DECOMPOSITION INTO SUBSPACES PRECONDITIONING: ABSTRACT FRAMEWORK

JAKUB HRNČÍŘ *, IVANA PULTAROVÁ †, AND ZDENĚK STRAKOŠ ‡

Abstract. The paper re-considers the abstract infinite-dimensional framework of operator preconditioning based on decomposition into subspaces. Such framework has been developed in early 90's in the works of Nepomnyaschikh, Matsokin, Oswald, Griebel, Dahmen, Kunoth, Rüde and others, with inspiration from particular applications, e.g. to fictitious domains, additive Schwarz methods, multilevel methods etc. In our exposition we aim at the simplest possible framework that is on one side general and avoids using features that are specific to particular methods or applications formulated in finite-dimensional spaces, and on the other side allows to cover most of the widely used approaches. Motivated by the work of Faber, Manteuffel and Parter published in 1990 we will use the concepts of norm equivalence and spectral equivalence of infinite-dimensional operators, which goes beyond the works mentioned above. Although the emphasize is more on the exposition that uses minor modifications of results published previously than on new results, we believe that the text also clarifies and strengthens several points that might be of general interest. The second part with the subtitle Applications will show how the presented framework is used for comparison of different methods. It will be published separately.

Key words. Decomposition into infinite-dimensional subspaces, operator preconditioning, stable splitting, norm and spectral equivalence of operators, additive Schwarz methods, multilevel methods.

1. Introduction. Numerical solution of boundary value problems formulated via partial differential equations (PDEs) consists of several tightly interconnected steps. First the mathematical model is analyzed, which leads to the appropriate concept of solution of the infinite-dimensional problem, such as the weak solution using the associated function spaces. Then the problem is discretized, giving a finite-dimensional matrix-vector representation, and subsequently an approximate solution of the discretized problem is computed. Although it is of no particular importance in this text, we emphasize that the discretized problem is not solved exactly, apart from trivial cases. In solving large discretized problems, an approximate solution is typically computed iteratively. In order to ensure computational efficiency (in the sense of computing time or energy consumption), the discretized problem is typically transformed into a problem that is easier to solve via the given iterative process. Such transformation is historically called preconditioning.

In the recent book [31] it is argued that formulation of the infinite-dimensional problem using function spaces, its discretization, preconditioning and computation of an approximate solution using appropriate stopping criteria should be considered as inseparable parts of *a single effort*. As argued by many authors, it is useful to link preconditioning considered in algebraic matrix computations with the infinite-dimensional operator formulation of the problem and with its discretization using the concept of *operator preconditioning*.

1.1. Operator preconditioning. The ideas of operator preconditioning were developed in the 90's independently by several authors; see, e.g., Klawonn [26, 27] and Arnold, Falk and Winther [1, 2]. They were immediately used and further developed in many works. Even before that, a seminal paper by Faber, Manteuffel and Parter [15] analyzed closely related concepts of norm equivalence and spectral equivalence of operators, with references to the early papers of D'Yakonov [13, 14] and Gunn [20, 21]. Another line of development can be represented by the works of Matsokin and Nepomnyaschikh [33, 37, 36, 35], Oswald [39, 40, 41] and Dahmen and Kunoth [11], which are closely related to the multilevel methods and multilevel preconditioning; see the summary and the list of references in the paper by Axelsson and Karátson [5] and in the introduction to Chapter 2 of the book [44]. Classical related references are, e.g.,

*Faculty of Mathematics and Physics, Charles University, Prague, Sokolovská 83, 186 75 Praha 8, Czech Republic. E-mail: hrncir.jakub@gmail.com. Supported by the Grant Agency of the Czech Republic under the contract No. 17-04150J.

† Department of Mathematics, Faculty of Civil Engineering, Czech Technical University in Prague, Thákurova 7, 166 29 Prague 6, Czech Republic. E-mail: ivana.pultarova@cvut.cz. Supported by the Grant Agency of the Czech Republic under the contract No. 17-04150J.

‡Faculty of Mathematics and Physics, Charles University, Prague, Sokolovská 83, 186 75 Praha 8, Czech Republic. E-mail: strakos@karlin.mff.cuni.cz. Supported by the Grant Agency of the Czech Republic under the contract No. 17-04150J.

[6, 7, 8, 16, 19, 25, 32, 47, 49]. This paper will build upon [31] and, motivated by [15], [44, Chapter 2], [43, Section 3], and [42, Chapter 4], it will revisit an abstract formulation of operator preconditioning based on the idea of decomposition of a Hilbert space into a finite number of infinite-dimensional subspaces.

We will now outline the main ideas, with detailed descriptions (including references to the literature) provided further in the text. Using a real (infinite-dimensional) Hilbert space V and its dual $V^\#$ consisting of all linear bounded functionals from V to $\mathbb{R}$, we will consider the functional equation in $V^\#$

$$\mathcal{A}u = b, \quad \text{where} \quad \mathcal{A} : V \to V^\#,\ b \in V^\#,\ u \in V. \tag{1.1}$$

We will assume that $\mathcal{A}$ is linear, bounded, coercive, and self-adjoint

$$\langle \mathcal{A}u, v\rangle = \langle \mathcal{A}v, u\rangle \quad \text{for all } u, v \in V.$$

Some statements given throughout the text allow for a more general setting. By the Lax-Milgram lemma the solution $u \in V$ of (1.1) always exists and it continuously depends on the right-hand side $b \in V^\#$. The given setting represents, e.g., the weak formulation of linear second-order elliptic PDEs that generate self-adjoint operators; see, e.g., [31, Chapters 1–3]. It is worth noting that although the original differential operator is in the classical formulation typically *unbounded*, the representation (1.1) using the appropriate Sobolev spaces uses *bounded* operators $\mathcal{A} : V \to V^\#$ and *bounded* functionals $b \in V^\#$.

Operator preconditioning can in its general form be formulated using the Riesz representation theorem. Considering *any* inner product $(\cdot,\cdot)_* : V \times V \to \mathbb{R}$ on V (that is, in general, different from the inner product $(\cdot,\cdot)_V$ that is associated with the definition of the Hilbert space V) and the associated Riesz map $\tau_* : V^\# \to V$, it is possible to write the problem (1.1) as an equation in the solution space V:

$$\tau_*\mathcal{A}u = \tau_* b, \quad \tau_*\mathcal{A} : V \to V, \quad \tau_* b \in V, \quad u \in V. \tag{1.2}$$

Since τ_* represents an isometry, the operator $\tau_*\mathcal{A}$ on V is bounded and coercive, and it is self-adjoint with respect to the inner product $(\cdot,\cdot)_*$.

Equivalently, operator preconditioning can be formulated using a linear, bounded, coercive, and self-adjoint operator $\mathcal{B} : V \to V^\#$ that defines the $\mathcal{B}$-inner product

$$(\cdot,\cdot)_{\mathcal{B}} : V \times V \to \mathbb{R}, \quad (w, v)_{\mathcal{B}} := \langle \mathcal{B}w, v\rangle \quad \text{for all } w, v \in V, \tag{1.3}$$

where $\langle\cdot,\cdot\rangle : V^\# \times V$ is the duality pairing associated with V and $V^\#$. Using the Riesz map $\tau_{\mathcal{B}}$ determined by $(\cdot,\cdot)_{\mathcal{B}}$ and the easily derived equality $\tau_{\mathcal{B}} = \mathcal{B}^{-1} : V^\# \to V$, the problem (1.2) is written as

$$\mathcal{B}^{-1}\mathcal{A}u = \mathcal{B}^{-1}b, \quad \mathcal{B}^{-1}\mathcal{A} : V \to V, \quad \mathcal{B}^{-1}b \in V, \quad u \in V. \tag{1.4}$$

The question to be addressed next is which relationship between the operators $\mathcal{A}$ and $\mathcal{B}$ can ensure that the transformed (preconditioned) problem (1.4) can be easily solved by a particular iterative method.

1.2. Norm and spectral equivalence, condition and spectral number. This question is, in general, very difficult to handle. For stationary iterative methods (and, more generally, for methods based on contraction) the question can be addressed by an appropriate single-number characteristic, such as the condition number. This is also where the term *preconditioning* founds its origin. For highly nonlinear iterations such as Krylov subspace methods, any single-number characteristic is insufficient for describing convergence behavior and its use can be highly misleading; see, e.g., [31, Chapter 11], [30, Chapter 5], and [18]. In relation to the abstract Schwarz theory the point is made very clear in [42, Section 4.1, pp. 83–84]. Nevertheless, single-number characteristics can even in such cases be useful as first indicators, and for powerful preconditioners it can even provide the desired information whenever the guaranteed number of the resulting iterations is very small. We are well-aware of the limitations of single-number characteristic descriptions, as can be documented by the front cover of the monograph [30] where the presented figure symbolizes several misconceptions related to condition number bounds for the conjugate gradient method widespread in literature. For the reasons mentioned above we nevertheless use single

number characteristics throughout this paper. In particular, we will use the concepts of norm equivalence and spectral equivalence of operators as presented in [15], and the related condition number and spectral number characteristics of the preconditioned operators. The presented abstract framework goes therefore beyond, e.g., [45, Section 4] and other works where the investigation boils down to the study of spectral equivalence.

Consider the operators $\mathcal{A}$, $\mathcal{B}$ given above. The operators $\mathcal{A}$ and $\mathcal{B}$ are called $V^{\#}$-norm equivalent on V if there exist constants $0 < \alpha \leq \beta < \infty$ such that

$$\alpha \leq \frac{\|\mathcal{A}w\|_{V^{\#}}}{\|\mathcal{B}w\|_{V^{\#}}} \leq \beta, \quad \text{for all } \ w \in V, w \neq 0, \tag{1.5}$$

and they are called spectrally equivalent on V if there exist constants $0 < \gamma \leq \delta < \infty$ such that

$$\gamma \leq \frac{\langle \mathcal{A}w, w\rangle}{\langle \mathcal{B}w, w\rangle} \leq \delta, \quad \text{for all } \ w \in V, w \neq 0, \tag{1.6}$$

see [15, Section 1.1, relation (1.16) and Section 1.2, relation (1.20)]. If α is close to β respectively γ is close to δ, then (1.5) respectively (1.6) represent a strong (geometric) relationship between the operators $\mathcal{A}$ and $\mathcal{B}$, and we can expect that this will positively affect properties of the preconditioned operator $\mathcal{B}^{-1}\mathcal{A}$. Such properties are in literature on operator preconditioning typically characterized by the *condition number*

$$\kappa(\mathcal{B}^{-1}\mathcal{A}) := \|\mathcal{B}^{-1}\mathcal{A}\|_{\mathcal{L}(V,V)}\|\mathcal{A}^{-1}\mathcal{B}\|_{\mathcal{L}(V,V)}. \tag{1.7}$$

Motivated by algebraic preconditioning of linear algebraic systems with finite matrices (see also [15, Section 1.1, in particular relations (1.12) and (1.13)]), we will introduce the *spectral number* of the pair $\mathcal{A}$, $\mathcal{B}$ that is linked with another view to preconditioning (1.1) using the operator $\mathcal{B}$. With the Riesz map $\tau : V^{\#} \to V$ defined by the inner product $(\cdot,\cdot)_V$, $\tau\mathcal{A}$ and $\tau\mathcal{B}$ are linear, bounded and coercive operators on V that are self-adjoint with respect to $(\cdot,\cdot)_V$. Taking the (uniquely determined) square root (see, e.g., [17, Theorem 6.6.4])

$$(\tau\mathcal{B})^{1/2} : V \to V, \tag{1.8}$$

the preconditioned system (1.4) can be rewritten as

$$(\tau\mathcal{B})^{-1/2}\,\tau\mathcal{A}\,(\tau\mathcal{B})^{-1/2}w = (\tau\mathcal{B})^{-1/2}\,\tau b, \tag{1.9}$$

where $w = (\tau\mathcal{B})^{1/2}u$. This substantiates the introduction of the spectral number of the pair $\mathcal{A}$, $\mathcal{B}$, related to the preconditioned operator $Q := (\tau\mathcal{B})^{-1/2}\tau\mathcal{A}(\tau\mathcal{B})^{-1/2} : V \to V$,

$$\hat{\kappa}(\mathcal{A},\mathcal{B}) := \frac{\sup_{z\in V,\,\|z\|_V=1}\left((\tau\mathcal{B})^{-1/2}\tau\mathcal{A}(\tau\mathcal{B})^{-1/2}z, z\right)_V}{\inf_{v\in V,\,\|v\|_V=1}\left((\tau\mathcal{B})^{-1/2}\tau\mathcal{A}(\tau\mathcal{B})^{-1/2}v, v\right)_V} = \frac{\sup_{z\in V,\,\|z\|_V=1}(Qz,z)_V}{\inf_{v\in V,\,\|v\|_V=1}(Qv,v)_V}, \tag{1.10}$$

which is determined by the shortest interval that contains the spectrum of Q, with more details given in the next section. We will also prove that (1.10) can be rewritten in terms of norms as

$$\hat{\kappa}(\mathcal{A},\mathcal{B}) = \frac{\sup_{z\in V,\,\|z\|_V=1}\|Qz\|_V}{\inf_{v\in V,\,\|v\|_V=1}\|Qv\|_V}, \tag{1.11}$$

which does not seem entirely obvious and the proof does not seem to be present in literature (see Theorem 2.1 in Section 2 and its proof given in Appendix, in particular relations (7.4) and (7.6)).

We point out that the condition number $\kappa(\mathcal{B}^{-1}\mathcal{A})$ should not be confused with the spectral number $\hat{\kappa}(\mathcal{A},\mathcal{B})$. Since $\mathcal{B}^{-1}\mathcal{A}$ is not generally self-adjoint, there is no simple relationship between these two characteristics. Even more important, it should be emphasized that the concepts of norm and spectral equivalence of *infinite-dimensional* operators are not equivalent. For finite matrices norm equivalence deals with the singular values while the spectral equivalence with the eigenvalues (or, more generally, with

the field of values). For linear compact infinite-dimensional operators that are positive and self-adjoint the norm equivalence implies spectral equivalence but not vice versa; see [15]. For more general setting the relationship between these two concepts seems unresolved; see [23]. Even within the setting with linear, bounded, coercive and self-adjoint operators it is useful to investigate both concepts[1]. This allows generalizations that can be used in a wider context.

In this paper we will not investigate in full the relationship between (1.5), (1.6), (1.7), and (1.10). There seem to be much to be done in that direction and therefore such investigation is beyond the scope of this text; see [23]. We will use (1.5) and (1.6) for stating some basic results about (1.7) and (1.10).

1.3. Decomposition into subspaces. The outlined setting will be used for the description of preconditioners based on decomposing the Hilbert spaces V and $V^\#$ into a *finite number of infinite-dimensional* subspaces. This enables construction of preconditioners for each individual subspaces with using them subsequently for assembling the global preconditioner.

The bounds derived using the norm and spectral equivalence on the infinite-dimensional operator level are *independent of any discretization* (see, e.g., [22]) and, as reproduced below, they carry over (in the norm equivalence case with an additional technical term depending on the discretization basis) to Galerkin discretizations using finite-dimensional subspaces. This is important because the bounds on convergence of iterative methods for solving algebraic problems that can be subsequently developed are automatically independent of discretization (apart from the technicality mentioned above). In most of the published literature the investigation proceeds from particular discretizations and preconditioning, and the generality of the bounds has to be proved for each approach separately. In other words, infinite-dimensional results are essentially identified with their finite-dimensional applications, and the information on whether the spaces are finite or infinite-dimensional is not used. Infinite-dimensional results are applied directly with the constants associated with the finite dimensional problems. Therefore it can be more difficult to see whether the possible ill-conditioning is negatively influenced by the inappropriate discretization.

The presented setting is minimalistic in assumptions yet it covers many traditional approaches. It can be used beyond the finite element method (FEM) and structures. While using infinite-dimensional Hilbert spaces is essential, we do not consider decompositions to infinite many subspaces. The reason is twofold. First, no practical method used in computations can benefit from such setting. Second, even from the purely mathematical point of view we consider infinite decompositions artificial with no substantial mathematical contribution. The difficulties of infinite decompositions in comparison with finite decompositions are in literature simply resolved by strong assumptions on unconditional convergence of the associated infinite series.

In the presented general infinite-dimensional setting we avoid features such as fictitious spaces, additional mappings and projections, which helps in simplicity of the whole exposition and clarity of the statements. The fictitious space lemma from [33], [37, Lemma 2.3], [35, Lemma 2.2] that is used, e.g., as a base of derivation in [42, Chapter 4], within our notation states the equivalence of (1.6) with (see [15, 23])

$$\gamma \le \frac{\langle f, \mathcal{B}^{-1} f\rangle}{\langle f, \mathcal{A}^{-1} f\rangle} \le \delta, \quad \text{for all } f \in V^\#,\ f \neq 0, \tag{1.12}$$

i.e., the operators $\mathcal{A}$ and $\mathcal{B}$ given above are spectrally equivalent with the constants $\gamma \le \delta$ if and only if the inverses $\mathcal{B}^{-1}$ and $\mathcal{A}^{-1}$ are spectrally equivalent with the same constants. As mentioned above, the simplicity of the chosen framework and notation will help in the subsequent comparison of various approaches.

1.4. Structure of the paper. Section 2 presents the description of the basic setting and notation. Section 3 recalls the concept of operator preconditioning and gives the bounds on the condition number

[1] Spectral equivalence (1.6) is in literature called also *norm equivalence*, which refers to the fact that for $\mathcal{A}$, $\mathcal{B}$ as above it represents estimating the ratio of the energy norms defined by the operators $\mathcal{A}$ and $\mathcal{B}$. Without making a distinction from the concept of norm equivalence (1.5), this ambiguity of notation can lead to further misunderstandings.

and the spectral number of the infinite-dimensional preconditioned operator. It is pointed out in detail that the condition number of the preconditioned problem should in general be distinguished from its spectral number. Section 4 gives consequences for the matrix formulations of the discretized problem when the general Galerkin discretization is used. Abstract splitting-based preconditioning is described and investigated in Section 5. This section also presents error bounds based on the residual of the preconditioned problem and on the locally preconditioned residuals. The link to the well-known context of stable splitting is briefly outlined in Section 6. The paper closes with conclusions. The appendix presents the proof of the characterization of the coercivity constant of the operator via the norm of its inverse, which seems to be absent in the literature. We have chosen the proof that illustrates the difference between the finite-dimensional and infinite-dimensional setting.

Within the paper we consider linear operators on real Hilbert spaces (i.e. real complete inner product spaces). Whenever the results on the infinite-dimensional operators $\mathcal{A}$ and $\mathcal{B}$ presented in this paper are linked with the results on their finite-dimensional analogues (matrices) that arise from discretization, it is understood that within our setting $\mathcal{A}$ and $\mathcal{B}$ are bounded operators on infinite-dimensional Hilbert spaces that have bounded inverses. Therefore, by standard functional analysis results (see, e.g. [31, p. 63], [3, p. 282], [28], [10, p. 174], [12, p. 486], and [4, p. 98]), $\mathcal{A}$ and $\mathcal{B}$ cannot be considered as limits of their discretized counterparts in any norm (a sequence of compact operators can converge in norm only to a compact operator).

For algebraic vectors $\mathbf{v}$ respectively for matrices $\mathbf{A}$ we will always denote by $\|\mathbf{v}\|$ respectively $\|\mathbf{A}\|$ the Euclidean norm respectively the associated induced (operator) matrix norm equal to the largest singular value of the matrix $\mathbf{A}$.

2. Basic setting. The following notation is mostly adopted from [31]. Let V be a real Hilbert space with the inner product $(\cdot,\cdot)_V : V \times V \to \mathbb{R}$ and the associated norm $\|\cdot\|_V := \sqrt{(\cdot,\cdot)_V}$. Let further $V^\#$ denote the dual space of bounded (continuous) linear functionals on V with the duality pairing

$$\langle \cdot,\cdot \rangle : V^\# \times V \to \mathbb{R} \tag{2.1}$$

and the dual norm

$$\|f\|_{V^\#} = \sup_{v\in V,\, \|v\|_V=1} \langle f, v\rangle.$$

The Riesz representation theorem associated with the inner product $(\cdot,\cdot)_V$ provides an isometric isomorphism between V and $V^\#$ given through the *Riesz map* $\tau : V^\# \to V$. For each $f \in V^\#$ there exists a unique $\tau f \in V$ such that

$$(\tau f, v)_V := \langle f, v\rangle \quad \text{for all } v \in V, \tag{2.2}$$

with

$$\|\tau f\|_V = \|f\|_{V^\#}. \tag{2.3}$$

Throughout this text we will consider the equation (1.1) or, equivalently,

$$\text{to find } u \in V: \quad \langle \mathcal{A}u, v\rangle = \langle b, v\rangle \quad \text{for all } \ v \in V. \tag{2.4}$$

In terms of an associated symmetric bounded (continuous) bilinear form

$$a(\cdot,\cdot) : V \times V \to \mathbb{R}, \qquad a(u,v) := \langle \mathcal{A}u, v\rangle \quad \text{for all } \ u, v \in V \tag{2.5}$$

the equation (2.4) is expressed as

$$\text{to find } u \in V: \quad a(u,v) = \langle b, v\rangle \quad \text{for all } \ v \in V. \tag{2.6}$$

As mentioned above, $\mathcal{A}$ is assumed to be linear, bounded, coercive and self-adjoint, with the associated boundedness and coercivity constants defined as

$$C_{\mathcal{A}} := \sup_{v\in V,\, \|v\|_V=1} \|\mathcal{A}v\|_{V^\#} < \infty, \tag{2.7}$$

and

$$c_{\mathcal{A}} := \inf_{v\in V,\, \|v\|_V=1} \langle \mathcal{A}v, v\rangle > 0; \tag{2.8}$$

note that under the given assumptions $\mathcal{A}$ represents an isomorfism between V and $V^\#$ (by the Lax-Milgram theorem) and therefore $\mathcal{A}^{-1}$ exists and represents an isomorfism between $V^\#$ and V. Obviously

$$\begin{aligned} &a(v,v) \geq c_{\mathcal{A}}\|v\|_V^2 \quad \text{for all } \ v\in V,\\ &|a(w,v)| \leq C_{\mathcal{A}}\|w\|_V\|v\|_V \quad \text{for all } \ w,v\in V. \end{aligned}$$

We will further use well known results from the spectral theory of self-adjoint operators in Hilbert spaces; see, e.g. [17, Section 6.5]. Because they are formulated (using our notation) for the operators from V to V, we will use them for the operator $\tau\mathcal{A}$. From the self-adjointness of $\mathcal{A}$ with respect to the dual map $\langle\cdot,\cdot\rangle$ we deduce the self-adjointness of $\tau\mathcal{A}$ with respect to the inner product $(\cdot,\cdot)_V$, and from the fact that τ is an isometric isomorfism from $V^\#$ to V we have

$$\sup_{u\in V,\, \|u\|_V=1} \|\tau\mathcal{A}u\|_V = \|\tau\mathcal{A}\|_{\mathcal{L}(V,V)} = \|\mathcal{A}\|_{\mathcal{L}(V,V^\#)} = \sup_{u\in V,\, \|u\|_V=1} \|\mathcal{A}u\|_{V^\#}. \tag{2.9}$$

The coercivity of $\mathcal{A}$ allows us to restrict further considerations regarding the spectrum of $\tau\mathcal{A}$ to the positive part of the real line. The spectrum of $\tau\mathcal{A}$ lies in the closed interval $[m_{\mathcal{A}}, M_{\mathcal{A}}]$,

$$0 < m_{\mathcal{A}} := \inf_{u\in V,\, \|u\|_V=1} \langle \mathcal{A}u, u\rangle \leq \langle \mathcal{A}u, u\rangle = (\tau\mathcal{A}u, u)_V \leq M_{\mathcal{A}} := \sup_{u\in V,\, \|u\|_V=1} \langle \mathcal{A}u, u\rangle. \tag{2.10}$$

Moreover, the lower bound $m_{\mathcal{A}}$ and the upper bound $M_{\mathcal{A}}$ belong to the spectrum of the operator $\tau\mathcal{A}$ but they need not be eigenvalues of $\tau\mathcal{A}$; see [17, Theorem 6.5.9].

It is worth noticing that while the coercivity constant $c_{\mathcal{A}}$ in (2.8) is expressed as the lower extremal point of the spectral interval determined by (2.10), i.e. $c_{\mathcal{A}} = m_{\mathcal{A}}$, the boundedness constant $C_{\mathcal{A}}$ is expressed in terms of the norms $C_{\mathcal{A}} = \|\tau\mathcal{A}\|_{\mathcal{L}(V,V)} = \|\mathcal{A}\|_{\mathcal{L}(V,V^\#)}$. We will therefore complete the description by relating $C_{\mathcal{A}}$ to the upper extremal point $M_{\mathcal{A}}$ in (2.10) and by relating $c_{\mathcal{A}}$ to the norm of the inverse operator $\|\mathcal{A}^{-1}\|_{\mathcal{L}(V^\#,V)}$. This relationship is used in literature but we have not been able to locate its proof (it is not difficult but it does not seem to be a one-line observation). The statement is formulated as the following theorem. Its proof is included in Appendix.

THEOREM 2.1. *Let $\mathcal{A} : V \to V^\#$ be a linear, bounded, coercive and self-adjoint operator. Using the standard definition of the operator norm, the boundedness constant $C_{\mathcal{A}}$ and the coercivity constant $c_{\mathcal{A}}$ can be expressed as*

$$C_{\mathcal{A}} = \|\mathcal{A}\|_{\mathcal{L}(V,V^\#)} = \sup_{u\in V,\, \|u\|_V=1} \langle \mathcal{A}u, u\rangle = M_{\mathcal{A}}, \tag{2.11}$$

$$c_{\mathcal{A}} = m_{\mathcal{A}} = \inf_{v\in V,\, \|v\|_V=1} \langle \mathcal{A}v, v\rangle = \frac{1}{\sup_{f\in V^\#,\, \|f\|_{V^\#}=1} \|\mathcal{A}^{-1}f\|_V} = \left\{\|\mathcal{A}^{-1}\|_{\mathcal{L}(V^\#,V)}\right\}^{-1}. \tag{2.12}$$

Using this result,

$$\|\mathcal{A}^{-1}\|^{-1}_{\mathcal{L}(V^\#,V)}\|v\|_V^2 \leq a(v,v) \leq \|\mathcal{A}\|_{\mathcal{L}(V,V^\#)}\|v\|_V^2 \quad \text{for all } \ v\in V. \tag{2.13}$$

Now consider a linear, bounded, coercive, and self-adjoint operator $\mathcal{B} : V \to V^\#$ that will play within our setting the role of a $\mathcal{B}$-preconditioner for the functional equation (1.1), with $C_{\mathcal{B}}$ and $c_{\mathcal{B}}$ defined analogously to (2.7) and (2.8), respectively. Using the operator $\mathcal{B}$, we introduce the $\mathcal{B}$-inner product (1.3)

$$(\cdot,\cdot)_{\mathcal{B}} : V\times V \to \mathbb{R}, \quad (w,v)_{\mathcal{B}} := \langle \mathcal{B}w, v\rangle \quad \text{for all } w,v\in V$$

and the associated Riesz map

$$\tau_{\mathcal{B}} : V^{\#} \to V, \quad f \in V^{\#} \mapsto \tau_{\mathcal{B}} f \in V$$

defined by

$$(\tau_{\mathcal{B}} f, v)_{\mathcal{B}} := \langle f, v \rangle \quad \text{for all } f \in V^{\#}, v \in V. \tag{2.14}$$

Using this and the definition of the $\mathcal{B}$-inner product,

$$(\tau_{\mathcal{B}} f, v)_{\mathcal{B}} = \langle \mathcal{B}\tau_{\mathcal{B}} f, v \rangle = \langle f, v \rangle \ \Big(= (\tau f, v)_V \Big),$$

and therefore the Riesz map $\tau_{\mathcal{B}}$ associated with $\mathcal{B}$ is given simply by

$$\tau_{\mathcal{B}} = \mathcal{B}^{-1} : V^{\#} \to V. \tag{2.15}$$

3. Norm and spectral equivalence in operator preconditioning. Operator preconditioning can be introduced in several ways. We prefer using the relationship with the Riesz map. Considering any inner product $(.,.)_* : V \times V \to \mathbb{R}$ and the associated Riesz map $\tau_* : V^{\#} \to V$ defined by

$$(\tau_* f, v)_* := \langle f, v \rangle \quad \text{for all } v \in V,$$

the formulation (2.4) of (1.1)

$$\langle \mathcal{A}u - b, v \rangle = 0 \quad \text{for all } v \in V$$

(the weak formulation of the PDE problem) can be equivalently written as

$$(\tau_*(\mathcal{A}u) - \tau_* b, v)_* = 0 \quad \text{for all } v \in V,$$

and, consequently, as *transformation* of the equation $\mathcal{A}u = f$ in the space $V^{\#}$ of bounded linear functionals on V into the equation in the solution space V,

$$\tau_* \mathcal{A}\, u = \tau_* b, \quad \tau_* \mathcal{A} : V \to V, \quad u \in V, \quad \tau_* b \in V. \tag{3.1}$$

This transformation is called *operator preconditioning*. It can motivate or directly lead to the construction of acceleration techniques used in order to improve the behavior of iterative methods for solving associated discretized problems.

With the choice of the inner product $(.,.)_* = (.,.)_{\mathcal{B}}$ determined via the operator $\mathcal{B}$ as above, the transformed problem (3.1) can simply be written as

$$\mathcal{B}^{-1}\mathcal{A}\, u = \mathcal{B}^{-1} b, \quad \mathcal{B}^{-1}\mathcal{A} : V \to V, \quad u \in V, \quad \mathcal{B}^{-1} b \in V, \tag{3.2}$$

which resembles the standard algebraic preconditioning of linear algebraic systems. It is worth recalling in this context the bounds on the condition number (1.7)[2]

$$\kappa(\mathcal{B}^{-1}\mathcal{A}) := \|\mathcal{B}^{-1}\mathcal{A}\|_{\mathcal{L}(V,V)} \|\mathcal{A}^{-1}\mathcal{B}\|_{\mathcal{L}(V,V)}.$$

Since

$$\begin{aligned} \|\mathcal{B}^{-1}\mathcal{A}\|_{\mathcal{L}(V,V)} &= \sup_{z \in V,\, \|z\|_V = 1} \|\mathcal{B}^{-1}\mathcal{A}z\|_V = \sup_{z \in V,\, \|z\|_V = 1} \left\| \mathcal{B}^{-1} \frac{\mathcal{A}z}{\|\mathcal{A}z\|_{V^{\#}}} \|\mathcal{A}z\|_{V^{\#}} \right\|_V \\ &\le \sup_{f \in V^{\#},\, \|f\|_{V^{\#}} = 1} \|\mathcal{B}^{-1} f\|_V \sup_{z \in V,\, \|z\|_V = 1} \|\mathcal{A}z\|_{V^{\#}} = \frac{C_{\mathcal{A}}}{c_{\mathcal{B}}} \end{aligned} \tag{3.3}$$

[2]We point out that in the literature motivated by preconditioning, the condition number $\kappa(\mathcal{B}^{-1}\mathcal{A})$ is often confused with the spectral number $\hat{\kappa}(\mathcal{A}, \mathcal{B})$; see (1.7) and (1.10).

and, analogously,

$$\|\mathcal{A}^{-1}\mathcal{B}\|_{\mathcal{L}(V,V)} \le \frac{C_{\mathcal{B}}}{c_{\mathcal{A}}}, \tag{3.4}$$

we get an upper bound

$$\kappa(\mathcal{B}^{-1}\mathcal{A}) \le \frac{C_{\mathcal{A}}}{c_{\mathcal{B}}}\frac{C_{\mathcal{B}}}{c_{\mathcal{A}}} = \kappa(\mathcal{A})\kappa(\mathcal{B}). \tag{3.5}$$

THEOREM 3.1 (Norm equivalence and condition number). *Assuming that the linear, bounded, coercive and self-adjoint operators $\mathcal{A}$ and $\mathcal{B}$ are $V^{\#}$-*norm equivalent *on V, i.e. there exist $0 < \alpha \le \beta < \infty$ such that*

$$\alpha \le \frac{\|\mathcal{A}w\|_{V^{\#}}}{\|\mathcal{B}w\|_{V^{\#}}} \le \beta, \quad \textit{for all } \ w \in V, w \neq 0, \tag{3.6}$$

then

$$\|\mathcal{B}^{-1}\mathcal{A}\|_{\mathcal{L}(V,V)} \le \beta, \tag{3.7}$$

$$\|\mathcal{A}^{-1}\mathcal{B}\|_{\mathcal{L}(V,V)} \le \frac{1}{\alpha}. \tag{3.8}$$

Consequently,

$$\kappa(\mathcal{B}^{-1}\mathcal{A}) := \|\mathcal{B}^{-1}\mathcal{A}\|_{\mathcal{L}(V,V)}\|\mathcal{A}^{-1}\mathcal{B}\|_{\mathcal{L}(V,V)} \le \frac{\beta}{\alpha}. \tag{3.9}$$

Proof. For the Riesz map τ defined by (2.2) we have, using (3.6) and (2.3), that

$$\alpha \le \frac{\|\tau\mathcal{A}w\|_V}{\|\tau\mathcal{B}w\|_V} \le \beta, \quad \text{for all } \ w \in V, w \neq 0. \tag{3.10}$$

Substituting $w = (\tau\mathcal{A})^{-1}u$ and $w = (\tau\mathcal{B})^{-1}v$, we get

$$\alpha \le \frac{\|u\|_V}{\|\tau\mathcal{B}(\tau\mathcal{A})^{-1}u\|_V} \le \beta \quad \text{and} \quad \alpha \le \frac{\|\tau\mathcal{A}(\tau\mathcal{B})^{-1}v\|_V}{\|v\|_V} \le \beta, \tag{3.11}$$

respectively, for all $u, v \in V$, $u \neq 0$, $v \neq 0$, and thus

$$\|\tau\mathcal{B}(\tau\mathcal{A})^{-1}\|_{\mathcal{L}(V,V)} \le \frac{1}{\alpha} \quad \text{and} \quad \|\tau\mathcal{A}(\tau\mathcal{B})^{-1}\|_{\mathcal{L}(V,V)} \le \beta. \tag{3.12}$$

Denote by Q^* the adjoint operator to $Q : V \to V$; and recall that $\|Q^*\|_{\mathcal{L}(V,V)} = \|Q\|_{\mathcal{L}(V,V)}$. From the self-adjointness of $\tau\mathcal{A}$ and $(\tau\mathcal{B})^{-1}$ we have for all $u, v \in V$,

$$(((\tau\mathcal{B})^{-1}\tau\mathcal{A})^*u, v)_V = ((\tau\mathcal{B})^{-1}\tau\mathcal{A}v, u)_V = ((\tau\mathcal{B})^{-1}u, \tau\mathcal{A}v)_V = (v, \tau\mathcal{A}(\tau\mathcal{B})^{-1}u)_V = (\tau\mathcal{A}(\tau\mathcal{B})^{-1}u, v)_V,$$

and thus $((\tau\mathcal{B})^{-1}\tau\mathcal{A})^* = \tau\mathcal{A}(\tau\mathcal{B})^{-1}$, which results in

$$\|(\tau\mathcal{B})^{-1}\tau\mathcal{A}\|_{\mathcal{L}(V,V)} = \|((\tau\mathcal{B})^{-1}\tau\mathcal{A})^*\|_{\mathcal{L}(V,V)} = \|\tau\mathcal{A}(\tau\mathcal{B})^{-1}\|_{\mathcal{L}(V,V)}. \tag{3.13}$$

Similarly,

$$\|(\tau\mathcal{A})^{-1}\tau\mathcal{B}\|_{\mathcal{L}(V,V)} = \|\tau\mathcal{B}(\tau\mathcal{A})^{-1}\|_{\mathcal{L}(V,V)}. \tag{3.14}$$

Considering an arbitrary $w \in V$, $w \neq 0$, (3.13) and (3.11), we get

$$\frac{\|\mathcal{B}^{-1}\mathcal{A}w\|_V}{\|w\|_V} = \frac{\|\mathcal{B}^{-1}\tau^{-1}\tau\mathcal{A}w\|_V}{\|w\|_V} = \frac{\|(\tau\mathcal{B})^{-1}\tau\mathcal{A}w\|_V}{\|w\|_V} \le \beta, \tag{3.15}$$

which proves (3.7). Similarly, for arbitrary $w \in V$, $w \neq 0$, using (3.14) and (3.11) we get

$$\frac{\|\mathcal{A}^{-1}\mathcal{B}w\|_V}{\|w\|_V} = \frac{\|\mathcal{A}^{-1}\tau^{-1}\tau\mathcal{B}w\|_V}{\|w\|_V} = \frac{\|(\tau\mathcal{A})^{-1}\tau\mathcal{B}w\|_V}{\|w\|_V} \leq \frac{1}{\alpha}, \tag{3.16}$$

which proves (3.8). Relation (3.9) then trivially follows. □

For β close to α the bound (3.9) proves that the condition number $\kappa(\mathcal{B}^{-1}\mathcal{A})$ is small irrespectively of the values of the constants $c_{\mathcal{A}}$, $C_{\mathcal{A}}$, $c_{\mathcal{B}}$ and $C_{\mathcal{B}}$.

COROLLARY 3.2. *Inequalities (3.7) and (3.8) in Theorem 3.1 mean*

$$\frac{\|\mathcal{B}^{-1}\mathcal{A}v\|_V}{\|v\|_V} \leq \beta, \quad \frac{\|\mathcal{A}^{-1}\mathcal{B}w\|_V}{\|w\|_V} \leq \frac{1}{\alpha}, \quad \textit{for all } v, w \in V,\ v \neq 0,\ w \neq 0.$$

Substituting $v = \mathcal{A}^{-1}f$ and $w = \mathcal{B}^{-1}g$, we get

$$\frac{\|\mathcal{B}^{-1}f\|_V}{\|\mathcal{A}^{-1}f\|_V} \leq \beta, \quad \frac{\|\mathcal{A}^{-1}g\|_V}{\|\mathcal{B}^{-1}g\|_V} \leq \frac{1}{\alpha}, \quad \textit{for all } f, g \in V^{\#},\ f \neq 0,\ g \neq 0$$

or, equivalently

$$\alpha \leq \frac{\|\mathcal{B}^{-1}f\|_V}{\|\mathcal{A}^{-1}f\|_V} \leq \beta, \quad \textit{for all } f \in V^{\#},\ f \neq 0. \tag{3.17}$$

We have just shown that (3.6) implies (3.17). Analogously, (3.17) implies (3.6). Thus the $V^{\#}$-norm equivalence of $\mathcal{A}$ and $\mathcal{B}$ on V with constants α and β in the form (3.6) is equivalent to the V-norm equivalence of $\mathcal{B}^{-1}$ and $\mathcal{A}^{-1}$ on $V^{\#}$ with the same constants in the form (3.17).

REMARK 3.1. *The self-adjointness of $\mathcal{A}$ and $\mathcal{B}$ is used in the proof only for the commutativity argument in (3.15). Theorem 3.1 and Corollary 3.2 therefore hold also for commuting non self-adjoint operators $\mathcal{A}, \mathcal{B} : V \to V^{\#}$ (i.e., within our setting, satisfying $\tau\mathcal{A}\tau\mathcal{B} = \tau\mathcal{B}\tau\mathcal{A}$) that are bounded and coercive.*

THEOREM 3.3 (Spectral equivalence and spectral number). *Assuming that the operators $\mathcal{A}$ and $\mathcal{B}$ are* spectrally equivalent *on V, i.e. there exist $0 < \gamma \leq \delta < \infty$ such that*

$$\gamma \leq \frac{\langle \mathcal{A}w, w\rangle}{\langle \mathcal{B}w, w\rangle} \leq \delta, \quad \textit{for all } \ w \in V, w \neq 0, \tag{3.18}$$

then

$$\hat{\kappa}(\mathcal{A}, \mathcal{B}) := \frac{\sup_{z\in V,\, \|z\|_V=1} \left((\tau\mathcal{B})^{-1/2}\tau\mathcal{A}(\tau\mathcal{B})^{-1/2}z, z\right)_V}{\inf_{v\in V,\, \|v\|_V=1} \left((\tau\mathcal{B})^{-1/2}\tau\mathcal{A}(\tau\mathcal{B})^{-1/2}v, v\right)_V} \leq \frac{\delta}{\gamma}. \tag{3.19}$$

Proof. From (3.18) we have for all $w \in V$, $w \neq 0$

$$\gamma \leq \frac{(\tau\mathcal{A}w, w)_V}{(\tau\mathcal{B}w, w)_V} \leq \delta. \tag{3.20}$$

For $\tau\mathcal{B} : V \to V$ consider the uniquely determined linear, bounded, coercive and self-adjoint square root $(\tau\mathcal{B})^{1/2} : V \to V$ such that $(\tau\mathcal{B})^{1/2}(\tau\mathcal{B})^{1/2} = \tau\mathcal{B}$. Thus $(\tau\mathcal{B}w, w)_V = ((\tau\mathcal{B})^{1/2}w, (\tau\mathcal{B})^{1/2}w)_V$ for $w \in V$. Substituting $w = (\tau\mathcal{B})^{-1/2}v$ in (3.20), we get for all $v \in V$, $v \neq 0$

$$\gamma \leq \frac{(\tau\mathcal{A}(\tau\mathcal{B})^{-1/2}v, (\tau\mathcal{B})^{-1/2}v)_V}{(v, v)_V} \leq \delta$$

and, using the self-adjointness of $(\tau\mathcal{B})^{-1/2}$

$$\gamma \leq \frac{((\tau\mathcal{B})^{-1/2}\tau\mathcal{A}(\tau\mathcal{B})^{-1/2}v, v)_V}{(v, v)_V} \leq \delta.$$

This leads to

$$\sup_{z\in V,\,\|z\|_V=1}\Big((\tau\mathcal{B})^{-1/2}\tau\mathcal{A}(\tau\mathcal{B})^{-1/2}z,z\Big)_V\le\delta,$$
$$\inf_{v\in V,\,\|v\|_V=1}\Big((\tau\mathcal{B})^{-1/2}\tau\mathcal{A}(\tau\mathcal{B})^{-1/2}v,v\Big)_V\ge\gamma,$$

which yields (3.19). □

We note that within our setting we have trivially

$$\frac{c_{\mathcal{A}}}{C_{\mathcal{B}}}\le\frac{\langle\mathcal{A}w,w\rangle}{\langle\mathcal{B}w,w\rangle}\le\frac{C_{\mathcal{A}}}{c_{\mathcal{B}}},\quad\text{for all } w\in V,\,w\neq0,\tag{3.21}$$

which, however, does not consider a possible link between $\mathcal{A}$ and $\mathcal{B}$, and it can be therefore impractical. In the following section we will examine the condition and the spectral numbers of the preconditioned discretized system matrix that aroses from the general Galerkin discretization without any reference to a specific construction of the discretization basis.

4. Condition and spectral numbers of the matrix representations of discretized operators. In order to perform numerical computations, the problem (1.1) must first be discretized. Using an N-dimensional subspace $V_h\subset V$, the abstract Galerkin discretization looks for the approximation $u_h\in V_h, u_h\approx u\in V$ satisfying

$$\langle\mathcal{A}u_h-b,v\rangle=0\quad\text{for all } v\in V_h\,.\tag{4.1}$$

In other words, the discretized approximation u_h gives the residual $b-\mathcal{A}u_h\in V^\#$ that is orthogonal to the subspace V_h with respect to the duality pairing $\langle\cdot,\cdot\rangle$. This property is called *Galerkin orthogonality*. The same residual restricted to $V_h^\#$ is *identically zero*, which results in the discretized functional equation below. Considering the restriction $\mathcal{A}_h:V_h\to V_h^\#$ of the operator $\mathcal{A}$ such that

$$\langle\mathcal{A}_hw,v\rangle=\quad\langle\mathcal{A}w,v\rangle\quad\text{for all } w,v\in V_h,\tag{4.2}$$

and the restriction $b_h:V_h\to\mathbb{R}$ of the functional b to $V_h^\#$, (4.1) is written as

$$\langle\mathcal{A}_hu_h-b_h,v\rangle=0\quad\text{for all } v\in V_h\tag{4.3}$$

or, in the operator form, as the equation in the N-dimensional functional space

$$\mathcal{A}_hu_h=b_h,\quad u_h\in V_h,\quad b_h\in V_h^\#,\quad\mathcal{A}_h:V_h\to V_h^\#\,.\tag{4.4}$$

Considering further the inner product $(.,.)_{\mathcal{B}}$ and the associated restricted Riesz map $\tau_{\mathcal{B},h}:V_h^\#\to V_h$, we finally get the abstract form of the preconditioned discretized problem

$$\tau_{\mathcal{B},h}\mathcal{A}_hu_h=\tau_{\mathcal{B},h}b_h,\quad u_h\in V_h,\quad b_h\in V_h^\#,\quad\mathcal{A}_h:V_h\to V_h^\#.\tag{4.5}$$

We note that the subscript h is used for convenience of notation in possible mesh-based implementations (using, e.g., the finite element method, where it characterizes the size of the mesh elements). The abstract formulation used here is, however, more general and it is independent of any notion of mesh or mesh-related discretization.

4.1. Matrix representations of the discretized problem. The matrix formulation of the discretized problems is obtained in a standard way. Consider a basis $\Phi_h=(\phi_1,\dots,\phi_N)$ of V_h and the canonical dual basis $\Phi_h^\#=(\phi_1^\#,\dots,\phi_N^\#)$ of $V_h^\#$,[3]

$$\langle\phi_i^\#,\phi_j\rangle=\delta_{ij},\quad i,j=1,\dots,N,\quad\text{or, using matrix notation,}\quad(\Phi_h^\#)^*\Phi_h=\mathbf{I}_N,$$

[3]Here for simplicity of notation we omit the subscript h in the individual basis functions.

where $\mathbf{I}_N$ denotes the $N \times N$ identity matrix. We wish to construct a linear algebraic system

$$\mathbf{M}_h^{-1}\mathbf{A}_h\mathbf{x}_h = \mathbf{M}_h^{-1}\mathbf{b}_h, \quad \mathbf{A}_h, \mathbf{M}_h \in \mathbb{R}^{N\times N},\ \mathbf{x}_h \in \mathbb{R}^N,\ \mathbf{b}_h \in \mathbb{R}^N, \tag{4.6}$$

where $\mathbf{A}_h$ represents the discretized operator $\mathcal{A}_h$, $\mathbf{M}_h^{-1}$ the discretized preconditioner $\tau_{\mathcal{B},h}$, $\mathbf{b}_h$ the discretized right-hand side functional b_h, and $\mathbf{x}_h$ the coordinates of the approximate solution u_h in the basis Φ_h, (recalling that $\mathbf{z}^*$ means the transpose of the vector $\mathbf{z}$)

$$\mathbf{x}_h = (\langle \phi_1^\#, u_h\rangle, \dots, \langle \phi_N^\#, u_h\rangle)^*.$$

This algebraic system is obtained using the following equalities

$$\mathcal{A}_h u_h = \mathcal{A}_h\Phi_h\mathbf{x}_h = \Phi_h^\#\mathbf{A}_h\mathbf{x}_h,$$

where

$$\mathcal{A}_h\Phi_h =: \Phi_h^\#\mathbf{A}_h, \quad \mathbf{A}_h = \Big(a(\phi_j^{(h)}, \phi_i^{(h)})\Big)_{i,j=1,\dots,N} = \Big(\langle \mathcal{A}\phi_j^{(h)}, \phi_i^{(h)}\rangle\Big)_{i,j=1,\dots,N}, \tag{4.7}$$

or, using symbolic notation,

$$\mathbf{A}_h = (\mathcal{A}\Phi_h)^*\Phi_h, \tag{4.8}$$

and

$$\begin{aligned}\tau_{\mathcal{B},h}\mathcal{A}_h u_h = \tau_{\mathcal{B},h}\mathcal{A}_h\Phi_h\mathbf{x}_h = \tau_{\mathcal{B},h}\Phi_h^\#\mathbf{A}_h\mathbf{x}_h = \Phi_h\mathbf{M}_h^{-1}\mathbf{A}_h\mathbf{x}_h,\\ \tau_{\mathcal{B},h}b_h = \tau_{\mathcal{B},h}\Phi_h^\#\mathbf{b}_h = \Phi_h\mathbf{M}_h^{-1}\mathbf{b}_h,\end{aligned}$$

where

$$\tau_{\mathcal{B},h}\Phi_h^\# = \Phi_h\mathbf{M}_h^{-1}, \quad \mathbf{M}_h := \Big(\langle \mathcal{B}\phi_j^{(h)}, \phi_i^{(h)}\rangle\Big)_{i,j=1,\dots,N}, \tag{4.9}$$

or, using symbolic notation,

$$\mathbf{M}_h = (\mathcal{B}\Phi_h)^*\Phi_h. \tag{4.10}$$

Here the representation of the restricted Riesz map $\tau_{\mathcal{B},h}$ is based on the equalities that hold for any N-dimensional vectors $\mathbf{v}$ and $\mathbf{f}$, with $f = \Phi_h^\#\mathbf{f}$, $v = \Phi_h\mathbf{v}$, $\tau_{\mathcal{B},h}\Phi_h^\# = \Phi_h\mathbf{M}_\tau$ for some $\mathbf{M}_\tau \in \mathbb{R}^{N\times N}$,

$$\mathbf{v}^*\mathbf{f} = \langle f, v\rangle = (\tau_{\mathcal{B},h}f, v)_{\mathcal{B}} = (\tau_{\mathcal{B},h}\Phi_h^\#\mathbf{f}, \Phi_h\mathbf{v})_{\mathcal{B}} = (\Phi_h\mathbf{M}_\tau\mathbf{f}, \Phi_h\mathbf{v})_{\mathcal{B}} = \langle \mathcal{B}\Phi_h\mathbf{M}_\tau\mathbf{f}, \Phi_h\mathbf{v}\rangle = \mathbf{v}^*\mathbf{M}_h\mathbf{M}_\tau\mathbf{f},$$

and therefore $\mathbf{M}_\tau = \mathbf{M}_h^{-1}$,

$$\tau_{\mathcal{B},h}\Phi_h^\# = \Phi_h\mathbf{M}_h^{-1}. \tag{4.11}$$

Finally, the preconditioned algebraic system can indeed be written in the form (4.6)

$$\mathbf{M}_h^{-1}\mathbf{A}_h\mathbf{x}_h = \mathbf{M}_h^{-1}\mathbf{b}_h,$$

or, using the factorization $\mathbf{M}_h = \mathbf{M}_h^{1/2}\mathbf{M}_h^{1/2}$, as

$$\mathbf{M}_h{}^{-1/2}\mathbf{A}_h\mathbf{M}_h^{-1/2}(\mathbf{M}_h^{1/2}\mathbf{x}_h) = \mathbf{M}_h^{-1/2}\mathbf{b}_h \tag{4.12}$$

or

$$\widetilde{\mathbf{A}}_{t,h}\widetilde{\mathbf{x}}_h^t = \widetilde{\mathbf{b}}_h^t, \tag{4.13}$$

where

$$\widetilde{\mathbf{A}}_{t,h} := {\mathbf{M}_h}^{-1/2}\mathbf{A}_h\mathbf{M}_h^{-1/2}, \quad \widetilde{\mathbf{x}}_h^t := \mathbf{M}_h^{1/2}\mathbf{x}_h, \quad \widetilde{\mathbf{b}}_h^t := \mathbf{M}_h^{-1/2}\mathbf{b}_h.$$

It is worth noticing that the discretized form of the problem (4.6) allows many different factorizations of $\mathbf{M}_h$. Instead of the square root of the operator $\mathbf{M}_h$, we can consider an arbitrary decomposition $\mathbf{M}_h = \mathbf{L}_h\mathbf{L}_h^*$, which can be more practical computationally. Then we can write

$$\mathbf{A}_{t,h}\mathbf{x}_h^t = \mathbf{b}_h^t, \tag{4.14}$$

having in this case

$${\mathbf{L}_h}^{-1}\mathbf{A}_h(\mathbf{L}_h^*)^{-1}(\mathbf{L}_h^*\mathbf{x}_h) = \mathbf{L}_h^{-1}\mathbf{b}_h, \quad \mathbf{A}_{t,h} := {\mathbf{L}_h}^{-1}\mathbf{A}_h(\mathbf{L}_h^*)^{-1}, \quad \mathbf{x}_h^t := \mathbf{L}_h^*\mathbf{x}_h, \quad \mathbf{b}_h^t := \mathbf{L}_h^{-1}\mathbf{b}_h.$$

Due to

$$\mathbf{L}_h^{-1}\mathbf{M}_h(\mathbf{L}_h^*)^{-1} = \mathbf{L}_h^{-1}\mathbf{M}_h^{1/2}\left(\mathbf{M}_h^{1/2}(\mathbf{L}_h^*)^{-1}\right) = \mathbf{I}_N \tag{4.15}$$

we have

$$\left(\mathbf{L}_h^{-1}\mathbf{M}_h^{1/2}\right)^{-1} = \left(\mathbf{L}_h^{-1}\mathbf{M}_h^{1/2}\right)^*,$$

i.e. $\mathbf{L}_h^{-1}\mathbf{M}_h^{1/2}$ is an orthogonal matrix. For any given $\mathbf{L}_h$ there is therefore an orthogonal transformation

$$\mathbf{M}_h^{1/2} \to \mathbf{M}_h^{1/2}\left(\mathbf{M}_h^{-1/2}\mathbf{L}_h\right) = \mathbf{L}_h$$

from $\mathbf{M}_h^{1/2}$ to $\mathbf{L}_h$.

The transformed system (4.14) can moreover be obtained *mathematically equivalently* (this term is used in order to indicate that mathematical equivalence does not necessarily mean equivalent computational efficiency or accuracy in practical computations) by first orthogonalizing the discretization basis with respect to the $\mathcal{B}$-inner product

$$\Phi_{t,h} = \Phi_h(\mathbf{L}_h^*)^{-1}, \quad \Phi_{t,h}^{\#} = \Phi_h^{\#}\mathbf{L}_h, \quad \Phi_{t,h} = (\phi_1^t, \dots, \phi_N^t), \quad \Phi_{t,h}^{\#} = (\phi_1^{t\#}, \dots, \phi_N^{t\#})$$

which indeed gives (using the symbolic notation $\mathbf{M}_h = (\mathcal{B}\Phi_h)^*\Phi_h$)

$$(\mathcal{B}\Phi_{t,h})^*\Phi_{t,h} = \mathbf{L}_h^{-1}(\mathcal{B}\Phi_h)^*\Phi_h(\mathbf{L}_h^*)^{-1} = \mathbf{L}_h^{-1}\mathbf{M}_h(\mathbf{L}_h^*)^{-1} = \mathbf{I}_N,$$

and subsequently forming the matrix of the algebraic system (4.13) using (4.7) with the basis Φ_h replaced by $\Phi_{t,h}$; cf. [31, Chapter 8].

In summary, there is a deep connection between discretization of the infinite-dimensional problem and preconditioning of the discretized algebraic system. In addition, any algebraic preconditioning can be viewed as orthogonalization of the discretization basis with respect to the appropriate inner product; for details see [31].

4.2. Condition and spectral number of the preconditioned system matrix. The question of the rate of convergence of an iterative method applied to the preconditioned algebraic system (4.13) is typically reduced to estimates based on the condition number of the preconditioned system matrix. We will leave aside the question when such an approach leads to descriptive results and which (more or less restrictive) assumptions must be considered whenever it is applied to practical problems; for a detailed discussion of these topics we refer to [31, Section 5.2 and Chapter 11] and [30, Section 3.5 and Chapter 5]. In the rest of this section we will describe bounds on the condition and spectral numbers of the matrices $\mathbf{A}_{t,h}$ (that include also the special choice $\widetilde{\mathbf{A}}_{t,h}$) and $\mathbf{M}_h^{-1}\mathbf{A}_h$ in terms of the properties of the operators $\mathcal{A}$ and $\mathcal{B}$. The following theorem that generalizes the results from [15, Section 3, in particular Theorem 3.10] is a finite-dimensional analogue to Theorem 3.1.

THEOREM 4.1 (Norm equivalence and condition number). *Consider the assumptions of Theorem 3.1. Let* $\mathbf{S}_h$ *be the Gram matrix of the discretization basis* $\Phi_h = (\phi_1, \dots, \phi_N)$ *of* $V_h \subset V$, $(\mathbf{S}_h)_{ij} = (\phi_i, \phi_j)_V$, *and* $\mathbf{A}_h$, $\mathbf{M}_h$ *be determined by (4.8) and (4.10), respectively. Then the condition number of the matrix* $\mathbf{M}_h^{-1}\mathbf{A}_h$ *is bounded as*

$$\kappa(\mathbf{M}_h^{-1}\mathbf{A}_h) := \|\mathbf{M}_h^{-1}\mathbf{A}_h\|\,\|\mathbf{A}_h^{-1}\mathbf{M}_h\| \le \frac{\beta}{\alpha}\,\kappa(\mathbf{S}_h). \tag{4.16}$$

Proof. For $w = \Phi_h\mathbf{y}$, $\mathbf{y} \in \mathbb{R}^N$, we have

$$\begin{aligned}\|\mathcal{A}w\|_{V^\#} = \|\Phi_h^\#\mathbf{A}_h\mathbf{y}\|_{V^\#} &= \sup_{u\in V_h,\,u\neq 0} \frac{\langle \Phi_h^\#\mathbf{A}_h\mathbf{y}, u\rangle}{\|u\|_V} = \sup_{\mathbf{z}\in\mathbb{R}^N,\,\mathbf{z}\neq 0} \frac{\langle \Phi_h^\#\mathbf{A}_h\mathbf{y}, \Phi_h\mathbf{z}\rangle}{\|\Phi_h\mathbf{z}\|_V}\\ &= \sup_{\mathbf{z}\in\mathbb{R}^N,\,\mathbf{z}\neq 0} \frac{\mathbf{z}^*\mathbf{A}_h\mathbf{y}}{\|\Phi_h\mathbf{z}\|_V} = \sup_{\mathbf{z}\in\mathbb{R}^N,\,\mathbf{z}\neq 0} \frac{\mathbf{z}^*\mathbf{A}_h\mathbf{y}}{\sqrt{\mathbf{z}^*\mathbf{S}_h\mathbf{z}}}.\end{aligned}$$

Setting $\mathbf{z} = \mathbf{S}_h^{-1/2}\mathbf{v}$ and using $(\mathbf{S}_h^{-1/2})^* = \mathbf{S}_h^{-1/2}$ leads to

$$\|\mathcal{A}w\|_{V^\#} = \sup_{\mathbf{v}\in\mathbb{R}^N,\,\mathbf{v}\neq 0} \frac{\mathbf{v}^*\mathbf{S}_h^{-1/2}\mathbf{A}_h\mathbf{y}}{\|\mathbf{v}\|} = \|\mathbf{S}_h^{-1/2}\mathbf{A}_h\mathbf{y}\|.$$

Analogously

$$\|\mathcal{B}w\|_{V^\#} = \|\Phi_h^\#\mathbf{M}_h\mathbf{y}\|_{V^\#} = \|\mathbf{S}_h^{-1/2}\mathbf{M}_h\mathbf{y}\|.$$

Then

$$\begin{aligned}\frac{\beta}{\alpha} &\ge \sup_{w\in V,\,w\neq 0} \frac{\|\mathcal{A}w\|_{V^\#}}{\|\mathcal{B}w\|_{V^\#}} \sup_{v\in V,\,v\neq 0} \frac{\|\mathcal{B}v\|_{V^\#}}{\|\mathcal{A}v\|_{V^\#}}\\ &\ge \sup_{w\in V_h,\,w\neq 0} \frac{\|\mathcal{A}w\|_{V^\#}}{\|\mathcal{B}w\|_{V^\#}} \sup_{v\in V_h,\,v\neq 0} \frac{\|\mathcal{B}v\|_{V^\#}}{\|\mathcal{A}v\|_{V^\#}}\\ &= \sup_{\mathbf{w}\in\mathbb{R}^N,\,\mathbf{w}\neq 0} \frac{\|\mathbf{S}_h^{-1/2}\mathbf{A}_h\mathbf{w}\|}{\|\mathbf{S}_h^{-1/2}\mathbf{M}_h\mathbf{w}\|} \sup_{\mathbf{v}\in\mathbb{R}^N,\,\mathbf{v}\neq 0} \frac{\|\mathbf{S}_h^{-1/2}\mathbf{M}_h\mathbf{v}\|}{\|\mathbf{S}_h^{-1/2}\mathbf{A}_h\mathbf{v}\|}\\ &= \sup_{y\in\mathbb{R}^N,\,y\neq 0} \frac{\|\mathbf{S}_h^{-1/2}\mathbf{A}_h\mathbf{M}_h^{-1}\mathbf{S}_h^{1/2}\mathbf{y}\|}{\|\mathbf{y}\|} \sup_{\mathbf{z}\in\mathbb{R}^N,\,\mathbf{z}\neq 0} \frac{\|\mathbf{S}_h^{-1/2}\mathbf{M}_h\mathbf{A}_h^{-1}\mathbf{S}_h^{1/2}\mathbf{z}\|}{\|\mathbf{z}\|}\\ &= \|\mathbf{S}_h^{-1/2}\mathbf{A}_h\mathbf{M}_h^{-1}\mathbf{S}_h^{1/2}\|\,\|\mathbf{S}_h^{-1/2}\mathbf{M}_h\mathbf{A}_h^{-1}\mathbf{S}_h^{1/2}\|.\end{aligned} \tag{4.17}$$

Since for any $\mathbf{G} \in \mathbf{R}^{N\times N}$ we have

$$\begin{aligned}\|\mathbf{G}\mathbf{S}_h^{1/2}\| = \sup_{\mathbf{w}\in\mathbb{R}^N,\,\mathbf{w}\neq 0} \frac{\|\mathbf{G}\mathbf{S}_h^{1/2}\mathbf{w}\|}{\|\mathbf{w}\|} &= \sup_{\mathbf{w}\in\mathbb{R}^N,\,\mathbf{w}\neq 0} \frac{(\lambda_{\min}(\mathbf{S}_h))^{1/2}\|\mathbf{G}\mathbf{S}_h^{1/2}\mathbf{w}\|}{(\lambda_{\min}(\mathbf{S}_h))^{1/2}\|\mathbf{w}\|}\\ &\ge (\lambda_{\min}(\mathbf{S}_h))^{1/2} \sup_{\mathbf{w}\in\mathbb{R}^N,\,\mathbf{w}\neq 0} \frac{\|\mathbf{G}\mathbf{S}_h^{1/2}\mathbf{w}\|}{\|\mathbf{S}_h^{1/2}\mathbf{w}\|} = (\lambda_{\min}(\mathbf{S}_h))^{1/2}\|\mathbf{G}\|,\end{aligned} \tag{4.18}$$

and, using $\|\mathbf{S}_h^{-1/2}\mathbf{G}\| = \|\mathbf{G}^*\mathbf{S}_h^{-1/2}\|$, we get analogously

$$\|\mathbf{S}_h^{-1/2}\mathbf{G}\| \ge (\lambda_{\max}(\mathbf{S}_h^{-1}))^{1/2}\|\mathbf{G}^*\| = (\lambda_{\max}(\mathbf{S}_h))^{-1/2}\|\mathbf{G}\|. \tag{4.19}$$

Finally, applying (4.18) and (4.19) to (4.17) yields

$$\begin{aligned}\frac{\beta}{\alpha} &\ge \|\mathbf{S}_h^{-1/2}\mathbf{A}_h\mathbf{M}_h^{-1}\mathbf{S}_h^{1/2}\|\,\|\mathbf{S}_h^{-1/2}\mathbf{M}_h\mathbf{A}_h^{-1}\mathbf{S}_h^{1/2}\|\\ &\ge \frac{\lambda_{\min}(\mathbf{S}_h)}{\lambda_{\max}(\mathbf{S}_h)}\,\|\mathbf{A}_h\mathbf{M}_h^{-1}\|\,\|\mathbf{M}_h\mathbf{A}_h^{-1}\|\\ &= \frac{\lambda_{\min}(\mathbf{S}_h)}{\lambda_{\max}(\mathbf{S}_h)}\,\|\mathbf{M}_h^{-1}\mathbf{A}_h\|\,\|\mathbf{A}_h^{-1}\mathbf{M}_h\|\end{aligned}$$

which finishes the proof. □

Using the coordinates in the transformed basis $\Phi_{t,h}$, for any $z \in V_h$ we have the following useful equality

$$\|z\|_{\mathcal{B}}^2 = (z,z)_{\mathcal{B}} = (\Phi_{t,h}\mathbf{z}, \Phi_{t,h}\mathbf{z})_{\mathcal{B}} = \|\mathbf{z}\|^2.$$

We will now turn to the spectral number

$$\hat{\kappa}(\mathbf{A}_h, \mathbf{M}_h) := \frac{\sup_{\mathbf{z}\in\mathbb{R}^N,\, \|\mathbf{z}\|=1} \left(\mathbf{M}_h^{-1/2}\mathbf{A}_h\mathbf{M}_h^{-1/2}\mathbf{z}, \mathbf{z}\right)}{\inf_{\mathbf{v}\in\mathbb{R}^N,\, \|\mathbf{v}\|=1} \left(\mathbf{M}_h^{-1/2}\mathbf{A}_h\mathbf{M}_h^{-1/2}\mathbf{v}, \mathbf{v}\right)} = \frac{\lambda_{\max}(\mathbf{M}_h^{-1}\mathbf{A}_h)}{\lambda_{\min}(\mathbf{M}_h^{-1}\mathbf{A}_h)} = \kappa(\mathbf{A}_{t,h}). \tag{4.20}$$

Clearly, the spectra of the matrices $\mathbf{M}_h^{-1}\mathbf{A}_h$ and $\mathbf{L}_h^{-1}\mathbf{A}_h(\mathbf{L}_h^*)^{-1}$ are identical, and therefore the spectral number $\hat{\kappa}(\mathbf{A}_h, \mathbf{M}_h)$ is determined via the extremal eigenvalues of $\mathbf{L}_h^{-1}\mathbf{A}_h(\mathbf{L}_h^*)^{-1}$. While for the symmetric positive definite matrix the condition number is given as a ratio of extremal eigenvalues, the same is not in general true for the nonsymmetric matrix. Analogously to the derivation in [31, Chapter 8],

$$\begin{aligned}
\hat{\kappa}(\mathbf{A}_h, \mathbf{M}_h) &= \frac{\max_{\|\mathbf{u}\|=1} \mathbf{u}^* \mathbf{A}_{t,h}\mathbf{u}}{\min_{\|\mathbf{v}\|=1} \mathbf{v}^* \mathbf{A}_{t,h}\mathbf{v}} \\
&= \frac{\max_{\|\mathbf{u}\|=1} \mathbf{u}^* \left(\langle \mathcal{A}\phi_j^t, \phi_i^t\rangle\right)_{i,j=1,\dots,N} \mathbf{u}}{\min_{\|\mathbf{v}\|=1} \mathbf{v}^* \left(\langle \mathcal{A}\phi_j^t, \phi_i^t\rangle\right)_{i,j=1,\dots,N} \mathbf{v}} \\
&= \frac{\max_{u\in V_h,\, \|u\|_{\mathcal{B}}=1} \langle \mathcal{A}u, u\rangle}{\min_{v\in V_h,\, \|v\|_{\mathcal{B}}=1} \langle \mathcal{A}v, v\rangle} \\
&= \frac{\langle \mathcal{A}\tilde{u}, \tilde{u}\rangle}{\langle \mathcal{A}\tilde{v}, \tilde{v}\rangle},
\end{aligned} \tag{4.21}$$

where $\tilde{u}, \|\tilde{u}\|_{\mathcal{B}} = 1$ gives the maximum and $\tilde{v}, \|\tilde{v}\|_{\mathcal{B}} = 1$ the minimum, respectively. Since

$$\begin{aligned}
\|\tilde{v}\|_{\mathcal{B}}^2 &= \langle \mathcal{B}\tilde{v}, \tilde{v}\rangle \le C_{\mathcal{B}}\|\tilde{v}\|_V^2, \\
\|\tilde{u}\|_{\mathcal{B}}^2 &= \langle \mathcal{B}\tilde{u}, \tilde{u}\rangle \ge c_{\mathcal{B}}\|\tilde{u}\|_V^2,
\end{aligned}$$

we get

$$\begin{aligned}
\hat{\kappa}(\mathbf{A}_h, \mathbf{M}_h) = \kappa(\mathbf{A}_{t,h}) &= \frac{\langle \mathcal{A}\tilde{u}, \tilde{u}\rangle}{\langle \mathcal{A}\tilde{v}, \tilde{v}\rangle} \\
&= \frac{\|\tilde{u}\|_V^2}{\|\tilde{v}\|_V^2} \frac{\langle \mathcal{A}\tilde{u}/\|\tilde{u}\|_V, \tilde{u}/\|\tilde{u}\|_V\rangle}{\langle \mathcal{A}\tilde{v}/\|\tilde{v}\|_V, \tilde{v}/\|\tilde{v}\|_V\rangle} \\
&\le \frac{C_{\mathcal{B}}}{c_{\mathcal{B}}} \frac{\langle \mathcal{A}z, z\rangle}{\langle \mathcal{A}w, w\rangle} \le \frac{C_{\mathcal{B}}}{c_{\mathcal{B}}} \frac{C_{\mathcal{A}}}{c_{\mathcal{A}}},
\end{aligned} \tag{4.22}$$

where $z = \tilde{u}/\|\tilde{u}\|_V$, $\|z\|_V = 1$, $w = \tilde{v}/\|\tilde{v}\|_V$, $\|w\|_V = 1$. Summarizing, we get independently of the discretization parameter h the following analogue of (3.5):

$$\hat{\kappa}(\mathbf{A}_h, \mathbf{M}_h) = \kappa(\mathbf{A}_{t,h}) \le \kappa(\mathcal{B})\kappa(\mathcal{A}). \tag{4.23}$$

For related statements (in a more general setting) we refer, e.g., to [22, Theorem 2.1 and relation (3.2)]. The following theorem is a finite-dimensional analogue to Theorem 3.3.

THEOREM 4.2 (Spectral equivalence and spectral number). *Consider the assumptions of Theorem 3.3, and* $\mathbf{A}_h$, $\mathbf{M}_h$ *determined by (4.8) and (4.10), respectively. Then the spectral number* $\hat{\kappa}(\mathbf{A}_h, \mathbf{M}_h)$, *which is equal to the condition number of the symmetric matrix* $\mathbf{A}_{t,h} = \mathbf{L}_h^{-1}\mathbf{A}_h(\mathbf{L}_h^*)^{-1}$ *for any* $\mathbf{L}_h$ *such that* $\mathbf{M}_h = \mathbf{L}_h\mathbf{L}_h^*$, *is bounded as*

$$\hat{\kappa}(\mathbf{A}_h, \mathbf{M}_h) = \kappa(\mathbf{A}_{t,h}) \le \frac{\delta}{\gamma}. \tag{4.24}$$

Proof. From (4.21) and (4.15), considering $\|\tilde{u}\|_{\mathcal{B}} = 1$, $\|\tilde{v}\|_{\mathcal{B}} = 1$,

$$\hat{\kappa}(\mathbf{A}_h, \mathbf{M}_h) = \kappa(\mathbf{A}_{t,h}) = \frac{\langle \mathcal{A}\widetilde{u}, \widetilde{u}\rangle}{\langle \mathcal{B}\widetilde{u}, \widetilde{u}\rangle} \frac{\langle \mathcal{B}\widetilde{v}, \widetilde{v}\rangle}{\langle \mathcal{A}\widetilde{v}, \widetilde{v}\rangle} \leq \frac{\delta}{\gamma}, \tag{4.25}$$

yielding the assertion. □

This can give a much stronger bound than (4.23). For related early results that can further illustrate the difference between (4.23) and (4.24) we refer, e.g., to [15] and [48, Sections 4.1 and 4.2].

5. Abstract description of the splitting-based preconditioning. We will now use the operator preconditioning framework of the previous sections in order to describe splitting-based preconditioning. Here we will not consider particular approaches developed for particular problems using various specific assumptions. Following the ideas in [42, Section 4.1], [44, Section 2.1], [43, 19], and the motivation decsribed in Section 1.3, the goal is to present as simple as possible abstract framework that will underline the common basic principles for a variety of different approaches published in literature. For specific problems and using specific assumptions, the abstract framework can be used for deriving properties of specific methods. This can contribute towards easier description of the relationship between various methods and towards their easier comparison.

We will use the setting of the problem (1.1) and (2.4)–(2.6), i.e.

$$\mathcal{A}u = b \tag{5.1}$$

in the functional space $V^{\#}$, or, using the bilinear form,

$$a(u, v) = \langle b, v\rangle \quad \text{for all } v \in V, \quad \langle \mathcal{A}u, v\rangle = a(u, v).$$

We are now going to transform (5.1) into the form (cf. (3.2))

$$\mathcal{M}^{-1}\mathcal{A}u = \mathcal{M}^{-1}b, \quad \mathcal{M}^{-1}\mathcal{A} : V \to V, \quad u \in V, \quad \mathcal{M}^{-1}b \in V, \tag{5.2}$$

where the *preconditioning* $\mathcal{M}$ is constructed using a *decomposition (splitting)* of the space V into a finite[4] collection of (nontrivial) subspaces $\{V_j\}_{j\in J}$ that are *not necessarily nested,* $V_j \subset V$, each complete with respect to its own inner product $(\cdot,\cdot)_j : V_j \times V_j \to \mathbb{R}$ and the associated norm $\|\cdot\|_j$, such that

$$V = \sum_{j\in J} V_j, \quad \text{i.e.,} \quad v = \sum_{j\in J} v_j,\ v_j \in V_j, \quad \text{for all } v \in V. \tag{5.3}$$

For each V_j we can consider its dual $V_j^{\#}$ with the duality pairing identical to (2.1) and the norm $\|\cdot\|_j^{\#}$ induced by $\|\cdot\|_j$. We will assume the continuous embedding $V_j \hookrightarrow V$, see, e.g., [9, Section 6.6]

$$c_{V_j}\|u\|_V^2 \leq \|u\|_j^2 \quad \text{for all } u \in V_j,\ 0 < c_{V_j},\ j \in J. \tag{5.4}$$

For V_j finite-dimensional, (5.4) always holds true (all norms on finite-dimensional V_j are trivially *topologically* equivalent). Thus (5.4) is nontrivial only in the case of V_j (and thus V) infinite-dimensional. Then the assumption (5.4) avoids a possible pathological situation when a converging sequence of elements from $V_j \subset V$ may diverge in V. Moreover, the assumption (5.4) guarantees that any functional from $V^{\#}$ restricted to V_j belongs to $V_j^{\#}$. Indeed, let $f \in V^{\#}$, then

$$\|f\|_j^{\#} = \sup_{u\in V_j,\, u\neq 0} \frac{\langle f, u\rangle}{\|u\|_j} = \sup_{u\in V_j,\, u\neq 0} \frac{\langle f, u\rangle}{\|u\|_V} \frac{\|u\|_V}{\|u\|_j} \leq \frac{1}{\sqrt{c_{V_j}}} \sup_{u\in V,\, u\neq 0} \frac{\langle f, u\rangle}{\|u\|_V} \leq \frac{1}{\sqrt{c_{V_j}}}\|f\|_{V^{\#}}. \tag{5.5}$$

[4]Since this text is motivated by numerical methods and, in particular, by the construction of preconditioning, with no loss of generality it is sufficient to consider splitting of the Hilbert space V into a finite number of subspaces that can be infinite-dimensional. This setting is convenient since it simplifies the exposition of the abstract splitting-based preconditioning. As mentioned in Section 1.3, we also consider a finite number of subspaces, which is the choice fully justified, in our opinion, also from the purely mathematical reason.

The necessity of (5.4) for $V^\# \subset V_j^\#$ is an open question.

The splitting-based preconditioning $\mathcal{M}$ will be composed of the individual preconditionings at the subspaces V_j, $j \in J$. Let $\mathcal{B}_j$ be a linear, bounded, coercive, and self-adjoint operator

$$\mathcal{B}_j : V_j \to V_j^\#, \quad \langle \mathcal{B}_j u, v\rangle = \langle \mathcal{B}_j v, u\rangle \quad \text{for all } \ u, v \in V_j, \tag{5.6}$$

with the associated bilinear form $ß_j : V_j \times V_j \to \mathcal{R}$

$$ß_j(u, v) := \langle \mathcal{B}_j u, v\rangle, \quad \text{for all } u, v \in V_j.$$

Analogously to (2.7), (2.8) and Theorem 2.1, for $j \in J$

$$C_{\mathcal{B}_j} := \sup_{v\in V_j,\, \|v\|_j=1} \|\mathcal{B}_j v\|_j^\# < \infty, \tag{5.7}$$

$$c_{\mathcal{B}_j} := \inf_{v\in V_j,\, \|v\|_j=1} \langle \mathcal{B}_j v, v\rangle = \frac{1}{\sup_{f\in V_j^\#,\, \|f\|_j^\#=1} \|\mathcal{B}_j^{-1} f\|_j} > 0, \tag{5.8}$$

and

$$c_{\mathcal{B}_j} \|u\|_j^2 \le ß_j(u, u) \le C_{\mathcal{B}_j} \|u\|_j^2. \tag{5.9}$$

In other words, $\mathcal{B}_j$ is coercive and bounded on V_j, $j \in J$. The operator $\mathcal{B}_j$ (the bilinear form $ß_j$) defines on V_j the inner product[5]

$$(\cdot,\cdot)_{\mathcal{B}_j} : V_j \times V_j \to \mathbb{R}, \quad (w, v)_{\mathcal{B}_j} := ß_j(w, v) = \langle \mathcal{B}_j w, v\rangle \quad \text{for all } w, v \in V_j, \tag{5.10}$$

with the corresponding Riesz map

$$\tau_{\mathcal{B}_j} : V_j^\# \to V_j, \quad f \in V_j^\# \mapsto \tau_{\mathcal{B}_j} f \in V_j$$

defined by

$$(\tau_{\mathcal{B}_j} f, v)_{\mathcal{B}_j} := \langle f, v\rangle \quad \text{for all } f \in V_j^\#,\, v \in V_j. \tag{5.11}$$

Clearly, analogously to the construction presented in Section 2,

$$(\tau_{\mathcal{B}_j} f, v)_{\mathcal{B}_j} = \langle \mathcal{B}_j \tau_{\mathcal{B}_j} f, v\rangle = \langle f, v\rangle, \quad \text{for all } f \in V_j^\#,\, v \in V_j$$

and therefore

$$\tau_{\mathcal{B}_j} = {\mathcal{B}_j}^{-1} : V_j^\# \to V_j. \tag{5.12}$$

We will now construct a *splitting-based preconditioning* $\mathcal{M}^{-1}$ in (5.2). For any $u \in V$ and $j \in J$ we have

$$\langle \mathcal{A}u, v\rangle = (\mathcal{B}_j^{-1}\mathcal{A}u, v)_{\mathcal{B}_j} \quad \text{for all } v \in V_j,$$

and

$$\langle b, v\rangle = (\mathcal{B}_j^{-1} b, v)_{\mathcal{B}_j} \quad \text{for all } v \in V_j;$$

under the assumption (5.4) we have $V^\# \subset V_j^\#$ and therefore $\mathcal{B}_j^{-1}\mathcal{A}u$ and $\mathcal{B}_j^{-1}b$ are well-defined. Combining the last two equations gives

$$\langle \mathcal{A}u - b, v\rangle = (\mathcal{B}_j^{-1}\mathcal{A}u - \mathcal{B}_j^{-1} b, v)_{\mathcal{B}_j} \quad \text{for all } v \in V_j,$$

[5] Here we do not need the form $ß_i(\cdot,\cdot)$. We introduce this notation for convenience. Part of the literature uses the bilinear form formulation instead of the operator formulation.

and therefore on each subspace V_j, $j \in J$, we can formulate the preconditioned equation

$$\mathcal{B}_j^{-1}\mathcal{A}u = \mathcal{B}_j^{-1}b, \tag{5.13}$$

that must be satisfied by the solution $u \in V$ of $\mathcal{A}u = b$. Consequently, from $\mathcal{A}u = b$ we get

$$\left(\sum_{j\in J}\mathcal{B}_j^{-1}\right)\mathcal{A}u = \left(\sum_{j\in J}\mathcal{B}_j^{-1}\right)b,$$

or, equivalently,

$$\mathcal{M}^{-1}\mathcal{A}u = \mathcal{M}^{-1}b, \quad \mathcal{M}^{-1} := \sum_{j\in J}\mathcal{B}_j^{-1}. \tag{5.14}$$

Using the properties of the operator $\mathcal{A}$, of the particular decomposition $V = \sum_{j\in J} V_j$, and of the particular preconditioning operators $\mathcal{B}_j, j \in J$, the goal is to prove the equivalence of (5.14) and (5.1) and, in addition, prove results that are as strong as possible on the conditioning and other relevant properties of the preconditioned problem (5.14) and of its matrix representations obtained by discretization.[6]

We start with proving the equivalence of (5.14) and (5.1). By construction, the unique solution $u = \mathcal{A}^{-1}b$ of (5.1) solves also (5.14). It remains to prove that $u = \mathcal{A}^{-1}b$ is the only solution of (5.14).

THEOREM 5.1. *Let the splitting of the Hilbert space V satisfy (5.3) and (5.4), and let the splitting-based preconditioning $\mathcal{M}^{-1}$ be defined by (5.6)–(5.14). Then (5.14) has the unique solution $u = \mathcal{A}^{-1}b$, and for $\mathcal{M}^{-1}$ we have*

$$\|\mathcal{M}^{-1}f\|_V \le \sum_{j\in J}\frac{1}{c_{\mathcal{B}_j}c_{V_j}}\|f\|_{V^\#} \quad \textit{for all } f \in V^\#.$$

Proof. Let (5.14) have two different solutions, i.e., there exists $g \in V^\#$, $g \neq 0$, such that $\mathcal{M}^{-1}g = 0$. Then

$$0 = \langle g, \mathcal{M}^{-1}g\rangle = \left\langle g, \sum_{j\in J}\mathcal{B}_j^{-1}g\right\rangle = \sum_{j\in J}\langle g, \mathcal{B}_j^{-1}g\rangle = \sum_{j\in J}\left(\mathcal{B}_j^{-1}g, \mathcal{B}_j^{-1}g\right)_{\mathcal{B}_j} = \sum_{j\in J}\|\mathcal{B}_j^{-1}g\|^2_{\mathcal{B}_j},$$

i.e.,

$$\|\mathcal{B}_j^{-1}g\|^2_{\mathcal{B}_j} = 0 \quad \text{for all } j \in J. \tag{5.15}$$

Since $g \neq 0$, there exists a $z \in V$ such that $\langle g, z\rangle \neq 0$. Consider a decomposition $z = \sum_{j\in J} z_j$, $z_j \in V_j$, $j \in J$. Then

$$0 \neq \left\langle g, \sum_{j\in J} z_j\right\rangle = \sum_{j\in J}\langle g, z_j\rangle = \sum_{j\in J}\left(\mathcal{B}_j^{-1}g, z_j\right)_{\mathcal{B}_j},$$

and thus at least one term in the last sum, say $\left(\mathcal{B}_k^{-1}g, z_k\right)_{\mathcal{B}_k}$, must be non-zero. This contradicts (5.15) and completes the proof of the first statement. Using (5.8) and (5.4) (and thus (5.5)), we have

$$\|\mathcal{B}_j{}^{-1}f\|_j \le \frac{1}{c_{\mathcal{B}_j}}\|f\|_j^\# \le \frac{1}{c_{\mathcal{B}_j}\sqrt{c_{V_j}}}\|f\|_{V^\#},$$

and thus,

$$\|\mathcal{M}^{-1}f\|_V = \left\|\sum_{j\in J}\mathcal{B}_j^{-1}f\right\|_V \le \sum_{j\in J}\frac{1}{\sqrt{c_{V_j}}}\|\mathcal{B}_j^{-1}f\|_j \le \sum_{j\in J}\frac{1}{c_{\mathcal{B}_j}c_{V_j}}\|f\|_{V^\#},$$

[6] This text does not deal with particular matrix representations that are in practice based on further specific assumptions.

which completes the proof. □
Theorem 5.1 proves that $\mathcal{M}^{-1}$ is bounded with

$$\|\mathcal{M}^{-1}\|_{\mathcal{L}(V^\#,V)} = \sup_{f\in V^\#,\,\|f\|_{V^\#}=1} \|\mathcal{M}^{-1}f\|_V \le C_{\mathcal{M}^{-1}} := \sum_{j\in J} \frac{1}{c_{\mathcal{B}_j} c_{V_j}} < \infty. \tag{5.16}$$

We will now show that $\mathcal{M}^{-1}$ is also coercive and define its bounded and coercive inversion

$$\mathcal{M} := \left(\mathcal{M}^{-1}\right)^{-1} : V \to V^\# \tag{5.17}$$

(see the analogy with the operator $\mathcal{B}$ in Sections 2 and 3). In order to accomplish this, we will assume there exists a $C_{\mathrm{S}} < \infty$ such that[7]

$$\|u\|_{\mathrm{S}}^2 := \inf_{u_j\in V_j,\, u=\sum_{j\in J} u_j} \left\{ \sum_{j\in J} \|u_j\|_j^2 \right\} \le C_{\mathrm{S}} \|u\|_V^2 \quad \text{for all } u \in V. \tag{5.18}$$

Remark 5.1. *Let (5.18) be replaced by a stronger assumption that there exists a positive constant C such that*

$$\sum_{j\in J} \|u_j\|_j^2 \le C \|u\|_V^2, \quad \textit{for all } u = \sum_{j\in J} u_j,\ u_j \in V_j,\ j \in J. \tag{5.19}$$

Then for any $u \in V$ the decomposition $u = \sum_{j\in J} u_j$, $u_j \in V_j$, $j \in J$, is unique. Indeed, let $u = \sum_{j\in J} u_j = \sum_{j\in J} v_j$, $u_j, v_j \in V_j$, and let there exist at least one $m \in J$ such that $u_m \ne v_m$. Then $0 = \sum_{j\in J}(u_j - v_j)$, and one has

$$0 < \sum_{j\in J} \|u_j - v_j\|_j^2 \le C \left\| \sum_{j\in J} (u_j - v_j) \right\|_V^2 = C \left\| \sum_{j\in J} u_j - \sum_{j\in J} v_j \right\|_V^2 = 0,$$

which contradicts $u_m \ne v_m$. The assumption (5.19) is, however, too strong and in the consequence too restrictive. Therefore it is not further considered.

Theorem 5.2. *Let the splitting of the Hilbert space V satisfy (5.3), (5.4) and (5.18), and let the splitting-based preconditioning $\mathcal{M}^{-1}$ be defined by (5.6)–(5.14). Then*

$$\|f\|_{V^\#}^2 \le C_{\mathrm{S}} \sum_{j\in J} \left(\|f\|_j^\# \right)^2 \quad \textit{for all } \ f \in V^\#,$$

and for $\mathcal{M}^{-1}$ we have

$$\langle f, \mathcal{M}^{-1} f \rangle \ge \frac{1}{C_{\mathrm{S}} \max_{j\in J} C_{\mathcal{B}_j}} \|f\|_{V^\#}^2 \quad \textit{for all } \ f \in V^\#.$$

Proof. In order to prove the first statement, we consider for $u \in V$ its arbitrary fixed decomposition $u = \sum_{j\in J} u_j$, $u_j \in V_j$, $j \in J$. Then

$$\begin{aligned} \langle f, u\rangle^2 = \left\langle f, \sum_{j\in J} u_j \right\rangle^2 = \left(\sum_{j\in J} \langle f, u_j\rangle \right)^2 \le \left(\sum_{j\in J} |\langle f, u_j\rangle| \right)^2 \le \left(\sum_{j\in J} \|f\|_j^\# \|u_j\|_j \right)^2 \\ \le \sum_{k\in J} \left(\|f\|_k^\# \right)^2 \sum_{j\in J} \|u_j\|_j^2. \end{aligned}$$

[7] In, e.g., [44, Definition 2.1.1], [19, 43] the norm $\|u\|_{\mathrm{S}}$ defined in (5.18) is called the additive Schwarz norm in V with respect to the splitting (5.3); see Section 6 below.

This must hold for any decomposition of u, therefore also for those with $\sum_{j\in J}\|u_j\|_j^2$ arbitrarily close to $\|u\|_{\mathrm{S}}^2$. Consequently, for all $u\in V$,

$$\langle f,u\rangle^2 \le C_{\mathrm{S}} \sum_{k\in J} \left(\|f\|_k^{\#}\right)^2 \|u\|_V^2,$$

and

$$\|f\|_{V^\#}^2 = \sup_{u\in V,\,\|u\|_V=1} \langle f,u\rangle^2 \le C_{\mathrm{S}} \sum_{k\in J} \left(\|f\|_k^{\#}\right)^2.$$

For proving the second statement we use the inequality

$$\left\langle f, \mathcal{B}_j^{-1} f\right\rangle \ge \frac{1}{C_{\mathcal{B}_j}} \left(\|f\|_j^{\#}\right)^2, \quad \text{for all } f\in V^\#. \tag{5.20}$$

It follows from

$$\begin{aligned}\sup_{f\in V_j^\#,\,f\neq 0} \frac{\left(\|f\|_j^{\#}\right)^2}{\left\langle f,\mathcal{B}_j^{-1}f\right\rangle} &= \sup_{u\in V_j,\,u\neq 0} \frac{\left(\|\mathcal{B}_j u\|_j^{\#}\right)^2}{\langle \mathcal{B}_j u,u\rangle} = \sup_{u\in V_j,\,u\neq 0} \left(\frac{1}{\langle \mathcal{B}_j u,u\rangle} \sup_{v\in V_j,\,v\neq 0} \frac{\langle \mathcal{B}_j u,v\rangle^2}{\|v\|_j^2}\right)\\ &\le \sup_{u\in V_j,\,u\neq 0}\sup_{v\in V_j,\,v\neq 0} \frac{\langle \mathcal{B}_j u,u\rangle\langle \mathcal{B}_j v,v\rangle}{\langle \mathcal{B}_j u,u\rangle\,\|v\|_j^2} = \sup_{v\in V_j,\,v\neq 0} \frac{\langle \mathcal{B}_j v,v\rangle}{\|v\|_j^2} = C_{\mathcal{B}_j},\end{aligned}$$

where we used the Cauchy-Schwarz inequality $\langle \mathcal{B}_j u,v\rangle^2 = (u,v)_{\mathcal{B}_j}^2 \le \|u\|_{\mathcal{B}_j}^2\|v\|_{\mathcal{B}_j}^2$. With (5.20)

$$\begin{aligned}\langle f,\mathcal{M}^{-1}f\rangle = \left\langle f,\sum_{j\in J}\mathcal{B}_j^{-1}f\right\rangle &= \sum_{j\in J}\left\langle f,\mathcal{B}_j^{-1}f\right\rangle \ge \sum_{j\in J}\frac{1}{C_{\mathcal{B}_j}}\left(\|f\|_j^{\#}\right)^2 \ge \frac{1}{\max_{j\in J} C_{\mathcal{B}_j}}\sum_{j\in J}\left(\|f\|_j^{\#}\right)^2\\ &\ge \frac{1}{C_{\mathrm{S}}\max_{j\in J} C_{\mathcal{B}_j}}\|f\|_{V^\#}^2,\end{aligned}$$

which finishes the proof. □

Theorem 5.2 proves that $\mathcal{M}^{-1}$ is coercive with

$$\inf_{f\in V^\#,\,\|f\|_{V^\#}=1} \langle f,\mathcal{M}^{-1}f\rangle \ge c_{\mathcal{M}^{-1}} := \frac{1}{C_{\mathrm{S}}\max_{j\in J} C_{\mathcal{B}_j}} > 0. \tag{5.21}$$

We note the little ambiguity in notation. Here the definition of $C_{\mathcal{M}^{-1}}$ (see (5.16)) and $c_{\mathcal{M}^{-1}}$ (see (5.21)) anticipate the particular construction of $\mathcal{M}^{-1}$ and they are not defined as the boundedness and coercivity constants for a general operator $\mathcal{M}^{-1}$. For simplicity of notation we use this and do not introduce another symbols.

Corollary 5.3. *Let the splitting of the Hilbert space V satisfy (5.3), (5.4) and (5.18), and let the splitting-based preconditioning $\mathcal{M}^{-1}$ be defined by (5.6)–(5.14). Then the operator*

$$\mathcal{M} := \left(\mathcal{M}^{-1}\right)^{-1} : V\to V^\# \tag{5.22}$$

is bounded and coercive with

$$C_{\mathcal{M}} := \sup_{v\in V,\,\|v\|_V=1} \|\mathcal{M}v\|_{V^\#} \le \frac{1}{c_{\mathcal{M}^{-1}}}, \tag{5.23}$$

$$c_{\mathcal{M}} := \inf_{v\in V,\,\|v\|_V=1} \langle \mathcal{M}v,v\rangle \ge \frac{1}{C_{\mathcal{M}^{-1}}}. \tag{5.24}$$

Proof. The existence of the bounded operator $\mathcal{M}$ follows from the Lax-Milgram lemma applied to $\mathcal{M}^{-1}$. The bound (5.23) follows from (5.21) using the substitution $f = \mathcal{M}v/\|\mathcal{M}v\|_{V^\#}$,

$$c_{\mathcal{M}^{-1}} \le \inf_{f\in V^\#,\, \|f\|_{V^\#}=1} \langle f, \mathcal{M}^{-1} f\rangle = \inf_{v\in V,\, v\neq 0} \frac{\langle \mathcal{M}v, v\rangle}{\|\mathcal{M}v\|^2_{V^\#}} \le \inf_{v\in V,\, v\neq 0} \frac{\|\mathcal{M}v\|_{V^\#}\|v\|_V}{\|\mathcal{M}v\|^2_{V^\#}}$$
$$= \inf_{v\in V,\, v\neq 0} \frac{\|v\|_V}{\|\mathcal{M}v\|_{V^\#}} = \frac{1}{\sup_{v\in V,\, v\neq 0} \frac{\|\mathcal{M}v\|_{V^\#}}{\|v\|_V}} = \frac{1}{\|\mathcal{M}\|_{\mathcal{L}(V,V^\#)}}.$$

The bound (5.24) is a consequence of (2.12) used for $\mathcal{M}$, and of (5.16); see also Theorem 5.1. □

Up to now, we have studied the properties of $\mathcal{M}^{-1}$ and of $\mathcal{M}$ that plays the role of the preconditioning operator $\mathcal{B}$ from Sections 2 and 3. In the following, our aim is to prove the norm and spectral equivalence, and some two-sided error bounds using the properties of $\mathcal{M}^{-1}\mathcal{A}$. We will use the norms of $\mathcal{M}^{-1}\mathcal{A}$ and $\mathcal{A}^{-1}\mathcal{M}$ defined in the standard way

$$\|\mathcal{M}^{-1}\mathcal{A}\|_{\mathcal{L}(V,V)} = \sup_{u\in V,\, \|u\|_V=1} \|\mathcal{M}^{-1}\mathcal{A}u\|_V, \qquad \|\mathcal{A}^{-1}\mathcal{M}\|_{\mathcal{L}(V,V)} = \sup_{u\in V,\, \|u\|_V=1} \|\mathcal{A}^{-1}\mathcal{M}u\|_V.$$

Obviously,

$$\|\mathcal{M}^{-1}\mathcal{A}\|_{\mathcal{L}(V,V)} \le \|\mathcal{M}^{-1}\|_{\mathcal{L}(V^\#,V)}\|\mathcal{A}\|_{\mathcal{L}(V,V^\#)}, \qquad \|\mathcal{A}^{-1}\mathcal{M}\|_{\mathcal{L}(V,V)} \le \|\mathcal{A}^{-1}\|_{\mathcal{L}(V^\#,V)}\|\mathcal{M}\|_{\mathcal{L}(V,V^\#)},$$

and correspondingly to (3.3), (3.4) and (3.5), we have the V-norm equivalence of $\mathcal{A}^{-1}$ and $\mathcal{M}^{-1}$ on $V^\#$ stated in the following theorem. Corollary 3.2 shows that within our setting the lower and upper bounds on $\|\mathcal{A}w\|_{V^\#}/\|\mathcal{M}w\|_{V^\#}$ for $w \in V$, $w \neq 0$, and on $\|\mathcal{M}^{-1}f\|_V/\|\mathcal{A}^{-1}f\|_V$ for $f \in V^\#$, $f \neq 0$, hold simultaneously. In other words, the $V^\#$-norm equivalence of $\mathcal{A}$ and $\mathcal{M}$ on V and the V-norm equivalence of $\mathcal{M}^{-1}$ and $\mathcal{A}^{-1}$ on $V^\#$ represent equivalent properties of the pair of operators $\mathcal{A}$ and $\mathcal{M}$. This allows to consider any of these two forms of norm equivalence appropriately to the specific context. In the case of the splitting-based preconditioning the form using $\mathcal{M}^{-1}$ seems more appropriate, because $\mathcal{M}^{-1}$ is constructed as the primary object using the operators $\mathcal{B}_j^{-1}$, $j \in J$; see (5.14). The following theorem just reformulates the inequality (3.5).

THEOREM 5.4 (Norm equivalence). *Let the linear operator $\mathcal{A}$ satisfy (2.7) and (2.8). Let the splitting of the Hilbert space V satisfy (5.3), (5.4) and (5.18), and let the splitting-based preconditioning $\mathcal{M}^{-1}$ be defined by (5.6)–(5.14). Then $\mathcal{A}^{-1}$ and $\mathcal{M}^{-1}$ are V-norm equivalent on $V^\#$,*

$$\|\mathcal{M}^{-1}\mathcal{A}\|_{\mathcal{L}(V,V)} = \sup_{f\in V^\#,\, f\neq 0} \frac{\|\mathcal{M}^{-1}f\|_V}{\|\mathcal{A}^{-1}f\|_V} \le \frac{C_\mathcal{A}}{c_\mathcal{M}}, \tag{5.25}$$

$$\|\mathcal{A}^{-1}\mathcal{M}\|_{\mathcal{L}(V,V)} = \sup_{f\in V^\#,\, f\neq 0} \frac{\|\mathcal{A}^{-1}f\|_V}{\|\mathcal{M}^{-1}f\|_V} \le \frac{C_\mathcal{M}}{c_\mathcal{A}}, \tag{5.26}$$

$$\kappa(\mathcal{M}^{-1}\mathcal{A}) \le \frac{C_\mathcal{A}}{c_\mathcal{M}}\frac{C_\mathcal{M}}{c_\mathcal{A}} = \kappa(\mathcal{A})\kappa(\mathcal{M}) \tag{5.27}$$

and

$$\left(C_\mathcal{A}\sum_{j\in J}\frac{1}{c_{\mathcal{B}_j}c_{V_j}}\right)^{-1} \le \frac{c_\mathcal{M}}{C_\mathcal{A}} \le \frac{\|\mathcal{A}^{-1}f\|_V}{\|\mathcal{M}^{-1}f\|_V} \le \frac{C_\mathcal{M}}{c_\mathcal{A}} \le \frac{C_\mathrm{S}}{c_\mathcal{A}}\max_{j\in J} C_{\mathcal{B}_j} \quad \textit{for all } f \in V^\#,\, f \neq 0. \tag{5.28}$$

Proof. The statement follows from the previous considerations; see also Theorem 3.1. □

The two-sided error bounds introduced in the next theorem hold for an arbitrary approximate solution $v \in V$ of (5.1), (5.14); see also [44, Theorem 2.6.1] that uses the finite-dimensional setting.

THEOREM 5.5. *Let the splitting of the Hilbert space V satisfy (5.3), (5.4) and (5.18), and let the splitting-based preconditioning $\mathcal{M}^{-1}$ be defined by (5.6)–(5.14). Let u be the solution of (5.1), (5.14).*

Then for any $v \in V$ we have

$$\left(C_{\mathcal{A}} \sum_{j\in J} \frac{1}{c_{\mathcal{B}_j} c_{V_j}} \right)^{-1} \left\| \sum_{j\in J} (\mathcal{B}_j^{-1}\mathcal{A}v - \mathcal{B}_j^{-1}b) \right\|_V \le \|v-u\|_V \le \frac{C_{\mathrm{S}}}{c_{\mathcal{A}}} \max_{j\in J} C_{\mathcal{B}_j} \left\| \sum_{j\in J} (\mathcal{B}_j^{-1}\mathcal{A}v - \mathcal{B}_j^{-1}b) \right\|_V .$$

Proof. The statement follows from

$$\|\mathcal{M}^{-1}(\mathcal{A}v-b)\|_V = \|\mathcal{M}^{-1}\mathcal{A}(v-u)\|_V \le \|\mathcal{M}^{-1}\mathcal{A}\|_{\mathcal{L}(V,V)} \|v-u\|_V, \quad \text{for all } v \in V$$

and

$$\|v-u\|_V = \|\mathcal{A}^{-1}\mathcal{M}(\mathcal{M}^{-1}\mathcal{A}(v-u))\|_V \le \|\mathcal{A}^{-1}\mathcal{M}\|_{\mathcal{L}(V,V)} \|(\mathcal{M}^{-1}(\mathcal{A}v-b))\|_V,$$

which give

$$\frac{1}{\|\mathcal{M}^{-1}\mathcal{A}\|_{\mathcal{L}(V,V)}} \left\| \sum_{j\in J} (\mathcal{B}_j^{-1}\mathcal{A}v - \mathcal{B}_j^{-1}b) \right\|_V \le \|v-u\|_V \le \|\mathcal{A}^{-1}\mathcal{M}\|_{\mathcal{L}(V,V)} \left\| \sum_{j\in J} (\mathcal{B}_j^{-1}\mathcal{A}v - \mathcal{B}_j^{-1}b) \right\|_V .$$

Using (5.25), (5.26), and (5.28) finishes the proof. □

The following theorem states the spectral equivalence of $\mathcal{A}$ and $\mathcal{M}$ without using any specific relationship between $\mathcal{A}$ and $\mathcal{M}$. The result therefore reduces to (3.21) in Section 3.

THEOREM 5.6 (Spectral equivalence). *Let the linear self-adjoint operator $\mathcal{A}$ satisfy (2.7) and (2.8). Let the splitting of the Hilbert space V satisfy (5.3), (5.4) and (5.18), and let the splitting-based preconditioning $\mathcal{M}^{-1}$ be defined by (5.6)–(5.14). Then $\mathcal{A}$ and $\mathcal{M}$ are spectrally equivalent and*

$$\frac{c_{\mathcal{A}}}{C_{\mathrm{S}} \max_{j\in J} C_{\mathcal{B}_j}} \le \frac{\langle \mathcal{A}z, z\rangle}{\langle \mathcal{M}z, z\rangle} \le C_{\mathcal{A}} \sum_{j\in J} \frac{1}{c_{\mathcal{B}_j} c_{V_j}} \quad \textit{for all } z \in V,\ z \neq 0. \tag{5.29}$$

Proof. The statement follows from (3.21) using (5.16), (5.21), (5.23), and (5.24). □

Let u be the solution of (5.1), (5.14). Motivated by [44, Chapter 2], we consider the *locally preconditioned residual* associated with $v \in V$

$$\bar{r}_j := \mathcal{B}_j^{-1}\mathcal{A}v - \mathcal{B}_j^{-1}b = \mathcal{B}_j^{-1}\mathcal{A}(v-u) \in V_j, \quad j \in J. \tag{5.30}$$

Clearly, for all $v_j \in V_j$,

$$(\bar{r}_j, v_j)_{\mathcal{B}_j} = \langle \mathcal{A}(v-u), v_j\rangle = a(v-u, v_j) = a(v, v_j) - \langle b, v_j\rangle. \tag{5.31}$$

As a consequence of splitting the problem (5.1) into the set of problems (5.13)–(5.14), we have an (a posteriori) error estimate based on the norms of the locally preconditioned residuals, which is motivated by [44, Theorem 2.6.2]. Before introducing the theorem, we prove a useful lemma.

LEMMA 5.7. *Let the linear self-adjoint operator $\mathcal{A}$ satisfy (2.7) and (2.8). Let the splitting of the Hilbert space V satisfy (5.3), (5.4) and (5.18), and let the splitting-based preconditioning $\mathcal{M}^{-1}$ be defined by (5.6)–(5.14). Then*

$$a(\mathcal{M}^{-1}\mathcal{A}z, \mathcal{M}^{-1}\mathcal{A}z) \le C_{\mathcal{A}} \sum_{k\in J} \frac{1}{c_{\mathcal{B}_k} c_{V_k}} a(\mathcal{M}^{-1}\mathcal{A}z, z), \tag{5.32}$$

and

$$\frac{c_{\mathcal{A}}}{C_{\mathrm{S}} \max_{j\in J} C_{\mathcal{B}_j}} a(z,z) \le a(\mathcal{M}^{-1}\mathcal{A}z, z). \tag{5.33}$$

Proof. We have

$$
\begin{aligned}
a\left(\mathcal{M}^{-1}\mathcal{A}z,\mathcal{M}^{-1}\mathcal{A}z\right) &\le C_{\mathcal{A}}\|\mathcal{M}^{-1}\mathcal{A}z\|_V^2 = C_{\mathcal{A}}\left\|\sum_{j\in J}\mathcal{B}_j^{-1}\mathcal{A}z\right\|_V^2 \le C_{\mathcal{A}}\left(\sum_{j\in J}\left\|\mathcal{B}_j^{-1}\mathcal{A}z\right\|_V\right)^2\\
&\le C_{\mathcal{A}}\left(\sum_{j\in J}\frac{1}{\sqrt{c_{V_j}}}\left\|\mathcal{B}_j^{-1}\mathcal{A}z\right\|_j\right)^2 \le C_{\mathcal{A}}\left(\sum_{j\in J}\frac{1}{\sqrt{c_{\mathcal{B}_j}c_{V_j}}}\left\|\mathcal{B}_j^{-1}\mathcal{A}z\right\|_{\mathcal{B}_j}\right)^2\\
&\le C_{\mathcal{A}}\sum_{k\in J}\frac{1}{c_{\mathcal{B}_k}c_{V_k}}\sum_{j\in J}(\mathcal{B}_j^{-1}\mathcal{A}z,\mathcal{B}_j^{-1}\mathcal{A}z)_{\mathcal{B}_j}\\
&= C_{\mathcal{A}}\sum_{k\in J}\frac{1}{c_{\mathcal{B}_k}c_{V_k}}\sum_{j\in J}\langle\mathcal{A}z,\mathcal{B}_j^{-1}\mathcal{A}z\rangle = C_{\mathcal{A}}\sum_{k\in J}\frac{1}{c_{\mathcal{B}_k}c_{V_k}}\left\langle\mathcal{A}z,\sum_{j\in J}\mathcal{B}_j^{-1}\mathcal{A}z\right\rangle\\
&= C_{\mathcal{A}}\sum_{k\in J}\frac{1}{c_{\mathcal{B}_k}c_{V_k}}\left\langle\mathcal{A}z,\mathcal{M}^{-1}\mathcal{A}z\right\rangle = C_{\mathcal{A}}\sum_{k\in J}\frac{1}{c_{\mathcal{B}_k}c_{V_k}}\,a(z,\mathcal{M}^{-1}\mathcal{A}z),
\end{aligned}
$$

which yields (5.32). For proving (5.33) we consider an arbitrary decomposition of $z\in V$, $z=\sum_{j\in J}z_j$, $z_j\in V_j$, $j\in J$. Then

$$
\begin{aligned}
a(z,z) = a\left(z,\sum_{j\in J}z_j\right) &= \sum_{j\in J}a\left(z,z_j\right) = \sum_{j\in J}\langle\mathcal{A}z,z_j\rangle = \sum_{j\in J}(\mathcal{B}_j^{-1}\mathcal{A}z,z_j)_{\mathcal{B}_j}\\
&\le \left(\sum_{j\in J}(\mathcal{B}_j^{-1}\mathcal{A}z,\mathcal{B}_j^{-1}\mathcal{A}z)_{\mathcal{B}_j}\right)^{1/2}\left(\sum_{j\in J}(z_j,z_j)_{\mathcal{B}_j}\right)^{1/2}\\
&\le \left(\sum_{j\in J}\langle\mathcal{A}z,\mathcal{B}_j^{-1}\mathcal{A}z\rangle\right)^{1/2}\left(\sum_{j\in J}C_{\mathcal{B}_j}\|z_j\|_j^2\right)^{1/2}\\
&\le \max_{j\in J}\sqrt{C_{\mathcal{B}_j}}\,\langle\mathcal{A}z,\mathcal{M}^{-1}\mathcal{A}z\rangle^{1/2}\left(\sum_{j\in J}\|z_j\|_j^2\right)^{1/2}.
\end{aligned}
$$

Considering $\sum_{j\in J}\|z_j\|_j^2$ arbitrarily close to its infimum over all possible decompositions of z,

$$
\begin{aligned}
a(z,z) &\le \max_{j\in J}\sqrt{C_{\mathcal{B}_j}}\,a(z,\mathcal{M}^{-1}\mathcal{A}z)^{1/2}\|z\|_{\mathrm{S}} \le \sqrt{C_{\mathrm{S}}}\max_{j\in J}\sqrt{C_{\mathcal{B}_j}}\,a(z,\mathcal{M}^{-1}\mathcal{A}z)^{1/2}\|z\|_V\\
&\le \frac{\sqrt{C_{\mathrm{S}}}\,\max_{j\in J}\sqrt{C_{\mathcal{B}_j}}}{\sqrt{c_{\mathcal{A}}}}\,a(z,\mathcal{M}^{-1}\mathcal{A}z)^{1/2}a(z,z)^{1/2},
\end{aligned}
$$

which yields (5.33). □

THEOREM 5.8. *Let the splitting of the Hilbert space V satisfy (5.3), (5.4) and (5.18). Let the linear self-adjoint operator $\mathcal{A}$ satisfy (2.7) and (2.8), and let the splitting-based preconditioning $\mathcal{M}^{-1}$ be defined by (5.6)–(5.14). Let u be the solution of (5.1), (5.14). Then*

$$
a\left(v-u,\mathcal{M}^{-1}\mathcal{A}(v-u)\right) = \sum_{j\in J}\|\bar{r}_j\|_{\mathcal{B}_j}^2
$$

and

$$
\frac{\min_{j\in J}c_{\mathcal{B}_j}}{C_{\mathcal{A}}^2}\left(\sum_{k\in J}\frac{1}{c_{\mathcal{B}_k}c_{V_k}}\right)^{-1}\sum_{j\in J}\|\bar{r}_j\|_j^2 \le \|v-u\|_V^2 \le \frac{C_{\mathrm{S}}\max_{j\in J}C_{\mathcal{B}_j}^2}{c_{\mathcal{A}}^2}\sum_{j\in J}\|\bar{r}_j\|_j^2.
$$

Proof. We have for $v \in V$, $\bar{r}_j = \mathcal{B}_j^{-1}\mathcal{A}(v-u)$,

$$\|\bar{r}_j\|_{\mathcal{B}_j}^2 = (\bar{r}_j, \bar{r}_j)_{\mathcal{B}_j} = \langle \mathcal{A}(v-u), \bar{r}_j \rangle = a(v-u, \mathcal{B}_j^{-1}\mathcal{A}(v-u))$$

and thus

$$\sum_{j\in J} \|\bar{r}_j\|_{\mathcal{B}_j}^2 = a\left(v-u, \mathcal{M}^{-1}\mathcal{A}(v-u)\right).$$

Then, using (5.33),

$$\begin{aligned}\|v-u\|_V^2 &\le \frac{1}{c_{\mathcal{A}}}\, a(v-u, v-u) \le \frac{C_{\mathrm{S}} \max_{j\in J} C_{\mathcal{B}_j}}{c_{\mathcal{A}}^2}\, a(v-u, \mathcal{M}^{-1}\mathcal{A}(v-u)) \\ &= \frac{C_{\mathrm{S}} \max_{j\in J} C_{\mathcal{B}_j}}{c_{\mathcal{A}}^2} \sum_{j\in J} \|\bar{r}_j\|_{\mathcal{B}_j}^2 \le \frac{C_{\mathrm{S}} \max_{j\in J} C_{\mathcal{B}_j}}{c_{\mathcal{A}}^2} \sum_{j\in J} C_{\mathcal{B}_j} \|\bar{r}_j\|_j^2 \\ &\le \frac{C_{\mathrm{S}} \max_{j\in J} C_{\mathcal{B}_j}^2}{c_{\mathcal{A}}^2} \sum_{j\in J} \|\bar{r}_j\|_j^2,\end{aligned}$$

which gives the upper bound. A straightforward calculation gives

$$\begin{aligned}L &:= \frac{\min_{j\in J} c_{\mathcal{B}_j}}{C_{\mathcal{A}}^2} \left(\sum_{k\in J} \frac{1}{c_{\mathcal{B}_k} c_{V_k}}\right)^{-1} \sum_{j\in J} \|\bar{r}_j\|_j^2 \le \frac{\min_{j\in J} c_{\mathcal{B}_j}}{C_{\mathcal{A}}^2} \left(\sum_{k\in J} \frac{1}{c_{\mathcal{B}_k} c_{V_k}}\right)^{-1} \sum_{j\in J} \frac{1}{c_{\mathcal{B}_j}} \|\bar{r}_j\|_{\mathcal{B}_j}^2 \\ &\le \frac{1}{C_{\mathcal{A}}^2} \left(\sum_{k\in J} \frac{1}{c_{\mathcal{B}_k} c_{V_k}}\right)^{-1} \sum_{j\in J} \|\bar{r}_j\|_{\mathcal{B}_j}^2 = \frac{1}{C_{\mathcal{A}}^2} \left(\sum_{k\in J} \frac{1}{c_{\mathcal{B}_k} c_{V_k}}\right)^{-1} a(v-u, \mathcal{M}^{-1}\mathcal{A}(v-u)).\end{aligned}$$

Using

$$a(v-u, \mathcal{M}^{-1}\mathcal{A}(v-u))^2 \le a(\mathcal{M}^{-1}\mathcal{A}(v-u), \mathcal{M}^{-1}\mathcal{A}(v-u))\, a(v-u, v-u)$$

and (5.32) gives

$$a(v-u, \mathcal{M}^{-1}\mathcal{A}(v-u)) \le C_{\mathcal{A}} \sum_{k\in J} \frac{1}{c_{\mathcal{B}_k} c_{V_k}} a(v-u, v-u)$$

and finally

$$L \le \frac{1}{C_{\mathcal{A}}}\, a(v-u, v-u) \le \|v-u\|_V^2,$$

which completes the proof. □

The bound for $\|v-u\|_V$ of Theorem 5.5 is given in terms of the norm of the sum of the locally preconditioned residuals $\left\|\sum_{j\in J} \bar{r}_j\right\|_V$, while the bound of Theorem 5.8 is in terms of the sum of squares of the local norms $\sum_{j\in J} \|\bar{r}_j\|_j^2$. In particular, from Theorem 5.5 we have

$$L_1 := \left(C_{\mathcal{A}} \sum_{j\in J} \frac{1}{c_{\mathcal{B}_j} c_{V_j}}\right)^{-2} \left\|\sum_{j\in J} \bar{r}_j\right\|_V^2 \le \|v-u\|_V^2 \le \frac{C_{\mathrm{S}}^2}{c_{\mathcal{A}}^2} \max_{j\in J} C_{\mathcal{B}_j}^2 \left\|\sum_{j\in J} \bar{r}_j\right\|_V^2 =: R_1,$$

while from Theorem 5.8 we obtain

$$L_2 := \frac{\min_{j\in J} c_{\mathcal{B}_j}}{C_{\mathcal{A}}^2} \left(\sum_{k\in J} \frac{1}{c_{\mathcal{B}_k} c_{V_k}}\right)^{-1} \sum_{j\in J} \|\bar{r}_j\|_j^2 \le \|v-u\|_V^2 \le \frac{C_{\mathrm{S}} \max_{j\in J} C_{\mathcal{B}_j}^2}{c_{\mathcal{A}}^2} \sum_{j\in J} \|\bar{r}_j\|_j^2 =: R_2.$$

For the upper bounds R_1 and R_2 we get

$$\frac{R_1}{R_2} = \frac{C_{\mathrm{S}} \left\| \sum_{j\in J} \bar{r}_j \right\|_V^2}{\sum_{j\in J} \|\bar{r}_j\|_j^2} \le \frac{C_{\mathrm{S}} \left(\sum_{j\in J} \|\bar{r}_j\|_V \right)^2}{\sum_{j\in J} \|\bar{r}_j\|_j^2} \le \frac{C_{\mathrm{S}} \left(\sum_{j\in J} \|\bar{r}_j\|_j / \sqrt{c_{V_j}} \right)^2}{\sum_{j\in J} \|\bar{r}_j\|_j^2} \le C_{\mathrm{S}} \sum_{j\in J} \frac{1}{c_{V_j}},$$

and $R_1 \ge R_2$, i.e. Theorem 5.8 gives at least as good an upper bound as Theorem 5.5, if and only if

$$\sum_{j\in J} \|\bar{r}_j\|_j^2 \le C_{\mathrm{S}} \left\| \sum_{j\in J} \bar{r}_j \right\|_V^2; \tag{5.34}$$

cf. (5.18). If the residual splitting satisfies (5.34), then we get for the lower bounds L_1 and L_2

$$\frac{L_1}{L_2} = \frac{1}{\min_{j\in J} c_{\mathcal{B}_j}} \left(\sum_{k\in J} \frac{1}{c_{\mathcal{B}_k} c_{V_k}} \right)^{-1} \frac{\left\| \sum_{j\in J} \bar{r}_j \right\|_V^2}{\sum_{j\in J} \|\bar{r}_j\|_j^2} \ge \frac{1}{\min_{j\in J} c_{\mathcal{B}_j}} \left(\sum_{k\in J} \frac{1}{c_{\mathcal{B}_k} c_{V_k}} \right)^{-1} \frac{1}{C_{\mathrm{S}}} \ge \left(C_{\mathrm{S}} \sum_{k\in J} \frac{1}{c_{V_k}} \right)^{-1}$$

and

$$\begin{aligned} \frac{L_1}{L_2} &\le \frac{1}{\min_{j\in J} c_{\mathcal{B}_j}} \left(\sum_{k\in J} \frac{1}{c_{\mathcal{B}_k} c_{V_k}} \right)^{-1} \frac{\left(\sum_{j\in J} \|\bar{r}_j\|_V \right)^2}{\sum_{j\in J} \|\bar{r}_j\|_j^2} \le \frac{1}{\min_{j\in J} c_{\mathcal{B}_j}} \left(\sum_{k\in J} \frac{1}{c_{\mathcal{B}_k} c_{V_k}} \right)^{-1} \frac{\left(\sum_{j\in J} \|\bar{r}_j\|_j / \sqrt{c_{V_j}} \right)^2}{\sum_{j\in J} \|\bar{r}_j\|_j^2} \\ &\le \frac{1}{\min_{j\in J} c_{\mathcal{B}_j}} \left(\sum_{k\in J} \frac{1}{c_{\mathcal{B}_k} c_{V_k}} \right)^{-1} \sum_{j\in J} \frac{1}{c_{V_j}} \le \frac{\max_{k\in J} c_{\mathcal{B}_k}}{\min_{j\in J} c_{\mathcal{B}_j}} \left(\sum_{k\in J} \frac{1}{c_{V_k}} \right)^{-1} \sum_{j\in J} \frac{1}{c_{V_j}} = \frac{\max_{k\in J} c_{\mathcal{B}_k}}{\min_{j\in J} c_{\mathcal{B}_j}}. \end{aligned}$$

Finally,

$$\left(\frac{R_1}{L_1} \right) \Big/ \left(\frac{R_2}{L_2} \right) = \frac{R_1 L_2}{L_1 R_2} = C_{\mathrm{S}} \min_{j\in J} c_{\mathcal{B}_j} \sum_{j\in J} \frac{1}{c_{\mathcal{B}_j} c_{V_j}}.$$

Summarizing, comparison of the bounds of Theorem 5.5 and Theorem 5.8 is problem- and mesh-dependent.

6. Stable splitting. The splitting of V defined by (5.3) is in literature called *stable* providing that there are constants $c_{\mathrm{S}} > 0$ and $C_{\mathrm{S}} > 0$ such that

$$c_{\mathrm{S}} \|u\|_V^2 \le \|u\|_{\mathrm{S}}^2 \le C_{\mathrm{S}} \|u\|_V^2 \quad \text{for all } u \in V. \tag{6.1}$$

As pointed out in [44, Remark 2.1.3], for V finite-dimensional all its splittings are trivially stable. The issue is then not the existence but the value of the constants c_{S} and C_{S}. The stable splitting assumption (6.1) can be easily linked with the assumptions (5.4) and (5.18) above (the last one coincides with the right inequality in (6.1)). This gives unique solvability of (5.14)[8] and it will allow to apply results formulated in the previous sections.

LEMMA 6.1. *The left inequality of (6.1) is fulfilled if and only if (5.4) holds.*

Proof. Assuming $c_{\mathrm{S}} \|u\|_V^2 \le \|u\|_{\mathrm{S}}^2$ for $u \in V$, we have for $u_j \in V_j$, $j \in J$,

$$\|u_j\|_V^2 \le \frac{1}{c_{\mathrm{S}}} \|u_j\|_{\mathrm{S}}^2 \le \frac{1}{c_{\mathrm{S}}} \|u_j\|_j^2. \tag{6.2}$$

Thus setting $c_{V_j} := c_{\mathrm{S}}$, $j \in J$, we get (5.4). Here (6.2) shows that if $c_{\mathrm{S}} \|u\|_V^2 \le \|u\|_{\mathrm{S}}^2$ for all $u \in V$, then (5.4) is satisfied with the same universal constant c_{S} valid for all $j \in J$, which does not exclude the

[8] Using our notation, the operator equation (2.18) from [44, Theorem 2.1.1] is identical to the transformed system (5.14) from Section 5.

option that (5.4) is also satisfied for some constants c_{V_j} larger than c_{S}. On the other hand, assuming (5.4), we get for any $u \in V$ and for any decomposition $u = \sum_{j\in J} u_j$, $u_j \in V_j$, $j \in J$,

$$\|u\|_V^2 = \left\|\sum_{j\in J} u_j\right\|_V^2 \le \left(\sum_{j\in J} \|u_j\|_V\right)^2 \le \left(\sum_{j\in J} \frac{1}{\sqrt{c_{V_j}}}\|u_j\|_j\right)^2 \le \sum_{j\in J}\frac{1}{c_{V_j}} \sum_{k\in J}\|u_k\|_k^2. \tag{6.3}$$

Since (6.3) holds for any decomposition of u, we get by considering $\sum_{k\in J}\|u_k\|_k^2$ arbitrarily close to its infimum

$$\|u\|_V^2 \le \sum_{j\in J}\frac{1}{c_{V_j}} \inf_{u_k\in V_k,\, u=\sum_{k\in J} u_k} \sum_{k\in J}\|u_k\|_k^2 = \sum_{j\in J}\frac{1}{c_{V_j}}\|u\|_{\mathrm{S}}^2. \tag{6.4}$$

Thus setting $c_{\mathrm{S}} := \left(\sum_{j\in J} c_{V_j}^{-1}\right)^{-1}$ yields the first inequality of (6.1). It is worth noting that if for some constant c we have $c_{V_j} = c$, $j \in J$ in (5.4), then the value of c_{S} derived for (6.4) is $c/|J|$, and we are unable to deduce (5.4) from the left inequality of (6.1) with the same constant c but with the much weaker $c/|J|$. Here we denote by $|J|$ the size of the index set J. □

Using (6.1) instead of (5.4) and (5.18), the statements of Theorems 5.1, 5.4, 5.6, and 5.8 can be easily modified.

REMARK 6.1. *In some published works the setting corresponds to that used above, and the subspaces V_j, $j \in J$, are not required to be nested. In most of the hierarchical approaches, however, it is additionally assumed that the splitting is based on nested subspaces*

$$V_1 \subset V_2 \subset \cdots \subset V_{k-1} \subset V_k = V, \quad J = \{1, 2, \ldots, k\}. \tag{6.5}$$

In addition to that, some works define also the subspaces W_j, $j \in J$ such that (with $V_0 := \{0\}$)

$$V_{j-1} \oplus W_j := V_j,\ j \in J,\ j \neq 0, \tag{6.6}$$

giving an equivalent splitting representation

$$V = \sum_{j\in J} V_j = \sum_{j\in J} W_j. \tag{6.7}$$

The individual preconditioners can then be constructed by the subtraction of projectors onto the individual hierarchical levels; see, e.g. [46, Section 13.2.2] and the references given there.

7. Conclusions. In the presented construction of the splitting-based preconditioning $\mathcal{M}$ we have not used any specific information about the operator $\mathcal{A}$ except of being bounded, coercive and self-adjoint. As in the variety of approaches, methods and theoretical results published in literature, we therefore can not expect to prove, in general, that the condition number $\kappa(\mathcal{M}^{-1}\mathcal{A})$ of the operator $\mathcal{M}^{-1}\mathcal{A}$ in the operator equation (5.14) (see Theorem 5.4) is small. Similarly, we can not expect to prove that the constants determining the spectral equivalence of the operators $\mathcal{A}$ and $\mathcal{M}$ are close to each other (see Theorem 5.6), with implications to the discretized problem, cf. Section 4.2. Apart from the condition number of the Gram matrix $\mathbf{S}_h$ in (4.16), Theorems 4.1 and 4.2 give the bounds for the condition number and the spectral number of the discretized preconditioned operator, respectively, that are independent of the discretization, but *not more*.

We believe however, that the presented generally formulated results can serve as a basis for an easier comparison of existing approaches that can be put into the given framework. Incorporating an appropriate information about the operator $\mathcal{A}$ into the construction of the preconditioning $\mathcal{M}$ can lead to stronger results on the condition number and/or the spectral number of the preconditioned operators and of their discretizations; for recent examples see, e.g. [24, 29, 34, 38, 45]. Results of further work in this direction will be reported in the subsequent part of this work.

Within the given framework we concentrate on the condition number $\kappa(\mathbf{M}_h^{-1}\mathbf{A}_h)$ and on the spectral number $\hat{\kappa}(\mathbf{A}_h, \mathbf{M}_h)$ defined by (4.16) and (4.20), respectively; see also (4.24). As emphasized in the introduction, one should always be aware that, in general, these single number characteristics are (as any other single number characteristics) insufficient for describing convergence behavior of Krylov subspace methods. In this context we note that an arbitrary decomposition $\mathbf{M}_h = \mathbf{L}_h\mathbf{L}_h^*$ leads to the uniquely determined spectral number $\hat{\kappa}(\mathbf{A}_h, \mathbf{M}_h)$, and different choices of $\mathbf{L}_h$, which are all related via orthogonal transformations, see Section 4.1, result in the same convergence behavior of the preconditioned conjugate gradient method despite the fact that they can be associated with different transformations of the discretization bases. It is worth noticing that here the same convergence behavior does not necessarily mean the same computational cost as the computational cost per iteration can be different for different choices of $\mathbf{L}_h$.

A work on the inner structure of the spectrum of the preconditioned operator in relation to the particular problem and its preconditioning is in progress and the results will be reported elsewhere. Such work will extend our investigation of the efficiency of operator preconditioning beyond single-number characteristics.

Appendix. In the Appendix we give the proof of the following theorem.

THEOREM 2.1. *Let $\mathcal{A} : V \to V^\#$ be a linear, bounded, coercive and self-adjoint operator. Using the standard definition of the operator norm, the boundedness constant $C_\mathcal{A}$ and the coercivity constant $c_\mathcal{A}$ can be expressed as*

$$C_\mathcal{A} = \|\mathcal{A}\|_{\mathcal{L}(V,V^\#)} = \sup_{u\in V,\, \|u\|_V=1} \langle \mathcal{A}u, u\rangle = M_\mathcal{A}, \tag{7.1}$$

$$c_\mathcal{A} = m_\mathcal{A} = \inf_{v\in V,\, \|v\|_V=1} \langle \mathcal{A}v, v\rangle = \frac{1}{\sup_{f\in V^\#,\, \|f\|_{V^\#}=1} \|\mathcal{A}^{-1}f\|_V} = \left\{\|\mathcal{A}^{-1}\|_{\mathcal{L}(V^\#,V)}\right\}^{-1}. \tag{7.2}$$

Proof. The equality (7.1) is well known. It follows from the following sequence of equalities

$$C_\mathcal{A} = \|\mathcal{A}\|_{\mathcal{L}(V,V^\#)} = \|\tau\mathcal{A}\|_{\mathcal{L}(V,V)} = \sup_{u\in V,\, \|u\|_V=1} (\tau\mathcal{A}u, u)_V = \sup_{u\in V,\, \|u\|_V=1} \langle \mathcal{A}u, u\rangle = M_\mathcal{A}. \tag{7.3}$$

Here we used the fact that for any self-adjoint operator S in a Hilbert space V

$$\|S\|_{\mathcal{L}(V,V)} = \sup_{z\in V,\, \|z\|_V=1} \|Sz\|_V = \sup_{z\in V,\, \|z\|_V=1} (Sz, Sz)_V^{1/2} = \sup_{z\in V,\, \|z\|_V=1} |(Sz, z)_V|; \tag{7.4}$$

see [9, Theorem 4.10.1, p. 220], [17, Theorem 6.5.1]. The second statement (7.2) was published without proof in [31, Section 3.3]. Since $c_\mathcal{A} = m_\mathcal{A}$, it remains to prove that

$$m_\mathcal{A} = \frac{1}{\sup_{f\in V^\#,\, \|f\|_{V^\#}=1} \|\mathcal{A}^{-1}f\|_V} = \inf_{u\in V,\, \|u\|_V=1} \|\mathcal{A}u\|_{V^\#}, \tag{7.5}$$

where the second equality results from the substitution $f = \mathcal{A}u/\|\mathcal{A}u\|_{V^\#}$, $u \in V$. Equivalently, it remains to prove that

$$m_\mathcal{A} := \inf_{u\in V,\, \|u\|_V=1} (\tau\mathcal{A}u, u)_V = \inf_{u\in V,\, \|u\|_V=1} \|\tau\mathcal{A}u\|_V. \tag{7.6}$$

Clearly $(\tau\mathcal{A}u, u)_V \le \|\tau\mathcal{A}u\|_V\|u\|_V$, therefore the inequality

$$m_\mathcal{A} \le \inf_{u\in V,\, \|u\|_V=1} \|\tau\mathcal{A}u\|_V$$

is trivial. In order to prove the opposite inequality

$$m_\mathcal{A} \ge \inf_{u\in V,\, \|u\|_V=1} \|\tau\mathcal{A}u\|_V,$$

we use the fact that $m_{\mathcal{A}}$ belongs to the spectrum of $\tau\mathcal{A}$ and therefore there exists a sequence $\{v_k\}_{k=1,2,\dots}$ in V, $\|v_k\|_V = 1$, such that

$$\lim_{k\to\infty} \|\tau\mathcal{A}v_k - m_{\mathcal{A}}v_k\|_V = 0; \tag{7.7}$$

see [17, Corollary 6.5.6]. We will finish the proof by contradiction. Assume that

$$m_{\mathcal{A}} < \inf_{u\in V,\, \|u\|_V=1} \|\tau\mathcal{A}u\|_V - \triangle$$

for some $\triangle > 0$. Using the Cauchy-Schwarz inequality,

$$\begin{aligned}\|\tau\mathcal{A}v_k - m_{\mathcal{A}}v_k\|_V^2 &= \|\tau\mathcal{A}v_k\|_V^2 + m_{\mathcal{A}}^2 - 2m_{\mathcal{A}}(\tau\mathcal{A}v_k, v_k)_V \\ &\geq \|\tau\mathcal{A}v_k\|_V^2 + m_{\mathcal{A}}^2 - 2m_{\mathcal{A}}\|\tau\mathcal{A}\|_V = (\|\tau\mathcal{A}v_k\|_V - m_{\mathcal{A}})^2.\end{aligned}$$

Then

$$\|\tau\mathcal{A}v_k - m_{\mathcal{A}}v_k\|_V^2 \geq \triangle^2 \quad \text{for all } k = 1, 2, \dots,$$

which gives the contradiction with (7.7) and completes the proof. □

Acknowledgement. The work was supported by the Grant Agency of the Czech Republic under the contract No. 17-04150J and by the Charles University, project GA UK No. 172915. The authors thank Miroslav Bulíček, Vít Dolejší, Josef Málek, Endre Süli, and Jan Zeman for their careful reading the manuscript and for many helpful suggestions and comments.

REFERENCES

[1] D. N. Arnold, R. S. Falk, and R. Winther. Preconditioning discrete approximations of the Reissner-Mindlin plate model. *RAIRO Modél. Math. Anal. Numér.*, 31(4):517–557, 1997.

[2] D. N. Arnold, R. S. Falk, and R. Winther. Preconditioning in $H(\mathrm{div})$ and applications. *Math. Comp.*, 66(219):957–984, 1997.

[3] D. N. Arnold, R. S. Falk, and R. Winther. Finite element exterior calculus: from Hodge theory to numerical stability. *Bull. Amer. Math. Soc. (N.S.)*, 47:281–354, 2010.

[4] K. Atkinson and W. Han. *Theoretical Numerical Analysis. A Functional Analysis Framework*, volume 39 of *Texts in Applied Mathematics*. Springer, Dordrecht, third edition, 2009.

[5] O. Axelsson and J. Karátson. Equivalent operator preconditioning for elliptic problems. *Numer. Algorithms*, 50(3):297–380, 2009.

[6] O. Axelsson and P. S. Vassilevski. Algebraic multilevel preconditioning methods. I. *Numer. Math.*, 56(2-3):157–177, 1989.

[7] J. Bramble, J. Pasciak, and J. Xu. Parallel multilevel preconditioners. *Math. Comp.*, 55(191):1–22, 1990.

[8] J. H. Bramble, J. E. Pasciak, J. P. Wang, and J. Xu. Convergence estimates for multigrid algorithms without regularity assumptions. *Math. Comp.*, 57(195):23–45, 1991.

[9] P. G. Ciarlet. *Linear and Nonlinear Functional Analysis with Applications*. SIAM, Philadelphia, 2013.

[10] J. B. Conway. *A Course in Functional Analysis*, volume 96 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 1990.

[11] W. Dahmen and A. Kunoth. Multilevel preconditioning. *Numer. Math.*, 63(3):315–344, 1992.

[12] N. Dunford and J. T. Schwartz. *Linear Operators. I. General Theory*. With the assistance of W. G. Bade and R. G. Bartle. Pure and Applied Mathematics, Vol. 7. Interscience Publishers, Inc., New York, 1958.

[13] E. G. D'Yakonov. On an iterative method for the solution of a system of finite-difference equations. *Dokl. Akad. Nauk*, 138:522–526, 1961.

[14] E. G. D'Yakonov. The construction of iterative methods based on the use of spectrally equivalent operators. *U.S.S.R Comput. Math. and Math. Phys.*, 6(1):14–46, 1966.

[15] V. Faber, T. A. Manteuffel, and S. V. Parter. On the theory of equivalent operators and application to the numerical solution of uniformly elliptic partial differential equations. *Adv. in Appl. Math.*, 11(2):109–163, 1990.

[16] Ch. Farhat, J. Mandel, and F.-X. Roux. Optimal convergence properties of the FETI domain decomposition method. *Comput. Methods Appl. Mech. Engrg.*, 115(3-4):365–385, 1994.

[17] A. Friedman. *Foundations of Modern Analysis*. Dover Publications, New York, 1982.

[18] T. Gergelits and Z. Strakoš. Composite convergence bounds based on Chebyshev polynomials and finite precision conjugate gradient computations. *Numer. Algorithms*, 65(4):759–782, 2014.

[19] M. Griebel and P. Oswald. On the abstract theory of additive and multiplicative Schwarz algorithms. *Numer. Math.*, 70(2):163–180, 1995.

[20] J. E. Gunn. The numerical solution of $\nabla \cdot a \nabla u = f$ by a semi-explicit alternating-direction iterative technique. *Numer. Math.*, 6:181–184, 1964.

[21] J. E. Gunn. The solution of elliptic difference equations by semi-explicit iterative techniques. *J. Soc. Indust. Appl. Math. Ser. B Numer. Anal.*, 2:24–45, 1965.

[22] R. Hiptmair. Operator preconditioning. *Comput. Math. Appl.*, 52(5):699–706, 2006.

[23] J. Hrnčíř and Z. Strakoš. Norm and spectral equivalence in operator preconditioning. *In preparation*, 2018.

[24] E. Keilegavlen and J. M. Nordbotten. Inexact linear solvers for control volume discretizations in porous media. *Comput. Geosci.*, 19(1):159–176, 2014.

[25] R. C. Kirby. From functional analysis to iterative methods. *SIAM Rev.*, 52(2):269–293, 2010.

[26] A. Klawonn. An optimal preconditioner for a class of saddle point problems with a penalty term, Part II: general theory. Technical report, 14/95, Westflische Wilhelms-Universität Münster, Germany, 04 1995.

[27] A. Klawonn. *Preconditioners for indefinite problems*. PhD thesis, Universität Münster, 1996.

[28] A. N. Kolmogorov and S. V. Fomin. *Elements of the Theory of Functions and Functional Analysis. Vol. 1. Metric and Normed Spaces*. Graylock Press, Rochester, N. Y., 1957. Translated from the first Russian edition by Leo F. Boron.

[29] R. Kornhuber and H. Yserentant. Numerical homogenization of elliptic multiscale problems by subspace decomposition. *Multiscale Model. Simul.*, 14(3):1017–1036, 2016.

[30] J. Liesen and Z. Strakoš. *Krylov Subspace Methods: Principles and Analysis*. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, 2013.

[31] J. Málek and Z. Strakoš. *Preconditioning and the Conjugate Gradient Method in the Context of Solving PDEs*, volume 1 of *SIAM Spotlights*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2015.

[32] K.-A. Mardal and R. Winther. Preconditioning discretizations of systems of partial differential equations. *Numer. Linear Algebra Appl.*, 18:1–40, 2011.

[33] A. M. Matsokin and S. V. Nepomnyaschikh. Schwarz alternating method in a subspace. *Sov. Math. (Izv. vuz.)*, 29:78–84, 1985.

[34] A. Napov and Y. Notay. Algebraic multigrid for moderate order finite elements. *SIAM J. Sci. Comput.*, 36(4):A1678–A1707, 2014.

[35] S. V. Nepomnyaschikh. Mesh theorems on traces, normalization of function traces and their inversion. *Sov. J. Numer. Anal. Math. Modelling*, 6:1–25, 1991.

[36] S. V. Nepomnyaschikh. Method of splitting into subspaces for solving elliptic boundary value problems in complex-form domains. *Sov. J. Numer. Anal. Math. Modelling*, 6:151–168, 1991.

[37] S. V. Nepomnyaschikh. Decomposition and fictitious domain methods for elliptic boundary value problems. In *Chan, T. F.; Keyes, D. E.; Meurant, G. A.; Scroggs, J. S.; Voigt, R. G. (eds.): 5th Conference on Domain Decomposition Methods for PDE*, pages 62–72. SIAM Publ., Philadelphia, 1992.

[38] B. F. Nielsen, A. Tveito, and W. Hackbusch. Preconditioning by inverting the Laplacian: an analysis of the eigenvalues. *IMA Journal of Numerical Analysis*, 29(1):24–42, 2009.

[39] P. Oswald. On function spaces related to finite element approximation theory. *Z. Anal. Anwendungen*, 9:43–64, 1990.

[40] P. Oswald. Norm equivalencies and multilevel Schwarz preconditioning for variational problems. *Forschungsergebnisse Math/92/01, Friedrich Schiller Universität, Jena*, 1992.

[41] P. Oswald. Stable splittings of Sobolev spaces and fast solution of variational problems. *Forschungsergebnisse Math/92/05, Friedrich Schiller Universität, Jena*, 1992.

[42] P. Oswald. *Multilevel Finite Element Approximation, Theory and Applications*. Teubner Skripten zur Numerik. B. G. Teubner Stuttgart, Stuttgart, 1994.

[43] P. Oswald. Stable subspace splittings for Sobolev spaces and domain decomposition algorithms. In *Domain decomposition methods in scientific and engineering computing (University Park, PA, 1993)*, volume 180 of *Contemp. Math.*, pages 87–98. Amer. Math. Soc., Providence, RI, 1994.

[44] U. Rüde. *Mathematical and Computational Techniques for Multilevel Adaptive Methods*, volume 13 of *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1993.

[45] J. Sogn and W. Zulehner. Schur complement preconditioners for multiple saddle point problems of block tridiagonal form with application to optimization problems. *ArXiv:1708.09245v1 [math.NA]*, 2017.

[46] O. Steinbach. *Numerical Approximation Methods for Elliptic Boundary Value Problems, Finite and Boundary Elements*. Springer Science + Bussiness Media, LLC, New York, NY, 2008.

[47] P. S. Vassilevski. *Multilevel Block Factorization Preconditioners. Matrix-based Analysis and Algorithms for Solving Finite Element Equations*. Springer, New York, 2008.

[48] J. Xu. Iterative methods by space decomposition and subspace correction. *SIAM Rev.*, 34(4):581–613, 1992.

[49] W. Zulehner. Nonstandard norms and robust estimates for saddle point problems. *SIAM J. Matrix Anal. Appl.*, 32(2):536–560, 2011.