

Frames and Space Splittings in Hilbert Spaces

Peter Oswald

Bell Labs, Lucent Technologies, Rm. 2C403

600 Mountain Av.

Murray Hill, NJ 07974-0636

e-mail: poswald@research.bell-labs.com

www: <http://cm.bell-labs.com/who/poswald>

Abstract

This is the first part of lecture notes based on short courses held by the author at the Universities of Bonn and Lancaster in 1997. We review the abstract theory of frames and their generalization - the so-called stable space splittings - in a Hilbert space setting. While the frame concept was developed as part of (non-harmonic) Fourier analysis and mainly in connection with signal processing applications, the latter theory of stable subspace splitting has led to a better understanding of iterative solvers (multigrid/multilevel resp. domain decomposition methods) for large-scale discretizations of elliptic operator equations.

1 Hilbert space notation

For the basics of Hilbert space terminology, you may consult any of your favorite textbooks, additional sources [1, 2, 3, 4, 5] can be found in the reference list. Throughout the paper, V will denote a *separable* or *finite-dimensional* real Hilbert space (the occasional appearance of examples of complex Hilbert spaces or the use of results valid for complex Hilbert spaces in the real case via complexification will not be commented on). By default, $(f, g) \equiv (f, g)_V$ denotes the scalar product of two elements $f, g \in V$, and $\|f\| \equiv \|f\|_V = \sqrt{(f, f)_V}$ the norm of $f \in V$. Since often one and the same linear space V will be equipped with several scalar products, we will occasionally use the notation $V = \{V; (\cdot, \cdot)\}$ to indicate the specific scalar product. Recall that knowing the norm on a Hilbert space V is equivalent to knowing the scalar product since

$$4(f, g) = \|f + g\|^2 - \|f - g\|^2 .$$

A subset $V_1 \subset V$ is called *subspace* of V if it is a linear manifold in V (i.e., $f, g \in V_1$ and $\lambda \in \mathbb{R}$ always imply $f + g, \lambda f \in V_1$) and closed with respect to the Hilbert space

topology. Thus, any subspace equipped with a scalar product obtained by restriction from V turns into a Hilbert space on its own. The orthogonal projection $P_{V_1} : V \rightarrow V_1$ is then well-defined by

$$\|u - P_{V_1}u\| = \inf_{v_1 \in V_1} \|u - v_1\|$$

(i.e., $u_1 = P_{V_1}u$ is the unique element in V_1 of minimal distance to u) or, equivalently, by the orthogonality requirement

$$(u - P_{V_1}u, v_1) = 0 \quad \forall v_1 \in V_1 .$$

The set

$$V_1^\perp = \{u \in V : (u, v_1) = 0 \quad \forall v_1 \in V_1\} = \text{ran}(\text{Id} - P_{V_1}) .$$

is a subspace in V , the *orthogonal complement* of V_1 . Clearly, $P_{V_1^\perp} = \text{Id} - P_{V_1}$, where Id denotes the identity operator.

For an arbitrary subset $F \subset V$ we define its *span* (denoted by $[F]$) as the smallest subspace in V containing F . Equivalently,

$$[F] = \text{clos}_V \left\{ \sum_{i=1}^n c_i f_i : f_i \in F, c_i \in \mathbb{R}, n \geq 1 \right\} .$$

While the intersection $V_1 \cap V_2$ of two subspaces V_1, V_2 of V is automatically a subspace of V , their (*closed*) *sum* needs to be defined as

$$V_1 + V_2 = \text{clos}_V \{f_1 + f_2 : f_1 \in V_1, f_2 \in V_2\} = [V_1 \cup V_2] ,$$

the latter definition applies also to the case of arbitrarily many subspaces. The closure with respect to V can be dropped only in special cases (e.g., if all but one of the subspaces V_j in a sum are finite-dimensional). Counterexamples can be found in [1, p.110] or [2, p. 26/27], [3, p. 39/40]. Another case of interest is when the subspaces are mutually *orthogonal*, in this case we write $V_1 \oplus V_2 \oplus \dots$ instead of $V_1 + V_2 + \dots$. The closure operation can be dropped if the orthogonal sum has finitely many terms.

There are many more generic constructions of interest for Hilbert spaces. We note the *Hilbert sum* [4] (which is also called *product space* [1]) of a family of Hilbert spaces $\{V_j; (\cdot, \cdot)_j\}, j \in I$. It is defined as the Hilbert space $\tilde{V} = \times_I V_j$ of all sequences $\tilde{v} = (v_j)$ where $v_j \in V_j, j \in I$, such that

$$\|\tilde{v}\|_{\tilde{V}}^2 \equiv \sum_{j \in I} \|v_j\|_j^2 < \infty .$$

Applications will be considered in Section 3 and 4. Another possibility [5, Section 3.4] is the *tensor product* of Hilbert spaces. A formal way to produce $V_1 \otimes V_2$ is to take complete orthonormal systems (CONS) $\{e_i^1\}$ and $\{e_j^2\}$ in V_1 and V_2 , respectively, and to define

$$V_1 \otimes V_2 = \left\{ f = \sum_{i,j} c_{ij} e_i^1 \otimes e_j^2 : \|f\|_{V_1 \otimes V_2}^2 = \sum_{i,j} c_{ij}^2 \right\} .$$

Partly, the motivation for this construction comes from considering the special case of Hilbert spaces V_k of functions defined on domains Ω_k . Then, by identifying $e_i^1(\cdot) \otimes e_j^2(\cdot)$ with the usual product $e_i^1(x_1)e_j^2(x_2)$, the resulting space $V_1 \otimes V_2$ can be interpreted as space of functions on the product domain $\Omega_1 \times \Omega_2$. E.g., $L_2(\Omega_1 \times \Omega_2) \cong L_2(\Omega_1) \otimes L_2(\Omega_2)$ (see [5]).

Let us conclude with the variational setting of operator equations. Let V' denote the dual space to the Hilbert space V (i.e., the set of all bounded linear functionals $\Phi : V \rightarrow \mathbb{R}$, equipped with the norm

$$\|\Phi\|_{V'} = \sup_{\|v\|_V=1} |\Phi(v)| .$$

By the Riesz-Fischer theorem, V' is also a Hilbert space (which could be identified with V). Set

$$\langle \Phi, v \rangle_{V' \times V} = \Phi(v) \quad \forall \Phi \in V', v \in V .$$

Then any bounded linear operator $A : V \rightarrow V'$ generates a bounded bilinear form

$$a(u, v) = \langle Au, v \rangle_{V' \times V}, \quad u, v \in V, \quad (1)$$

vice versa. *Boundedness* of the bilinear form $a(\cdot, \cdot)$ means the existence of a constant $C_1 < \infty$ such that

$$|a(u, v)| \leq C_1 \|u\| \|v\|, \quad u, v \in V, \quad (2)$$

which together with *coercivity* (i.e., the existence of another constant $C_2 > 0$ such that

$$a(u, u) \geq C_2 \|u\|^2, \quad u \in V) \quad (3)$$

guarantees unique solvability of the following variational problem: Given any $\Phi \in V'$, find $u = u_\Phi \in V$ such that

$$a(u, v) = \Phi(v) \quad \forall v \in V. \quad (4)$$

This follows from the Lax-Milgram theorem. In other words, the operator equation $Au = \Phi$ has a unique solution in V , for any $\Phi \in V'$, i.e., $A^{-1} : V' \rightarrow V$ exists. According to (1-3), we have

$$\|A\|_{V \rightarrow V'} \leq C_1, \quad \|A^{-1}\|_{V' \rightarrow V} \leq C_2^{-1} .$$

A bilinear form which satisfies both (2) and (3) is called *V-elliptic*, it is *symmetric* if

$$a(u, v) = a(v, u) \quad \forall u, v \in V .$$

The optimal constants C_1, C_2 are called ellipticity constants.

As an immediate consequence of these properties we mention that any *V-elliptic symmetric bilinear form* $a(\cdot, \cdot)$ generates an equivalent scalar product in V , i.e.,

$$\{V; (\cdot, \cdot)\} \cong \{V; a(\cdot, \cdot)\}, \quad C_2 \|u\|^2 \leq \|u\|_a^2 \equiv a(u, u) \leq C_1 \|u\|^2 ,$$

for all $u \in V$. Thus, on a theoretical level and up to constants, we can always switch from the energy norm $\|u\|_a$ associated with the variational problem (4) resp. with A to the canonical norm $\|u\|$ in V if a is *V-elliptic* and *symmetric*.

2 Frames

The abstract notion of a *frame* (or, in other words, *stable representation system*) in a Hilbert space was introduced in [6]. A first survey with emphasis on frames was [8], see also [10, Chapter 3], [11, Chapter 3]. A more recent and comprehensive source is the collection [12] which we recommend for further reading.

For the purpose of this section, we will consider the following definitions. Let $F \equiv \{f_k\} \subset V$ be an at most countable system of elements in V .

- We will call F a *frame system* in V if there are two constants $0 < A \leq B < \infty$ such that

$$A\|f\|^2 \leq \sum_k |(f, f_k)|^2 \leq B\|f\|^2 \quad \forall f \in [F]. \quad (5)$$

A frame system F is called *frame* in V if it is dense, i.e., if $[F] = V$. The optimal constants A, B in (5) are the *lower and upper frame bounds*, respectively, their ratio B/A defines the *condition* of F and will be denoted by $\kappa(F)$. A frame (system) F is called *tight* if $A = B$, i.e., if $\kappa(F) = 1$. Finally, a system F is *minimal* if $f_k \notin [F \setminus \{f_k\}]$ for all k , i.e., if the deletion of any f_k from the system reduces the span.

- F is a *Riesz system* in V if there are constants $0 < \tilde{A} \leq \tilde{B} < \infty$ such that for all finite linear combinations $f = \sum_k c_k f_k$

$$\tilde{A}\|f\|^2 \leq \sum_k c_k^2 \leq \tilde{B}\|f\|^2. \quad (6)$$

If $[F] = V$ then a Riesz system is a Riesz basis (the reader should check that in this case indeed *any* $f \in V$ possesses a *unique decomposition*

$$f = \sum_k c_k f_k, \quad (7)$$

which V -converges unconditionally (with respect to rearrangements) to f). Again, the optimal constants \tilde{A}, \tilde{B} , and $\kappa(F) = \tilde{B}/\tilde{A}$ are called lower/upper Riesz bounds and condition of the Riesz system, respectively.

Any finite set in V is a frame system, any finite set of linearly independent elements of V is a Riesz system (the question reduces then just to the size of the frame bounds). Therefore, most of the following discussion is substantial only for infinite F and $\dim V = \infty$.

Orthonormal systems (ONS) in V are obviously frame systems and Riesz systems at the same time (if we have a CONS then it is a frame and Riesz basis), use the orthogonality and Bessel equality. Transformations of a CONS $\{e_j\} \subset V$ give rise to more examples ('real-world' examples are given later), cf. [8, 11].

Example 1. After scaling, a CONS (Riesz or frame system) may lose these properties. E.g.,

$$\left\{e_1, \frac{1}{2}e_2, \dots, \frac{1}{k}e_k, \dots\right\}$$

is not a frame system ($A = 0$!) nor a Riesz system ($\tilde{B} = \infty$!) although it remains an (orthogonal) Schauder basis. More generally, $\{\alpha_k e_k\}$ is a Riesz basis (a frame) in V if and only if $0 < A = \inf_k \alpha_k \leq \sup_k \alpha_k = B < \infty$. Scaled CONS are clearly minimal, as frames they are tight only if $\{\alpha_k\}$ is a constant sequence.

Example 2. The system

$$\left\{e_1; \frac{1}{\sqrt{2}}e_2, \frac{1}{\sqrt{2}}e_2; \frac{1}{\sqrt{3}}e_3, \frac{1}{\sqrt{3}}e_3, \frac{1}{\sqrt{3}}e_3; \dots\right\}$$

is a first example of a tight frame with $A = B = 1$ which is not minimal (e.g., the ‘copies’ of scaled e_k can be deleted without changing the density in V). This is the typical feature of a frame: it contains *redundancy*. The subsystem $\{e_1, \frac{1}{\sqrt{2}}e_2, \frac{1}{\sqrt{3}}e_3, \dots\}$ is not a frame. This is a bad property, and in sharp contrast to Riesz systems: It is obvious from definition (6) that *any subsystem of a Riesz system is again a Riesz system, with the same (or better) Riesz bounds*.

Example 3 [15]. Set

$$F = \left\{f_1 = e_1, f_2 = e_1 + \frac{1}{2}e_2, \dots, f_k = e_{k-1} + \frac{1}{k}e_k, \dots\right\}.$$

Since

$$\frac{1}{2}(f, e_k)^2 - \frac{1}{(k+1)^2}(f, e_{k+1})^2 \leq (f, f_k)^2 \leq 2((f, e_k)^2 + \frac{1}{(k+1)^2}(f, e_{k+1})^2), \quad k \geq 1,$$

this is a frame (Exercise: find good frame bounds!). The crazy thing about this frame is that for any increasing sequence of finite subsystems F_n such that $F_n \rightarrow F$ (i.e., any f_k belongs to all F_n with sufficiently large $n \geq n_0(k)$) the lower frame bound A_n deteriorates: $A_n \rightarrow 0$. To see this in the specific case of $F_n = \{f_1, \dots, f_n\}$, take

$$f = e_1 - 2!e_2 + 3!e_3 - \dots + (-1)^{n+1}n!e_n \in [F_n]$$

as test function in (5). Show that $\kappa(F_n) \geq (n!)^2$ which is a dramatic blow-up. This example is important to have in mind in connection with discrete algorithms based on a frame (compare [15] and our considerations below).

In general, one may consider the whole class of systems F generated by suitable linear transformations T from a CONS (i.e., $f_k = \sum_j t_{kj}e_j$) and study various properties of a system in a systematic way. This is not too hard on an abstract level (compare [8, 16]) and left upon the reader. Without proof, let us formulate the following theorem on the connections between CONS, Riesz bases, and frames (for historical references and more details, see [8] and the article by Benedetto/Walnut in [13]).

Theorem 1 a) F is a CONS (ONS) in V if and only if it is a tight frame (frame system) with $A = B = 1$ and $\|f_k\| = 1$.

b) F is a Riesz basis (Riesz system) in V if and only if it is a minimal frame (frame system). In this case, $\tilde{A} = 1/B$, $\tilde{B} = 1/A$, and the definitions of frame and Riesz condition are consistent.

We turn to the operators associated with a frame which allow us to formulate the main properties and applications of frames for representation purposes. Proofs will be given in a more general context in Section 3. Let F be a frame in V (the consideration of frame systems is similar). Then the *synthesis operator* R given by

$$c = (c_k) \in \ell^2 \quad \longmapsto \quad Rc = \sum_k c_k f_k \in V$$

is well-defined on the sequence space ℓ^2 (with index set inherited from F) and a bounded as linear operator from ℓ^2 to V . Its adjoint $R^* : V \rightarrow \ell^2$ takes the form

$$f \in V \quad \longmapsto \quad R^* f = ((f, f_k)) \in \ell^2$$

and is called *analysis operator*. The boundedness of R and R^* follows exclusively from the upper estimate in the definition (5). The two-sided inequality (5) can be rephrased as

$$A(f, f) \leq \|R^* f\|_{\ell^2}^2 = (RR^* f, f) \leq B(f, f) \quad \forall f \in V,$$

which shows that the symmetric operator $\mathcal{P} = RR^* : V \rightarrow V$ is boundedly invertible with

$$A \text{Id} \leq \mathcal{P} \leq B \text{Id}, \quad \|\mathcal{P}\|_{V \rightarrow V} = B, \quad \frac{1}{B} \text{Id} \leq \mathcal{P}^{-1} \leq \frac{1}{A} \text{Id}, \quad \|\mathcal{P}^{-1}\|_{V \rightarrow V} = \frac{1}{A}. \quad (8)$$

It is assumed in (8) that A, B are the best possible constants in (5), i.e. the frame bounds of F . As a consequence, the spectral condition number of \mathcal{P} coincides with the frame condition:

$$\kappa(\mathcal{P}) = \|\mathcal{P}\|_{V \rightarrow V} \|\mathcal{P}^{-1}\|_{V \rightarrow V} = \kappa(F). \quad (9)$$

The operator \mathcal{P} is called *frame operator* and is central to all applications. Obviously,

$$f = \sum_k (f, f_k) \mathcal{P}^{-1} f_k = \sum_k (\mathcal{P}^{-1} f, f_k) f_k = \sum_k (f, \mathcal{P}^{-1} f_k) f_k \quad \forall f \in V. \quad (10)$$

The system $\tilde{F} = \{\tilde{f}_k = \mathcal{P}^{-1} f_k\}$ is called *dual frame*. It is easy to see that \tilde{F} is indeed a frame, with frame operator \mathcal{P}^{-1} . Since for tight frames $\mathcal{P} = A \text{Id} (= B \text{Id})$, in this (and only this) case \tilde{F} coincides with F up to a constant multiple $1/A$ which means that tight frames are essentially *self-dual* (up to constant scaling).

The equation (10) shows why frames deserve the name *stable representation system*: There is an ℓ^2 -stable procedure of computing coefficients c_k by linear functionals (more precisely, $c_k = (f, \tilde{f}_k)$) for representing arbitrary $f \in V$ by a series with respect to F .

Linearity and stability of representations are essential for many applications. Clearly, uniqueness of representation (which holds only if F is a Riesz basis) is an additional desire but not essential in some other application areas.

The frame decomposition is optimal in a certain sense as was already observed in the paper [6].

Theorem 2 *Let F be a frame in V . The representation of an arbitrary $f \in V$ in (10) is optimal in the sense that if $f = \sum_k c_k f_k$ for some coefficient sequence c then*

$$\sum_k c_k^2 \geq \sum_k |(f, \tilde{f}_k)|^2 .$$

Thus,

$$\| \|f\| \|^2 \equiv \inf_{c: f = \sum_k c_k f_k} \|c\|_{\ell^2}^2 = \sum_k |(f, \tilde{f}_k)|^2 = (\mathcal{P}^{-1}f, f) \quad \forall f \in V .$$

Theorem 2 and (8) show that in the definition of a frame the basic inequalities (5) can be replaced by the requirement

$$\frac{1}{B} \|f\|^2 \leq \| \|f\| \|^2 \leq \frac{1}{A} \|f\|^2 \quad \forall f \in V , \quad (11)$$

which makes no use of the specific scalar product, shows the *robustness* (up to constants) of the frame property with respect to spectrally equivalent changes of the Hilbert space structure on V , and is easier to compare with the definition (6) of a Riesz basis (in the latter case, due to the uniqueness of representation the infimum in the definition of the triple bar norm is superfluous). Note that $\| \|f\| \| = +\infty$ if f has no representation with ℓ^2 -coefficients or has no representation at all (by definition of the infimum taken over an empty set). Thus, density (and a bit more) of F is implicitly assumed in (11).

Computing with frame representations means, in one or the other way, to compute \mathcal{P}^{-1} on certain elements of V , or equivalently, to solve the operator equation

$$\mathcal{P}g = f$$

for given f . It was already proposed in [6] (and repeated by other authors, compare, e.g., [13, Section 8.2]) that simple *Richardson iteration*

$$g^{(n+1)} = g^{(n)} - \omega(\mathcal{P}g^{(n)} - f), \quad n \geq 0 ,$$

with parameter $\omega = 2/(A + B)$ and arbitrary starting element $g^{(0)}$ could be used. This gives (in the best case) a convergence rate

$$\rho_R = \rho(\text{Id} - \omega\mathcal{P}) = 1 - \frac{2}{1 + \kappa(F)} .$$

Note that this convergence rate is exclusively depending on the frame condition (we have used (9)!), which makes tight frames particularly interesting. The method heavily

depends on knowledge about good bounds for A, B which is considered a nontrivial task. Application of the slightly more involved *conjugate gradient method* is possible since \mathcal{P} is symmetric, and would overcome this difficulty resulting in a even better rate if the frame is far from being a tight one. Other iterative methods might be tried as well. There is another tricky point. In many applications, the theoretical investigations are for infinite frames (in infinite-dimensional V) while the real algorithms work with sections of the frame. Example 3 above shows that one has to care about the conditioning of these finite sections (here, using a Riesz basis would simplify the matter).

Finally, note that there is another interesting operator

$$\tilde{\mathcal{P}} = R^*R : \ell^2 \rightarrow \ell^2 ,$$

which is also symmetric (in ℓ^2) but not necessarily invertible. Its matrix representation (with respect to the index set of F) is

$$\tilde{\mathcal{P}} = (\tilde{\mathcal{P}}_{k,j} = (f_j, f_k)) ,$$

which makes the name *Gramian of F* plausible for $\tilde{\mathcal{P}}$. Again, $\tilde{\mathcal{P}}$ can be used for characterizing properties of a frame (see [16] for this kind of analysis in a special case).

We conclude with a list of practically important examples of frames appearing in the literature. All of them are *function frames*, and we will present their simplest one-dimensional versions (with respect to (subspaces of) $L_2(\mathbb{R})$). Generalizations to more general Hilbert and Banach spaces (atomic decompositions, ϕ -transform) can be found in [9] and [13, Chapters 3, 6, and 16].

- **Irregular sampling.** Frames have originated [6] from sampling bandlimited signals at general, irregularly spaced locations. A comprehensive survey of the mathematical and computational aspects of this problem is given by Groechenig and Feichtinger [13, Chapter 8]. The problem is as follows: For which sequences of locations x_n are there constants A, B such that

$$A\|f\|^2 \leq \sum_n |f(x_n)|^2 \leq B\|f\|_2^2$$

for all $f \in B_\omega$ where B_ω is the subspace of all functions $f \in L_2(\mathbb{R})$ such that $\text{supp } \hat{f} \subset [-\omega, \omega]$ (in other words, B_ω can be obtained by taking the inverse Fourier transform of all functions in $L_2(-\omega, \omega)$)? For $f \in B_\omega$ we have

$$f(x) = \frac{1}{2\pi} \int_{\mathbb{R}} \hat{f}(\xi) e^{ix\xi} \chi_{[-\omega, \omega]} d\xi = \int_{\mathbb{R}} f(t) \frac{\sin(\omega(t-x))}{\pi(t-x)} dt ,$$

where \hat{f} denotes the Fourier transform of f . Thus, setting $f_n(t) = \sin(\omega(t-x_n))/(\pi(t-x_n))$ and comparing with the frame definition, we see that the above question of stable reconstruction of $f \in B_\omega$ from the sample values $(f(x_n))$ is equivalent to $F = \{f_n\}$ being a frame in B_ω . A necessary and sufficient condition

for this has recently be found by Jaffard whose work is based on previous contributions by Duffin/Schaeffer and Landau, for details, see [13, Chapter 8]. Going to the Fourier transform domain, we see that there is yet another equivalent formulation (in terms of non-harmonic Fourier series) of the above irregular sampling problem which goes back to Wiener/Paley (see [6]): Under which conditions on the frequencies x_n is the system $\{e^{-ix_n t}\}$ a frame in $L_2(-\omega, \omega)$?

- **Gabor frames.** Gabor frames are function systems of the form

$$g_{mn}(t) = e^{2\pi imbt} g(t - na) , \quad m, n \in \mathbb{Z} ,$$

which are generated by *translation and modulation* from a single function $g \in L_2(\mathbb{R})$. Gabor's original proposal was to use the Gaussian

$$g(t) = C e^{-st^2} .$$

where $s > 0$ is a fixed parameter, and $C > 0$ some scaling constant. The point is that then g_{mn} is a 'minimizer' (i.e., realizes equality) of the following uncertainty principle for localization in time-frequency domain: For any $f \in L_2(\mathbb{R})$,

$$\|f\|^2 \leq 4\pi \|(t - na)f(t)\| \|(\xi - mb)\hat{f}(\xi)\| .$$

Thus, looking for decompositions

$$f(t) = \sum_{m,n} c_{mn} g_{mn}(t)$$

means to look for a decomposition with respect to functions which are *best localized* with respect to the different points in a lattice $\{(na, mb)\}$ in the time-frequency plane. Again, stable decomposition is equivalent to the frame property of the systems $F_{a,b,g} = \{g_{mn}\}$. A detailed discussion of Gabor frames can be found in [13, Chapter 3 and 7]. E.g., Theorem 7.8 there states that for the Gaussian $g(t)$ as defined above $F_{a,b,g}$ is a frame in $L_2(\mathbb{R})$ if and only if $ab < 1$. In addition, [13, Chapter 5] introduces to a class of similar systems, the so-called *Malvar-Wilson bases* or local Fourier bases which have found applications in signal analysis and, more recently, to operator equations (compare work by the Coifman group).

- **Wavelet frames.** If modulation is replaced by *dilation* we arrive at wavelet systems. More precisely, given a normalized function $\psi \in L_2(\mathbb{R})$, we define

$$\psi_{j,i}(t) = 2^{j/2} \psi(2^j t - i) , \quad j, i \in \mathbb{Z} .$$

The classical counterparts of this construction are the Haar and the Faber-Schauder system. These are obtained if the functions depicted in Figure 1 a), b) are used as the basic ψ . Both choices lead to minimal systems. In the Haar case, the resulting wavelet system $F_\psi = \{\psi_{j,i}\}$ is even a CONS in $L_2(\mathbb{R})$. The other system is not a

frame in $L_2(\mathbb{R})$. The system F_ψ resulting from the hat function in Figure 1 c) is the prototype of a *multilevel frame* in Sobolev spaces which has generalizations to finite element multigrid schemes in higher dimensions. Since wavelet systems will be discussed in later sections in connection with efficient adaptive solution strategies for operator equations, we will stop with these examples. More information can be found in [11, 10, 12, 13].

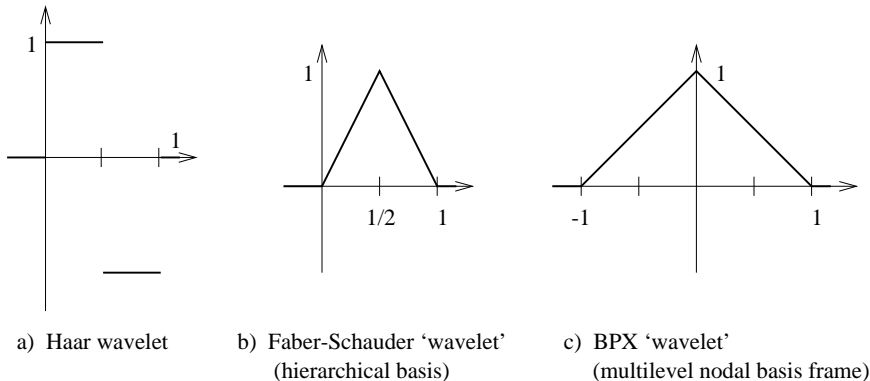


Figure 1: Three basic 'wavelets' ψ

3 Stable space splittings

We will start with some notation which will be consistently used throughout the exposition. Again, V is the basic Hilbert space, with (\cdot, \cdot) resp. $\langle \cdot, \cdot \rangle \equiv \langle \cdot, \cdot \rangle_{V' \times V}$ as basic scalar product resp. duality pairing. Consider a symmetric V -elliptic variational problem (4) to be solved. As discussed in Section 1, $\{V; a(\cdot, \cdot)\}$ is an isomorphic copy of V . Let V_j , $j = 1, 2, \dots$, be an at most countable family of Hilbert spaces, with $(\cdot, \cdot)_j$, $\langle \cdot, \cdot \rangle_j$ introduced similarly. To each V_j we assign its own symmetric V_j -elliptic bilinear form $b_j(\cdot, \cdot) : V_j \times V_j \rightarrow \mathbb{R}$ which in particular means that $\{V_j; b_j(\cdot, \cdot)\}$ are Hilbert spaces. The V_j and $b_j(\cdot, \cdot)$ will be used to create simpler (both in structure and size) auxiliary problems and to compose from their solution operators an approximate inverse to A . The latter is then used as a preconditioner in an iterative method for solving (4), see Section 4 for the details. It is not assumed that the V_j are subspaces of V (but it is implicit that they correspond to certain portions of V , see below).

Denote the Hilbert sum of this family by \tilde{V} , i.e., for

$$\tilde{u} = (u_j), \quad \tilde{v} = (v_j), \quad u_j, v_j \in V_j \quad \forall j$$

set

$$\tilde{a}(\tilde{u}, \tilde{v}) = \sum_j b_j(u_j, v_j)$$

which makes sense as a scalar product on

$$\tilde{V} = \{ \tilde{u} : \tilde{a}(\tilde{u}, \tilde{u}) < \infty \} .$$

Finally, consider bounded linear mappings $R_j : V_j \rightarrow V$. Formally, they can be considered as the components of an operator $R : \tilde{V} \rightarrow V$ given by $R\tilde{u} = \sum_j R_j u_j$.

Definition 3 ([25]) *The system $\{ \{V_j; b_j\}, R_j \}$ gives rise to a **stable splitting** of $\{V; a\}$ which will be expressed by the short-hand notation*

$$\{V; a\} = \sum_j R_j \{V_j; b_j\} , \quad (12)$$

if there are two constants $0 < \tilde{A} \leq \tilde{B} < \infty$ such that

$$\tilde{A}a(u, u) \leq \|u\|^2 \equiv \inf_{\tilde{u} \in \tilde{V}: u=R\tilde{u}} \tilde{a}(\tilde{u}, \tilde{u}) \leq \tilde{B}a(u, u) \quad \forall u \in V . \quad (13)$$

The optimal constants \tilde{A}, \tilde{B} in (13) will be called *lower and upper stability constants*, and their ratio $\kappa = \tilde{B}/\tilde{A}$ *condition of the splitting* (12).

It should be noted that (13) implicitly requires that R makes sense (convergence of the sum if infinitely many V_j are involved) and is surjective, i.e., $\text{ran}(R) = V$. A sufficient condition for R being well-defined and bounded on all of \tilde{V} is that on the set of all *finite* sums

$$u = \sum_{j=1}^n R_j u_j, \quad u_j \in V_j, \quad j = 1, \dots, n, \quad n \geq 1,$$

the following replacement for the lower estimate in (13) is satisfied:

$$a\left(\sum_{j=1}^n R_j u_j, \sum_{j=1}^n R_j u_j\right) \leq \frac{1}{\tilde{A}} \sum_{j=1}^n b_j(u_j, u_j) . \quad (14)$$

Indeed, if $\tilde{u} \in \tilde{V}$ is arbitrary then for the sections $u^{m,n} = \sum_{j=m}^n R_j u_j$ of the series under consideration we have from (14)

$$a(u^{m,n}, u^{m,n}) \leq \frac{1}{\tilde{A}} \sum_{j=m}^n b_j(u_j, u_j) ,$$

where the right-hand side converges to zero for $m, n \rightarrow \infty$ since the infinite sum converges due to $\tilde{u} \in \tilde{V}$. Thus, the partial sums of the series $\sum_j R_j u_j$ form a Cauchy sequence in V and, hence, are convergent to some $u = R\tilde{u} \in V$. Taking $m = 1$ and letting $n \rightarrow \infty$ the boundedness of $R : \tilde{V} \rightarrow V$ comes out: $\|R\|_{\tilde{V} \rightarrow V}^2 \leq 1/\tilde{A}$.

By definition, the adjoint $R^* : V \rightarrow \tilde{V}$ is defined as

$$R^* : u \in V \quad \mapsto R^*u = (R_1^*u, R_2^*u, \dots) \in \tilde{V} ,$$

where the components $R_j^* : V \rightarrow V_j$ are determined by solving the auxiliary variational problems:

$$b_j(R_j^*u, v_j) = a(u, R_j v_j) \quad \forall v_j \in V_j . \quad (15)$$

We postpone the derivation of the familiar representations of these operators (which follow if the duality pairings are used) to the beginning of Section 4 and concentrate on the pure functional-analytic setting. Clearly, R^* is bounded as well ($\|R^*\|_{V \rightarrow \tilde{V}}^2 \leq 1/\tilde{A}$), and we can introduce the two bounded linear operators

$$\mathcal{P} = RR^* : u \in V \mapsto \mathcal{P}u = \sum_j T_j u \in V \quad (T_j = R_j R_j^* : V \rightarrow V) \quad (16)$$

and

$$\tilde{\mathcal{P}} = R^*R : \tilde{u} \rightarrow \tilde{\mathcal{P}}\tilde{u} \in \tilde{V} \quad (17)$$

where $\tilde{\mathcal{P}}$ can be considered as operator matrix which acts according to

$$(\tilde{\mathcal{P}}\tilde{u})_j = \sum_k \overbrace{R_j^* R_k}^{\tilde{P}_{jk}} u_k \quad \forall j .$$

Following some tradition [18, 20], \mathcal{P} is called *Schwarz operator* associated with the stable splitting (12) while the operator matrix associated with $\tilde{\mathcal{P}}$ will be called *extended Schwarz operator* (it is nothing but the generalization of the Gramian for frames discussed in Section 2, and the abstract analog of the matrix of the semi-definite system [26]). We next prove an analog of Theorem 2.

Theorem 4 *The Schwarz operator (16) associated with a stable splitting (12) is symmetric positive definite and has a bounded inverse. Moreover,*

$$\| \|u\| \|^2 = a(\mathcal{P}^{-1}u, u) \quad \forall u \in V ,$$

and

$$\frac{1}{\tilde{B}} \text{Id} \leq \mathcal{P} \leq \frac{1}{\tilde{A}} \text{Id} , \quad \kappa(\mathcal{P}) = \kappa .$$

Proof. See [25, pp.73-75], where the proof is reduced to using Nepomnyashchich's Fictitious Space Lemma. To be self-contained, here is the argument. First, symmetry follows from

$$a(\mathcal{P}u, v) = \tilde{a}(R^*u, R^*v) = \sum_j b_j(R_j^*u, R_j^*v) .$$

As a by-product, note the formula

$$a(\mathcal{P}u, u) = \tilde{a}(R^*u, R^*u) = \sum_j b_j(R_j^*u, R_j^*u) \geq 0 , \quad (18)$$

which also shows that \mathcal{P} is non-negative.

Next consider any $\tilde{v} \in \tilde{V}$ such that $\mathcal{P}u = R\tilde{v}$ (one such \tilde{v} can be given explicitly: $\tilde{v}_0 = R^*u$). Then

$$a(\mathcal{P}u, u)^2 = a(R\tilde{v}, u)^2 = \tilde{a}(\tilde{v}, R^*u)^2 \leq \tilde{a}(\tilde{v}, \tilde{v})\tilde{a}(R^*u, R^*u) = \tilde{a}(\tilde{v}, \tilde{v})a(\mathcal{P}u, u) ,$$

with equality attained for \tilde{v}_0 . Thus,

$$a(\mathcal{P}u, u) = \inf_{\mathcal{P}u=R\tilde{v}} \tilde{a}(\tilde{v}, \tilde{v}) = \|\mathcal{P}u\|^2 .$$

With this intermediate result at hand, we can establish the invertibility of \mathcal{P} . Indeed,

$$a(\mathcal{P}u, u)^2 \leq a(\mathcal{P}u, \mathcal{P}u)a(u, u) \leq \frac{1}{\tilde{A}}\|\mathcal{P}u\|^2 a(u, u) = \frac{1}{\tilde{A}}a(\mathcal{P}u, u)a(u, u) ,$$

where the lower stability estimate has been incorporated. This gives $\mathcal{P} \leq 1/\tilde{A} \cdot \text{Id}$. On the other hand, consider any $\tilde{v} \in \tilde{V}$ such that $u = R\tilde{v}$ (the existence of such \tilde{v} follows from the upper stability estimate). Then we have in the same way as above

$$a(u, u)^2 = \tilde{a}(\tilde{v}, R^*u)^2 \leq \tilde{a}(\tilde{v}, \tilde{v})a(\mathcal{P}u, u) .$$

After taking the infimum with respect to all admissible \tilde{v} and using the upper stability bound, we see that

$$a(u, u)^2 \leq \|u\|a(\mathcal{P}u, u) \leq \tilde{B}a(u, u)a(\mathcal{P}u, u) ,$$

from which $1/\tilde{B} \cdot \text{Id} \leq \mathcal{P}$ follows. This argument includes the proof of positive definiteness, and the invertibility of \mathcal{P} follows from writing

$$\mathcal{P} = \frac{1}{\omega}(\text{Id} - (\text{Id} - \omega\mathcal{P})) \equiv \frac{1}{\omega}(\text{Id} - B) ,$$

where $B = \text{Id} - \omega\mathcal{P}$ is symmetric and satisfies

$$(1 - \frac{\omega}{\tilde{A}})a(u, u) \leq a(Bu, u) \leq (1 - \frac{\omega}{\tilde{B}})a(u, u) .$$

Thus, taking $\omega = 2\tilde{A}\tilde{B}/(\tilde{A} + \tilde{B})$, we get

$$\|B\|_a \leq \frac{\tilde{B} - \tilde{A}}{\tilde{B} + \tilde{A}} < 1 .$$

Now, the invertibility of \mathcal{P} comes from applying a Neumann series argument:

$$\mathcal{P}^{-1} = \omega \sum_{k=0}^{\infty} B^k .$$

Clearly, $\tilde{A} \leq \mathcal{P}^{-1} \leq \tilde{B}$. The arguments show that all estimates are sharp which gives

$$\kappa(\mathcal{P}) = \|\mathcal{P}\|_a \|\mathcal{P}^{-1}\|_a = \frac{\tilde{B}}{\tilde{A}} = \kappa .$$

Finally, to get the expression for $\|u\|^2$, substitute $\mathcal{P}^{-1}u$ for u in the above expression of $\|\mathcal{P}u\|^2$.

Example 4. Let us first make precise why frame theory is a special case (which is already expressed by the notation). Let $F = \{f_j\}$ be a frame in V , set $a(\cdot, \cdot) = (\cdot, \cdot)$. Without loss of generality, we can assume that all $f_j \neq 0$. Denote by V_j the one-dimensional subspace spanned by f_j , and set

$$b_j(u_j, v_j) = \frac{(u_j, v_j)}{(f_j, f_j)} \quad \forall u_j, v_j \in V_j, \quad j \geq 1.$$

As R_j we take the natural embeddings. Then the space \tilde{V} is formed by all sequences

$$\tilde{u} = (u_j) = (c_j f_j) \longleftrightarrow c = (c_j)$$

with

$$\tilde{a}(\tilde{u}, \tilde{u}) = \sum_j c_j^2.$$

which means that \tilde{V} is an isometric copy of the sequence space ℓ^2 used in the frame context. On the basis of (15) one computes

$$\frac{(R_j^* u, f_j)}{(f_j, f_j)} = (u, f_j) \implies R_j^* u = (u, f_j) f_j$$

for all j and after substitution into (18) one arrives at the familiar equation for frames:

$$a(\mathcal{P}u, u) = (\mathcal{P}u, u) = \sum_j |(u, f_j)|^2 \quad \forall u \in V.$$

R and R^* coincide (up to isometry) with the synthesis and analysis operator for frames while the Schwarz operator is nothing but the frame operator. Thus, stable splittings in the sense of the above definition are a generalization of frames in two directions. First, instead of decomposing with respect to a fixed system of elements, best decompositions $u = \sum_j R_j u_j$ are considered where the terms are chosen from *auxiliary spaces* V_j of *arbitrary dimension*. This adds a greater flexibility, especially in connection with parallelization of computations. Secondly, the auxiliary spaces need not be subspaces which is of interest for a number of applications (e.g., for outer approximation schemes).

On the other hand, since frames (and Riesz bases) provide in some sense the most detailed decomposition of a space they are particularly useful in adaptive computations. Also, by clustering techniques (see Section 5) frames may be successfully used to construct more complicated subspace splittings. Last but not least, we may try to introduce methods and notions which have shown their usefulness in the frame case to the case of space splittings. E.g., one might call

$$\{V; a\} = \sum_j \hat{R}_j \{V_j; b_j\}, \quad \hat{R}_j = \mathcal{P}^{-1} R_j,$$

dual space splitting since with this new set of operators \hat{R}_j the generalization of (10)

$$Id = \hat{R}R^* = R\hat{R}^*$$

holds. It has not yet been investigated which of the frame concepts give rise to interesting applications in the area of space splittings vice versa.

Example 5. This example provides a bridge with the material of the next section. Let $V = \mathbb{R}^n$, and A be a real $n \times n$ spd matrix. The scalar product is the Euclidean vector product in \mathbb{R}^n . Set $a(x, y) = (Ax, y)$, $x, y \in V$. Let $V_j \subset V$ denote the j -th coordinate space, i.e., the set of all \mathbb{R}^n vectors where only the j -th coordinate is nonvanishing. Introduce

$$b_j(x, y) = b_j x_j y_j, \quad x, y \in V_j \quad (b_j > 0).$$

Again, R_j are the natural embeddings for subspaces. Since we have a splitting of a finite-dimensional space into a finite sum of subspaces, (13) is guaranteed (with some constants). Direct computation yields

$$(R_i^* x)_i = \frac{1}{b_i} \sum_{j=1}^n a_{ij} x_j, \quad i = 1, \dots, n.$$

Thus, the action of \mathcal{P} (as well as the “operator” matrix $\tilde{\mathcal{P}}$) is given by the matrix $D^{-1}A$ where $D = \text{diag}(b)$. Formally, the introduction of the above splitting has led to a Schwarz operator whose representation is a *preconditioned version* of the initial matrix A . This is a general fact, and the theory of stable splittings can be viewed as a theoretical tool to control the condition number of the preconditioned operator \mathcal{P} via the stability constants \tilde{A}, \tilde{B} of the splitting. Moreover, classical iterative methods (Richardson/Jacobi, Gauss-Seidel/SOR) can be recovered as subspace correction methods [20]. E.g., $b_i = 1/\omega > 0$ leads to the Richardson iteration (with relaxation parameter ω), and $b_i = a_{ii}$ to the Jacobi method. In turn, when applied to the operator matrix $\tilde{\mathcal{P}}$, classical methods such as Richardson iteration or SOR result in modern algorithms for large linear systems such as domain decomposition and multigrid methods.

Example 6. This is the core example from the field of *domain decomposition methods* which has influenced the appearance of the present theory of stable splittings very much. Consider the Poisson equation for Laplace’s equation

$$-\Delta u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega. \quad (19)$$

If we seek for weak solutions, this problem is turned into the form (4) with $V = H_0^1(\Omega)$. As usual,

$$a(u, v) = \int_{\Omega} \nabla u \cdot \nabla v, \quad u, v \in V.$$

Domain decomposition ideas come in if one tries to split this problem into a finite number of similar problems with respect to some subdomains Ω_j . Figure 2 shows the case of two

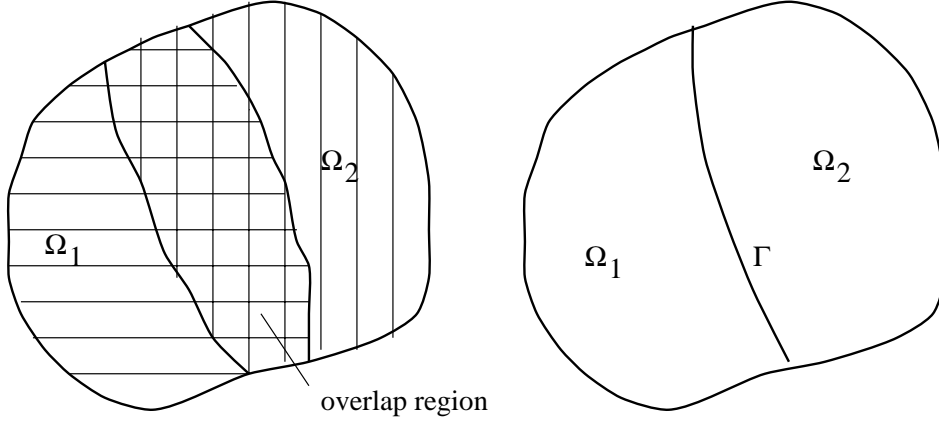


Figure 2: Overlapping and non-overlapping domain decomposition

subdomains Ω_1 and Ω_2 . The left picture corresponds to the case of sufficient *overlap*, the other one to the case of *nonoverlapping domains*.

Set $V_j = H_0^1(\Omega_j)$ and

$$b_j(u, v) = \int_{\Omega_j} \nabla u \cdot \nabla v, \quad u, v \in V_j, \quad j = 1, 2.$$

For functions from V_j defined on Ω_j , extension by zero to the whole domain Ω is a simple choice for R_j , $j = 1, 2$. Then the auxiliary problems (15) are essentially Poisson problems on the domains Ω_j . It turns out that

$$\{H_0^1(\Omega); a\} = R_1\{H_0^1(\Omega_1); b_1\} + R_2\{H_0^1(\Omega_2); b_2\}$$

is a stable splitting in the sense of our definition (use a smooth partition of unity adapted to the domain splitting to see the non-trivial upper inequality (13)). The constants depend on the regularity and the shape of the domains involved. The most critical parameter is the “thickness” of the corridor of overlap $\Omega_1 \cap \Omega_2$, reducing it means blowing up \tilde{B} .

The famous alternating method by H. Schwarz is essentially a (mathematical) algorithm for solving (19) which fits into the algorithms explained in the next Section. Formally, it starts with an arbitrary initial guess $u^{(0)} \in H_0^1(\Omega)$ and does alternately the following:

$$\begin{aligned} v^{(k)} &= u^{(k)} - R_1 B_1^{-1} R_1' (-\Delta u^{(k)} - f) \\ u^{(k+1)} &= v^{(k)} - R_2 B_2^{-1} R_2' (-\Delta v^{(k)} - f) \end{aligned} \quad k = 0, 1, \dots,$$

until a stopping criteria is fulfilled. In each iteration step, the residual $r(u) = -\Delta u - f$ has to be computed twice, for $u^{(k)}$ and the intermediate iterate $v^{(k)}$. R_j' is the restriction operator to Ω_j , R_j the extension-by-zero operator from above, and B_j^{-1} is the solution operator for the Poisson problem in $H_0^1(\Omega_j)$ (with homogeneous Dirichlet boundary

values). An alternative to the alternating method of Schwarz would be to work with the simpler iteration

$$u^{(k+1)} = u^{(k)} - R_1 B_1^{-1} R_1' r(u^{(k)}) - R_2 B_2^{-1} R_2' r(u^{(k)}), \quad k \geq 0.$$

Now only one residual has to be computed, and part of the computation can be done in parallel. The disadvantage is that in analogy to the Jacobi and Gauss-Seidel method, the second method is usually slower (it needs roughly twice as many iterations to reach the same error reduction).

In the nonoverlapping case (right picture in Figure 2), a third space which serves the data on the interface Γ is needed. Its design is related to the trace operator. Set $V_3 = \tilde{H}^{1/2}(\Gamma)$, and take for $R_3 : V_3 \rightarrow V$ some kind of harmonic extension of functions defined on Γ to the whole domain Ω , satisfying

$$\|R_3 f\|_{H^1(\Omega)} \leq C \|f\|_{\tilde{H}^{1/2}(\Gamma)}, \quad R_3 f|_{\Gamma} = f.$$

The exact meaning of $\tilde{H}^{1/2}(\Omega)$ and the construction of a suitable auxiliary form $b_3(\cdot, \cdot)$ which should be V_3 -elliptic and symmetric, will not be explained here (Sobolev spaces of this type incorporate the zero values at the endpoints of Γ and can be defined by interpolation). Under suitable assumptions on the regularity of Γ and Ω , the repaired splitting is again stable. In the practical application, the construction of an (approximate) inverse B_3^{-1} (the so-called interface solver) is now the challenge. For more details and historical references, see [19].

Example 7. Stable splittings have a long tradition in connection with approximation processes. For the understanding of the present example, the booklet [24, Chapter 1-2] gives the necessary background (only the case $q = 2$, $X_1 = X = H$, $X_2 = V \subset H$, $\theta_1 = 0$, $\theta_2 = \gamma > 0$ of [24, Satz 2.4.1] is of interest in the Hilbert space context).

Let

$$V_0 \subset V_1 \subset \dots \subset V_j \subset \dots \subset V = \text{clos}_V(\cup_j V_j) \quad (20)$$

be an increasing sequence of subspaces of V . Assume that V is continuously embedded into another Hilbert space H , and that $H = \text{clos}_H(V)$. To avoid any problems, we also require all V_j to be subspaces of H , too. Fix some $a > 1$ and $\gamma > 0$. We say that, with respect to the family $\{V_j\}$, the pair (H, V) satisfies a *Jackson inequality* if

$$e_j(u)_H \equiv \inf_{v_j \in V_j} \|u - v_j\| \leq C a^{-\gamma j} \|u\|_V \quad \forall u \in V, \quad (21)$$

resp. a *Bernstein inequality* if

$$\|v_j\|_V \leq C a^{\gamma j} \|v_j\|_H \quad \forall v_j \in V_j, \quad (22)$$

uniformly in $j \geq 0$. Often, one uses this abstract setting with $H = L_2(\Omega) \supset V = H^\gamma(\Omega)$. Typically (e.g., for finite element spaces V_j obtained by regular dyadic refinement or in the case of dyadic multiresolution analysis), we have $a = 2$ while the choice of γ

is connected with the degree of approximation obtainable from $\{V_j\}$ resp. with the smoothness of the elements $v_j \in V_j$.

From [24, Satz 2.4.1] it follows that

$$X^s \equiv [H, V]_{s/\gamma, 2} = \sum_j \{V_j; a^{2sj}(\cdot, \cdot)_H\}, \quad 0 < s < \gamma, \quad (23)$$

is a stable splitting for the intermediate interpolation spaces $V \subset X^s \subset H$ which are obtained by the K -method. We have dropped the R_j in the notation which indicates natural embedding for subspaces. The power of this result lies in the fact that these interpolation spaces X^s are often nontrivial spaces (e.g., Sobolev spaces of smoothness parameter $0 < s < \gamma$) while the spaces V_j in the splitting are equipped with scaled $H(= L_2(\Omega))$ -scalar products which means that the auxiliary problems associated with the forms $b_j(\cdot, \cdot)$ are simpler and s -independent (up to a scaling factor).

A useful consequence of (23) is as follows. Let $a_s(\cdot, \cdot)$ denote a symmetric X^s -elliptic bilinear form for some fixed $s \in (0, \gamma)$. Then

$$\{V_J; a_s(\cdot, \cdot)\} = \sum_{j=0}^J \{V_j; a^{2sj}(\cdot, \cdot)_H\} \quad (24)$$

is a stable splitting for each J , with condition κ_J that is uniformly bounded in J : $\kappa_J \leq C\kappa(s)$, where $\kappa(s)$ is the condition of the splitting (23) for that particular s . The constant C only depends on a , s , and the ellipticity constants of $a_s(\cdot, \cdot)$.

Let us prove this result (since this is essentially a result on the conditioning of subsplittings of (23), it deserves a proof due to the counterexamples of Section 2). Denote by $\tilde{A}(s), \tilde{B}(s)$ the stability constants for the splitting (23) (with respect to a fixed scalar product in X^s). The lower estimate is trivial since by definition of the triple bar norms $||| \cdot |||$ (23) resp. $||| \cdot |||_J$ for (24) obviously

$$|||u_J|||_J^2 \geq |||u_J|||^2 \geq \tilde{A}(s)||u_J||_{X^s}^2 \geq c\tilde{A}(s)a_s(u_J, u_J) \quad \forall u_J \in V_J.$$

For the upper bound, by definition of the infimum for any $u_J \in V_J$ there are $v_j \in V_j$, $j \geq 0$, such that $u_J = \sum_j v_j$ and

$$\sum_j a^{2sj} \|v_j\|_H^2 \leq 2\tilde{B}(s) \|u_J\|_{X^s}^2 \leq C\tilde{B}(s) a_s(u_J, u_J).$$

Again, the last step is the ellipticity assumption for $a_s(\cdot, \cdot)$. Define

$$\hat{v}_j = v_j, \quad j < J, \quad \hat{v}_J = \sum_{j \geq J} v_j = u_J - \sum_{j=0}^{J-1} v_j \in V_J.$$

Thus, $u_J = \sum_{j=0}^J \hat{v}_j$ and since

$$\|\hat{v}_J\|^2 \leq \left(\sum_{j \geq J} a^{-js} \cdot (a^{js} \|v_j\|_H) \right)^2$$

$$\begin{aligned}
&\leq \left(\sum_{j \geq J} a^{-2js} \right) \left(\sum_{j \geq J} a^{2js} \|v_j\|_H^2 \right) \\
&\leq C a^{-2js} \sum_{j \geq J} a^{2js} \|v_j\|_H^2
\end{aligned}$$

we immediately obtain

$$\|u_J\|_J^2 \leq \sum_{j=0}^J a^{2js} \|\hat{v}_j\|^2 \leq C \sum_j a^{2js} \|v_j\|_H^2 \leq C \tilde{B}(s) a_s(u_J, u_J) .$$

This proves the assertion.

In the same way one shows that the stability of (23) implies the stability of

$$X^s = \sum_j \{W_j; a^{2sj}(\cdot, \cdot)_H\} , \quad 0 < s < \gamma , \quad (25)$$

where $W_0 = V_0$, $W_j = \text{ran}(P_j - P_{j-1})$, $j \geq 1$, are “*difference*” spaces generated by an arbitrary sequence of *uniformly bounded projectors* $P_j : H \rightarrow V_j$. Note that in this case X^s is stably decomposed into a *direct sum* of subspaces which is the counterpart to a Riesz basis construction.

A special case of a family of projectors with the above properties are *orthogonal projectors*. With this, a possible advantage of direct sum decompositions becomes transparent: they may be even good for H and some of the dual spaces $X^{-s} = (X^s)'$. Indeed, let the linear operators $Q_j : H \rightarrow V_j$ be such that $Q_j v_j = v_j$ (projection property) and

$$(v - Q_j v, v_j)_H = 0 \quad \forall v_j \in V_j \quad \forall v \in V \quad (\text{orthogonality}) .$$

Then obviously

$$v = \underbrace{Q_0 v}_{w_0} + \underbrace{(Q_1 - Q_0)v}_{w_1} + \dots + \underbrace{(Q_j - Q_{j-1})v}_{w_j} + \dots$$

with mutually orthogonal terms $w_j \in W_j$ as defined above and

$$\|v\|_H^2 = \sum_j \|w_j\|_H^2 .$$

The latter equality can be reinterpreted as stability assertion for the case $s = 0$ in (25) and $H = X^0$.

Analogous results can be obtained for $-\gamma < s < 0$ by a duality argument which we leave upon the reader (Hint: Use

$$\|\tilde{v}\|_{X^{-s}} := \sup_{v \neq 0} \frac{(\tilde{v}, v)_H}{\|v\|_{X^s}} , \quad \tilde{v} \in H ,$$

the mutual H -orthogonality of the components of the decompositions of \tilde{v} and v , together with the two-sided stability estimates for the splitting (25). Note that X^{-s} is defined as the closure of H under the above norm.)

The corresponding assertions are not true for the splitting (23). E.g., putting $s = 0$ in the right-hand side of (23) does not lead to a stable splitting for H . Indeed, if one takes $u = u_0 \in V_0$ then due to the monotonicity assumption (20), the decompositions

$$u_0 = \frac{1}{J+1} \underbrace{(u_0 + u_0 + \dots + u_0)}_{J+1 \text{ times}}$$

are admissible in the definitions for the triple bar norms associated with both (23) and (24). Thus,

$$|||u_0|||^2 \leq |||u_0|||_J^2 \leq (J+1)^{-1} \|u_0\|^2 \quad \forall J \geq 0.$$

This shows the non-stability of the splitting for H .

What concerns the splitting of $\{V_J; (\cdot, \cdot)_H\}$ given by (24) for $s = 0$, we have

$$\|u_J\|_H^2 \leq \left(\sum_{j=0}^J \|v_j\|_H \right)^2 \leq (J+1) \sum_{j=0}^J \|v_j\|_H^2 \quad \forall u_J = \sum_{j=0}^J v_j$$

which shows (together with the above opposite inequality for a particular function $u_0 \in V_0 \subset V_J$) that $A_J = 1/(J+1)$. The reader can easily verify that, on the other hand, $\hat{B}_J = 1$ if $V_{J-1} \subset V_J$ is strict. Thus, normally the condition of the finite splittings (24) is exactly $\kappa_J = J+1$, i.e., exhibits moderate growth in the number of levels.

Further examples will be given later.

4 Iterative solvers

In this section we come to some consequences of the notion of stable space splittings for the construction of iterative solution methods for solving variational problems such as (4) and its discretizations. Throughout this section, assume that the splitting (12) is stable. We will use the notation introduced in Section 3. Recall that

$$a(u, v) = \langle Au, v \rangle, \quad b_j(u_j, v_j) = \langle B_j u_j, v_j \rangle_j,$$

defines invertible operators $A : V \rightarrow V'$, $B_j : V_j \rightarrow V'_j$, and introduce the dual operators $R'_j : V' \rightarrow V'_j$ by

$$\langle R'_j \Phi, v_j \rangle_j = \langle \Phi, R_j v_j \rangle \quad \forall \Phi \in V', v_j \in V_j.$$

For given $\Phi \in V'$, define $\phi_j \in V_j$ by solving the auxiliary variational problems

$$b_j(\phi_j, v_j) = \Phi(R_j v_j) = \langle \Phi, R_j v_j \rangle = \langle R'_j \Phi, v_j \rangle_j \quad \forall v_j \in V_j,$$

$j \geq 1$. Observe that $\tilde{\phi} = (\phi_j) \in \tilde{V}$ (to this end, represent $\Phi = Au_\Phi$ and check that $\tilde{\phi} = R^* u_\Phi$), and set $\phi = R\tilde{\phi} = \sum_j R_j \phi_j$.

Then we have the following obvious

Theorem 5 *The unique solution $u = u_\Phi \in V$ of the variational problem (4) (or, what is the same, of the operator equation $Au = \Phi$ in V') is also the unique solution of the operator equation*

$$\mathcal{P}u = \phi \quad (26)$$

in V . Moreover, we have $u = R\tilde{u}$ for any solution $\tilde{u} \in \tilde{V}$ of the operator equation

$$\tilde{\mathcal{P}}\tilde{u} = \tilde{\phi} \quad (27)$$

in \tilde{V} . In addition, we have the representations

$$\mathcal{P} = \left(\sum_j R_j B_j^{-1} R_j' \right) A \equiv CA \quad (28)$$

and

$$\tilde{\mathcal{P}} = (\tilde{\mathcal{P}}_{ij}), \quad \tilde{\mathcal{P}}_{ij} = B_i^{-1} R_i' A R_j : V_j \rightarrow V_i.$$

Thus, the introduction of a stable space splitting allows us to switch to several equivalent formulations, with the advantage that a proper choice of the space splitting (i.e., of V_j, B_j, R_j) leads via Theorem 4 to a well-conditioned operator \mathcal{P} . Note that the operator

$$C = \sum_j R_j B_j^{-1} R_j' \equiv \sum_j \hat{T}_j : V' \rightarrow V \quad (29)$$

is symmetric with respect to $\langle \cdot, \cdot \rangle$ and can be considered as *preconditioner* or *approximate inverse* for A . Recall that $T_j = \hat{T}_j A$ in comparison with (16).

Following this setup, several iterative methods for solving (4) based on auxiliary subproblems associated with the given stable splitting can be introduced and analyzed (see [20, 22, 21, 25, 27]). To avoid discussions about situations which are of no direct practical use, we will assume from now on that the number of spaces V_j is finite, i.e., we consider a finite stable splitting

$$\{V; a\} = \sum_{j=1}^J R_j \{V_j; b_j\}. \quad (30)$$

(Clearly, real computer implementations will also require finite-dimensionality of the V_j). In their abstract form, the following algorithms have been formulated in [20]:

(AS) Additive Schwarz method. Starting with an initial guess $u^{(0)} \in V$, repeat

$$u^{(n+1)} = u^{(n)} - \omega \sum_{j=1}^J (T_j u^{(n)} - \phi_j) \equiv u^{(n)} - \omega \sum_{j=1}^J \hat{T}_j (A u^{(n)} - \Phi),$$

until a stopping criteria is satisfied.

(MS) Multiplicative Schwarz method. Starting with an initial guess $u^{(0)} \in V$, repeat

$$\begin{aligned} v^{(0)} &= u^{(n)} , \\ v^{(j)} &= v^{(j-1)} - \omega(T_j v^{(j-1)} - \phi_j) \equiv v^{(j-1)} - \omega \hat{T}_j(Av^{(j-1)} - \Phi) , \quad j = 1, \dots, J , \\ u^{(n+1)} &= v^{(J)} , \end{aligned}$$

until a stopping criteria is satisfied.

(SMS) Symmetric multiplicative Schwarz method. Starting with an initial guess $u^{(0)} \in V$, repeat

$$\begin{aligned} v^{(0)} &= u^{(n)} , \\ v^{(j)} &= v^{(j-1)} - \omega \hat{T}_j(Av^{(j-1)} - \Phi) , \quad j = 1, \dots, J , \\ v^{(2J-j+1)} &= v^{(2J-j)} - \omega \hat{T}_j(Av^{(2J-j)} - \Phi) , \quad j = J, \dots, 1 , \\ u^{(n+1)} &= v^{(2J)} , \end{aligned}$$

until a stopping criteria is satisfied.

Note that the ordering of the subproblems has impact only on the multiplicative methods **(MS)** and **(SMS)**, and that one iteration step with **(SMS)** is composed of two iteration steps of **(MS)** where the second visits the subproblems in reverse order. The *relaxation parameter* $\omega > 0$ has to be chosen appropriately (e.g., in dependence on knowledge on \tilde{A}, \tilde{B}). If the choice of ω seems to be a problem, the symmetry of C can be explored and a preconditioned conjugate gradient method used (cf. [23, Section 9.4.4]):

(SCG) Schwarz-preconditioned conjugate gradient method. Choose an initial $u^{(0)} \in V$, and set

$$r^{(0)} = Au^{(0)} - \Phi \in V' , \quad p^{(0)} = Cr^{(0)} = \sum_{j=1}^J \hat{T}_j r^{(0)} \in V , \quad \rho^{(0)} = \langle r^{(0)}, p^{(0)} \rangle .$$

Repeat

$$\begin{aligned} q &= Ap^{(n)} \in V' , & \lambda &= \rho^{(n)} / \langle q, p^{(0)} \rangle \\ u^{(n+1)} &= u^{(n)} - \lambda p^{(n)} \\ r^{(n+1)} &= r^{(n)} - \lambda q \\ q &= Cr^{(n+1)} = \sum_{j=1}^J \hat{T}_j r^{(n+1)} \in V , & \rho^{(n+1)} &= \langle r^{(n+1)}, q \rangle \\ p^{(n+1)} &= q + (\rho^{(n)} / \rho^{(n+1)}) p^{(n)} \end{aligned}$$

until a stopping criteria is satisfied.

A further alternative is to derive from **(SMS)** a preconditioner to be used within a pcg algorithm (again, the motivation might be to avoid problems with ω or to improve the convergence rate, i.e., to add robustness to the multiplicative solver).

An elegant way to analyze the above iterations is to rewrite them in terms of classical iterative methods applied to the operator matrix $\tilde{\mathcal{P}}$ (which is now of size J) as proposed in [26, 27]. Write $\tilde{\mathcal{P}}$ as a sum of strictly lower triangular, diagonal and strictly upper triangular parts

$$\tilde{\mathcal{P}} = \tilde{\mathcal{L}} + \tilde{\mathcal{D}} + \tilde{\mathcal{U}},$$

and denote by $\tilde{\text{Id}}$ the identity matrix (operator) in \tilde{V} . The reader may verify that (with respect to $\{\tilde{V}; \tilde{a}\}$)

$$\tilde{\mathcal{D}}^* = \tilde{\mathcal{D}}, \tilde{\mathcal{U}}^* = \tilde{\mathcal{L}} \iff b_i(\tilde{\mathcal{P}}_{ij}u_j, v_i) = b_j(u_j, \tilde{\mathcal{P}}_{ji}v_i) \quad \forall u_j \in V_j, v_i \in V_i.$$

In the considerations below, we set $u^{(n)} = R\tilde{u}^{(n)}$, $n \geq 0$, and consider a linear iteration scheme in \tilde{V} (which is now a space of J -vectors $\tilde{v} = (v_1, \dots, v_J)^T$) to generate the sequence $(\tilde{u}^{(n)})$:

$$\tilde{u}^{(n+1)} = \tilde{u}^{(n)} - \tilde{\mathcal{N}}(\tilde{\mathcal{P}}\tilde{u}^{(n)} - \tilde{\phi}), \quad n \geq 0. \quad (31)$$

Lemma 6 a) *Damped Richardson iteration: If $\tilde{\mathcal{N}} \equiv \tilde{\mathcal{N}}_{AS} = \omega\tilde{\text{Id}}$ then the iteration (31) is equivalent to **(AS)**.*

b) *Damped Jacobi iteration: Assume that $\tilde{\mathcal{D}}$ is invertible in \tilde{V} . If $\tilde{\mathcal{N}} \equiv \omega\tilde{\mathcal{D}}^{-1}$ then the iteration (31) is equivalent to **(AS)** for the modified stable splitting*

$$\{V; a\} = \sum_{j=1}^J R_j \{V_j; a(R_j \cdot, R_j \cdot)\}.$$

c) *Richardson-SOR method: If $\tilde{\mathcal{N}} \equiv \tilde{\mathcal{N}}_{MS} = (\frac{1}{\omega}\tilde{\text{Id}} + \tilde{\mathcal{L}})^{-1}$ then the iteration (31) is equivalent to **(MS)**. In analogy to part b) above, the original SOR-method defined by $\tilde{\mathcal{N}} = (\frac{1}{\omega}\tilde{\mathcal{D}} + \tilde{\mathcal{L}})^{-1}$ is a particular case.*

d) *Richardson-SSOR method: If*

$$\tilde{\mathcal{N}} \equiv \tilde{\mathcal{N}}_{SMS} = (\frac{1}{\omega}\tilde{\text{Id}} + \tilde{\mathcal{U}})^{-1}(\frac{2}{\omega}\tilde{\text{Id}} - \tilde{\mathcal{D}})(\frac{1}{\omega}\tilde{\text{Id}} + \tilde{\mathcal{L}})^{-1}$$

*then the iteration (31) is equivalent to **(SMS)**.*

Proof. We give some hints, details of the algebraic computations are left upon the reader. To see the stated equivalences, one has to compare the iteration operator

$$M = \text{Id} - R\tilde{\mathcal{N}}R^*$$

for the sequence $u^{(n)}$ in V resulting from a certain choice of $\tilde{\mathcal{N}}$ with the iteration operator of the corresponding abstract method.

a) Here,

$$M_{AS} = \text{Id} - \omega RR^* = \text{Id} - \omega \mathcal{P}.$$

Now, compare with the definition of **(AS)**.

b) \tilde{D} is invertible in \tilde{V} if (and only if)

$$\omega_0 b_j(u_j, u_j) \leq \tilde{b}_j(u_j, u_j) \equiv a(R_j u_j, R_j u_j) \leq \omega_1 b_j(u_j, u_j) \quad \forall u_j \in V_j, j = 1, \dots, J, \quad (32)$$

holds with two constants $0 < \omega_0 \leq \omega_1 < \infty$. The upper estimate is obvious from (12): $\omega_1 \leq 1/\tilde{A}$. The lower estimate does not follow from (12) and needs to be assumed (if all V_j are finite-dimensional it is certainly true with some positive ω_0). By the two-sided estimate in (32) and the definition of stable splittings it follows that the modified splitting is also stable (with other \tilde{A}, \tilde{B}). Check that the Schwarz operator associated with this new splitting takes the form $R\tilde{D}^{-1}R^*$, and compare with the result of part a).

c) First of all, since $\tilde{\mathcal{L}}$ is strictly lower triangular, we can compute $\tilde{\mathcal{N}}_{MS}$ by the formula

$$\tilde{\mathcal{N}}_{MS} = \left(\frac{1}{\omega}\tilde{\text{Id}} + \tilde{\mathcal{L}}\right)^{-1} = \omega \sum_{k=0}^{J-1} (-\omega\tilde{\mathcal{L}})^k .$$

Thus,

$$M_{MS} = \text{Id} + \sum_{k=1}^J (-\omega)^k R\tilde{\mathcal{L}}^{k-1}R^*$$

which has to be compared with the iteration matrix for **(MS)**:

$$(\text{Id} - \omega T_J) \dots (\text{Id} - \omega T_1) = \text{Id} + \sum_{k=1}^J (-\omega)^k \sum_{J \geq j_k > \dots > j_1 \geq 1} T_{j_k} \dots T_{j_1} .$$

Now, prove by induction that the coefficients in the two operator polynomials (with respect to the variable $(-\omega)$) coincide. For $k = 1$ this is certainly true since

$$\sum_{j=1}^J T_j = \mathcal{P} = RR^*$$

(this also shows that the ‘linear’ part of the iteration **(MS)** coincides with the iteration operator M_{AS} of **(AS)**). For $k = 2$,

$$R\tilde{\mathcal{L}}R^* = \sum_{i=1}^J R_i \sum_{j=1}^{i-1} (R_i^* R_j) R_j^* = \sum_{1 \leq j < i \leq J} T_i T_j ,$$

and so on. Compare [25, p.78].

d) As in part c), one has

$$M_{MS}^* = (\text{Id} - \omega T_1) \dots (\text{Id} - \omega T_J) = \text{Id} - R\left(\frac{1}{\omega}\tilde{\text{Id}} + \tilde{\mathcal{U}}\right)^{-1}R^* .$$

Therefore, we have to check that

$$\begin{aligned} M_{MS}^* M_{MS} &= \left(\text{Id} - R\left(\frac{1}{\omega}\tilde{\text{Id}} + \tilde{\mathcal{U}}\right)^{-1}R^*\right) \left(\text{Id} - R\left(\frac{1}{\omega}\tilde{\text{Id}} + \tilde{\mathcal{L}}\right)^{-1}R^*\right) \\ &= \text{Id} - R\left(\frac{1}{\omega}\tilde{\text{Id}} + \tilde{\mathcal{U}}\right)^{-1} \left(\frac{2}{\omega}\tilde{\text{Id}} + \tilde{\mathcal{L}} + \tilde{\mathcal{U}} - \tilde{\mathcal{P}}\right) \left(\frac{1}{\omega}\tilde{\text{Id}} + \tilde{\mathcal{L}}\right)^{-1}R^* \\ &= \text{Id} - R\tilde{\mathcal{N}}_{MSM}R^* = M_{SMS} \end{aligned}$$

This completes the proof of the lemma.

We can now state the main result of this section. Let us denote the *asymptotic convergence rate* of an iterative method governed by an iteration operator M by its spectral radius:

$$\rho = \lim_{n \rightarrow \infty} \|M^n\|^{1/n} .$$

Below, we choose the operator norm $\|\cdot\|_a$ in $\{V; a\}$ in this definition. If M is symmetric $\{V; a\}$ then the above formula simplifies to

$$\rho = \|M\|_a = \sup_{a(u,u) \leq 1} a(Mu, u) .$$

This simplifying argument applies to **(AS)** and **(SMS)** directly, for **(MS)** we use the obvious relationship

$$\rho_{MS} \leq \|M_{MS}\|_a = \sqrt{\|M_{MS}^* M_{MS}\|_a} = \sqrt{\|M_{SMS}\|_a} = \rho_{SMS}^{1/2} .$$

Theorem 7 *Let (12) be a stable space splitting into finitely many auxiliary spaces V_j , $j = 1, \dots, J$, with stability constants \tilde{A}, \tilde{B} , and condition κ . condition κ .*

a) *The additive method **(AS)** converges for $0 < \omega < 2\tilde{A}$, at rate*

$$\rho_{AS,\omega} = \max\{|1 - \omega/\tilde{A}|, |1 - \omega/\tilde{B}|\} .$$

The optimal convergence rate is achieved for $\omega^ = 2\tilde{A}\tilde{B}/(\tilde{A} + \tilde{B})$:*

$$\rho_{AS}^* = \rho_{AS,\omega^*} = \inf_{0 < \omega < 2\tilde{A}} \rho_{AS,\omega} = 1 - \frac{2}{1 + \kappa} . \quad (33)$$

*For the related method **(SCG)**, we have the estimate*

$$\|u - u^{(n)}\|_a \leq 2 \left(1 - \frac{2}{1 + \sqrt{\kappa}}\right)^n \|u - u^{(0)}\|_a , \quad n \geq 1 , \quad (34)$$

for the guaranteed error reduction in the energy norm.

b) *For the multiplicative algorithms **(MS)** and **(SMS)**, convergence is guaranteed if $0 < \omega < 2/\omega_1$, where $\omega_1 \leq 1/\tilde{A}$ is defined by (32). The convergence rate can be estimated by*

$$\rho_{MS,\omega}^2 \leq \rho_{SMS} \leq 1 - \frac{\omega(2 - \omega\omega_1)}{\tilde{B}\|\text{Id} + \omega\tilde{\mathcal{L}}\|_{\tilde{a}}^2} .$$

The optimal rates satisfy

$$(\rho_{MS,\omega}^*)^2 \leq \rho_{SMS}^* \leq 1 - \frac{1}{\tilde{B}(2\|\tilde{\mathcal{L}}\|_{\tilde{a}} + \omega_1)} . \quad (35)$$

If no additional assumptions on the splitting are available, we can use

$$\|\tilde{\mathcal{L}}\|_a \leq \frac{\lceil \log_2(2J) \rceil}{2\tilde{A}} \quad (36)$$

to arrive at

$$(\rho_{MS,\omega}^*)^2 \leq \rho_{SMS}^* \leq 1 - \frac{1}{\log_2(4J) \cdot \kappa}. \quad (37)$$

Proof (see [27]) for more details). Part a) is just a repetition of the convergence rate estimates for the Richardson resp. conjugate gradient iteration applied to the equation (26) with the spd operator \mathcal{P} (sharp spectral bounds for the latter are contained in Theorem 4). The reader is recommended to look at the classical results in [23, Sections 4.4 and 9.4].

We concentrate on part b). Together with the above observations on computing convergence rates ρ , Lemma 6 allows us to reduce the statements on the convergence behavior for the multiplicative Schwarz methods to norm estimates in \tilde{V} for the case of the symmetric iteration (**SMS**). Indeed,

$$\begin{aligned} a(M_{SMS}u, u) &= a(u, u) - a(R\tilde{\mathcal{N}}_{SMS}R^*u, u) \\ &= a(u, u) - \tilde{a}\left(\left(\frac{2}{\omega}\tilde{\text{Id}} - \tilde{\mathcal{D}}\right)\tilde{w}, \tilde{w}\right) \\ &\leq a(u, u) - \left(\frac{2}{\omega} - \omega_1\right)\tilde{a}(\tilde{w}, \tilde{w}), \quad \tilde{w} = \left(\frac{1}{\omega}\tilde{\text{Id}} + \tilde{\mathcal{L}}\right)^{-1}R^*u. \end{aligned}$$

Here, we have used Lemma 6 d), $\left(\left(\frac{1}{\omega}\tilde{\text{Id}} + \tilde{\mathcal{U}}\right)^{-1}\right)^* = \left(\frac{1}{\omega}\tilde{\text{Id}} + \tilde{\mathcal{L}}\right)^{-1}$, and $\tilde{\mathcal{D}} \geq \omega_1\tilde{\text{Id}}$ (see (32)). Now we continue with

$$\tilde{a}(\tilde{w}, \tilde{w}) \geq \left\|\frac{1}{\omega}\tilde{\text{Id}} + \tilde{\mathcal{L}}\right\|_{\tilde{a}}^{-2}\tilde{a}(R^*u, R^*u) \geq \frac{\omega^2}{(1 + \omega\|\tilde{\mathcal{L}}\|_{\tilde{a}})^2}\tilde{a}(R^*u, R^*u)$$

and

$$\tilde{a}(R^*u, R^*u) = a(\mathcal{P}u, u) \geq \frac{1}{B}a(u, u).$$

Putting things together, we get the estimate for $\rho_{SMS,\omega}$. Minimization with respect to $0 < \omega < 2/\omega_1$ gives (35).

The proof of (36) is given in detail in [27, p. 174–176], and will not be repeated here. The estimate (37) follows if (36) and $\omega_1 \leq 1/\tilde{A}$ are substituted into (35). This concludes the proof of Theorem 7.

The above theorem leads to a number of questions. First, from the estimates it becomes clear that the multiplicative algorithms are never much worse than the additive Schwarz method. That the logarithmical factor $\log_2(4J) \asymp \log_2 J$ is sharp shows the following example [29] of a sequence of stable space splittings. The splittings are (in the spirit of Example 5) inherited from a sequence of real spd matrices $A_n = L_n + D_n + L_n^T$, where $D_n = \text{Id}_n$. The spectrum of the A_n is uniformly bounded away from 0 and ∞ :

$$0 < \lambda_{\min} \leq \lambda_{\min}(A_n) \leq \lambda_{\max}(A_n) \leq \lambda_{\max} < \infty,$$

On the other hand, there exist unit vectors $x_n \in \mathbb{C}^n$ such that for a sequence of real numbers $\alpha_n \asymp \log_2 n$ the equality $(L_n x_n, x_n) = i\alpha_n$ holds for $n \rightarrow \infty$ (throughout this example, (\cdot, \cdot) denotes the scalar product in \mathbb{C}^n).

For this matrix setting, it is easy to see (compare Example 5 and [23]) that **(SMS)** coincides with the usual SSOR-method for A_n . By the above assumptions, using the identity

$$\|M_{SMS}\|_a = 1 - \frac{2/\omega - 1}{\|A_n^{-1/2}(1/\omega \text{Id}_n + L_n)\|^2}, \quad n \geq 1,$$

(see [23, Lemma 4.4.23]), one is lead to estimating

$$\|A_n^{-1/2}(1/\omega \text{Id}_n + L_n)\|^2 \geq \lambda_{\max}^{-1} \|1/\omega \text{Id}_n + L_n\|^2 \geq \lambda_{\max}^{-1} r(1/\omega \text{Id}_n + L_n)^2,$$

where $r(B)$ denotes the numerical radius of B ([23, Section 2.9.5]). Using the vectors x_n , we see that

$$r(1/\omega \text{Id}_n + L_n) = \sup_{0 \neq x} \frac{|((1/\omega \text{Id}_n + L_n)x, x)|}{(x, x)} \geq \left| \frac{1}{\omega} + i\alpha_n \right|,$$

which yields

$$\|A^{-1/2}(1/\omega \text{Id}_n + L_n)\|^2 \geq c(\omega^{-2} + (\log_2 n)^2), \quad n \rightarrow \infty,$$

for some positive constant c . Now substitute into the formula for $\|M_{SMS}\|_a$ and take the infimum with respect to $0 < \omega < 2$.

A family of matrices which have the above properties are the finite sections

$$A_n = (a_{ij} = a_{i-j} : i, j = 1, \dots, n), \quad a_k = \begin{cases} 1 & , \quad k = 0 \\ \frac{c_0 \sin(k\pi/2)}{k} & , \quad k \neq 0 \end{cases},$$

of a ℓ^2 -bounded and boundedly ℓ^2 -invertible Toeplitz operator A_∞ ($|c_0| < 2/\pi$). The auxiliary vectors are given by

$$x_n = \frac{1}{\sqrt{n}}(i, i^2, \dots, i^n), \quad n \geq 1.$$

The reader may consult [29] for details and references on Toeplitz matrices. Challenge: The above argument does not show that the convergence rate for **(MS)** deteriorates in the same manner (even though numerical evidence provided in [29] shows that this is the case). Is there an argument to close this little gap?

On the other hand, in most of the classical cases it is well known that SOR-like methods outperform the simpler Richardson iteration. Artificial examples show that the multiplicative algorithms may be arbitrarily fast compared to the additive method **(AS)**. Many sufficient conditions for convergence of multiplicative schemes have been discussed (see, e.g., [20, 22, 21]). Some of them are, at the same time, methods to establish the stability of a given splitting. The reader is recommended to learn from the above sources about the so-called strengthened Cauchy-Schwarz inequalities associated with a splitting which, in their simplest form,

$$a(R_i u_i, R_j v_j) \leq \gamma_{ij} \sqrt{b_i(u_i, u_i)} \sqrt{b_j(v_j, v_j)} \quad \forall u_i \in V_i, v_j \in V_j, \quad (38)$$

are equivalent to estimates for the norms of the operator entries $\tilde{\mathcal{P}}_{ij}$ of the extended Schwarz operator $\tilde{\mathcal{P}}$. Indeed, (38) is equivalent to $\|\tilde{\mathcal{P}}_{ij}\|_{V_j \rightarrow V_i} = \|\tilde{\mathcal{P}}_{ji}\|_{V_i \rightarrow V_j} \leq \gamma_{ij}$. Exercise: Consider in detail the case $J = 2$ and give the best possible estimates. This would also cover the two iterative methods given in connection with Example 6.

The theory has been developed so far exclusively for *linear spd* problems. This is certainly a drawback. Attempts to slightly remove these assumptions will be discussed later.

5 Modifications of space splittings

This very short section emphasizes the fact that abstract concepts could pay off if used in a systematic way. Due to the theorems of the previous two sections we now know that various iterative solution schemes are governed by the availability of two-sided stability estimates such as (13) for a space splitting (12). This makes it possible to introduce some general operations on stable space splittings which allow us to modify a given one, in order to adapt it to a specific application or to optimize the implementation with respect to a given hardware platform. We list a few of them.

Refinement and clustering. The method of *refinement* consists in splitting in a given stable splitting (say, in (30)) the auxiliary spaces further:

$$\{V_j; b_j\} = \sum_{k=1}^{I_j} R_{jk} \{V_{jk}; b_{jk}\}, \quad j = 1, \dots, J, \quad (39)$$

Now, denote the stability constants of the j -th splitting in (39) by \tilde{A}_j, \tilde{B}_j . Then the *refined splitting*

$$\{V; a\} = \sum_{j=1}^J \sum_{k=1}^{I_j} R_j R_{jk} \{V_{jk}; b_{jk}\}, \quad j = 1, \dots, J. \quad (40)$$

again satisfies the stability definition, with constants

$$\tilde{A}_{ref} \geq \tilde{A} \min_{j=1, \dots, J} \tilde{A}_j, \quad \tilde{B}_{ref} \leq \tilde{B} \max_{j=1, \dots, J} \tilde{B}_j.$$

Even though these are very rough estimates (not equalities for the optimal stability constants), they give first hints on which modifications are useful. If $I_j = 1$, the statement tells us that spectrally equivalent changes of the bilinear forms in the spaces V_j are admissible as long as they are uniformly tight and correctly scaled. This was implicitly used in the argument for Lemma 6, b).

Clustering is the inverse operation. Given a stable splitting into many component spaces (such as (40)), we will build several groups which we replace then by a single auxiliary space. Let us assume that the grouping is according to (39). Then the original

splitting turns into the *clustered splitting* of (40). Having access to the stability constants of the splittings (40) and (39) we get now

$$\tilde{A} \equiv (\tilde{A}_{ref})_{clust} \geq \tilde{A}_{ref} (\max_{j=1,\dots,J} \tilde{B}_j)^{-1} ,$$

$$\tilde{B} \equiv (\tilde{B}_{ref})_{clust} \leq \tilde{B}_{ref} (\min_{j=1,\dots,J} \tilde{A}_j)^{-1} .$$

Clustering is very useful if one starts with frame decompositions and clusters them such that larger spaces with certain properties result. Load balancing on parallel architectures with a small to moderate number of processors might be one motivation. Below we will show some applications of this technique. The refinement/clustering schemes are schematically illustrated in Figure 3.

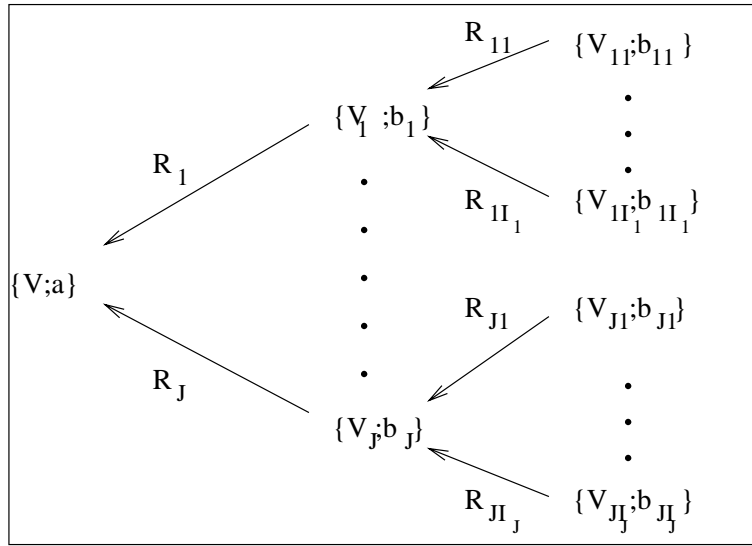


Figure 3: Refinement and clustering

Selection is a critical but very important procedure. It is visualized in Figure 4. Each space V_j in (30) is replaced by its subspace \bar{V}_j while the bilinear form is preserved by restriction (taking the trivial subspace $\{0\}$ is allowed, in this case V_j is simply removed from the original splitting and the index j may be dropped from the set of indices). Consider the subspace

$$\bar{V} = \sum_j R_j \bar{V}_j$$

in V . We may hope for having produced a new stable *selected space splitting* from (30):

$$\{\bar{V}; a\} = \sum_{j=1}^J R_j \{\bar{V}_j; b_j\} . \quad (41)$$

The difficulty with this procedure is that it is almost equivalent to selecting *subsplittings* of a given splitting which can make, in analogy with the situation for frames, the

condition of a splitting much worse. However, in the multilevel applications (as was shown in Example 7) some of the most natural selections (the *finite sections* of an infinite splitting) are uniformly stable. Further examples of the selection procedure will be discussed later, in connection with adaptivity (see also [25, Section 4.2.2], be aware of inconsistencies there). Recall again that for subsplittings of stable splittings into direct sums of auxiliary spaces a deterioration of condition numbers is not possible.

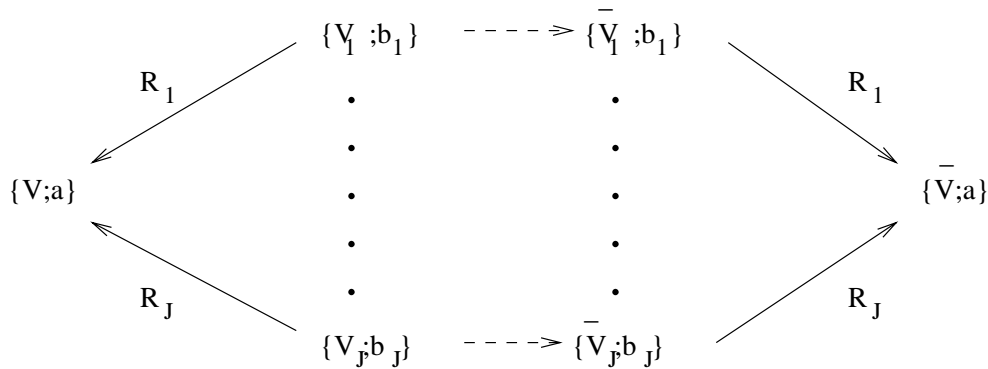


Figure 4: Selection

Let us finally mention that we have successfully used **tensor products** of stable space splittings of univariate schemes in connection with sparse grid spaces and anisotropic constant coefficient operators in higher dimensions [28]. With the above techniques, a lot of practical situations can be covered (i.e., seemingly independent applications can be derived from one and the same basic splitting). However, characterizations of all useful, i.e., stability-preserving, transformations of stable splittings and frames have not yet been studied in a general and systematic way.

References

- [1] P. R. Halmos, Introduction to Hilbert space and the theory of spectral multiplicity, Chelsea, New York, 1951.
- [2] P. R. Halmos, A Hilbert space problem book, Van Nostrand, Princeton, 1967.
- [3] G. Helmberg, Introduction to spectral theory in Hilbert space, North-Holland, Amsterdam 1969 (p. 36–47).
- [4] J. Wloka, Funktionalanalysis mit Anwendungen, de Gruyter, Berlin, 1971.
- [5] J. Weidmann, Lineare Operatoren in Hilberträumen, Teubner, Stuttgart, 1976.
- [6] R. Duffin, A. Schaeffer, A class of nonharmonic Fourier series, TAMS 72 (1952), 341–366.

- [7] R. Young, An introduction to nonharmonic Fourier series, Academic Press, New York, 1980.
- [8] C. Heil, D. F. Walnut, Continuous and discrete wavelet transform, SIAM Review 31 (1989), 628–666.
- [9] K. Groechenig, Describing functions: atomic decompositions versus frames, Monatsh. Math. 112 (1992), 1–42.
- [10] C. K. Chui, An introduction to wavelets, Academic Press, Boston, 1992 (Chapter 1 + 3).
- [11] I. Daubechies, Ten lectures on wavelets, CBMS-NSF Reg. Conf. Ser. Appl. Math. v. 61, SIAM, Philadelphia, 1992 (Chapter 3).
- [12] Wavelets and Their Applications (M. R. Ruskai, G. Beylkin, R. Coifman, I. Daubechies, S. Mallat, Y. Meyer, L. Raphael, eds.), Jones & Bartlett, Boston, 1992 (Chapter V).
- [13] Wavelets: Mathematics and Applications (J. J. Benedetto, M. W. Frazier, eds.), CRC Press, Boca Raton, 1994 (up-to-date information on frames, related issues, and their applications).
- [14] C. K. Chui, X. Shi, Bessel sequences and affine frames. J. Appl. Comput. Harm. Anal. 1 (1993), 29–49.
- [15] O. Christensen, Frames and the projection method. J. Appl. Comput. Harm. Anal. 1 (1993), 50–53.
- [16] A. Ron, Z. Shen, Affine systems in $L_2(\mathbb{R}^d)$: the analysis of the analysis operator, J. Funct. Anal. (to appear), Preprint December 1995, Univ. of Wisconsin, Madison.
- [17] P. L. Lions, On the Schwarz alternating method, In: Proc. 1st Int. Symp. on DDM for PDE (R. Glowinski, G. H. Golub, G. A. Meurant, J. Periaux, eds.), SIAM, Philadelphia, 1987.
- [18] M. Dryja, O. Widlund, Towards a unified theory of domain decomposition for elliptic problems. In: Proc. 3rd Int. Symp. on DDM for PDE (T. Chan, R. Glowinski, J. Periaux, O. Widlund, eds.), SIAM, Philadelphia, 1990, .
- [19] T. Chan, T. Mathew, Domain Decomposition Methods, Acta Numerica 94, Cambr. Univ. Press, 1994, 61–143.
- [20] J. Xu, Iterative methods by space decomposition and subspace correction, SIAM Review 34 (1992), 581–613.

- [21] J. H. Bramble, Multigrid methods, Pitman Research Notes in Mathematical Sciences v. 294, Longman Sci.&Techn., Harlow, Essex, 1993.
- [22] H. Yserentant, Old and new convergence proofs for multigrid methods, Acta Numerica 93, Cambr. Univ. Press, 1993, 285–326.
- [23] W. Hackbusch, Iterative solution of large sparse systems of equations. Appl. Math. Sci. vol. 95, Springer, New York, 1994.
- [24] P. L. Butzer, K. Scherer, Approximationsprozesse und Interpolationsmethoden, Bibliogr. Institut, Mannheim, 1968.
- [25] P. Oswald, Multilevel Finite Element Approximation: Theory and Application, Teubner Skripten zur Numerik, Teubner, Stuttgart, 1994 (Chapter 4.1).
- [26] M. Griebel, Multilevelverfahren als Iterationsverfahren über Erzeugendensystemen, Teubner Skripten zur Numerik, Teubner, Stuttgart, 1994.
- [27] M. Griebel and P. Oswald, Remarks on the abstract theory of additive and multiplicative Schwarz methods, Numer. Math. 70 (1995), 163–180.
- [28] M. Griebel and P. Oswald, Tensor-product-type subspace splittings and multilevel iterative methods for anisotropic problems, Adv. Comput. Math. 4 (1995), 171–206.
- [29] P. Oswald, On the convergence rate of SOR: A worst case estimate. Computing 52 (1994), 245–255.