \le M_a\|\mathbf{x}\|_X, \qquad \|A^T\mathbf{x}\|_F \le M_a\|\mathbf{x}\|_X, \tag{3.4.18}$$

$$\|B\mathbf{x}\|_G \le M_b\|\mathbf{x}\|_X, \qquad \|B^T\mathbf{y}\|_F \le M_b\|\mathbf{y}\|_Y, \tag{3.4.19}$$

$$\mathbf{x}^T A\mathbf{z} \le M_a\|\mathbf{x}\|_X \, \|\mathbf{z}\|_X, \qquad \mathbf{x}^T B^T\mathbf{y} \le M_b\|\mathbf{x}\|_X \, \|\mathbf{y}\|_Y, \tag{3.4.20}$$

and

$$\|\mathbf{f}\|_{K'} \le \|\mathbf{f}\|_F. \tag{3.4.21}$$

If moreover A *is symmetric and positive semi-definite, then* (3.4.18) *can be improved to*

$$\|A\mathbf{x}\|_F \le M_a^{1/2}(\mathbf{x}^T A\mathbf{x})^{1/2}. \tag{3.4.22}$$

Proof. The proof of (3.4.17) is immediate. For instance, the first inequality follows from the fact that for every fixed $\tilde{\mathbf{x}} \in \mathbf{X} \setminus \{0\}$ we obviously have

$$\frac{\tilde{\mathbf{x}}^T\mathbf{f}}{\|\tilde{\mathbf{x}}\|_X} \le \sup_{\mathbf{x}\in\mathbf{X}} \frac{\mathbf{x}^T\mathbf{f}}{\|\mathbf{x}\|_X} \equiv \|\mathbf{f}\|_F, \tag{3.4.23}$$

which multiplied by $\|\tilde{\mathbf{x}}\|_X$ gives $\tilde{\mathbf{x}}^T\mathbf{f} \le \|\tilde{\mathbf{x}}\|_X \, \|\mathbf{f}\|_F$. The second one can be proven in exactly the same way. The proof of (3.4.18) and (3.4.19) is also immediate, as is the proof of (3.4.21) (in the right-hand side we take the supremum over a bigger set). Let us see for instance the proof of (3.4.18) (as the proofs of the other two are identical): using first (3.4.12) and then (3.4.15), we have:

$$\|A\mathbf{x}\|_F \le \|A\| \, \|\mathbf{x}\|_{\mathbf{X}} \le M_a \, \|\mathbf{x}\|_{\mathbf{X}}. \tag{3.4.24}$$

Property (3.4.20) will then follow immediately from (3.4.18) and (3.4.19), and the proof of (3.4.21) is immediate. Finally, for the proof of (3.4.22), we can first use Lemma 3.3.1, which, for every $\mathbf{x}, \mathbf{z} \in \mathbf{X}$, gives

$$|\mathbf{z}^T A\mathbf{x}| \le (\mathbf{z}^T A\mathbf{z})^{1/2}(\mathbf{x}^T A\mathbf{x})^{1/2}. \tag{3.4.25}$$

Then, we use (3.4.10), (3.4.25), and (3.4.20) to get

$$\begin{aligned} \|A\mathbf{x}\|_F = \sup_{\mathbf{z}\in\mathbf{X}} \frac{\mathbf{z}^T A\mathbf{x}}{\|\mathbf{z}\|_X} &\le \sup_{\mathbf{z}\in\mathbf{X}} \frac{(\mathbf{z}^T A\mathbf{z})^{1/2}(\mathbf{x}^T A\mathbf{x})^{1/2}}{\|\mathbf{z}\|_X} \\ &\le \sup_{\mathbf{z}\in\mathbf{X}} \frac{(M_a\|\mathbf{z}\|_X^2)^{1/2}(\mathbf{x}^T A\mathbf{x})^{1/2}}{\|\mathbf{z}\|_X} = M_a^{1/2}(\mathbf{x}^T A\mathbf{x})^{1/2}. \end{aligned} \tag{3.4.26}$$

□

From now on, in this chapter, the Euclidean norm will be denoted by $\|\cdot\|_E$, that is

$$\|\mathbf{z}\|_E^2 := \mathbf{z}^T\mathbf{z}. \tag{3.4.27}$$

The following proposition is an elementary consequence of Corollary 3.1.4.

Proposition 3.4.3. *Let B be an $m \times n$ matrix, and set $K := KerB$ (as usual) and $H := KerB^T$. Then, there exists a positive constant $\tilde{\beta}$ such that*

$$\inf_{\mathbf{y}\in H^\perp}\sup_{\mathbf{x}\in K^\perp}\frac{\mathbf{x}^T B^T\mathbf{y}}{\|\mathbf{x}\|_X\,\|\mathbf{y}\|_Y} = \inf_{\mathbf{x}\in K^\perp}\sup_{\mathbf{y}\in H^\perp}\frac{\mathbf{y}^T B\mathbf{x}}{\|\mathbf{x}\|_X\,\|\mathbf{y}\|_Y} = \tilde{\beta} > 0. \tag{3.4.28}$$

Moreover, with the notation of Proposition 3.1.1, we have exactly

$$\frac{1}{\tilde{\beta}} \equiv \|L_B\| \equiv \|L_{B^T}\|. \tag{3.4.29}$$

Proof. Corollary (3.1.4) implies that B is one-to-one from $K^\perp$ to $H^\perp$ and B^T is one-to-one from $K^\perp$ to $H^\perp$. It is not difficult to see that $\tilde{\beta}$ in (3.4.28) is exactly the value of the norms of L_B and L_{B^T} (that are equal to each other). See also Examples 3.1.5 and 3.1.8.

We are now ready to introduce a *precise definition of stability*.

Definition of stability. *Given a numerical method that produces a sequence of matrices A and B when applied to a given sequence of meshes (with the mesh-size h going to zero), we choose norms $\|\cdot\|_X$ and $\|\cdot\|_Y$ that satisfy the continuity condition (3.4.20), and dual norms $\|\cdot\|_F$ and $\|\cdot\|_G$ according to (3.4.10). Then, we say that* ***the method is stable*** *if there exists a constant c,* ***independent of the mesh size****, such that for all vectors $\mathbf{x}, \mathbf{y}, \mathbf{f}, \mathbf{g}$ satisfying the general system (3.2.1) and (3.2.2), it holds*

$$\|\mathbf{x}\|_X + \|\mathbf{y}\|_Y \le c(\|\mathbf{f}\|_F + \|\mathbf{g}\|_G). \tag{3.4.30}$$

Remark 3.4.3. We recall (as we have also seen in Remark 3.1.7) that for a square matrix, we have unique solvability for every right-hand side if and only if the only solution of the homogeneous system is the zero solution. We note here that (3.4.30) implies that, whenever $\mathbf{f}$ and $\mathbf{g}$ are zero, the only possible solution of (3.2.1) and (3.2.2) is $\mathbf{x} = \mathbf{0}$ and $\mathbf{y} = \mathbf{0}$. Hence, we deduce that (3.4.30) implies the unique solvability of (3.2.1) and (3.2.2). This is the reason why, on several occasions in this section, we will state theorems that ensure the stability (3.4.30) without mentioning explicitly that we have unique solvability for every right-hand side $\mathbf{f}$ and $\mathbf{g}$. □

Having now a precise definition of stability, we can look for suitable assumptions on the matrices A and B that may provide the stability result (3.4.30). In Sect. 3.2,

we started with the basic Theorem 3.2.1, giving the necessary and sufficient conditions for solvability, and then we discussed possible variants with stronger assumptions which gave only sufficient conditions but were easier to deal with. In the present section, we shall follow somehow the opposite path: we shall start with stronger assumptions (allowing an easier proof) and move progressively towards weaker assumptions.

In particular, as we did in the previous sections, we will consider essentially three possible situations, with three different levels of generality. In all three cases, we shall assume an *inf-sup* condition on the matrix B. On the other hand, for the matrix A, we shall consider the three cases: ellipticity on the whole space V, ellipticity only on the kernel K, and a non-singularity condition on A_{KK} of the type of (3.2.18).

Different assumptions on the *symmetry* of A will often affect the dependence of the final stability constants on the *inf-sup* and ellipticity constants.

As a first step, however, we shall discuss the basic assumption to be made on the matrix B (the *inf-sup* condition) that will be used in all the theorems of the Section. In several applications, checking whether the *inf-sup* condition holds or not will be *the main difficulty*. It is therefore necessary to try to have a good understanding of it.

3.4.2 *The* inf-sup *Condition for the Matrix* B*: An Elementary Discussion*

As we are going to see at the end of this subsection, with the definitions and the notation that we introduced in the previous part of this chapter, the so-called *inf-sup* condition can be expressed rather quickly.

However, as it is often one of the main difficulties (to check or to enforce) in many applications, we expect a certain number of readers to pick up the book and start reading this subsection first.

This, clearly, is not recommended, and, frankly speaking, cannot be done. Nevertheless, we tried, in the beginning of this subsection, to be softer than usual, rephrasing many concepts that were seen before, and (if not really restarting from scratch, that would be a total nonsense) to recover some concepts in a more heuristic way.

Let us start from one of its most common formulations.

Inf-sup condition on B. *There exists a positive constant β,* ***independent of the mesh-size*** *h, such that:*

$$\forall\, \mathbf{y} \in \mathbf{Y} \quad \exists\, \mathbf{x} \in \mathbf{X} \setminus \{\mathbf{0}\} \text{ such that } \mathbf{x}^T B^T \mathbf{y} \geq \beta \|\mathbf{x}\|_X \|\mathbf{y}\|_Y . \tag{3.4.31}$$

In order to understand it better, we start by rewriting condition (3.4.31) in different equivalent forms, which will also clarify the reason why it is called *inf-sup condition*.

Since, by assumption, $\mathbf{x}$ is different from zero, condition (3.4.31) can equivalently be written as:

$$\forall\, \mathbf{y} \in \mathbf{Y} \quad \exists\, \mathbf{x} \in \mathbf{X}\backslash\{\mathbf{0}\} \quad \text{such that} \quad \frac{\mathbf{x}^T B^T \mathbf{y}}{\|\mathbf{x}\|_X} \geq \beta \|\mathbf{y}\|_Y . \tag{3.4.32}$$

Given $\mathbf{y} \in \mathbf{Y}$, the most suitable $\mathbf{x} \in \mathbf{X}$ (for making the inequality in (3.4.32) hold) is clearly the one that makes the left-hand side of the inequality as big as possible. Hence, the best we can do is to take the *supremum* of the left-hand side when $\mathbf{x}$ varies among all possible $\mathbf{x} \in \mathbf{X}$ different from $\mathbf{0}$. Hence, recalling also the notation in (3.4.7), we may equivalently require that

$$\forall\, \mathbf{y} \in \mathbf{Y} \quad \sup_{\mathbf{x} \in \mathbf{X}} \frac{\mathbf{x}^T B^T \mathbf{y}}{\|\mathbf{x}\|_X} \geq \beta \|\mathbf{y}\|_Y . \tag{3.4.33}$$

In a sense, we got rid of the task of choosing $\mathbf{x}$. We observe that, making use of the notation of (3.4.10) for dual norms, we immediately have

$$\sup_{\mathbf{x} \in \mathbf{X}} \frac{\mathbf{x}^T B^T \mathbf{y}}{\|\mathbf{x}\|_X} \equiv \|B^T \mathbf{y}\|_F , \tag{3.4.34}$$

so that condition (3.4.33) could easily be rewritten as

$$\forall\, \mathbf{y} \in \mathbf{Y} \qquad \|B^T \mathbf{y}\|_F \geq \beta \|\mathbf{y}\|_Y . \tag{3.4.35}$$

We recall now that the usual condition required in the previous section for the matrix B (see (3.2.19)) was: *B is surjective or, equivalently, B^T is injective*. We also recall that the injectivity (3.1.11) could be written as

$$\{\|B^T \mathbf{y}\| = 0\} \;\Rightarrow\; \{\|\mathbf{y}\| = 0\}. \tag{3.4.36}$$

Looking back at the basic algebraic property (3.1.41) (that, in finite dimension, is always true), with $M = B^T$ we see that here we are first asking that the inequality holds for every $\mathbf{y} \in \mathbf{Y}$ (and not, as in (3.1.41), for every $\mathbf{y} \in (\mathrm{Ker} B^T)^\perp$). Hence, we require that, for every k in our sequence, $(\mathrm{Ker} B^T)^\perp = \{\mathbf{0}\}$. Moreover, we require that the constant μ that appears in (3.1.41) is uniformly bounded from below by a uniform constant β.

We also easily recognise that the *inf-sup* condition, in its equivalent form (3.4.35), easily implies (3.4.36). Hence, it can be seen as *a stronger form* of the plain injectivity (3.4.36), depending on the choice of the norms, and requiring *a uniform bound, β, independent of the mesh-sizes*.

However: why is it called *inf-sup* condition? We note that condition (3.4.35) still depends on $\mathbf{y}$. We also note that it clearly always holds for $\mathbf{y} = \mathbf{0}$, and therefore we can concentrate on the $\mathbf{y}$'s that are different from $\mathbf{0}$; in particular, for $\mathbf{y} \neq 0$, condition (3.4.35) can be also written as

$$\forall\, \mathbf{y} \in \mathbf{Y}\backslash\{\mathbf{0}\} \qquad \frac{\|B^T\mathbf{y}\|_F}{\|\mathbf{y}\|_Y} \geq \beta. \tag{3.4.37}$$

The worst possible $\mathbf{y}$ is therefore the one that makes the left-hand side of (3.4.37) as small as possible. If we want (3.4.37) to hold *for every* $\mathbf{y} \in \mathbf{Y} \setminus \{\mathbf{0}\}$, we might as well consider the worst case, looking directly at the *infimum* of the left-hand side of (3.4.37) among all possible $\mathbf{y}$'s, requiring that

$$\inf_{\mathbf{y}\in\mathbf{Y}} \frac{\|B^T\mathbf{y}\|_F}{\|\mathbf{y}\|_Y} \geq \beta, \tag{3.4.38}$$

(still following the notation (3.4.7)) that is, recalling (3.4.34),

$$\inf_{\mathbf{y}\in\mathbf{Y}} \sup_{\mathbf{x}\in\mathbf{X}} \frac{\mathbf{x}^T B^T\mathbf{y}}{\|\mathbf{x}\|_X\, \|\mathbf{y}\|_Y} \geq \beta, \tag{3.4.39}$$

which is possibly the most used equivalent presentation of the assumption, and which gave it its name. The advantage of formulation (3.4.39) over the original formulation (3.4.31), if any, is that we got rid of the dependence on $\mathbf{y}$ and $\mathbf{x}$. Indeed, condition (3.4.39) is now clearly a condition on *the matrix* B, on *the spaces* $\mathbf{X}$ and $\mathbf{Y}$ (together with their *norms*), as well as on the crucial *constant* β.

Remark 3.4.4. We point out once more that the *inf-sup* condition is *stronger* than the simple injectivity (3.4.36). Considering for simplicity the matrix

$$B_\theta := \begin{pmatrix} 1 & 0 & 0 \\ 0 & \theta & 0 \end{pmatrix} \tag{3.4.40}$$

and taking the Euclidean norm for all the spaces, we easily see that, for $0 < \theta < 1$,

$$\inf_{\mathbf{y}\in\mathbb{R}^2} \frac{\|B^T\mathbf{y}\|}{\|\mathbf{y}\|} = \inf_{\mathbf{y}\in\mathbb{R}^2} \frac{(y_1^2 + (\theta y_2)^2)^{1/2}}{(y_1^2 + y_2^2)^{1/2}} = \theta.$$

In a sequence of problems, sub-matrices as B_θ can appear, in crucial places, with smaller and smaller θ's. In these cases, for every single problem of the sequence, we shall have a positive infimum in (3.4.38), but there will **not** be a positive uniform β bounding them all from below. □

We collect the previous discussion in the following proposition.

Proposition 3.4.4. *Given a sequence of spaces* $\mathbf{X}$, $\mathbf{Y}$, *a sequence of matrices* A *and* B *and a single positive constant* β, *then the inf-sup condition* (3.4.31) *is equivalent to*

$$\beta\|\mathbf{y}\|_Y \leq \|B^T\mathbf{y}\|_F. \qquad \forall\, \mathbf{y} \in \mathbf{Y}. \tag{3.4.41}$$

Moreover, recalling Proposition 3.4.3, we have that the inf-sup condition (3.4.31) is also equivalent to

$$\exists L_B : \mathbf{G} \to \mathbf{X} \text{ such that } BL_B\mathbf{g} = \mathbf{g} \ \forall\, \mathbf{g} \in \mathbf{G} \tag{3.4.42}$$

with

$$\beta \|L_B\mathbf{g}\|_X \leq \|\mathbf{g}\|_G \qquad \forall\, \mathbf{g} \in \mathbf{G}. \tag{3.4.43}$$

Therefore, in particular, the inf-sup condition (3.4.31) implies that all the matrices B in the sequence are surjective and all the matrices B^T are injective. □

3.4.3 *The* inf-sup *Condition and the Singular Values*

Now we shall see that, using the definitions and the notation of the previous part of this chapter, the discussion of the previous subsection could be drastically shortened. However, first we recall some basic notion on the *singular value decomposition* (see e.g. [228]). Given an $m \times n$ matrix M, it is always possible to find an $n \times n$ unitary matrix U and an $m \times m$ unitary matrix V such that

$$M = V\, \Sigma\, U \tag{3.4.44}$$

where Σ is an $m \times n$ non-negative diagonal matrix. We recall that a rectangular matrix Σ is said to be a *non-negative diagonal matrix* if all its entries are non-negative and for all $i \neq j$ we have $\sigma_{ij} = 0$. On the other hand, an $r \times r$ matrix Λ is said to be a *unitary matrix* when the product $\Lambda^T \Lambda$ is equal to the identity $r \times r$ matrix $\mathbb{I}_r$. Note that this implies that $(\Lambda\mathbf{z})^T \Lambda\mathbf{z} = \mathbf{z}^T\mathbf{z}$ for all $\mathbf{z} \in \mathbb{R}^r$, so that Λ *does not change the Euclidean norm.*

In (3.4.44), the diagonal entries of Σ are known as the **singular values** of M. It can be shown that the non-zero singular values of M are the square roots of the non-zero eigenvalues of $M^T M$.

We now focus our attention on a fundamental example already considered in Sect. 3.1.

Example 3.4.1. Let us go back to the Example 3.1.5, and consider the matrix (that we now denote by Σ) given by

$$\Sigma = \begin{pmatrix} \mu_1 & 0 & \cdot & 0 & 0 & 0 & 0 & 0 \\ 0 & \mu_2 & \cdot & 0 & 0 & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \mu_k & 0 & 0 & 0 & 0 \\ 0 & 0 & \cdot & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \cdot & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \tag{3.4.45}$$

where again k is the dimension of $(\mathrm{Ker}\Sigma)^{\perp}$, which coincides with the dimension of $(\mathrm{Ker}\Sigma^T)^{\perp}$. Here we have $n = k + 4$ and $m = k + 2$. Assuming that the singular values μ_i have been ordered in decreasing order, that is

$$\mu_1 \geq \mu_2 \geq \ldots . \geq \mu_{k-1} \geq \mu_k, \tag{3.4.46}$$

we clearly have (referring to Corollary 3.1.4)

$$\sup_{\boldsymbol{\xi}\in\mathbb{R}^n} \frac{\|\Sigma\boldsymbol{\xi}\|_E}{\|\boldsymbol{\xi}\|_E} \equiv \mu_1 \qquad \text{and} \quad \sup_{\boldsymbol{\eta}\in\mathrm{Im}\Sigma} \frac{\|L_\Sigma\boldsymbol{\eta}\|_E}{\|\boldsymbol{\eta}\|_E} \equiv \mu_k^{-1}, \tag{3.4.47}$$

which, using Proposition 3.4.3, gives immediately

$$\inf_{\boldsymbol{\eta}\in(\mathrm{Ker}\Sigma^T)^{\perp}} \sup_{\boldsymbol{\xi}\in(\mathrm{Ker}\Sigma)^{\perp}} \frac{\boldsymbol{\eta}^T\Sigma\boldsymbol{\xi}}{\|\boldsymbol{\xi}\|_E\,\|\boldsymbol{\eta}\|_E} =: \tilde{\beta}_\Sigma \equiv \mu_k. \tag{3.4.48}$$

Now, we remark that in (3.4.48) there would be no gain and no loss in taking the supremum for $\boldsymbol{\xi} \in \mathbb{R}^n$ rather than for $\boldsymbol{\xi} \in (\mathrm{Ker}\Sigma)^{\perp} \subseteq \mathbb{R}^n$. In general, taking the supremum on a bigger set will provide a bigger (or equal) supremum. Here, for $\boldsymbol{\xi} \in \mathrm{Ker}\Sigma$, the numerator in (3.4.48) (that is $\boldsymbol{\eta}^T\Sigma\boldsymbol{\xi}$) will always be zero and therefore the supremum will not change. Hence,

$$\inf_{\boldsymbol{\eta}\in(\mathrm{Ker}\Sigma^T)^{\perp}} \sup_{\boldsymbol{\xi}\in\mathbb{R}^n} \frac{\boldsymbol{\eta}^T\Sigma\boldsymbol{\xi}}{\|\boldsymbol{\xi}\|_E\,\|\boldsymbol{\eta}\|_E} = \mu_k. \tag{3.4.49}$$

□

Now, given an $m \times n$ matrix B, we set (recalling assumption (3.4.9))

$$M := S_Y\, B\, S_X, \qquad \text{so that} \quad B = S_Y^{-1} M\, S_X^{-1}. \tag{3.4.50}$$

Taking the singular value decomposition (3.4.44) for M will correspond to writing B as

$$B = S_Y\, V\, \Sigma\, U\, S_X. \tag{3.4.51}$$

It is not difficult to check that writing $\mathbf{x} = S_X^{-1}U^T\boldsymbol{\xi}$ and $\mathbf{y} = S_Y^{-1}V\boldsymbol{\eta}$ yields

$$\frac{\mathbf{y}^T B\mathbf{x}}{\|\mathbf{x}\|_X\,\|\mathbf{y}\|_Y} = \frac{\boldsymbol{\eta}^T\, V^T\, S_Y^{-1}\, S_Y\, V\Sigma\, U\, S_X\, S_X^{-1}\, U^T\boldsymbol{\xi}}{\|S_X\, S_X^{-1}U^T\boldsymbol{\xi}\|_E\,\|S_Y\, S_Y^{-1}V\boldsymbol{\eta}\|_E} = \frac{\boldsymbol{\eta}^T\Sigma\boldsymbol{\xi}}{\|\boldsymbol{\xi}\|_E\,\|\boldsymbol{\eta}\|_E} \tag{3.4.52}$$

where, in the last step, we used the definition of the norms (3.4.9) and the fact that U and V are unitary.

Noting that, as it can be easily checked, for $\mathbf{y} = S_Y^{-1}V\boldsymbol{\eta}$ and B given by (3.4.51) (so that $B^T = S_X\, U^T\, \Sigma^T\, V^T\, S_Y$), we have

$$\mathbf{y} \in \mathrm{Ker} B^T \quad \textbf{iff} \quad \boldsymbol{\eta} \in \mathrm{Ker} \Sigma^T,$$

we conclude that

$$\inf_{\mathbf{y}\in(\mathrm{Ker}B^T)^\perp} \sup_{\mathbf{x}\in\mathbb{R}^n} \frac{\mathbf{y}^T B\mathbf{x}}{\|\mathbf{x}\|_X \, \|\mathbf{y}\|_Y} = \inf_{\boldsymbol{\eta}\in(\mathrm{Ker}\Sigma^T)^\perp} \sup_{\boldsymbol{\xi}\in\mathbb{R}^n} \frac{\boldsymbol{\eta}^T \Sigma\boldsymbol{\xi}}{\|\boldsymbol{\xi}\|_E \, \|\boldsymbol{\eta}\|_E} = \mu_k. \tag{3.4.53}$$

We collect the result in the following proposition.

Proposition 3.4.5. *Let B be an $m \times n$ matrix, let the norms in $\mathbf{X}$ and $\mathbf{Y}$ be defined as in (3.4.9) through the matrices S_X and S_Y, respectively, and let $\tilde{\beta}$ be defined as*

$$\inf_{\mathbf{y}\in H^\perp} \sup_{\mathbf{x}\in K^\perp} \frac{\mathbf{x}^T B^T\mathbf{y}}{\|\mathbf{x}\|_X \, \|\mathbf{y}\|_Y} \equiv \inf_{\mathbf{y}\in H^\perp} \sup_{\mathbf{x}\in \mathbf{X}} \frac{\mathbf{x}^T B^T\mathbf{y}}{\|\mathbf{x}\|_X \, \|\mathbf{y}\|_Y} =: \tilde{\beta}, \tag{3.4.54}$$

*where, as usual, $K := KerB$ and $H := KerB^T$. Then, $\tilde{\beta}$ coincides with **the smallest positive singular value** of the matrix $S_Y \, B \, S_X$. In particular, the inf-sup condition (3.4.31) is equivalent to say that "all the singular values of $S_Y \, B \, S_X$ are positive, and the smallest singular value $\tilde{\beta}$ is bounded from below by a fixed positive constant β, independent of the decomposition".* □

3.4.4 The Case of A Elliptic on the Whole Space

As we have seen when discussing solvability, the *inf-sup* condition alone cannot be sufficient for having stability for problems of the general form (3.2.1) and (3.2.2). In order to have sufficient conditions, we now introduce a further assumption on the matrix A. As discussed at the end of Sect. 3.4.1, we start considering a strong condition. More precisely, we make the following assumption.

Ellipticity condition. *There exists a positive constant α, independent of the mesh-size h, such that*

$$\alpha\|\mathbf{x}\|_X^2 \leq \mathbf{x}^T A\mathbf{x} \qquad \forall\, \mathbf{x} \in \mathbf{X}. \tag{3.4.55}$$

We immediately note that, from (3.4.20) and (3.4.55), we easily deduce that

$$\alpha \leq M_a. \tag{3.4.56}$$

We now have the following Theorem.

Theorem 3.4.1. *Let the assumptions (3.4.8)–(3.4.15) on spaces, norms and matrices be satisfied. Let $\mathbf{x}, \mathbf{y}, \mathbf{f}, \mathbf{g}$ satisfy the general system of equations (3.2.1) and (3.2.2). Assume moreover that the inf-sup condition (3.4.31) and the ellipticity (3.4.55) are satisfied. Then, we have*

$$\|\mathbf{x}\|_X \leq \frac{1}{\alpha}\|\mathbf{f}\|_F + \frac{M_a}{\alpha\beta}\|\mathbf{g}\|_G, \tag{3.4.57}$$

$$\|\mathbf{y}\|_Y \leq \frac{2M_a}{\alpha\beta}\|\mathbf{f}\|_F + \frac{M_a^2}{\alpha\beta^2}\|\mathbf{g}\|_G. \tag{3.4.58}$$

Proof. We shall prove the result by splitting $\mathbf{x} = \mathbf{x}_f + \mathbf{x}_g$ and $\mathbf{y} = \mathbf{y}_f + \mathbf{y}_g$, defined as the solutions of

$$\begin{cases} A\mathbf{x}_f + B^T\mathbf{y}_f = \mathbf{f}, \\ B\mathbf{x}_f = 0, \end{cases} \tag{3.4.59}$$

and

$$\begin{cases} A\mathbf{x}_g + B^T\mathbf{y}_g = 0, \\ B\mathbf{x}_g = \mathbf{g}. \end{cases} \tag{3.4.60}$$

We proceed in several steps.

- *Step 1 – Estimate of* $\mathbf{x}_f$ *and* $A\mathbf{x}_f$
  We multiply the first equation of (3.4.59) to the left by $\mathbf{x}_f^T$ and we note that $\mathbf{x}_f^T B^T \mathbf{y}_f \equiv \mathbf{y}^T B\mathbf{x}_f = 0$ (by the second equation). Hence,

  $$\mathbf{x}_f^T A\mathbf{x}_f = \mathbf{x}^T\mathbf{f} \tag{3.4.61}$$

  and, using the ellipticity condition (3.4.55), relation (3.4.61) and the first of the dual norm estimates (3.4.17), we have

  $$\alpha\|\mathbf{x}_f\|_X^2 \leq \mathbf{x}_f^T A\mathbf{x}_f = \mathbf{x}^T\mathbf{f} \leq \|\mathbf{x}_f\|_X \|\mathbf{f}\|_F, \tag{3.4.62}$$

  giving immediately

  $$\|\mathbf{x}_f\|_X \leq \frac{1}{\alpha}\|\mathbf{f}\|_F, \tag{3.4.63}$$

  and using (3.4.18),

  $$\|A\mathbf{x}_f\|_F \leq \frac{M_a}{\alpha}\|\mathbf{f}\|_F. \tag{3.4.64}$$

- *Step 2 – Estimate of* $\mathbf{y}_f$
  Using the equivalent form of the inf-sup condition (3.4.41), we have

  $$\beta\|\mathbf{y}_f\|_Y \leq \|B^T\mathbf{y}_f\|_F = \|\mathbf{f} - A\mathbf{x}_f\|_F. \tag{3.4.65}$$

  Then, using (3.4.65), (3.4.64) and (3.4.56), we obtain

  $$\|\mathbf{y}_f\|_Y \leq \frac{1}{\beta}\|\mathbf{f} - A\mathbf{x}_f\|_F \leq \frac{1}{\beta}\left(1 + \frac{M_a}{\alpha}\right)\|\mathbf{f}\|_F \leq \frac{2M_a}{\alpha\beta}\|\mathbf{f}\|_F. \tag{3.4.66}$$

- *Step 3 – Estimate of* $\|\mathbf{x}_g\|_X^2$ *by* $\|\mathbf{y}_g\|_Y$
  We use the ellipticity (3.4.55), then the first equation of (3.4.60), then (3.1.5), then the second equation of (3.4.60), and finally the second of the dual norm estimates (3.4.17):

$$\alpha\|\mathbf{x}_g\|_X^2 \leq \mathbf{x}_g^T A\mathbf{x}_g = -\mathbf{x}_g^T B^T\mathbf{y}_g \equiv -\mathbf{y}_g^T B\mathbf{x}_g = -\mathbf{y}_g^T\mathbf{g} \leq \|\mathbf{y}_g\|_Y\|\mathbf{g}\|_G. \quad (3.4.67)$$

- *Step 4 – Estimate of* $\|\mathbf{y}_g\|_Y$ *by* $\|\mathbf{x}_g\|_X$
  Using again the inf-sup condition in the from (3.4.41), the first equation of (3.4.60) and the continuity property (3.4.18), we have

$$\beta\|\mathbf{y}_g\|_Y \leq \|B^T\mathbf{y}_g\|_F = \|A\mathbf{x}_g\|_F \leq M_a\|\mathbf{x}_g\|_X. \quad (3.4.68)$$

- *Step 5 – Estimate of* $\|\mathbf{x}_g\|_X$ *and* $\|\mathbf{y}_g\|_Y$
  We combine (3.4.67) and (3.4.68) to obtain

$$\alpha\|\mathbf{x}_g\|_X^2 \leq \frac{M_a}{\beta}\|\mathbf{g}\|_G\|\mathbf{x}_g\|_X, \quad (3.4.69)$$

  which immediately implies

$$\|\mathbf{x}_g\|_X \leq \frac{M_a}{\alpha\beta}\|\mathbf{g}\|_G. \quad (3.4.70)$$

  Using this in (3.4.68), we therefore have

$$\|\mathbf{y}_g\|_X \leq \frac{M_a^2}{\alpha\beta^2}\|\mathbf{g}\|_G. \quad (3.4.71)$$

  The final estimate then follows by simply collecting the separate estimates (3.4.63), (3.4.66), (3.4.70) and (3.4.71).

□

Remark 3.4.5. In some applications (and in particular for the Stokes problem), the matrix A will always be symmetric and positive definite, essentially for all possible types of finite element discretisations, with an α easily bounded away from 0. In these cases, the only condition that we must check will be the *inf-sup* condition on B. This led some people to believe that the *inf-sup* condition for B is *the assumption* to be made for getting a good method when dealing with mixed formulations. This, however, is a superstition, based (as all superstitions) on a narrow horizon. We will see in Chap. 5, Sect. 5.2.4, some examples of discretisations of simple one-dimensional problems that illustrate this point. □

Remark 3.4.6. In some applications it might happen that the constants α and β either depend on h (and tend to zero as h tends to zero) or have a fixed value that is however very small. It is therefore important to keep track of the possible degeneracy of the constants in our estimates when α and/or β are very small. In particular, it is relevant to know whether our stability constants degenerate and tend to infinity, for example, as $1/\beta$ or $1/\beta^2$ or other powers of $1/\beta$ (and, similarly, of $1/\alpha$). In this respect, we point out that the behaviour indicated in (3.4.57) and (3.4.58) is optimal. This means that we cannot hope to find a better proof giving a better behaviour of the constants in terms of powers of $1/\alpha$ and $1/\beta$, as shown by the following example. Considering the system

$$\begin{pmatrix} 1 & -1 & b \\ 1 & a & 0 \\ b & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ y \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ g \end{pmatrix} \qquad 0 < a, b \ll 1, \tag{3.4.72}$$

one easily obtains

$$x_1 = \frac{g}{b}, \quad x_2 = \frac{f_2}{a} - \frac{g}{ab}, \quad y = \frac{f_1}{b} + \frac{f_2}{ab} - \frac{(1+a)g}{ab^2}. \tag{3.4.73}$$

Since $\alpha = a$ and $\beta = b$, from (3.4.73) we deduce that the bounds of Theorem 3.4.1 cannot be improved. □

The dependence of the stability constants on α and β *can* however be improved if we add as a further assumption the symmetry of the matrix A. We have indeed the following result.

Theorem 3.4.2. *Let the assumptions* (3.4.8)–(3.4.15) *on spaces, norms and matrices be satisfied. Let* $\mathbf{x}, \mathbf{y}, \mathbf{f}, \mathbf{g}$ *satisfy the general system of equations* (3.2.1) *and* (3.2.2)*. Assume moreover that the inf-sup condition* (3.4.31) *and the ellipticity* (3.4.55) *are satisfied, and assume moreover that A is symmetric. Then, we have*

$$\|\mathbf{x}\|_X \leq \frac{1}{\alpha}\|\mathbf{f}\|_F + \frac{M_a^{1/2}}{\alpha^{1/2}\beta}\|\mathbf{g}\|_G, \tag{3.4.74}$$

$$\|\mathbf{y}\|_Y \leq \frac{2M_a^{1/2}}{\alpha^{1/2}\beta}\|\mathbf{f}\|_F + \frac{M_a}{\beta^2}\|\mathbf{g}\|_G. \tag{3.4.75}$$

Proof. The following proof mimics rather closely the path of the previous one. In particular, it is done again analysing separately the two problems: (3.4.59), for $\mathbf{g} = 0$, and (3.4.60) for $\mathbf{f} = 0$. However, instead of just indicating the differences between the two proofs, we prefer to report also the second one in detail.

- *Step 1 – Estimate of* $\mathbf{x}_f$ *and* $A\mathbf{x}_f$
  We multiply the first equation of (3.4.59) to the left by $\mathbf{x}_f^T$ and we note that $\mathbf{x}_f^T B^T \mathbf{y}_f \equiv \mathbf{y}^T B \mathbf{x}_f = 0$ (by the second equation). Hence,

$$\mathbf{x}_f^T A \mathbf{x}_f = \mathbf{x}^T \mathbf{f} \tag{3.4.76}$$

and, using the ellipticity condition (3.4.55), relation (3.4.76) and the first of the dual norm estimates (3.4.17), we have

$$\alpha \|\mathbf{x}_f\|_X^2 \leq \mathbf{x}_f^T A \mathbf{x}_f = \mathbf{x}^T \mathbf{f} \leq \|\mathbf{x}_f\|_X \, \|\mathbf{f}\|_F,$$

giving immediately

$$\|\mathbf{x}_f\|_X \leq \frac{1}{\alpha} \|\mathbf{f}\|_F \tag{3.4.77}$$

as well as

$$\mathbf{x}_f^T A \mathbf{x}_f \leq \frac{1}{\alpha} \|\mathbf{f}\|_F^2. \tag{3.4.78}$$

Therefore, using (3.4.22), we also get

$$\|A\mathbf{x}_f\|_F \leq \frac{M_a^{1/2}}{\alpha^{1/2}} \|\mathbf{f}\|_F, \tag{3.4.79}$$

which improves estimate (3.4.64).

- *Step 2 – Estimate of* $\mathbf{y}_f$
  We now use the equivalent form of the inf-sup condition (3.4.41) with $\mathbf{y} = \mathbf{y}_f$. We have

$$\beta \|\mathbf{y}_f\|_Y \leq \|B^T \mathbf{y}_f\|_F = \|\mathbf{f} - A\mathbf{x}_f\|_F. \tag{3.4.80}$$

  Then, using (3.4.80), (3.4.79) and (3.4.56), we obtain

$$\|\mathbf{y}_f\|_Y \leq \frac{1}{\beta} \|\mathbf{f} - A\mathbf{x}_f\|_F \leq \left(\frac{1}{\beta} + \frac{M_a^{1/2}}{\alpha^{1/2}\beta} \right) \|\mathbf{f}\|_F \leq \frac{2M_a^{1/2}}{\alpha^{1/2}\beta} \|\mathbf{f}\|_F. \tag{3.4.81}$$

- *Step 3 – Estimate of* $\mathbf{x}_g^T A \mathbf{x}_g$ *by* $\|\mathbf{y}_g\|_Y$
  We multiply the first equation of (3.4.60) by $\mathbf{x}_g^T$. Using the second equation of (3.4.60) and the second of the dual norm estimates (3.4.17), we have

$$\mathbf{x}_g^T A \mathbf{x}_g = -\mathbf{x}_g^T B^T \mathbf{y}_g \equiv -\mathbf{y}_g^T B \mathbf{x}_g = -\mathbf{y}_g^T \mathbf{g} \leq \|\mathbf{y}_g\|_Y \|\mathbf{g}\|_G. \tag{3.4.82}$$

- *Step 4 – Estimate of* $\|\mathbf{y}_g\|_Y$ *by* $(\mathbf{x}_g^T A \mathbf{x}_g)^{1/2}$
  Using now the inf-sup condition in the form (3.4.31) with $\mathbf{y} = \mathbf{y}_g$, we get that there exists an $\tilde{\mathbf{x}} \neq 0$ such that $\tilde{\mathbf{x}}^T B^T \mathbf{y}_g \geq \beta \|\tilde{\mathbf{x}}\|_X \|\mathbf{y}_g\|_Y$. This relation, the first equation of (3.4.60) and the continuity property (3.4.25), yield

$$\beta \|\tilde{\mathbf{x}}\|_X \|\mathbf{y}_g\|_Y \leq \tilde{\mathbf{x}}^T B^T \mathbf{y}_g = -\tilde{\mathbf{x}}^T A \mathbf{x}_g \leq M_a^{1/2} \|\tilde{\mathbf{x}}\|_X \, (\mathbf{x}_g^T A \mathbf{x}_g)^{1/2}, \tag{3.4.83}$$

  giving (as $\tilde{\mathbf{x}} \neq 0$):

$$\|\mathbf{y}_g\|_Y \leq \frac{M_a^{1/2}}{\beta} (\mathbf{x}_g^T A \mathbf{x}_g)^{1/2}. \tag{3.4.84}$$

- *Step 5 – Estimate of* $\|\mathbf{x}_g\|_X$ *and* $\|\mathbf{y}_g\|_Y$
  We first combine (3.4.82) and (3.4.84) to obtain

$$\|\mathbf{y}_g\|_Y \leq \frac{M_a}{\beta^2}\|\mathbf{g}\|_G. \tag{3.4.85}$$

  Moreover, using the ellipticity assumption (3.4.55), then (3.4.82) and finally (3.4.85), we have

$$\alpha\|\mathbf{x}_g\|_X^2 \leq \mathbf{x}_g^T A\mathbf{x}_g \leq \|\mathbf{y}_g\|_Y\|\mathbf{g}\|_G \leq \frac{M_a}{\beta^2}\|\mathbf{g}\|_G^2,$$

  which can be rewritten as

$$\|\mathbf{x}_g\|_X \leq \frac{M_a^{1/2}}{\alpha^{1/2}\beta}\|\mathbf{g}\|_G. \tag{3.4.86}$$

  The final estimate follows then by simply collecting the separate estimates (3.4.77), (3.4.81), (3.4.86) and (3.4.85). □

Remark 3.4.7. We point out that the behaviour indicated in (3.4.74) and (3.4.75) is also optimal, in the sense that, as in the previous case, we cannot hope to find a better proof giving a better behaviour of the constants in terms of powers of $1/\alpha$ and $1/\beta$. Indeed, consider the system

$$\begin{pmatrix} 2 & \sqrt{a} & b \\ \sqrt{a} & a & 0 \\ b & 0 & 0 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ y \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ g \end{pmatrix} \qquad 0 < a, b \ll 1,$$

whose solution is

$$x_1 = \frac{g}{b}, \quad x_2 = \frac{f_2}{a} - \frac{g}{a^{1/2}b}, \quad y = \frac{f_1}{b} - \frac{f_2}{a^{1/2}b} - \frac{g}{b^2}. \tag{3.4.87}$$

Since the constants α and β are given by

$$\alpha = \frac{2 + a - \sqrt{a^2+4}}{2} = \frac{4a}{2\left(2 + a + \sqrt{a^2+4}\right)} \approx \frac{a}{2}$$

and

$$\beta = b,$$

we see from (3.4.87) that there are cases in which the actual stability constants behave exactly as predicted by the theory. □

3.4.5 The Case of A Elliptic on the Kernel of B

We now consider, together with the *inf-sup* condition on B, a condition on A that is weaker than the full ellipticity (3.4.55). In particular, we require the ellipticity of A to hold only in the kernel K of B.

More precisely, we make the following requirement.

***Elker* condition.** *There exists a positive constant α_0, independent of the mesh-size h, such that*

$$\alpha_0 \|\mathbf{x}\|_X^2 \leq \mathbf{x}^T A\mathbf{x} \qquad \forall \mathbf{x} \in K, \tag{3.4.88}$$

where K is the kernel of B.

The above condition is often called *elker* since it requires the ellipticity on the kernel.

We remark, for future use, that from (3.4.20) and (3.4.88) we get

$$\alpha_0 \leq M_a. \tag{3.4.89}$$

The following Theorem generalises Theorem 3.4.1.

Theorem 3.4.3. *Let the assumptions (3.4.8)–(3.4.15) on spaces, norms and matrices be satisfied. Let $\mathbf{x}, \mathbf{y}, \mathbf{f}, \mathbf{g}$ satisfy the general system of equations (3.2.1) and (3.2.2). Assume moreover that the inf-sup (3.4.31) and the elker condition (3.4.88) are satisfied. Then, we have*

$$\|\mathbf{x}\|_X \leq \frac{1}{\alpha_0}\|\mathbf{f}\|_F + \frac{2M_a}{\alpha_0\beta}\|\mathbf{g}\|_G, \tag{3.4.90}$$

$$\|\mathbf{y}\|_Y \leq \frac{2M_a}{\alpha_0\beta}\|\mathbf{f}\|_F + \frac{2M_a^2}{\alpha_0\beta^2}\|\mathbf{g}\|_G. \tag{3.4.91}$$

Proof. We first set $\mathbf{x}_g := \mathbf{L}\mathbf{g}$ where $\mathbf{L}$ is the lifting operator defined by Proposition 3.4.4. We also point out the following estimates on $\mathbf{x}_g$: from the continuity of the lifting $\mathbf{L}$ (3.4.43) we have

$$\beta\|\mathbf{x}_g\|_X \leq \|\mathbf{g}\|_G \tag{3.4.92}$$

and using (3.4.18) and (3.4.92) we obtain

$$\|A\mathbf{x}_g\|_F \leq M_a\|\mathbf{x}_g\|_X \leq \frac{M_a}{\beta}\|\mathbf{g}\|_G. \tag{3.4.93}$$

Then we set

$$\mathbf{x}_K := \mathbf{x} - \mathbf{x}_g = \mathbf{x} - \mathbf{L}\mathbf{g} \tag{3.4.94}$$

and we note that $\mathbf{x}_K \in K$. Moreover, $(\mathbf{x}_K, \mathbf{y})$ solves the linear system

$$\begin{cases} A\mathbf{x}_K + B^T\mathbf{y} = \mathbf{f} - A\mathbf{x}_g, \\ B\mathbf{x}_K = \mathbf{0}. \end{cases} \tag{3.4.95}$$

We can now proceed as in *Steps 1* and *2* of the proof of Theorem 3.4.1. We note that our weaker assumption *elker* (3.4.88) is sufficient for allowing the first step in (3.4.62). Proceeding as in the first part of *Step 1*, and using (3.4.93) at the end, we get

$$\|\mathbf{x}_K\|_X \leq \frac{1}{\alpha_0}\|\mathbf{f} - A\mathbf{x}_g\|_F \leq \frac{1}{\alpha_0}\left(\|\mathbf{f}\|_F + \frac{M_a}{\beta}\|\mathbf{g}\|_G\right). \tag{3.4.96}$$

This allows to reconstruct the estimate on $\mathbf{x}$:

$$\begin{aligned} \|\mathbf{x}\|_X = \|\mathbf{x}_K + \mathbf{x}_g\|_X &\leq \frac{1}{\alpha_0}\|\mathbf{f}\|_F + \left(\frac{M_a}{\alpha_0\beta} + \frac{1}{\beta}\right)\|\mathbf{g}\|_G \\ &\leq \frac{1}{\alpha_0}\|\mathbf{f}\|_F + \frac{2M_a}{\alpha_0\beta}\|\mathbf{g}\|_G, \end{aligned} \tag{3.4.97}$$

where we have used (3.4.89) in the last inequality. Combining (3.4.18) and (3.4.97), we also have

$$\|A\mathbf{x}\|_F \leq M_a\|\mathbf{x}\|_X \leq \frac{M_a}{\alpha_0}\|\mathbf{f}\|_F + \frac{2M_a^2}{\alpha_0\beta}\|\mathbf{g}\|_G. \tag{3.4.98}$$

Then, we proceed as in *Step 2* to obtain, as in (3.4.81),

$$\beta\|\mathbf{y}\|_Y \leq \|\mathbf{f} - A\mathbf{x}\|_F \tag{3.4.99}$$

and, using the above estimate (3.4.98) on $A\mathbf{x}$ in (3.4.99), we obtain

$$\|\mathbf{y}\|_Y \leq \left(\frac{1}{\beta} + \frac{M_a}{\alpha_0\beta}\right)\|\mathbf{f}\|_F + \frac{2M_a^2}{\alpha_0\beta^2}\|\mathbf{g}\|_G \leq \frac{2M_a}{\alpha_0\beta}\|\mathbf{f}\|_F + \frac{2M_a^2}{\alpha_0\beta^2}\|\mathbf{g}\|_G, \tag{3.4.100}$$

and the proof is concluded. □

Remark 3.4.8. In the spirit of Remark 3.4.6, we note that the dependence of the stability constants on α_0 and β is optimal. Indeed, the dependence is the same as the one proved in Theorem 3.4.1 under stronger assumptions. Hence, the optimality is again shown by example (3.4.72), for which we have $\alpha_0 = a$ and $\beta = b$. It is interesting to note that, contrary to the result of Theorem (3.4.2), **adding the assumption that A is symmetric would not improve the bounds** (!). Indeed, considering the system

$$\begin{pmatrix} 1 & 1 & b \\ 1 & a & 0 \\ b & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ y \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ g \end{pmatrix} \qquad 0 < a, b \ll 1, \tag{3.4.101}$$

one easily obtains

$$x_1 = \frac{g}{b}, \quad x_2 = \frac{f_2}{a} - \frac{g}{ab}, \quad y = \frac{f_1}{b} - \frac{f_2}{ab} + \frac{(1-a)g}{ab^2}. \tag{3.4.102}$$

Since $\alpha_0 = a$ and $\beta = b$, system (3.4.101) shows the same behaviour as the bounds of Theorem 3.4.3 (and not better), even though A is symmetric. □

In order to recover the better bounds found in Theorem 3.4.2, we have to assume that A, on top of satisfying the ellipticity in the kernel (3.4.88), is symmetric *and* positive semi-definite in the whole $\mathbb{R}^n$ (a property that the matrix A in (3.4.101) does not have for $a < 1$). This is because, in order to improve the bounds, one has to use (3.4.22) that requires A to be symmetric and positive semi-definite. We have indeed the following result, that we state without proof: indeed, we shall see in the next section that this result can be obtained as a particular case of a more general estimate (see Remark 3.6.4).

Theorem 3.4.4. *Let the assumptions* (3.4.8)–(3.4.15) *on spaces, norms and matrices be satisfied. Let* $\mathbf{x}, \mathbf{y}, \mathbf{f}, \mathbf{g}$ *satisfy the general system of equations* (3.2.1) *and* (3.2.2)*. Assume that the inf-sup* (3.4.31) *and the elker condition* (3.4.88) *are satisfied, and assume moreover that* A *is symmetric and positive semi-definite on the whole space* $\mathbf{X}$*. Then, we have*

$$\|\mathbf{x}\|_X \leq \frac{1}{\alpha_0}\|\mathbf{f}\|_F + \frac{2M_a^{1/2}}{\alpha_0^{1/2}\beta}\|\mathbf{g}\|_G, \tag{3.4.103}$$

$$\|\mathbf{y}\|_Y \leq \frac{2M_a^{1/2}}{\alpha_0^{1/2}\beta}\|\mathbf{f}\|_F + \frac{M_a}{\beta^2}\|\mathbf{g}\|_G. \tag{3.4.104}$$

3.4.6 *The Case of A Satisfying an* inf-sup *on the Kernel of B*

As we have seen in the previous sections, the ellipticity in the kernel for the matrix A is not the weakest condition we can use. Indeed, in order to get necessary and sufficient conditions for solvability, we used the surjectivity of B (here replaced with the *inf-sup* condition on B) and the non-singularity of A_{KK} on the kernel K of B. Hence, it is clear that we still have room to improve the result of Theorem 3.4.3 by assuming on A some property weaker than (3.4.88). In particular we can assume

Inf-sup condition on $\boldsymbol{A_{KK}}$**:** *There exists a positive constant* α_1*, independent of the mesh-size* h*, such that*

$$\inf_{\mathbf{x}\in K}\sup_{\mathbf{z}\in K}\frac{\mathbf{z}^T A\mathbf{x}}{\|\mathbf{z}\|_X\,\|\mathbf{x}\|_X}\geq\alpha_1. \tag{3.4.105}$$

We note that (3.4.105) can be equivalently written as

$$\alpha_1\|\mathbf{x}\|_X\leq\sup_{\mathbf{z}\in K}\frac{\mathbf{z}^T A\mathbf{x}}{\|\mathbf{z}\|_X}\qquad\forall\,\mathbf{x}\in K, \tag{3.4.106}$$

or

$$\alpha_1\|\mathbf{x}\|_X\leq\|A_{KK}\mathbf{x}\|_{K'}\qquad\forall\,\mathbf{x}\in K, \tag{3.4.107}$$

where we used the notation of (3.4.16).

We have then the following result.

Theorem 3.4.5. *Let the assumptions* (3.4.8)–(3.4.15) *on spaces, norms and matrices be satisfied. Let* $\mathbf{x},\mathbf{y},\mathbf{f},\mathbf{g}$ *satisfy the general system of equations* (3.2.1) *and* (3.2.2)*. Assume, moreover, that the inf-sup condition* (3.4.31) *on* B *and the bounding condition* (3.4.107) *on* A_{KK} *are satisfied. Then, we have*

$$\|\mathbf{x}\|_X\leq\frac{1}{\alpha_1}\|\mathbf{f}\|_F+\frac{2M_a}{\alpha_1\beta}\|\mathbf{g}\|_G, \tag{3.4.108}$$

$$\|\mathbf{y}\|_Y\leq\frac{2M_a}{\alpha_1\beta}\|\mathbf{f}\|_F+\frac{2M_a^2}{\alpha_1\beta^2}\|\mathbf{g}\|_G. \tag{3.4.109}$$

Proof. The proof is identical to that of Theorem 3.4.3. The only change is in the first inequality in (3.4.96). Using this time (3.4.107), and noting once more that from (3.2.5), we easily obtain

$$\alpha_1\|\mathbf{x}_K\|_X\leq\|A_{KK}\mathbf{x}_K\|_{K'}\leq\|\mathbf{f}-A_{KK}\mathbf{x}_g\|_{K'}\leq\|\mathbf{f}\|_F+\|A\mathbf{x}_g\|_F, \tag{3.4.110}$$

so that the first inequality of (3.4.96) still holds if we replace α_0 by α_1. The rest of the proof goes on unchanged. □

So far, for every type of bounding conditions on the matrix A (global ellipticity and ellipticity on K), we considered separately the special cases in which A had some additional property. In particular, after Theorem 3.4.1 (where A was assumed to be elliptic on the whole $\mathbf{X}$), we considered in Theorem 3.4.2 the case where A was also symmetric. Similarly, after Theorem 3.4.3 (where A was assumed to be elliptic on K), we considered in Theorem 3.4.4 the case where A was also symmetric and positive semi-definite on the whole $\mathbf{X}$. Now, after Theorem 3.4.5 (where A is supposed to satisfy the bounding condition (3.4.107) on K), we could ask ourselves what happens if we assume further that A is also symmetric and positive semi-definite on the whole $\mathbf{X}$. This, however, would bring us back to the case of Theorem 3.4.4, thanks to the following proposition.

Proposition 3.4.6. *Let A be an $n \times n$ matrix, and K a subspace of $\mathbb{R}^n$. Assume that A is symmetric, positive semi-definite, and verifies (3.4.107) on K. Then, A is elliptic on K.* □

Proof. Indeed, for $x \in K$, using (3.4.106) and then (3.3.1), we have

$$\alpha_1^2 \|\mathbf{x}\|_X^2 \leq \sup_{\mathbf{z} \in K} \frac{(\mathbf{z}^T A\mathbf{x})^2}{\|\mathbf{z}\|_X^2} \leq \sup_{\mathbf{z} \in K} \frac{\mathbf{x}^T A\mathbf{x}\, \mathbf{z}^T A\mathbf{z}}{\|\mathbf{z}\|_X^2} \leq M_a \mathbf{x}^T A\mathbf{x}, \tag{3.4.111}$$

and the result follows with $\alpha_0 = \alpha_1^2 / M_a$. □

3.5 Additional Results

In this section, we present some additional results concerning necessary conditions, modified problems and special cases.

3.5.1 Some Necessary Conditions

We see in this subsection that the above sufficient conditions for having existence and uniqueness of the solution, together with stability estimates, are indeed *necessary*.

Theorem 3.5.1. *Assume that there exists a constant C such that, for any quadruple $(\mathbf{x}, \mathbf{y}, \mathbf{f}, \mathbf{g})$ in $X \times Y \times F \times G$ solution of (3.2.1) and (3.2.2), we have*

$$\|\mathbf{x}\|_X + \|\mathbf{y}\|_Y \leq C(||\mathbf{f}||_F + ||\mathbf{g}||_G). \tag{3.5.1}$$

Then, (3.4.107) and (3.4.31) are verified with $\alpha_1 = \beta = 1/C$.

Proof. For every $\mathbf{y} \in \mathbf{Y}$, it is easy to see that $(0, \mathbf{y}, B^T\mathbf{y}, 0)$ satisfies (3.2.1) and (3.2.2). Hence, (3.5.1) shows that the *inf-sup* condition (3.4.31) is satisfied in the equivalent form (3.4.41), with $\beta = 1/C$. Then, for every $\mathbf{x} \in K = \mathrm{Ker}B$, set $\mathbf{f} := \pi_K A\mathbf{x} \equiv A_{KK}\mathbf{x}$. Note that $\pi_K(\mathbf{f} - A\mathbf{x}) = \mathbf{0}$, and hence $\mathbf{f} - A\mathbf{x}$ belongs to $K^\perp$. From (3.1.60) we have that there exists a $\mathbf{y} \in \mathbf{Y}$ such that $B^T\mathbf{y} = \mathbf{f} - A\mathbf{x}$, and since $\mathbf{x} \in K$, we have that $(\mathbf{x}, \mathbf{y}, A\mathbf{x}, 0)$ satisfies (3.2.1) and (3.2.2). Hence, inequality (3.5.1) gives now (3.4.107) with $\alpha_1 = 1/C$. □

Remark 3.5.1. Note that an inequality like (3.5.1) implies that the problem (3.2.1) and (3.2.2) has been adimensionalised. This is not the case for the results of the previous section. See also Remark 3.6.6 at the end of this chapter. □

Theorem 3.5.1 dealt with the necessity of the assumptions in Theorem 3.4.5. The following result deals with the necessity of the assumptions in Theorem 3.4.4.

Theorem 3.5.2. *Let A be symmetric and positive semi-definite. Assume that there exists a constant C such that for any quadruple $(\mathbf{x}, \mathbf{y}, \mathbf{f}, \mathbf{g})$ in $X \times Y \times F \times G$ solution of (3.2.1) and (3.2.2) we have that the bound (3.5.1) holds. Then, (3.4.88) and (3.4.31) are verified with $\alpha_0 = 1/(C^2 M_a)$ (where M_a is the continuity constant of A defined in (3.4.18)) and $\beta = 1/C$, respectively.*

Proof. The result is an immediate consequence of Theorem 3.5.1 and Proposition 3.4.6 □

Remark 3.5.2. As we have seen in Theorem 3.2.1, the estimate (3.5.1) implies the *inf-sup* condition (3.2.19) and the non singularity of A_{KK} on the kernel K (3.2.18). The purpose of Theorems 3.5.1 and 3.5.2 is mainly to show that a uniform bound for C implies uniform bounds for the constants α_1 (or α_0) and β. □

3.5.2 The Case of B Not Surjective. Modification of the Problem

Here, we come back, somehow, to the case of Remark 3.2.1. To start with, we observe that, proceeding as in Remark 3.2.1 we, immediately have the following result.

Proposition 3.5.1. *Assume that A_{KK} satisfies (3.4.105), and $\mathbf{g} \in \mathrm{Im}B$. Then, problem (3.2.1) and (3.2.2) has at least one solution $(\mathbf{x}, \mathbf{y})$. Moreover, $\mathbf{x}$ is uniquely determined and*

$$||\mathbf{x}||_X \leq \frac{1}{\alpha_1}(||\mathbf{f}||_F + \frac{M_a}{\tilde{\beta}}\|\mathbf{g}\|_G) \tag{3.5.2}$$

where $\tilde{\beta}$ is defined in (3.4.28). □

We note that (3.5.2) does not provide any estimate for the variable $\mathbf{y}$. This should be expected since in Proposition 3.5.1 we did not assume that the *inf-sup* condition (3.4.31) holds true. However, (3.4.28) will always hold so that for $\mathbf{g} \in \mathrm{Im}B$ we might consider the problem (3.2.1) and (3.2.2) in $\mathbf{X} \times H^{\perp}$ instead of $\mathbf{X} \times \mathbf{Y}$, keeping in H the same norm that we had in $\mathbf{Y}$. Hence, we can apply any of the previous theorems of this section (that is, one of the Theorems 3.4.1–3.4.5) and have an estimate in $\mathbf{X} \times H^{\perp}$ as a function of the norms of $\mathbf{f}$ and $\mathbf{g}$, of the constant α (or α_0, or α_1) and of the constant $\tilde{\beta}$ appearing in (3.4.28). For instance, applying Theorem 3.4.2, we have the following result.

Theorem 3.5.3. *Assume that the assumptions (3.4.8)–(3.4.15) on spaces, norms and matrices are satisfied. Let $\mathbf{x}, \mathbf{y}, \mathbf{f}, \mathbf{g}$ satisfy the general system of equations (3.2.1) and (3.2.2), with $\mathbf{y} \in H^{\perp}$. Assume moreover that A is symmetric and satisfies (3.4.55) and that the constant $\tilde{\beta}$ is defined by (3.4.28). Then, we have:*

$$\|\mathbf{x}\|_X \leq \frac{1}{\alpha}\|\mathbf{f}\|_F + \frac{M_a^{1/2}}{\alpha^{1/2}\tilde{\beta}}\|\mathbf{g}\|_G, \tag{3.5.3}$$

$$\|\mathbf{y}\|_{H^\perp} \leq \frac{2M_a^{1/2}}{\alpha^{1/2}\tilde{\beta}}\|\mathbf{f}\|_F + \frac{M_a}{\tilde{\beta}^2}\|\mathbf{g}\|_G. \tag{3.5.4}$$

3.5.3 Some Special Cases

In some applications, we shall encounter situations where the right-hand side has the special form $(\mathbf{f}, 0)$ or $(0, \mathbf{g})$. In fact, the proofs of the previous Theorems often used explicitly those special cases. We now consider them in more detail. For the sake of simplicity, we will restrict our attention to the case of A symmetric and positive semi-definite.

3.5.3.1 The case (f, 0)

From Proposition 3.5.1, we have immediately the following particular case.

Proposition 3.5.2. *Assume that A satisfies (3.4.105) and $\mathbf{g} = 0$. Then, problem (3.2.1) and (3.2.2) has at least one solution $(\mathbf{x}, \mathbf{y})$. Moreover, $\mathbf{x}$ is uniquely determined by $\mathbf{f}$ and*

$$||\mathbf{x}||_X \leq \frac{||\mathbf{f}||_F}{\alpha_1}. \tag{3.5.5}$$

Finally, $\mathbf{y}$ is unique up to an element in $H \equiv \mathrm{Ker}B^T$ and

$$||\pi_{H^\perp}\mathbf{y}||_Y \leq \frac{M_a||\mathbf{f}||_F}{\alpha_1\tilde{\beta}}. \tag{3.5.6}$$

□

Conversely, we have that Theorem 3.5.2 has two correspondents in the $(\mathbf{f}, 0)$ case.

Proposition 3.5.3. *Assume that A is symmetric and positive semi-definite, and assume that there exists a constant $C > 0$ such that, for every quadruple $(\mathbf{x}, \mathbf{y}, \mathbf{f}, 0) \in X \times Y \times F \times G$ satisfying (3.2.1) and (3.2.2), one has*

$$||\mathbf{x}||_X \leq C||\mathbf{f}||_F. \tag{3.5.7}$$

Then, the discrete ellipticity on the kernel (3.4.88) holds with $\alpha_0 = 1/(C^2 M_a)$, M_a being the continuity constant of A defined in (3.4.18). □

Proof. The proof is identical to the first part of the proof of Theorem 3.5.1, using Proposition 3.4.6. □

Proposition 3.5.4. *Assume that A is symmetric and positive semi-definite, and assume that there exists a constant $C > 0$ such that, for every quadruple $(\mathbf{x}, \mathbf{y}, \mathbf{f}, 0) \in X \times Y \times F \times G$ satisfying (3.2.1) and (3.2.2), one has*

$$||\mathbf{y}||_Y \leq C\,||\mathbf{f}||_F. \tag{3.5.8}$$

Then, the inf-sup condition (3.4.41) holds with $\beta = 1/C$. □

Proof. The proof is identical to the second part of the proof of Theorem 3.5.2. □

3.5.3.2 The case (0, g)

We begin with a simple lemma.

Lemma 3.5.1. *Assume that A is symmetric and positive semi-definite, and let Z be a subspace of $\mathbf{X}$. Then, $Ker(A_{ZZ}) \subset KerA$.*

Proof. If $\mathbf{z}$ is in the kernel of A_{ZZ}, we immediately have that

$$\mathbf{z}^T A\mathbf{z} = 0, \tag{3.5.9}$$

which, using (3.4.22), implies $A\mathbf{z} = 0$. □

We can now prove the following result.

Proposition 3.5.5. *Assume that A is symmetric and positive semi-definite and that the inf-sup condition (3.4.31) holds. Then, for every $\mathbf{g} \in G$ and $\mathbf{f} = 0$, problem (3.2.1) and (3.2.2) has at least one solution $(\mathbf{x}, \mathbf{y})$. Moreover, $\mathbf{y}$ is uniquely determined by $\mathbf{g}$ and we have the bound*

$$\|\mathbf{y}\|_Y \leq \frac{M_a}{\beta^2}\|\mathbf{g}\|_G. \tag{3.5.10}$$

□

Proof. Using Proposition 3.4.4, we have that, for every $\mathbf{g} \in G$, there exists at least one $\mathbf{x}_g \in X$ such that $B\mathbf{x}_g = \mathbf{g}$ and

$$\|\mathbf{x}_g\| \leq \frac{1}{\beta}\|\mathbf{g}\|_G. \tag{3.5.11}$$

Using Lemma 3.5.1 with $Z = K$, we see that $\mathrm{Ker}A_{KK} \subseteq \mathrm{Ker}A$. Then, using Proposition 3.1.3 with $r = s$ and $S = Z = K$, we have that $\pi_K \mathrm{Im}A \subseteq \mathrm{Im}A_{KK}$. Hence, the problem: *find $\mathbf{x}_K \in K$ such that*

$$A_{KK}\mathbf{x}_K = -\pi_K A\mathbf{x}_g \tag{3.5.12}$$

has at least one solution. Using (3.4.22), then using (3.5.12) (multiplied to the left by $\mathbf{x}_K$), and finally using the symmetry of A, one gets

$$\|A\mathbf{x}_K\|_F^2 \leq M_a \mathbf{x}_K^T A\mathbf{x}_K = M_a \mathbf{x}_K^T A\mathbf{x}_g \leq M_a \|\mathbf{x}_g\|_X \|A\mathbf{x}_K\|_F, \tag{3.5.13}$$

which, using (3.5.11), gives immediately

$$\|A\mathbf{x}_K\|_F \leq \frac{M_a}{\beta}\|\mathbf{g}\|_G. \tag{3.5.14}$$

Note that (3.5.12) implies that $A(\mathbf{x}_K + \mathbf{x}_g) \in K^\perp$, so that by (3.1.60) there exists a $\mathbf{y} \in Y$ such that $B^T\mathbf{y} = -(A\mathbf{x}_K + A\mathbf{x}_g)$, and by (3.4.41), (3.5.11), and (3.5.14) we have

$$\|\mathbf{y}\|_Y \leq \frac{1}{\beta}\|A(\mathbf{x}_K + \mathbf{x}_g)\|_F \leq \frac{M_a}{\beta^2}\|\mathbf{g}\|_G. \tag{3.5.15}$$

Finally, observe that $(\mathbf{x}_g + \mathbf{x}_K, \mathbf{y})$ solves (3.2.1) and (3.2.2) with $(0, \mathbf{g})$ as right-hand side.

To see the uniqueness, assume that $(\mathbf{x}^i, \mathbf{y}^i)$ $(i = 1, 2)$ are two solutions. Clearly, $\pi_K A(\mathbf{x}^1 - \mathbf{x}^2) = \pi_K B^T(\mathbf{y}^2 - \mathbf{y}^1) = 0$ and hence $\mathbf{x}^1 - \mathbf{x}^2$ is in the kernel of A_{KK}. Using Lemma 3.5.1, we see that $A(\mathbf{x}^1 - \mathbf{x}^2) = 0$ so that, from the first equations, $B^T(\mathbf{y}^2 - \mathbf{y}^1) = 0$ and the *inf-sup* condition (3.4.31) implies $\mathbf{y}^1 = \mathbf{y}^2$. □

Proposition 3.5.6. *Assume that A is symmetric and positive semi-definite, and that there exists a constant $C > 0$ such that, for every quadruple $(\mathbf{x}, \mathbf{y}, 0, \mathbf{g}) \in X \times Y \times F \times G$ satisfying (3.2.1) and (3.2.2), one has*

$$||\mathbf{y}||_Y \leq C||\mathbf{g}||_G. \tag{3.5.16}$$

Then, the inf-sup condition (3.4.31) holds. However, we cannot bound β in terms of the constant C appearing in (3.5.16). □

Proof. Let us first remark that assumption (3.5.16) implies that B^T is injective, and this implies (3.4.31). In order to see that the value of β cannot be deduced in general, consider the case when $A = 0$, $X = Y$ and B is γ times the identity. Then, the *inf-sup* condition holds with $\beta = |\gamma|$ and (3.5.16) holds with $C = 0$. □

Proposition 3.5.7. *Assume that A is symmetric and positive semi-definite, and that there exists a constant $C > 0$ such that for every quadruple $(\mathbf{x}, \mathbf{y}, 0, \mathbf{g}) \in X \times Y \times F \times G$ satisfying (3.2.1) and (3.2.2) one has,*

$$||\mathbf{x}||_X + ||\mathbf{y}||_Y \leq C||\mathbf{g}||_G, \tag{3.5.17}$$

then (3.2.1) and (3.2.2) has a solution for any $\mathbf{f} \in F$ and $\mathbf{g} \in G$, and (3.4.31) holds with $\beta = 1/C$. □

Proof. Clearly, (3.5.17) implies that (3.2.1) and (3.2.2) for $\mathbf{f} = 0$ and $\mathbf{g} = 0$ has only the zero solution. Hence, Corollary 3.1.3 implies the solvability of (3.2.1) and (3.2.2) for general $\mathbf{f}$ and $\mathbf{g}$, and then Theorem 3.2.1 gives us (3.2.18) and (3.2.19). Hence, we just have to deal with the estimate of β. Note that, now (as we already have the unique solvability), (3.5.17) ensures the existence of a lifting operator that associates to every $\mathbf{g} \in G$ the first component $\mathbf{x}$ of the unique solution of (3.2.1) and (3.2.2) with right-hand side $(0, \mathbf{g})$. Hence, the result follows from Proposition 3.4.4. □

3.5.4 Composite Matrices

In the previous section, we considered the case in which the matrix A has a block structure of the type

$$\mathbb{A} = \begin{pmatrix} C & D^T \\ D & 0 \end{pmatrix}, \tag{3.5.18}$$

and B has the structure $\mathbb{B} = (E \quad 0)$ or $\mathbb{B} = (0 \quad E)$, so that the whole matrix has the block structure

$$M = \begin{pmatrix} C & D^T & E^T \\ D & 0 & 0 \\ E & 0 & 0 \end{pmatrix} \tag{3.5.19}$$

or

$$M = \begin{pmatrix} C & D^T & 0 \\ D & 0 & E^T \\ 0 & E & 0 \end{pmatrix}, \tag{3.5.20}$$

respectively. We were also able to find necessary and sufficient conditions for the *solvability*, simply using in a reasonable way the conditions dictated by the basic Theorem 3.2.1.

Here, we would like to consider the associated *stability* properties. These again can be deduced from the general case. It is clear that we would need three sequences of spaces $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$, with norms that ensure the continuity of the quadratic forms associated with the matrices

- C (on $\mathbf{X} \times \mathbf{X}$),
- D (on $\mathbf{X} \times \mathbf{Y}$),
- E (on $\mathbf{X} \times \mathbf{Z}$ for (3.5.19) and $\mathbf{Y} \times \mathbf{Z}$ for (3.5.20)),

as we did in (3.4.20), together with dual norms as in (3.4.10). Then, we just have to change the non-singularity conditions into their corresponding *uniform bounds*.

For instance, in the case (3.5.19), we easily obtained the algebraic conditions (3.2.48), that we recall for convenience of the reader:

$$\begin{aligned} &\mathrm{Im}D^T \cap \mathrm{Im}E^T = \mathbf{0}_r, \\ &\pi_{\mathbb{K}} C \text{ is non-singular } \mathbb{K} \to \mathbb{K} \quad \text{where } \mathbb{K} = \mathrm{Ker}D \cap \mathrm{Ker}E. \end{aligned} \tag{3.5.21}$$

It is not difficult to verify that the corresponding stability conditions are:

$$\begin{aligned} &\inf_{(\mathbf{y},\mathbf{z})\in\mathbf{Y}\times\mathbf{Z}} \sup_{\mathbf{x}\in\mathbf{X}} \frac{\mathbf{x}^T D^T \mathbf{y} + \mathbf{x}^T E^T \mathbf{z}}{\|\mathbf{x}\|_{\mathbf{X}} \left(\|\mathbf{y}\|_{\mathbf{Y}} + \|\mathbf{z}\|_{\mathbf{Z}}\right)} \geq \delta > 0, \\ &\inf_{\tilde{\mathbf{x}}\in\mathbb{K}} \sup_{\mathbf{x}\in\mathbb{K}} \frac{\mathbf{x}^T C \tilde{\mathbf{x}}}{\|\tilde{\mathbf{x}}\|_{\mathbf{X}} \|\mathbf{x}\|_{\mathbf{Y}}} \geq \alpha > 0, \qquad \text{where } \mathbb{K} = \mathrm{Ker}D \cap \mathrm{Ker}E. \end{aligned} \tag{3.5.22}$$

Clearly, we could simplify the condition on C by requiring ellipticity on $\mathbb{K}$, or ellipticity on the whole $\mathbf{X}$.

For (3.5.20), we performed first an exchange of rows and columns, to reach the form

$$M = \begin{pmatrix} C & 0 & D^T \\ 0 & 0 & E \\ D & E^T & 0 \end{pmatrix},$$

and we found the following solvability conditions:

$$\begin{aligned} &\mathrm{Ker}D^T \cap \mathrm{Ker}E = 0, \\ &\mathrm{Ker}E^T = 0, \\ &\pi_{\tilde{\mathbb{K}}} C \text{ is non-singular } \tilde{\mathbb{K}} \to \tilde{\mathbb{K}}, \end{aligned} \tag{3.5.23}$$

where $\tilde{\mathbb{K}}$ (cfr. (3.2.52)) is given by

$$\tilde{\mathbb{K}} = \{\mathbf{x} \in \mathbf{X} \text{ such that } D\mathbf{x} \in (\mathrm{Ker}\, E)^{\perp}\}. \tag{3.5.24}$$

Again, it is not difficult to verify that the corresponding stability conditions are:

$$\begin{aligned} &\inf_{\mathbf{y}\in\mathbf{Y}} \sup_{(\mathbf{x},\mathbf{z})\in(\mathbf{X}\times\mathbf{Z})} \frac{\mathbf{y}^T D\mathbf{x} + \mathbf{y}^T E^T \mathbf{z}}{\left(\|\mathbf{x}\|_{\mathbf{X}} + \|\mathbf{z}\|_{\mathbf{Z}}\right) \|\mathbf{y}\|_{\mathbf{Y}}} \geq \delta > 0, \\ &\inf_{\mathbf{z}\in\mathbf{Z}} \sup_{\mathbf{y}\in\mathbf{Y}} \frac{\mathbf{y}^T E^T \mathbf{z}}{\|\mathbf{y}\|_{\mathbf{Y}} \|\mathbf{z}\|_{\mathbf{Z}}} \geq \eta > 0, \\ &\inf_{\tilde{\mathbf{x}}\in\tilde{\mathbb{K}}} \sup_{\mathbf{x}\in\tilde{\mathbb{K}}} \frac{\mathbf{x}^T C \tilde{\mathbf{x}}}{\|\tilde{\mathbf{x}}\|_{\mathbf{X}} \|\mathbf{x}\|_{\mathbf{Y}}} \geq \alpha > 0. \end{aligned} \tag{3.5.25}$$

Here too, the third condition could possibly by replaced by an ellipticity condition. Moreover, it is easy to see that, in order to get the first condition, it would be

sufficient to assume that one of the two matrices D or E^T satisfies an *inf-sup* condition by itself. However, this would often be an assumption too strong and difficult to obtain in practice. As we did in the previous section, we do not insist on these matters, and we shall not analyse the optimal dependence of the stability constants from δ, η and α appearing in (3.5.22) and (3.5.25).

3.6 Stability of Perturbed Matrices

We shall now discuss the case of problems of the type (3.3.1) where an additional matrix C is present. We assume that we are given, for each $k \in \mathbb{N}$, an $m(k) \times m(k)$ matrix C_k. Together with the matrices A_k and B_k, this will give us a sequence of perturbed problems

$$\begin{pmatrix} A_k & B_k^T \\ B_k & -C_k \end{pmatrix} \begin{pmatrix} \mathbf{x}_k \\ \mathbf{y}_k \end{pmatrix} = \begin{pmatrix} \mathbf{f}_k \\ \mathbf{g}_k \end{pmatrix}. \tag{3.6.1}$$

As we did for the unperturbed case (3.4.1), we *drop the index* k, and we just **remember** that we are actually dealing with a sequence of problems instead of a single one.

As a first step, we have to extend our assumptions (3.4.20) on the continuity of matrices A and B, requiring the continuity of C as well. Hence, we assume that there exists a constant M_c, independent of k, such that

$$\forall\, \mathbf{z} \in \mathbf{Y},\ \forall\, \mathbf{y} \in \mathbf{Y} \quad \mathbf{z}^T C \mathbf{y} \le M_c \, \|\mathbf{z}\|_Y \, \|\mathbf{y}\|_Y . \tag{3.6.2}$$

We note that, as in (3.4.18) and (3.4.19), we now have for every $\mathbf{y} \in \mathbf{Y}$:

$$\|C\mathbf{y}\|_G \equiv \sup_{\mathbf{z} \in \mathbb{R}^m} \frac{\mathbf{z}^T C \mathbf{y}}{\|\mathbf{z}\|_Y} \le M_c \|\mathbf{y}\|_Y . \tag{3.6.3}$$

We would like to extend the results of the previous subsection to the perturbed problem (3.6.1).

3.6.1 The Basic Estimate

Following Theorem 3.3.1 we are going to assume that A is symmetric and non-singular on $K = \operatorname{Ker} B$. It will therefore be useful, in order to reach optimal estimates in an easier way, to use directly (3.4.88), that we repeat for the convenience of the reader

$$\alpha_0 \|\mathbf{x}\|_X^2 \le \mathbf{x}^T A \mathbf{x} \qquad \forall \mathbf{x} \in K, \tag{3.6.4}$$

instead of (3.4.105). For technical reasons, it will also be easier to deal separately with the case $\mathbf{f} = 0$ and the case $\mathbf{g} = 0$, as we did, for instance, in the proofs of Theorems 3.4.1 and 3.4.2. This time, however, it will be more convenient to split the results in two different lemmata, and join them afterwards. We start therefore with the following lemma.

Lemma 3.6.1. *Let the assumptions* (3.4.8)–(3.4.15) *and* (3.6.2) *on spaces, norms and matrices be satisfied. Assume that the inf-sup condition* (3.4.31) *and the ellipticity requirement* (3.6.4) *are satisfied, and assume moreover that A is symmetric, and A and C are positive semi-definite. Then, if* $\mathbf{x}$, $\mathbf{y}$, *and* $\mathbf{g}$ *satisfy*

$$\begin{cases} A\mathbf{x} + B^T\mathbf{y} = 0 \\ B\mathbf{x} - C\mathbf{y} = \mathbf{g}, \end{cases} \tag{3.6.5}$$

we have the estimate

$$\|\mathbf{x}\|_X \leq \frac{2M_a^{1/2}(\beta^2 + M_c M_a)}{\alpha_0^{1/2}\beta^3}\|\mathbf{g}\|_G, \tag{3.6.6}$$

$$\|\mathbf{y}\|_Y \leq \frac{M_a}{\beta^2}\|\mathbf{g}\|_G. \tag{3.6.7}$$

Proof. Using the *inf-sup* condition in the form (3.4.41) together with the first equation of (3.6.5), we obtain

$$\beta\|\mathbf{y}\|_Y \leq \|B^T\mathbf{y}\|_F = \|A\mathbf{x}\|_F. \tag{3.6.8}$$

Now, we take the scalar product of the first equation of (3.6.5) times $\mathbf{x}$, we take the scalar product of the second equation of (3.6.5) times $\mathbf{y}$, and we take the difference, obtaining

$$\mathbf{x}^T A\mathbf{x} + \mathbf{y}^T C\mathbf{y} = -\mathbf{y}^T\mathbf{g}. \tag{3.6.9}$$

Using (3.4.22), then Eq. (3.6.9) with the assumption that C is positive semi-definite, and finally (3.4.17), we have

$$\|A\mathbf{x}\|_F^2 \leq M_a\mathbf{x}^T A\mathbf{x} \leq -M_a\mathbf{y}^T\mathbf{g} \leq M_a\|\mathbf{y}\|_Y\,\|\mathbf{g}\|_G, \tag{3.6.10}$$

which, combined with (3.6.8), yields

$$\|A\mathbf{x}\|_F \leq \frac{M_a}{\beta}\|\mathbf{g}\|_G, \tag{3.6.11}$$

so that, using again (3.6.8),

$$\|\mathbf{y}\|_Y \leq \frac{M_a}{\beta^2}\|\mathbf{g}\|_G, \tag{3.6.12}$$

which proves (3.6.7). The proof now becomes similar to that of Theorem 3.4.2. Using Proposition 3.4.4, we set

$$\tilde{\mathbf{x}} := \mathbf{L}(\mathbf{g} + C\mathbf{y}) \tag{3.6.13}$$

so that, from (3.4.43),

$$B\tilde{\mathbf{x}} = \mathbf{g} + C\mathbf{y}, \tag{3.6.14}$$

together with

$$\beta\|\tilde{\mathbf{x}}\|_X \le \|\mathbf{g} + C\mathbf{y}\|_G \le (1 + \frac{M_c M_a}{\beta^2})\|\mathbf{g}\|_G, \tag{3.6.15}$$

where, in the last step, we used (3.6.3) and (3.6.12). From (3.6.15), we have then immediately

$$\|\tilde{\mathbf{x}}\|_X \le \frac{\beta^2 + M_c M_a}{\beta^3}\|\mathbf{g}\|_G. \tag{3.6.16}$$

Setting now

$$\mathbf{x}_K := \mathbf{x} - \tilde{\mathbf{x}}, \tag{3.6.17}$$

we have from (3.6.14) and the second equation of (3.6.5) that $\mathbf{x}_K \in K$ (the kernel of B). We then note that, from the first equation of (3.6.1):

$$\mathbf{x}_K^T A\mathbf{x} = -\mathbf{x}_K^T B^T \mathbf{y} = -\mathbf{y}^T B\mathbf{x}_K = 0. \tag{3.6.18}$$

Moreover, using (3.6.17), (3.6.18), and then (3.3.5), we have

$$\mathbf{x}_K^T A\mathbf{x}_K = -\mathbf{x}_K^T A\tilde{\mathbf{x}} \le (\mathbf{x}_K^T A\mathbf{x}_K)^{1/2}(\tilde{\mathbf{x}}^T A\tilde{\mathbf{x}})^{1/2}, \tag{3.6.19}$$

which easily gives

$$\mathbf{x}_K^T A\mathbf{x}_K \le \tilde{\mathbf{x}}^T A\tilde{\mathbf{x}}. \tag{3.6.20}$$

Hence, we can use (3.6.4) and (3.6.20) to obtain

$$\alpha_0\|\mathbf{x}_K\|_X^2 \le \mathbf{x}_K^T A\mathbf{x}_K \le \tilde{\mathbf{x}}^T A\tilde{\mathbf{x}}, \tag{3.6.21}$$

and finally from (3.6.21) and (3.4.20)

$$\|\mathbf{x}_K\|_X \le \Big(\frac{M_a}{\alpha_0}\Big)^{1/2}\|\tilde{\mathbf{x}}\|_X. \tag{3.6.22}$$

Finally, we can collect (3.6.17), (3.6.22) and (3.6.16) and have an estimate for $\mathbf{x}$:

$$\begin{aligned}\|\mathbf{x}\|_X &\le \|\mathbf{x}_K\|_X + \|\tilde{\mathbf{x}}\|_X \le (1 + (\frac{M_a}{\alpha})^{1/2})\|\tilde{\mathbf{x}}\|_X \\ &\le (1 + (\frac{M_a}{\alpha_0})^{1/2})\,\frac{\beta^2 + M_c M_a}{\beta^3}\|\mathbf{g}\|_G.\end{aligned} \tag{3.6.23}$$

Using (3.4.89) in (3.6.23), we obtain (3.6.6) and the proof is completed. □

Remark 3.6.1. The dependence of the constants in (3.6.6) and (3.6.7) on α_0 and β cannot be improved. Indeed, considering for instance (for $0 < a, b \ll 1$) the problem

$$\begin{pmatrix} 2a & \sqrt{a} & -\sqrt{a} & 0 & 0 \\ \sqrt{a} & 2 & 1 & b & 0 \\ -\sqrt{a} & 1 & 2 & 0 & b \\ 0 & b & 0 & 0 & 1 \\ 0 & 0 & b & -1 & 0 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ -1 \\ -1 \end{pmatrix}, \tag{3.6.24}$$

we easily have, as unique solution,

$$x_1 = \frac{3}{b^3 a^{1/2}}, \quad x_2 = -\frac{3+b^2}{b^3}, \quad x_3 = \frac{3-b^2}{b^3}, \tag{3.6.25}$$

$$y_1 = \frac{3}{b^2}, \quad y_2 = \frac{3}{b^2}. \tag{3.6.26}$$

We can easily check that we have $\alpha_0 = 2a$ and $\beta = b$, and we verify the optimality of (3.6.12) and (3.6.23). □

We now consider the case where $\mathbf{g}$ is equal to zero and $\mathbf{f}$ is not.

Lemma 3.6.2. *Let the assumptions* (3.4.8)–(3.4.15) *and* (3.6.2) *on spaces, norms and matrices be satisfied. Assume that the inf-sup condition* (3.4.31) *and the ellipticity requirement* (3.6.4) *are satisfied, and assume moreover that A is symmetric, and A and C are positive semi-definite. Then, if $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{f}$ satisfy*

$$\begin{cases} A\mathbf{x} + B^T\mathbf{y} = \mathbf{f} \\ B\mathbf{x} - C\mathbf{y} = 0, \end{cases} \tag{3.6.27}$$

we have the estimates

$$\|\mathbf{x}\|_X \leq \frac{(\beta^2 + 2M_c M_a)^2 + 4(M_c M_a)^2}{\alpha_0 \beta^4} \|\mathbf{f}\|_F, \tag{3.6.28}$$

$$\|\mathbf{y}\|_Y \leq \frac{2M_a^{1/2}(2M_c M_a + \beta^2)}{\alpha_0^{1/2}\beta^3} \|\mathbf{f}\|_F. \tag{3.6.29}$$

Proof. As in the previous lemma, we take the scalar product of the first equation of (3.6.27) with $\mathbf{x}$, then we take the scalar product of the second equation of (3.6.27) with $\mathbf{y}$, and we take the difference, obtaining

$$\mathbf{x}^T A\mathbf{x} + \mathbf{y}^T C\mathbf{y} = \mathbf{x}^T \mathbf{f}. \tag{3.6.30}$$

Using then (3.4.22), Eq. (3.6.30) with the assumption that C is positive semi-definite, and finally (3.4.17), we have

$$\|A\mathbf{x}\|_F^2 \leq M_a \mathbf{x}^T A\mathbf{x} \leq M_a \mathbf{x}^T \mathbf{f} \leq M_a \|\mathbf{x}\|_X \, \|\mathbf{f}\|_F. \tag{3.6.31}$$

Next, we use the *inf-sup* condition in the form (3.4.41) to obtain, from the first equation of (3.6.27),

$$\begin{aligned} \beta \|\mathbf{y}\|_Y &\leq \|B^T \mathbf{y}\|_F \equiv \|\mathbf{f} - A\mathbf{x}\|_F \\ &\leq \|A\mathbf{x}\|_F + \|\mathbf{f}\|_F. \end{aligned} \tag{3.6.32}$$

We now consider, as we did before, the lifting operator $\mathbf{L}$ as defined in Proposition 3.4.4 and we set

$$\tilde{\mathbf{x}} := \mathbf{L}(C\mathbf{y}) \tag{3.6.33}$$

so that

$$B\tilde{\mathbf{x}} - C\mathbf{y} = 0. \tag{3.6.34}$$

Then, using (3.4.43) and (3.6.3),

$$\beta \|\tilde{\mathbf{x}}\|_X \leq \|C\mathbf{y}\|_G \leq M_c \, \|\mathbf{y}\|_Y. \tag{3.6.35}$$

We now set

$$\mathbf{x}_K := \mathbf{x} - \tilde{\mathbf{x}} \tag{3.6.36}$$

and we note that, clearly, $B\mathbf{x}_K = 0$, so that $\mathbf{x}_K \in K = \text{Ker}\, B$. Our next (and most delicate) step will be to estimate $\mathbf{x}_K$ in terms of $\tilde{\mathbf{x}}$. We first note that, using (3.6.4),

$$\alpha_0 \|\mathbf{x}_K\|_X^2 \leq \mathbf{x}_K^T A \mathbf{x}_K, \tag{3.6.37}$$

which implies that

$$\|\mathbf{x}_K\|_X \leq \Big(\frac{\mathbf{x}_K^T A \mathbf{x}_K}{\alpha_0} \Big)^{1/2}. \tag{3.6.38}$$

Then, we estimate $\mathbf{x}_K^T A \mathbf{x}_K$. We remember again that $\mathbf{x}_K^T B^T \mathbf{y} = 0$ (since $\mathbf{x}_K \in$ Ker B), so that, using (3.6.36) and the first equation of (3.6.27),

$$\mathbf{x}_K^T A \mathbf{x}_K = \mathbf{x}_K^T A \mathbf{x} - \mathbf{x}_K^T A \tilde{\mathbf{x}} = \mathbf{x}_K^T \mathbf{f} - \mathbf{x}_K^T A \tilde{\mathbf{x}}. \tag{3.6.39}$$

We now use (3.6.39) with (3.4.17) and (3.3.5), and then (3.6.38) to obtain

$$\begin{aligned} \mathbf{x}_K^T A \mathbf{x}_K &\leq \|\mathbf{f}\|_F \, \|\mathbf{x}_K\|_X + (\mathbf{x}_K^T A \mathbf{x}_K)^{1/2} (\tilde{\mathbf{x}}^T A \tilde{\mathbf{x}})^{1/2} \\ &\leq \|\mathbf{f}\|_F \Big(\frac{\mathbf{x}_K^T A \mathbf{x}_K}{\alpha_0} \Big)^{1/2} + (\mathbf{x}_K^T A \mathbf{x}_K)^{1/2} (\tilde{\mathbf{x}}^T A \tilde{\mathbf{x}})^{1/2} \\ &\leq (\mathbf{x}_K^T A \mathbf{x}_K)^{1/2} \Big(\frac{1}{\alpha_0^{1/2}} \|\mathbf{f}\|_F + (\tilde{\mathbf{x}}^T A \tilde{\mathbf{x}})^{1/2} \Big), \end{aligned} \tag{3.6.40}$$

implying

$$(\mathbf{x}_K^T A \mathbf{x}_K)^{1/2} \leq \frac{1}{{\alpha_0}^{1/2}} \|\mathbf{f}\|_F + (\tilde{\mathbf{x}}^T A \tilde{\mathbf{x}})^{1/2}. \tag{3.6.41}$$

Inserting (3.6.41) into (3.6.38), and then using (3.4.20), we now have

$$\|\mathbf{x}_K\|_X \leq \frac{1}{\alpha_0} \|\mathbf{f}\|_F + \Big(\frac{\tilde{\mathbf{x}}^T A \tilde{\mathbf{x}}}{\alpha_0}\Big)^{1/2} \leq \frac{1}{\alpha_0} \|\mathbf{f}\|_F + \frac{{M_a}^{1/2}}{{\alpha_0}^{1/2}} \|\tilde{\mathbf{x}}\|_X. \tag{3.6.42}$$

We can now collect (3.6.36), (3.6.42) and (3.6.35) to obtain an estimate for $\mathbf{x}$

$$\begin{aligned} \|\mathbf{x}\|_X \leq \|\mathbf{x}_K\|_X + \|\tilde{\mathbf{x}}\|_X &\leq \frac{1}{\alpha_0} \|\mathbf{f}\|_F + \Big(\frac{M_c {M_a}^{1/2}}{{\alpha_0}^{1/2}\beta} + \frac{M_c}{\beta}\Big) \|\mathbf{y}\|_Y \\ &\leq \frac{1}{\alpha_0} \|\mathbf{f}\|_F + \frac{2 M_c {M_a}^{1/2}}{{\alpha_0}^{1/2}\beta} \|\mathbf{y}\|_Y. \end{aligned} \tag{3.6.43}$$

Now, we take the square of both sides of (3.6.32), we use $(a+b)^2 \leq 2(a^2+b^2)$, we insert (3.6.31) and finally (3.6.43):

$$\begin{aligned} \beta^2 \|\mathbf{y}\|_Y^2 &\leq 2\|A\mathbf{x}\|_F^2 + 2\|\mathbf{f}\|_F^2 \leq 2M_a \|\mathbf{x}\|_X \, \|\mathbf{f}\|_F + 2\|\mathbf{f}\|_F^2 \\ &\leq 2\|\mathbf{f}\|_F \Big(\frac{2M_c {M_a}^{3/2}}{{\alpha_0}^{1/2}\beta} \|\mathbf{y}\|_Y + \frac{M_a}{\alpha_0} \|\mathbf{f}\|_F\Big) + 2\|\mathbf{f}\|_F^2. \end{aligned} \tag{3.6.44}$$

We now use the fact that, for positive real numbers t, a, and b, if $t^2 \leq at + b$, then $t \leq a + \sqrt{b}$. Applied to (3.6.44), this gives

$$\|\mathbf{y}\|_Y \leq \frac{4M_c {M_a}^{3/2}}{{\alpha_0}^{1/2}\beta^3} \|\mathbf{f}\|_F + \frac{(2M_a + 2\alpha_0)^{1/2}}{{\alpha_0}^{1/2}\beta} \|\mathbf{f}\|_F. \tag{3.6.45}$$

Using again the fact that $\alpha_0 \leq M_a$, we can rewrite (3.6.45) as

$$\|\mathbf{y}\|_Y \leq \frac{2{M_a}^{1/2}(2M_c M_a + \beta^2)}{{\alpha_0}^{1/2}\beta^3} \|\mathbf{f}\|_F. \tag{3.6.46}$$

Inserting (3.6.46) into (3.6.43), we obtain the corresponding estimate for $\mathbf{x}$:

$$\begin{aligned} \|\mathbf{x}\|_X &\leq \Big(\frac{8(M_c M_a)^2 + 4M_c M_a \beta^2}{\alpha_0 \beta^4} + \frac{1}{\alpha_0}\Big) \|\mathbf{f}\|_F \\ &= \frac{(\beta^2 + 2M_c M_a)^2 + 4(M_c M_a)^2}{\alpha_0 \beta^4} \|\mathbf{f}\|_F, \end{aligned} \tag{3.6.47}$$

which concludes the proof. □

Remark 3.6.2. The result (3.6.28) and (3.6.29) cannot be improved in its dependence from the constants α_0 and β. Indeed, if we consider, for $0 < a, b \ll 1$, the system

$$\begin{pmatrix} 2a & \sqrt{a} & -\sqrt{a} & 0 & 0 \\ \sqrt{a} & 2 & 1 & b & 0 \\ -\sqrt{a} & 1 & 2 & 0 & b \\ 0 & b & 0 & 0 & 1 \\ 0 & 0 & b & -1 & 0 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \tag{3.6.48}$$

we easily have, as unique solution,

$$x_1 = \frac{3 + b^4}{a\, b^4}, \quad x_2 = \frac{-3 - b^2}{a^{1/2}\, b^4}, \quad x_3 = \frac{3 - b^2}{a^{1/2}\, b^4}, \tag{3.6.49}$$

$$y_1 = \frac{3 - b^2}{a^{1/2}\, b^3}, \quad y_2 = \frac{3 + b^2}{a^{1/2}\, b^3}. \tag{3.6.50}$$

It is not difficult to check that $\alpha_0 = 2a$ and $\beta = b$. Hence, (3.6.49) and (3.6.50) shows the optimality of (3.6.28) and (3.6.29). □

We can now collect the results of the previous two lemmata.

Theorem 3.6.1. *Let the assumptions* (3.4.8)–(3.4.15) *and* (3.6.2) *on spaces, norms and matrices be satisfied. Assume that the inf-sup condition* (3.4.31) *and the ellipticity requirement* (3.6.4) *are satisfied, and assume moreover that A is symmetric and that A and C are positive semi-definite. Then, if* $\mathbf{x}$, $\mathbf{y}$, $\mathbf{f}$, *and* $\mathbf{g}$ *satisfy*

$$\begin{cases} A\mathbf{x} + B^T\mathbf{y} = \mathbf{f} \\ B\mathbf{x} - C\mathbf{y} = \mathbf{g}, \end{cases} \tag{3.6.51}$$

we have the estimates

$$\|\mathbf{x}\|_X \leq \frac{(\beta^2 + 2M_c M_a)^2 + 4(M_c M_a)^2}{\alpha_0 \beta^4} \|\mathbf{f}\|_F + \frac{2M_a^{1/2}(\beta^2 + M_c M_a)}{\alpha_0^{1/2} \beta^3} \|\mathbf{g}\|_G, \tag{3.6.52}$$

$$\|\mathbf{y}\|_Y \leq \frac{2M_a^{1/2}(2M_c M_a + \beta^2)}{\alpha_0^{1/2} \beta^3} \|\mathbf{f}\|_F + \frac{M_a}{\beta^2} \|\mathbf{g}\|_G. \tag{3.6.53}$$

The proof easily follows by linearity.

3.6.2 The Symmetric Case for Perturbed Matrices

The dependence of the constants in (3.6.52) and (3.6.53) on α_0 and β improves noticeably if we assume that C is symmetric as well. As an example, we can consider the particular case (relevant in applications) of systems still having the structure (3.6.1), where C is a symmetric and positive definite matrix verifying

$$\gamma \|\mathbf{y}\|_Y^2 \leq \mathbf{y}^T C \mathbf{y} \leq M_c \|\mathbf{y}\|_Y^2 \quad \forall\, \mathbf{y} \in \mathbf{Y}. \tag{3.6.54}$$

We note that our assumption (3.6.54) easily implies that

$$\frac{1}{M_c}\|\mathbf{z}\|_G^2 \leq \mathbf{z}^T C^{-1} \mathbf{z} \leq \frac{1}{\gamma}\|\mathbf{z}\|_G^2 \quad \forall\, \mathbf{z} \in \mathbf{Y}. \tag{3.6.55}$$

From (3.6.54), we easily obtain as well that

$$\|\mathbf{y}\|_Y \leq \frac{1}{\gamma}\|C\mathbf{y}\|_G \quad \forall\, \mathbf{y} \in \mathbf{Y} \tag{3.6.56}$$

and from (3.6.55)

$$\|\mathbf{z}\|_G \leq M_c \|C^{-1}\mathbf{z}\|_Y \quad \forall\, \mathbf{z} \in \mathbf{Y}. \tag{3.6.57}$$

We are now ready to prove our improved estimates.

Remark 3.6.3. We shall prove in the next chapter, Sect. 4.3, additional related results (in the infinite dimensional case) which may be considered as more elegant, but for which we have no example showing optimality. □

Theorem 3.6.2. *Let the assumptions* (3.4.8)–(3.4.15) *and* (3.6.2) *on spaces, norms and matrices be satisfied. Assume that the inf-sup condition* (3.4.31) *and the ellipticity requirement* (3.6.4) *are satisfied, and assume moreover that A is symmetric and positive semi-definite and that C is* ***symmetric*** *and satisfies* (3.6.54)*. Then, if* $\mathbf{x}$, $\mathbf{y}$, $\mathbf{f}$, *and* $\mathbf{g}$ *satisfy* (3.6.51)*, we have the estimate*

$$\|\mathbf{x}\|_X \leq \frac{\beta^2 + 4M_c M_a}{\alpha_0 \beta^2}\|\mathbf{f}\|_F + \frac{2M_a^{1/2} M_c}{\alpha_0^{1/2}\gamma\beta}\|\mathbf{g}\|_G \tag{3.6.58}$$

and

$$\|\mathbf{y}\|_Y \leq \frac{2M_c M_a^{1/2}}{\gamma \alpha_0^{1/2}\beta}\|\mathbf{f}\|_F + \frac{2\, M_a(M_c + \gamma)}{M_a\gamma^2 + (M_c + \gamma)\beta^2}\|\mathbf{g}\|_G. \tag{3.6.59}$$

Proof. As we are already used to, we shall split the two cases $\mathbf{f} = 0$ and $\mathbf{g} = 0$, and then combine the estimates by linearity.

Let us consider first the case $\mathbf{f} = 0$, and assume that $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{g}$ satisfy (3.6.5).

Following the notation of Lemma 3.6.1, we still have (3.6.11), (3.6.12) and (3.6.22). Our target is to improve (3.6.16), which is suboptimal in our (stronger)

assumptions. For this we restart by taking once more the scalar product of the first equation of (3.6.5) times $\mathbf{x}$, getting

$$\mathbf{x}^T A\mathbf{x} + \mathbf{x}^T B^T \mathbf{y} = 0 \tag{3.6.60}$$

and we substitute $\mathbf{y} = C^{-1}(B\mathbf{x} - \mathbf{g})$. Recalling that A is positive semi-definite, we obtain

$$\mathbf{x}^T B^T C^{-1} B\mathbf{x} \le \mathbf{x}^T B^T C^{-1}\mathbf{g} = \mathbf{g}^T C^{-1} B\mathbf{x}. \tag{3.6.61}$$

Using (3.6.55) with $\mathbf{z} = B\mathbf{x}$, then (3.6.61), then (3.4.17), and finally (3.6.56) with $\mathbf{y} = C^{-1}B\mathbf{x}$, we have

$$\begin{aligned} \|B\mathbf{x}\|_G^2 &\le M_c(\mathbf{x}^T B^T C^{-1} B\mathbf{x}) \le M_c(\mathbf{g}^T C^{-1} B\mathbf{x}) \\ &\le M_c \|\mathbf{g}\|_G \, \|C^{-1} B\mathbf{x}\|_Y \le \frac{M_c}{\gamma} \|\mathbf{g}\|_G \|B\mathbf{x}\|_G, \end{aligned} \tag{3.6.62}$$

which easily gives

$$\|B\mathbf{x}\|_G \le \frac{M_c}{\gamma} \|\mathbf{g}\|_G. \tag{3.6.63}$$

As in Lemma 3.6.1, we set again (see (3.6.13) and (3.6.14)) $\tilde{\mathbf{x}} := \mathbf{L}(\mathbf{g} + C\mathbf{y})$, getting $B\tilde{\mathbf{x}} = \mathbf{g} + C\mathbf{y} = B\mathbf{x}$. Using (3.4.43), we have therefore

$$\beta \|\tilde{\mathbf{x}}\|_X \le \|B\tilde{\mathbf{x}}\|_G = \|B\mathbf{x}\|_G \tag{3.6.64}$$

and combining (3.6.63) and (3.6.64), we obtain

$$\|\tilde{\mathbf{x}}\|_X \le \frac{M_c}{\gamma\beta} \|\mathbf{g}\|_G, \tag{3.6.65}$$

which is the required improvement of (3.6.16). We can now use this improved estimate in (3.6.22), and we obtain

$$\|\mathbf{x}_K\|_X \le \left(\frac{M_a}{\alpha}\right)^{1/2} \|\tilde{\mathbf{x}}\|_X \le \frac{M_c M_a^{1/2}}{\gamma\beta\alpha^{1/2}} \|\mathbf{g}\|_G. \tag{3.6.66}$$

We note at this point that we have another way to obtain an estimate for $\mathbf{y}$, apart from (3.6.12) that we keep from the previous analysis; actually, from (3.6.56) and the second equation of (3.6.5), and then (3.6.63):

$$\|\mathbf{y}\|_Y \le \frac{1}{\gamma} \|B\mathbf{x} - \mathbf{g}\|_G \le \left(\frac{1}{\gamma} + \frac{M_c}{\gamma^2}\right) \|\mathbf{g}\|_G = \frac{\gamma + M_c}{\gamma^2} \|\mathbf{g}\|_G. \tag{3.6.67}$$

With some manipulations, we see that (3.6.12) and (3.6.67) can be combined into

$$\|\mathbf{y}\|_Y \le \frac{2\, M_a(M_c + \gamma)}{M_a\gamma^2 + (M_c + \gamma)\beta^2} \|\mathbf{g}\|_G. \tag{3.6.68}$$

We collect the results for $\mathbf{f} = 0$:

$$\|\mathbf{x}\|_X \leq \|\mathbf{x}_K\|_X + \|\tilde{\mathbf{x}}\|_X \leq \Big(\Big(\frac{M_a}{\alpha_0}\Big)^{1/2} + 1\Big)\frac{M_c}{\gamma\beta}\|\mathbf{g}\|_G, \tag{3.6.69}$$

$$\|\mathbf{y}\|_Y \leq \frac{2\,M_a(M_c+\gamma)}{M_a\gamma^2 + (M_c+\gamma)\beta^2}\|\mathbf{g}\|_G. \tag{3.6.70}$$

We also note that, using $\alpha_0 \leq M_a$, the estimate (3.6.69) becomes

$$\|\mathbf{x}\|_X \leq \frac{2M_a^{1/2}M_c}{\alpha_0^{1/2}\gamma\beta}\|\mathbf{g}\|_G. \tag{3.6.71}$$

We consider now the case in which $\mathbf{g} = 0$ and assume that $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{f}$ satisfy (3.6.27). As before, we can keep part of the previous analysis, but we can improve it in several places. From the proof of Lemma 3.6.2, we keep the definition of $\tilde{\mathbf{x}}$ and $\mathbf{x}_K$, and the estimates (3.6.41) and (3.6.42). We now take the scalar product of the first equation of (3.6.27) times $\tilde{\mathbf{x}}$, and substitute $\mathbf{y} = C^{-1}B\mathbf{x}$:

$$\tilde{\mathbf{x}}^T A\mathbf{x} + \tilde{\mathbf{x}}^T B^T C^{-1} B\mathbf{x} = \tilde{\mathbf{x}}^T\mathbf{f}. \tag{3.6.72}$$

We now recall that $B\tilde{\mathbf{x}} = B\mathbf{x}$, and rewrite (3.6.72) as follows

$$\mathbf{x}^T B^T C^{-1} B\mathbf{x} = \tilde{\mathbf{x}}^T\mathbf{f} - \tilde{\mathbf{x}}^T A\mathbf{x}. \tag{3.6.73}$$

We now apply (3.6.55) with $\mathbf{z} = B\mathbf{x}$ and we use (3.6.73) to obtain:

$$\frac{1}{M_c}\|B\mathbf{x}\|_G^2 \leq \mathbf{x}^T B^T C^{-1} B\mathbf{x} = \tilde{\mathbf{x}}^T\mathbf{f} - \tilde{\mathbf{x}}^T A\mathbf{x}.$$

We then use (3.4.17) and the estimate $\beta\|\tilde{\mathbf{x}}\|_G \leq \|B\tilde{\mathbf{x}}\|_G = \|B\mathbf{x}\|_G$ as in (3.6.64) and we reach

$$\frac{1}{M_c}\|B\mathbf{x}\|_G^2 \leq \frac{1}{\beta}\|\mathbf{f}\|_F\,\|B\mathbf{x}\|_G - \tilde{\mathbf{x}}^T A\mathbf{x}. \tag{3.6.74}$$

We leave (3.6.74) for a while, and we estimate $-\tilde{\mathbf{x}}^T A\mathbf{x}$. Using the fact that $\mathbf{x} = \tilde{\mathbf{x}} + \mathbf{x}_K$, then (3.3.5), then (3.6.41), and finally some little algebra, we have

$$\begin{aligned}
-\tilde{\mathbf{x}}^T A\mathbf{x} &= -\tilde{\mathbf{x}}^T A\tilde{\mathbf{x}} - \tilde{\mathbf{x}}^T A\mathbf{x}_K \\
&\leq -\tilde{\mathbf{x}}^T A\tilde{\mathbf{x}} + (\tilde{\mathbf{x}}^T A\tilde{\mathbf{x}})^{1/2}(\mathbf{x}_K^T A\mathbf{x}_K)^{1/2} \\
&\leq -\tilde{\mathbf{x}}^T A\tilde{\mathbf{x}} + (\tilde{\mathbf{x}}^T A\tilde{\mathbf{x}})^{1/2}\Big(\frac{1}{{\alpha_0}^{1/2}}\|\mathbf{f}\|_F + (\tilde{\mathbf{x}}^T A\tilde{\mathbf{x}})^{1/2}\Big) \\
&= \frac{1}{{\alpha_0}^{1/2}}\|\mathbf{f}\|_F(\tilde{\mathbf{x}}^T A\tilde{\mathbf{x}})^{1/2},
\end{aligned} \tag{3.6.75}$$

which inserted in (3.6.74) gives

$$\frac{1}{M_c}\|B\mathbf{x}\|_G^2 \le \frac{1}{\alpha_0^{1/2}}\|\mathbf{f}\|_F(\tilde{\mathbf{x}}^T A\tilde{\mathbf{x}})^{1/2} + \frac{1}{\beta}\|\mathbf{f}\|_F\,\|B\mathbf{x}\|_G. \tag{3.6.76}$$

Using the continuity of A (3.4.18) and once more $\beta\|\tilde{\mathbf{x}}\|_X \le \|B\mathbf{x}\|_G$, inequality (3.6.76) gives:

$$\frac{1}{M_c}\|B\mathbf{x}\|_G^2 \le \frac{M_a^{1/2}}{\alpha_0^{1/2}\beta}\|\mathbf{f}\|_F\|B\mathbf{x}\|_G + \frac{1}{\beta}\|\mathbf{f}\|_F\,\|B\mathbf{x}\|_G, \tag{3.6.77}$$

so that we can divide both sides by $\|B\mathbf{x}\|_G$, obtaining

$$\frac{1}{M_c}\|B\mathbf{x}\|_G \le \frac{M_a^{1/2}}{\alpha_0^{1/2}\beta}\|\mathbf{f}\|_F + \frac{1}{\beta}\|\mathbf{f}\|_F \le \frac{M_a^{1/2}+\alpha_0^{1/2}}{\alpha_0^{1/2}\beta}\|\mathbf{f}\|_F, \tag{3.6.78}$$

which is the basis of our improved estimates. From (3.6.78), we first derive

$$\|\tilde{\mathbf{x}}\|_X \le \frac{1}{\beta}\|B\tilde{\mathbf{x}}\|_G \le \frac{M_c\,(M_a^{1/2}+\alpha_0^{1/2})}{\alpha_0^{1/2}\beta^2}\|\mathbf{f}\|_F, \tag{3.6.79}$$

and then we use it in (3.6.42)

$$\begin{aligned}\|\mathbf{x}_K\|_X &\le \frac{1}{\alpha_0}\|\mathbf{f}\|_F + \frac{M_a^{1/2}}{\alpha_0^{1/2}}\|\tilde{\mathbf{x}}\|_X \\ &\le \Big(\frac{1}{\alpha_0} + \frac{M_a^{1/2}}{\alpha_0^{1/2}}\,\frac{M_c\,(M_a^{1/2}+\alpha_0^{1/2})}{\alpha_0^{1/2}\beta^2}\Big)\|\mathbf{f}\|_F \\ &\le \Big(\frac{1}{\alpha_0} + \frac{M_cM_a + M_c(M_a\alpha_0)^{1/2}}{\alpha_0\beta^2}\Big)\|\mathbf{f}\|_F. \end{aligned}\tag{3.6.80}$$

From the second equation of (3.6.27), (3.6.56) and (3.6.78), we also derive our improved estimate for $\mathbf{y}$

$$\|\mathbf{y}\|_Y = \|C^{-1}B\mathbf{x}\|_Y \le \frac{1}{\gamma}\|B\mathbf{x}\|_G \le \frac{M_c}{\gamma}\,\frac{M_a^{1/2}+\alpha_0^{1/2}}{\alpha_0^{1/2}\beta}\|\mathbf{f}\|_F. \tag{3.6.81}$$

We collect the results for $\mathbf{g} = 0$, using the fact that $\alpha \le M_a$. From (3.6.79) and (3.6.80), we have the estimate on $\mathbf{x}$

$$\begin{aligned}\|\mathbf{x}\|_X &\le \|\tilde{\mathbf{x}}\|_X + \|\mathbf{x}_K\|_X \\ &\le \left(\frac{M_c\,(M_a^{1/2}+\alpha_0^{1/2})}{\alpha_0^{1/2}\beta^2} + \frac{1}{\alpha_0} + \frac{M_c M_a + M_c(M_a\alpha_0)^{1/2}}{\alpha_0\beta^2}\right)\|\mathbf{f}\|_F \\ &\le \frac{\beta^2 + 4M_c M_a}{\alpha_0\beta^2}\,\|\mathbf{f}\|_F, \quad (3.6.82)\end{aligned}$$

while, from (3.6.81), we have the estimate on $\mathbf{y}$

$$\|\mathbf{y}\|_Y \le \frac{2M_c M_a^{1/2}}{\gamma\alpha_0^{1/2}\beta}\|\mathbf{f}\|_F. \quad (3.6.83)$$

The final results can then be obtained collecting (3.6.70), (3.6.71), (3.6.82) and (3.6.83). □

Remark 3.6.4. We remark that in several applications we have $C = \varepsilon Identity$, so that $M_c = \gamma = \varepsilon$. In this case, the estimates (3.6.58) and (3.6.59) become

$$\|\mathbf{x}\|_X \le \frac{\beta^2 + 4\varepsilon M_a}{\alpha_0\beta^2}\|\mathbf{f}\|_F + \frac{2M_a^{1/2}}{\alpha_0^{1/2}\beta}\|\mathbf{g}\|_G \quad (3.6.84)$$

and

$$\|\mathbf{y}\|_Y \le \frac{2M_a^{1/2}}{\alpha_0^{1/2}\beta}\|\mathbf{f}\|_F + \frac{4M_a}{M_a\varepsilon + 2\beta^2}\|\mathbf{g}\|_G. \quad (3.6.85)$$

We also note that in the limit for $\varepsilon \to 0$ we recover the result of Theorem 3.4.4. □

Remark 3.6.5. We also point out that (3.6.84) and (3.6.85) are optimal, with respect to the dependency of the stability constants on the parameters α_0, β and ε. To see that, consider for $0 < a, b \ll 1$ the problem

$$\begin{pmatrix} 2a & \sqrt{a} & -\sqrt{a} & 0 & 0 \\ \sqrt{a} & 2 & 1 & b & 0 \\ -\sqrt{a} & 1 & 2 & 0 & b \\ 0 & b & 0 & -\varepsilon & 0 \\ 0 & 0 & b & 0 & -\varepsilon \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 2f \\ 0 \\ 0 \\ 0 \\ 2g \end{pmatrix}, \quad (3.6.86)$$

whose solution is given by

$$x_1 = \frac{f(b^2+\varepsilon)}{ab^2} + \frac{g}{ba^{1/2}}, \qquad x_2 = -\frac{f\varepsilon}{a^{1/2}b^2} - \frac{3g\varepsilon}{b(3\varepsilon+b^2)},$$

$$x_3 = \frac{f\varepsilon}{a^{1/2}b^2} + \frac{g(3\varepsilon+2b^2)}{b(3\varepsilon+b^2)}, \tag{3.6.87}$$

$$y_1 = -\frac{f}{a^{1/2}b} - \frac{3g}{3\varepsilon+b^2}, \qquad y_2 = \frac{f}{a^{1/2}b} - \frac{3g}{3\varepsilon+b^2}. \tag{3.6.88}$$

Indeed, it is easy to recognise that $\alpha_0 = 2a$ and $\beta = b$, and hence the optimality of estimates (3.6.84) and (3.6.85). □

Remark 3.6.6. It is of some interest to check that all our estimates are "dimensionally correct". Indeed, denoting by $[a]$, $[b]$, $[c]$, $[x]$, $[y]$, $[f]$ and $[g]$, respectively, the physical dimensions of A, B, C, $\mathbf{x}$, $\mathbf{y}$, $\mathbf{f}$ and $\mathbf{g}$, we have that M_a and α have the same dimensions as A (and hence, in particular, M_a/α is a pure number). Similarly, M_c and γ have the same dimension as $[c]$. Moreover, from the two equations of our system, we have

$$[x] = \frac{[f]}{[a]} = \frac{[g]}{[b]}, \quad [y] = \frac{[f]}{[b]} = \frac{[a][g]}{[b^2]} = \frac{[g]}{[c]} \tag{3.6.89}$$

from which we easily deduce that $[a]\,[c]$ equals to $[b^2]$, so that for instance $M_a M_c/\beta^2$ is also a pure number. Taking this into account, we can verify that, in every stability inequality, an $[x]$ is bounded by a $\dfrac{1}{[a]}[f]$ times a pure number or by a $\dfrac{1}{[b]}[g]$ times a pure number, while a $[y]$ is bounded by a $\dfrac{1}{[b]}[f]$ times a pure number or by a $\dfrac{[a]}{[b^2]}[g]$ times a pure number. □

http://www.springer.com/978-3-642-36518-8

Mixed Finite Element Methods and Applications
Boffi, D.; Brezzi, F.; Fortin, M.
2013, XIV, 685 p., Hardcover
ISBN: 978-3-642-36518-8