Chapter 2
Adaptive Multiscale Methods for the Numerical Treatment of Systems of PDEs

Angela Kunoth

Abstract These notes are concerned with numerical analysis issues arising in the solution of certain systems involving stationary and instationary linear variational problems. Standard examples are second order elliptic boundary value problems, where particular emphasis is placed on the treatment of essential boundary conditions, and linear parabolic equations. These operator equations serve as a core ingredient for control problems where in addition to the state, the solution of the PDE, a control is to be determined which together with the state minimizes a certain tracking-type objective functional. Having assured that the variational problems are well-posed, we discuss numerical schemes based on B-splines and B-spline-type wavelets as a particular multiresolution discretization methodology. The guiding principle is to devise fast and efficient solution schemes which are optimal in the number of arithmetic unknowns. We discuss optimal conditioning of the system matrices, numerical stability of discrete formulations, and adaptive approximations.

2.1 Introduction

Multilevel ingredients have for a variety of partial differential equations (PDEs) proved to achieve more efficient solution schemes than methods based on approximating the solution with respect to a fixed fine grid. The latter simple approach leads to the problem to solve a large ill-conditioned system of linear equations. The success of multilevel methods is due to the fact that solutions often exhibit a multiscale behaviour which one naturally wants to exploit. Among the first such schemes were multigrid methods. The basic idea of multigrid schemes is to successively solve smaller versions of the linear system which can be interpreted as discretizations with respect to coarser grids. Here 'efficiency of the scheme' means that one can solve the problem with respect to the finest grid with an amount

A. Kunoth (✉)
Mathematical Institute, University of Cologne, Cologne, Germany
e-mail: kunoth@math.uni-koeln.de

T. Lyche et al. (eds.), *Splines and PDEs: From Approximation Theory to Numerical Linear Algebra*, Lecture Notes in Mathematics 2219,
https://doi.org/10.1007/978-3-319-94911-6_2

of arithmetic operations which is proportional to the number of unknowns on this finest grid. Multigrid schemes provide an asymptotically optimal *preconditioner* for the original system on the finest grid. The search for such optimal preconditioners was one of the major topics in the solution of elliptic boundary value problems for many years. Another multiscale preconditioner which has this property is the BPX-preconditioner proposed first in [8]. It was proved to be asymptotically optimal with techniques from Approximation Theory in [27, 61]. In the context of isogeometric analysis, the BPX-preconditioner was further substantially optimized in [10]; this will be detailed in Sect. 2.2.

Wavelets as a particular example of a multiscale basis were constructed with compact support in the 1980s [36]. While mainly used for signal analysis and image compression, they were discovered to also provide optimal preconditioners in the above sense for elliptic boundary value problems [27, 47]. It was soon realized that biorthogonal spline-wavelets are better suited for the numerical solution of elliptic PDEs since they allow to work with piecewise polynomials instead of the only implicitly defined original wavelets [36], in addition to the fact that orthogonality of the Daubechies wavelets with respect to L_2 cannot really be exploited for elliptic PDEs. The principal ingredient that allows to prove optimality of the preconditioner are *norm equivalences* between Sobolev norms and sequence norms of weighted wavelet expansion coefficients. Optimal conditioning of the resulting linear system of equations can be achieved by applying the Fast Wavelet Transform together with a weighting in terms of an appropriate diagonal matrix. The terminology ‘wavelets’ here and in the sequel is to mean that these are classes of multiscale bases with three main properties: (R) Riesz basis property for the underlying function spaces, (L) locality of the basis functions, (CP) cancellation properties, all of which are detailed in Sect. 2.4.1.

After these initial results, research on using wavelets for numerically solving elliptic PDEs has gone into different directions. The original constructions in [18, 36] and many others are based on using the Fourier transform. Thus, these constructions provide bases for function spaces only on all of $\mathbb{R}$ or $\mathbb{R}^n$. In order for these tools to be applicable for the solution of PDEs which naturally live on a bounded domain $\Omega \subset \mathbb{R}^n$, there arose the need for having available constructions on bounded intervals without, of course, loosing the above mentioned properties (R), (L) and (CP). The first such systematic construction of biorthogonal spline-wavelets on [0, 1] (and, by tensor products, on $[0, 1]^n$) was provided in [34].

Aside from the investigations to provide appropriate bases, the built-in potential of *adaptivity* for wavelets has played a prominent role when solving PDEs, on account of the fact that wavelets provide a locally supported Riesz basis for a whole range of function spaces. The key issue is to approximate the solution of the variational problem on an infinite-dimensional function space by the least amount of degrees of freedom up to a certain prescribed accuracy. Many approaches use wavelet coefficients in a heuristic way, i.e., judging approximation quality by the size of the wavelet coefficients together with thresholding. In contrast, *convergence* of wavelet-based adaptive methods for stationary variational problems was investigated systematically in [19–21]. These schemes are particularly designed to provide *optimal complexity* of the schemes: they provide the solution in a total amount of arithmetic operations which is comparable to the wavelet-best N-term approximation of the solution. This means that, given a prescribed tolerance, to find a sparse representation of the solution by extracting the N largest expansion coefficients of the solution during the solution process.

As soon as one aims at numerically solving a variational problem which can no longer be formulated in terms of a single elliptic operator equation such as a saddle point problem, one is faced with the problem of numerical stability. This means that finite approximations of the continuous well-posed problem may be ill-posed, obstructing its efficient numerical solution. This issue will also be addressed below.

In these notes, I also would like to discuss the potential proposed by wavelet methods for the following classes of problems. First, we will be concerned with second order elliptic PDEs with a particular emphasis placed on treating essential boundary conditions. Another interesting class that will be covered are linear parabolic PDEs which are formulated in full weak space-time from [66]. Then *PDE-constrained control problems* guided by elliptic boundary value problems are considered, leading to a *system* of elliptic PDEs. The starting point for designing efficient solution schemes are wavelet representations of continuous well-posed problems in their variational form. Viewing the numerical solution of such a discretized, yet still infinite-dimensional operator equation as an approximation helps to discover multilevel preconditioners for elliptic PDEs which yield *uniformly bounded condition numbers*. *Stability issues* like the LBB condition for saddle point problems are also discussed in this context. In addition, the compact support of the wavelets allows for sparse representations of the implicit information contained in systems of PDEs, the *adaptive approximation* of their solution.

More information and extensive literature on applying wavelets for more general PDEs addressing, among other things, the connection between adaptivity and nonlinear approximation and the evaluation of nonlinearities may be found in [16, 24, 25].

These notes are structured as follows. In Sect. 2.2, we begin with a simple elliptic PDE in variational form in the context of isogeometric analysis. For this problem, we address additive, BPX-type preconditioners and provide the main ingredients for showing optimality of the scheme with respect to the grid spacing. In Sect. 2.3, several well-posed variational problem classes are compiled to which later several aspects of the wavelet methodology are applied. The simplest example is a linear elliptic boundary value problem for which we derive two forms of an operator equation, the simplest one consisting just of one equation for homogeneous boundary conditions and a more complicated one in form of a saddle point problem where nonhomogeneous boundary conditions are treated by means of Lagrange multipliers. In Sect. 2.3.4, we consider a full weak space-time form of a linear parabolic PDE. These three formulations are then employed for the following classes of PDE-constrained control problems. In the *distributed control problems* in Sect. 2.3.5 the control is exerted through the right hand side of the PDE, while in *Dirichlet boundary control problems* in Sect. 2.3.6 the Dirichlet boundary condition serves this purpose. The most potential for adaptive methods to be discussed below are control problems constrained by parabolic PDEs as formulated in Sect. 2.3.7.

Section 2.4 is devoted to assembling necessary ingredients and basic properties of wavelets which are required in the sequel. In particular, Sect. 2.4.4 collects the essential construction principles for wavelets on bounded domains which do not rely on Fourier techniques, namely, multiresolution analyses of function spaces and the concept of stable completions. In Sect. 2.5 we formulate the problem classes introduced in Sect. 2.3 in wavelet coordinates and derive in particular for the control problems the resulting systems of linear equations arising from the optimality conditions. Section 2.6 is devoted to the iterative solution of these systems. We investigate fully iterative schemes on uniform grids and show that the resulting systems can be solved in the wavelet framework together with a nested iteration strategy with an amount of arithmetic operations which is proportional to the total number of unknowns on the finest grid. Finally, in Sect. 2.6.2 a wavelet-based adaptive scheme for the distributed control problem constrained by elliptic or parabolic PDEs as in [29, 44] will be derived together with convergence results and complexity estimates, relying on techniques from Nonlinear Approximation Theory.

Throughout these notes we will employ the following notational convention: the relation $a \sim b$ will always stand for $a \lesssim b$ and $b \lesssim a$ where the latter inequality means that b can be bounded by some constant times a uniformly in all parameters on which a and b may depend. Norms and inner products are always indexed by the corresponding function space. $L_p(\Omega)$ are for $1 \leq p \leq \infty$ the usual Lebesgue spaces on a domain Ω, and $W_p^k(\Omega) \subset L_p(\Omega)$ denote for $k \in \mathbb{N}$ the Sobolev spaces of functions whose weak derivatives up to order k are bounded in $L_p(\Omega)$. For $p = 2$, we write as usual $H^k(\Omega) = W_2^k(\Omega)$.

2.2 BPX Preconditioning for Isogeometric Analysis

For a start, we consider linear elliptic PDEs in the framework of isogeometric analysis, combining modern techniques from computer aided design with higher order approximations of the solution. In this context, one exploits that the solution exhibits a certain *smoothness*. We treat the physical domain by means of a regular B-spline mapping from the parametric domain $\hat{\Omega} = (0, 1)^n$, $n \geq 2$, to the physical domain Ω. The numerical solution of the PDE is computed by means of tensor product B-splines mapped onto the physical domain. We will construct additive BPX-type multilevel preconditioners and show that they are asymptotically optimal. This means that the spectral condition number of the resulting preconditioned stiffness matrix is independent of the grid spacing h. Together with a nested iteration scheme, this enables an iterative solution scheme of optimal linear complexity. The theoretical results are substantiated by numerical examples in two and three space dimensions. The results of this section are essentially contained in [10].

We consider linear elliptic partial differential operators of order $2r = 2, 4$ on the domain Ω in variational form: for given $f \in H^{-r}(\Omega)$, find $u \in H_0^r(\Omega)$ such that

$$a(u, v) = \langle f, v \rangle \quad \text{for all } v \in H_0^r(\Omega) \tag{2.1}$$

holds. Here the energy space is $H_0^r(\Omega)$, a subset of the Sobolev space $H^r(\Omega)$, the space of square integrable functions with square integrable derivatives up to order r, containing homogeneous Dirichlet boundary conditions for $r = 1$ and homogeneous Dirichlet and Neumann derivatives for $r = 2$. The bilinear form $a(\cdot, \cdot)$ is derived from the linear elliptic PDE operator in a standard fashion, see, e.g., [7]. For example, the Laplacian is represented as $a(v, w) = \int_\Omega \nabla v \cdot \nabla w \, dx$. In order for the problem to be well-posed, we require the bilinear form $a(\cdot, \cdot) : H_0^r(\Omega) \times H_0^r(\Omega) \to \mathbb{R}$ to be symmetric, continuous and coercive on $H_0^r(\Omega)$. With $\langle \cdot, \cdot \rangle$, we denote on the right hand side of (2.1) the dual form between $H^{-r}(\Omega)$ and $H_0^r(\Omega)$. Our model problem (2.1) covers the second order Laplacian with homogeneous boundary conditions

$$-\Delta u = f \quad \text{on } \Omega, \qquad u|_{\partial\Omega} = 0, \tag{2.2}$$

as well as fourth order problems with corresponding homogeneous Dirichlet boundary conditions,

$$\Delta^2 u = f \quad \text{on } \Omega, \qquad u|_{\partial\Omega} = \mathbf{n} \cdot \nabla u|_{\partial\Omega} = 0 \tag{2.3}$$

where $\partial\Omega$ denotes the boundary of Ω and $\mathbf{n}$ the outward normal derivative at $\partial\Omega$. These PDEs serve as prototypes for more involved PDEs like Maxwell's equation or PDEs for linear and nonlinear elasticity. The reason we formulate these model problems of order $2r$ involving the parameter r is that this exhibits more clearly the order of the operator and the scaling in the subsequently used characterization of Sobolev spaces $H^r(\Omega)$. Thus, for the remainder of this section, the parameter $2r$ denoting the order of the PDE operator is fixed.

The assumptions on the bilinear form $a(\cdot, \cdot)$ entail that there exist constants $0 < c_A \le C_A < \infty$ such that the induced self-adjoint operator $\langle Av, w \rangle := a(v, w)$ satisfies the isomorphism relation

$$c_A \|v\|_{H^r(\Omega)} \le \|Av\|_{H^{-r}(\Omega)} \le C_A \|v\|_{H^r(\Omega)}, \quad v \in H_0^r(\Omega). \tag{2.4}$$

If the precise format of the constants in (2.4) does not matter, we abbreviate this relation as $\|v\|_{H^r(\Omega)} \lesssim \|Av\|_{H^{-r}(\Omega)} \lesssim \|v\|_{H^r(\Omega)}$, or shortly as

$$\|Av\|_{H^{-r}(\Omega)} \sim \|v\|_{H^r(\Omega)}. \tag{2.5}$$

Under these conditions, Lax-Milgram's theorem guarantees that, for any given $f \in H^{-r}(\Omega)$, the operator equation derived from (2.1)

$$Au = f \quad \text{in } H^{-r}(\Omega) \tag{2.6}$$

has a unique solution $u \in H_0^r(\Omega)$, see, e.g., [7].

In order to approximate the solution of (2.1) or (2.6), we choose a finite-dimensional subspace of $H_0^r(\Omega)$. We will construct these approximation spaces by using tensor products of B-splines as specified next.

2.2.1 B-Spline Discretizations

Our construction of optimal multilevel preconditioners will rely on tensor products so that principally any space dimension $n \in \mathbb{N}$ is permissible as long as storage permits; the examples cover the cases $n = 2, 3$. As discretization space, we choose in each spatial direction B-splines of the same degree p on uniform grids and with maximal smoothness. We begin with the univariate case and define B-splines on the interval [0, 1] recursively with respect to their degree p. Given this positive integer p and some $m \in \mathbb{N}$, we call $\Xi := \{\xi_1, \ldots, \xi_{m+p+1}\}$ a p-open knot vector if the knots are chosen such that

$$0 = \xi_1 = \ldots = \xi_{p+1} < \xi_{p+2} < \ldots < \xi_m < \xi_{m+1} = \ldots = \xi_{m+p+1} = 1, \tag{2.7}$$

i.e., the boundary knots 0 and 1 have multiplicity $p + 1$ and the interior knots are single. For Ξ, B-spline functions of degree p are defined following the well-known Cox-de Boor recursive formula, see [38]. Starting point are the piecewise constants for $p = 0$ (or characteristic functions)

$$N_{i,0}(\zeta) = \begin{cases} 1, & \text{if } 0 \leq \xi_i \leq \zeta < \xi_{i+1} < 1, \\ 0, & \text{otherwise,} \end{cases} \tag{2.8}$$

with the modification that the last B-spline $N_{m,0}$ is defined also for $\zeta = 1$. For $p \geq 1$ the B-splines are defined as

$$N_{i,p}(\zeta) = \frac{\zeta - \xi_i}{\xi_{i+p} - \xi_i} N_{i,p-1}(\zeta) + \frac{\xi_{i+p+1} - \zeta}{\xi_{i+p+1} - \xi_{i+1}} N_{i+1,p-1}(\zeta), \qquad \zeta \in [0, 1], \tag{2.9}$$

with the same modification for $N_{m,p}$. Alternatively, one can define the B-splines explicitly by applying divided differences to truncated powers [38]. This gives a set of m B-splines that form a basis for the space of *splines*, that is, piecewise polynomials of degree p with $p - 1$ continuous derivatives at the internal knots ξ_ℓ for $\ell = p + 2, \ldots, m$. (Of course, one can also define B-splines on a knot sequence with multiple internal knots which entails that the spline space is not of maximal smoothness.) For $p = 1$, the B-splines are at least $C^0([0, 1])$ which suffices for the discretization of elliptic PDEs of order 2, and for $p = 2$ they are $C^1([0, 1])$ which suffices for $r = 2$. By construction, the B-spline $N_{i,p}$ is supported in the interval $[\xi_i, \xi_{i+p+1}]$.

These definitions are valid for an arbitrary spacing of knots in Ξ (2.7). Recall from standard error estimates in the context of finite elements, see, e.g., [7], that *smooth* solutions of elliptic PDEs can be approximated best with discretizations on a uniform grid. Therefore, in this section, we assume from now on that the grid is *uniform*, i.e., $\xi_{\ell+1} - \xi_\ell = h$ for all $\ell = p + 1, \ldots, m$.

For n space dimensions, we employ tensor products of the one-dimensional B-splines. We take in each space dimension a p-open knot vector Ξ and define on the closure of the parametric domain $\hat{\Omega} = (0, 1)^n$ (which we also denote by $\hat{\Omega}$ for simplicity of presentation) the spline space

$$\begin{aligned} S_h(\hat{\Omega}) &:= \operatorname{span}\left\{ B_i(\mathbf{x}) := \prod_{\ell=1}^{n} N_{i_\ell,p}(x_\ell),\ i = 1, \ldots, N := mn,\ \mathbf{x} \in \hat{\Omega} \right\} \\ &=: \operatorname{span}\left\{ B_i(\mathbf{x}), i \in \mathscr{I},\ \mathbf{x} \in \hat{\Omega} \right\}. \end{aligned} \tag{2.10}$$

In the spirit of isogeometric analysis, we suppose that the computational domain Ω can also described in terms of B-splines. We assume that the computational domain Ω is the image of a mapping $\mathbf{F} : \hat{\Omega} \to \Omega$ with $\mathbf{F} := (F_1, \ldots, F_n)^T$ where each component F_i of $\mathbf{F}$ belongs to $S_{\bar{h}}(\hat{\Omega})$ for some given $\bar{h}$. In many applications, the geometry can be described in terms of a very coarse mesh, namely, $\bar{h} \gg h$. Moreover, we suppose that $\mathbf{F}$ is invertible and satisfies

$$\|D^\alpha \mathbf{F}\|_{L_\infty(\hat{\Omega})} \sim 1 \ \text{ for } \ |\alpha| \le r. \tag{2.11}$$

This assumption on the geometry can be weakened in the sense that the mapping $\mathbf{F}$ can be a piecewise C^∞ function on the mesh with respect to $\bar{h}$, independent of h, or the domain Ω may have a multi-patch representation. This means that one can allow Ω also to be the union of domains Ω_k where each one parametrized by a spline mapping of the parametric domain $\hat{\Omega}$.

We now define the approximation space for (2.6) as

$$V_h^r := \{v_h \in H_0^r(\Omega) :\ v_h \circ \mathbf{F} \in S_h(\hat{\Omega})\}. \tag{2.12}$$

We will formulate three important properties of this approximation space which will play a crucial role later for the construction of the BPX-type preconditioners. The first one is that we suppose from now on that the B-spline basis is *normalized* with respect to L_2, i.e.,

$$\|B_i\|_{L_2(\hat{\Omega})} \sim 1,\ \text{ and, thus, also } \|B_i \circ \mathbf{F}^{-1}\|_{L_2(\Omega)} \sim 1 \text{ for all } i \in \mathscr{I}. \tag{2.13}$$

Then one can derive the following facts [10].

Theorem 1 *Let $\{B_i\}_{i\in\mathscr{I}}$ be the B-spline basis defined in (2.10) and normalized as in (2.13), $N = \#\mathscr{I}$ and V_h^r as in (2.12). Then we have*

(S) Uniform stability with respect to $L_2(\Omega)$
For any $\mathbf{c} \in \ell_2(\mathscr{I})$,

$$\left\| \sum_{i=1}^{N} c_i \, B_i \circ \mathbf{F}^{-1} \right\|_{L_2(\Omega)}^2 \sim \sum_{i=1}^{N} |c_i|^2 =: \|\mathbf{c}\|_{\ell_2}^2, \qquad \mathbf{c} := (c_i)_{i=1,\ldots,N}; \tag{2.14}$$

(J) Direct or Jackson estimates

$$\inf_{v_h \in V_h^r} \|v - v_h\|_{L_2(\Omega)} \lesssim h^s \, |v|_{H^s(\Omega)} \text{ for any } v \in H^s(\Omega),\ 0 \le s \le r+1, \tag{2.15}$$

where $|\cdot|_{H^s(\Omega)}$ denotes the Sobolev seminorm of highest weak derivatives s;

(B) Inverse or Bernstein estimates

$$\|v_h\|_{H^s(\Omega)} \lesssim h^{-s} \|v_h\|_{L_2(\Omega)} \text{ for any } v_h \in V_h^r \text{ and } 0 \le s \le r. \tag{2.16}$$

In all these estimates, the constants are independent of h but may depend on $\mathbf{F}$*, i.e., Ω, on the polynomial degree p and on the spatial dimension n.*

In the next section, we construct BPX-type preconditioners for (2.6) in terms of approximations with (2.12) and show their optimality.

2.2.2 Additive Multilevel Preconditioners

The construction of optimal preconditioners are based on a *multiresolution analysis* of the underlying energy function space $H_0^r(\Omega)$. As before, $2r \in \{2, 4\}$ stands for the order of the PDEs we are solving and is always kept fixed.

We first describe the necessary ingredients within an abstract basis-free framework, see, e.g., [24]. Afterwards, we specify the realization for the parametrized tensor product spaces in (2.12).

Let $\mathscr{V}$ be a sequence of strictly nested spaces V_j, starting with some fixed coarsest index $j_0 > 0$, determined by the polynomial degree p which determines the support of the basis functions (which also depends on Ω), and terminating with a highest resolution level J,

$$V_{j_0} \subset V_{j_0+1} \subset \cdots \subset V_j \subset \cdots \subset V_J \subset H_0^r(\Omega). \tag{2.17}$$

The index j denotes the level of resolution defining approximations on a grid with dyadic grid spacing $h = 2^{-j}$, i.e., we use from now on the notation V_j instead of V_h to indicate different grid spacings. Then, V_J will be the space relative to the finest grid 2^{-J}. We associate with $\mathscr{V}$ a sequence of linear projectors $\mathscr{P} := \{P_j\}_{j \geq j_0}$ with the following properties.

Properties 1 We assume that

(P1) P_j maps $H_0^r(\Omega)$ onto V_j,
(P2) $P_j P_\ell = P_j$ for $j \leq \ell$,
(P3) $\mathscr{P}$ is uniformly bounded on $L_2(\Omega)$, i.e., $\|P_j\|_{L_2(\Omega)} \lesssim 1$ for any $j \geq j_0$ with a constant independent of j.

These conditions are satisfied, for example, for $L_2(\Omega)$-orthogonal projectors, or, in the case of splines, for the quasi-interpolant proposed and analyzed in [65, Chapter 4]. The second condition (P2) ensures that the differences $P_j - P_{j-1}$ are also projectors for any $j > j_0$. We define next a sequence $\mathscr{W} := \{W_j\}_{j \geq j_0}$ of complement spaces

$$W_j := (P_{j+1} - P_j)V_{j+1} \tag{2.18}$$

which then yields the direct (but not necessarily orthogonal) decomposition

$$V_{j+1} = V_j \oplus W_j. \tag{2.19}$$

Thus, for the finest level J, we can express V_J in its *multilevel decomposition*

$$V_J = \bigoplus_{j=j_0-1}^{J-1} W_j \tag{2.20}$$

upon setting $W_{j_0-1} := V_{j_0}$. Setting also $P_{j_0-1} := 0$, the corresponding multilevel representation of any $v \in V_J$ is then

$$v = \sum_{j=j_0}^{J} (P_j - P_{j-1})v. \tag{2.21}$$

We now have the following result which will be used later for the proof of the optimality of the multilevel preconditioners.

Theorem 2 *Let $\mathscr{P}, \mathscr{V}$ be as above where, in addition, we require that for each V_j, $j_0 \leq j \leq J$, a Jackson and Bernstein estimate as in Theorem 1 (J) and (B) hold with $h = 2^{-j}$. Then one has the function space characterization*

$$\|v\|_{H^r(\Omega)} \sim \left(\sum_{j=j_0}^{J} 2^{2rj} \|(P_j - P_{j-1})v\|_{L_2(\Omega)}^2 \right)^{1/2} \quad \text{for any } v \in V_J. \tag{2.22}$$

Such a result holds for much larger classes of function spaces, Sobolev or even Besov spaces which are subsets of $L_q(\Omega)$ for general q, possibly different from 2 and for any function $v \in H^r(\Omega)$, then with an infinite sum on the right hand side, see, e.g., [24]. The proof of Theorem 2 for such cases heavily relies on tools from approximation theory and can be found in [27, 61].

Next we demonstrate how to exploit the norm equivalence (2.22) in the construction of an optimal multilevel preconditioner. Define for any $v, w \in V_J$ the linear self-adjoint positive-definite operator $C_J : V_J \to V_J$ given by

$$(C_J^{-1}v, w)_{L_2(\Omega)} := \sum_{j=j_0}^{J} 2^{2rj} \big((P_j - P_{j-1})v, (P_j - P_{j-1})w\big)_{L_2(\Omega)}, \tag{2.23}$$

which we call a multilevel *BPX-type preconditioner*. Let $A_J : V_J \to V_J$ be the finite-dimensional operator defined by $(A_J v, w)_{L_2(\Omega)} := a(v, w)$ for all $v, w \in V_J$, the approximation of A in (2.6) with respect to V_J.

Theorem 3 *With the same prerequisites as in Theorem 2, C_J is an asymptotically optimal symmetric preconditioner for A_J, i.e., $\kappa_2(C_J^{1/2} A_J C_J^{1/2}) \sim 1$ with constants independent of J.*

Proof For the parametric domain $\hat{\Omega}$, the result was proved independently in [27, 61] and is based on the combination of (2.22) together with the well-posedness of the continuous problem (2.6). The result on the physical domain follows then together with (2.11). □

Realizations of the preconditioner defined in (2.23) based on B-splines lead to representations of the complement spaces W_j whose bases are called *wavelets*. For these, efficient implementations of optimal linear complexity involving the Fast Wavelet Transform can be derived explicitly, see Sect. 2.4.

However, since the order of the PDE operator r is positive, one can use here the argumentation from [8] which will allow to work with the same basis functions as for the spaces V_j. The first part of the argument relies on the assumption that the P_j are L_2- *orthogonal* projectors. For a clear distinction, we shall use the notation O_j for L_2-orthogonal projectors and reserve the notation P_j for the linear projectors with Properties 1. Then, the BPX-type preconditioner (2.23) (using the same symbol C_J for simplicity) reads as

$$C_J^{-1} := \sum_{j=j_0}^{J} 2^{2jr} (O_j - O_{j-1}), \tag{2.24}$$

which is by Theorem 3 a BPX-type preconditioner for the self-adjoint positive definite operator A_J. By the orthogonality of the projectors O_j, we can immediately derive from (2.24) that

$$C_J = \sum_{j=j_0}^{J} 2^{-2jr} (O_j - O_{j-1}). \tag{2.25}$$

Since $r > 0$, by rearranging the sum, the exponentially decaying scaling factors allow one to replace C_J by the spectrally equivalent operator

$$C_J = \sum_{j=j_0}^{J} 2^{-2jr} O_j \tag{2.26}$$

(for which we use the same notation C_J). Recall that two linear operators $\mathscr{A} : V_J \to V_J$ and $\mathscr{B} : V_J \to V_J$ are *spectrally equivalent* if they satisfy

$$(\mathscr{A}v, v)_{L_2(\Omega)} \sim (\mathscr{B}v, v)_{L_2(\Omega)}, \quad v \in V_J, \tag{2.27}$$

with constants independent of J. Thus, the realization of the preconditioner is reduced to a computation in terms of the bases of the spaces V_j instead of W_j. The orthogonal projector O_j can, in turn, be replaced by a simpler local operator which is spectrally equivalent to O_j, see [50] and the derivation below.

Up to this point, the introduction to multilevel preconditioners has been basis-free. We now show how this framework can be used to construct a BPX-preconditioner for the linear system (2.6). Based on the definition (2.12), we construct a sequence of spaces satisfying (2.17) such that $V_J = V_h^r$. In fact, we suppose that for each space dimension we have a sequence of p-open knot vectors $\Xi_{j_0,\ell}, \ldots, \Xi_{J,\ell}$, $\ell = 1, \ldots, n$, which provide a uniform partition of the interval $[0, 1]$ such that $\Xi_{j,\ell} \subset \Xi_{j+1,\ell}$ for $j = j_0, j_0 + 1, \ldots, J$. In particular, we assume that $\Xi_{j+1,\ell}$ is obtained from $\Xi_{j,\ell}$ by dyadic refinement, i.e., the grid spacing for $\Xi_{j,\ell}$ is proportional to 2^{-j} for each $\ell = 1, \ldots, n$. In view of the assumptions on the parametric mapping $\mathbf{F}$, we assume that $\bar{h} = 2^{-j_0}$, i.e., $\mathbf{F}$ can be represented in terms of B-splines on the coarsest level j_0. By construction, we have now achieved that

$$S_{j_0}(\hat{\Omega}) \subset S_{j_0+1}(\hat{\Omega}) \subset \ldots \subset S_J(\hat{\Omega}).$$

Setting $V_j^r := \{v \in H_0^r(\Omega) : \; v \circ \mathbf{F} \in S_j(\hat{\Omega})\}$, we arrive at a sequence of nested spaces

$$V_{j_0}^r \subset V_{j_0+1}^r \subset \ldots \subset V_J^r.$$

Setting $\mathscr{I}_j := \{1, \ldots, \dim S_j(\hat{\Omega})\}$, we denote by B_i^j, $i \in \mathscr{I}_j$, the set of L_2-normalized B-spline basis functions for the space $S_j(\hat{\Omega})$. Define now the positive definite operator $P_j : L_2(\Omega) \to V_j^r$ as

$$P_j := \sum_{i \in \mathscr{I}_j} (\,\cdot\,, B_i^j \circ \mathbf{F}^{-1})_{L_2(\Omega)} \, B_i^j \circ \mathbf{F}^{-1}. \tag{2.28}$$

Corollary 1 *For the basis $\{B_i^j \circ \mathbf{F}^{-1},\ i \in I\!I_j\}$, the operators P_j and the L_2-projectors O_j are spectrally equivalent for any j.*

Proof The assertion follows by combining (2.11), (2.14), with Remark 3.7.1 from [50], see [8] for the main ingredients. □

Finally, we obtain an explicit representation of the preconditioner C_J in terms of the mapped spline bases for V_j^r, $j = j_0, \ldots, J$,

$$C_J = \sum_{j=j_0}^{J} 2^{-2jr} \sum_{i \in \mathscr{I}_j} (\,\cdot\,, B_i^j \circ \mathbf{F}^{-1})_{L_2(\Omega)}\, B_i^j \circ \mathbf{F}^{-1} \tag{2.29}$$

(denoted again by C_J). Note that this preconditioner involves *all* B-splines from all levels j with an appropriate scaling, i.e., a properly scaled *generating system* for V_J^r.

Remark 1 The *hierarchical basis (HB) preconditioner* introduced for $n = 2$ in [71] for piecewise linear B-splines fits into this framework by choosing *Lagrangian interpolants* in place of the projectors P_j in (2.23). However, since these operators do not satisfy (P3) in Properties 1, they do not yield an asymptotically optimal preconditioner for $n \geq 2$. For $n = 3$, this preconditioner does not have an effect at all.

So far we have not explicitly addressed the dependence of the preconditioned system on p. Since all estimates in Theorem 1 which enter the proof of optimality depend on p, it is to be expected that the absolute values of the condition numbers, i.e., the values of the constants, depend on and increase with p. Indeed, in the next section, we show some numerical results which also aim at studying this dependence.

2.2.3 Realization of the BPX Preconditioner

Now we are in the position to describe the concrete implementation of the BPX preconditioner. Its main ingredient are linear *intergrid operators* which map vectors and matrices between different grids. Specifically, we need to define prolongation and restriction operators.

Since $V_j^r \subset V_{j+1}^r$, each B-spline B_i^j on level j can be represented by a linear combination of B-splines B_k^{j+1} on level $j+1$. Arranging the B-splines in the set $\{B_i^j, i \in \mathscr{I}_j\}$ into a vector $\mathbf{B}^j$ in a fixed order, this relation denoted as *refinement relation* can be written as

$$\mathbf{B}^j = \mathbf{I}_j^{j+1}\mathbf{B}^{j+1} \tag{2.30}$$

with *prolongation operator* $\mathbf{I}_j^{j+1}$ from the trial space V_j^r to the trial space V_{j+1}^r. The restriction $\mathbf{I}_{j+1}^j$ is then simply defined as the transposed operator, i.e., $\mathbf{I}_{j+1}^j = (\mathbf{I}_j^{j+1})^T$. In case of piecewise linear B-splines, this definition coincides with the well known prolongation and restriction operators from finite element textbooks obtained by interpolation, see, e.g., [7].

We will exemplify the construction in case of quadratic and cubic B-splines on the interval, see, e.g., [38], as follows. We equidistantly subdivide the interval [0, 1] into 2^j subintervals and obtain 2^j and $2^j + 1$, respectively, B-splines for $p = 2, 3$ and the corresponding quadratic and cubic spline space V_j^r which is given on this partition, respectively, see Fig. 2.1 for an illustration. Note that the two boundary functions which do not vanish at the boundary were removed in order to guarantee that $V_j^r \subset H_0^r(\Omega)$. Moreover, recall that the B-splines are L_2 normalized according to (2.13) which means that B_i^j is of the form $B_i^j(\zeta) = 2^{j/2}B(2^j\zeta - i)$ if B_i^j is an interior function, and correspondingly for the boundary functions.

In case of quadratic B-splines ($p = 2$), the restriction operator $\mathbf{I}_{j+1}^j$ reads

$$\mathbf{I}_{j+1}^j = 2^{-1/2} \begin{bmatrix} \frac{1}{2} & \frac{9}{8} & \frac{3}{8} & & & & & & & \\ & \frac{1}{4} & \frac{3}{4} & \frac{3}{4} & \frac{1}{4} & & & & & \\ & & & \frac{1}{4} & \frac{3}{4} & \frac{3}{4} & \frac{1}{4} & & & \\ & & & & & \ddots & & \ddots & & \\ & & & & & \frac{1}{4} & \frac{3}{4} & \frac{3}{4} & \frac{1}{4} & \\ & & & & & & & \frac{3}{8} & \frac{9}{8} & \frac{1}{2} \end{bmatrix} \in \mathbb{R}^{2^j \times 2^{j+1}}.$$

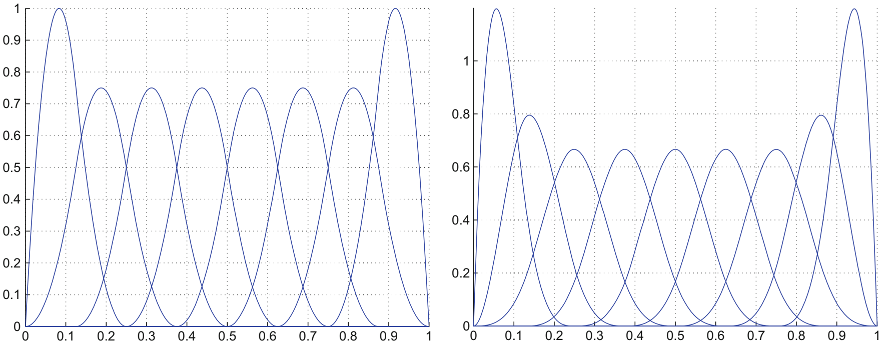


Fig. 2.1 Quadratic ($p = 2$) (left) and cubic ($p = 3$) (right) L_2-normalized B-splines (see (2.13)) on level $j = 3$ on the interval [0, 1], yielding basis functions for $V_j^r \subset H_0^r(\Omega)$

For cubic B-splines ($p = 3$), it has the form

$$\mathbf{I}_{j+1}^{j} = 2^{-1/2} \begin{bmatrix} \frac{1}{2} & \frac{9}{8} & \frac{3}{8} & & & & & & & & & \\ & \frac{1}{4} & \frac{11}{12} & \frac{2}{3} & \frac{1}{6} & & & & & & & \\ & & \frac{1}{8} & \frac{1}{2} & \frac{3}{4} & \frac{1}{2} & \frac{1}{8} & & & & & \\ & & & & \ddots & \ddots & \ddots & & & & & \\ & & & & & \frac{1}{8} & \frac{1}{2} & \frac{3}{4} & \frac{1}{2} & \frac{1}{8} & & \\ & & & & & & & \frac{1}{6} & \frac{2}{3} & \frac{11}{12} & \frac{1}{4} & \\ & & & & & & & & & \frac{3}{8} & \frac{9}{8} & \frac{1}{2} \end{bmatrix} \in \mathbb{R}^{(2^j+1)\times(2^{j+1}+1)}.$$

The normalization factor $2^{-1/2}$ stems from the L_2-normalization (2.13). The matrix entries are scaled in the usual fashion such that their rows sum to two. From these restriction operators for one dimensions, one obtains the related restriction operators on arbitrary unit cubes $[0,1]^n$ via tensor products. Finally, we set $\mathbf{I}_j^J := \mathbf{I}_{J-1}^J \mathbf{I}_{J-2}^{J-1} \cdots \mathbf{I}_j^{j+1}$ and $\mathbf{I}_J^j := \mathbf{I}_j^{j+1} \mathbf{I}_{j+1}^{j+2} \cdots \mathbf{I}_{J-1}^J$ to define prolongations and restrictions between arbitrary levels j and J.

In order to derive the explicit form of the discretized BPX-preconditioner, for given functions $u_J, v_J \in V_J$ with expansion coefficients $u_{J,k}$ and $v_{J,\ell}$, respectively, we conclude from (2.29) that

$$\begin{aligned}(C_J u_J, v_J)_{L_2(\Omega)} &= \sum_{k,\ell \in \mathscr{I}_J} u_{J,k} v_{J,\ell} (C_J(B_k^J \circ \mathbf{F}^{-1}), B_\ell^J \circ \mathbf{F}^{-1})_{L_2(\Omega)} \\ &= \sum_{k,\ell \in \mathscr{I}_J} u_{J,k} v_{J,\ell} \sum_{j=j_0}^{J} 2^{-2jr} \sum_{i \in \mathscr{I}_j} (B_k^J \circ \mathbf{F}^{-1}, B_i^j \circ \mathbf{F}^{-1})_{L_2(\Omega)} \\ &\quad \times (B_i^j \circ \mathbf{F}^{-1}, B_\ell^J \circ \mathbf{F}^{-1})_{L_2(\Omega)}.\end{aligned}$$

Next, one can introduce the mass matrix $\mathbf{M}_J = [(B_k^J \circ \mathbf{F}^{-1}, B_\ell^J \circ \mathbf{F}^{-1})_{L_2(\Omega)}]_{k,\ell}$ and obtains by the use of restrictions and prolongations

$$(C_J u_J, v_J)_{L_2(\Omega)} = \sum_{j=j_0}^{J} 2^{-2jr} \mathbf{u}_J^T \mathbf{M}_J \mathbf{I}_j^J \mathbf{I}_J^j \mathbf{M}_J \mathbf{v}_J.$$

The mass matrices which appear in this expression can be further suppressed since $\mathbf{M}_J$ is spectrally equivalent to the identity matrix. Finally, the *discretized BPX-preconditioner* to be implemented is of the simple form

$$\mathbf{C}_J = \sum_{j=j_0}^{J} 2^{-2jr} \mathbf{I}_j^J \mathbf{I}_J^j, \tag{2.31}$$

involving only restrictions and prolongations. A further simple improvement can be obtained by replacing the scaling factor 2^{-2jr} by $\operatorname{diag}(\mathbf{A}_j)^{-1}$, where $\operatorname{diag}(\mathbf{A}_j)$ denotes the diagonal matrix built from the diagonal entries of the stiffness matrix $\mathbf{A}_j$. This diagonal scaling has the same effect as the levelwise scaling by 2^{-2jr} but improves the condition numbers considerably, particularly if parametric mappings are involved. Thus, the discretized BPX-preconditioner takes on the form

$$\mathbf{C}_J = \sum_{j=j_0}^{J} \mathbf{I}_j^J \operatorname{diag}(\mathbf{A}_j)^{-1} \mathbf{I}_J^j \tag{2.32}$$

which we will use in the subsequent computations presented in Tables 2.1 and 2.2. If the condition number $\kappa(\mathbf{A}_{j_0})$ is already high in absolute numbers on the coarsest level j_0, it is worth to use its exact inverse on the coarse grid, i.e., to apply instead of (2.32) the operator

$$\mathbf{C}_J = \mathbf{I}_{j_0}^J \mathbf{A}_{j_0}^{-1} \mathbf{I}_J^{j_0} + \sum_{j=j_0+1}^{J} \mathbf{I}_j^J \operatorname{diag}(\mathbf{A}_j)^{-1} \mathbf{I}_J^j,$$

see [11, 62]. Another substantial improvement of the BPX-preconditioner can be achieved by replacing the diagonal scaling on each level by, e.g., a SSOR preconditioning as follows. We decompose the system matrix as $\mathbf{A}_j = \mathbf{L}_j + \mathbf{D}_j + \mathbf{L}_j^T$ with the diagonal matrix $\mathbf{D}_j$, the lower triangular part $\mathbf{L}_j$, and the upper triangular part $\mathbf{L}_j^T$. Then we replace the diagonal scaling on each level of the BPX-preconditioner (2.32) by the SSOR preconditioner, i.e., instead of (2.32) we apply the preconditioner

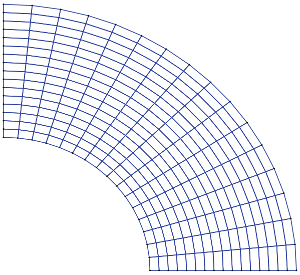
$$\mathbf{C}_J = \sum_{j=j_0}^{J} \mathbf{I}_j^J (\mathbf{D}_j + \mathbf{L}_j)^{-T} \mathbf{D}_j (\mathbf{D}_j + \mathbf{L}_j)^{-1} \mathbf{I}_J^j. \tag{2.33}$$

Table 2.1 Condition numbers of the BPX-preconditioned Laplacian on $\hat{\Omega} = (0,1)^n$ for $n = 1, 2, 3$

Level	Interval ($n=1$)				Square ($n=2$)				Cube ($n=3$)			
	$p=1$	$p=2$	$p=3$	$p=4$	$p=1$	$p=2$	$p=3$	$p=4$	$p=1$	$p=2$	$p=3$	$p=4$
3	7.43	3.81	7.03	5.93	5.93	7.31	22.8	133	3.49	39.5	356	5957
4	8.87	4.40	9.47	7.81	5.00	9.03	40.2	225	4.85	50.8	624	9478
5	10.2	4.67	11.0	9.36	5.70	9.72	51.8	293	5.75	56.6	795	11,887
6	11.3	4.87	12.1	10.7	6.27	10.1	58.7	340	6.40	59.7	895	13,185
7	12.2	5.00	12.7	11.5	6.74	10.4	63.1	371	6.91	61.3	961	13,211
8	13.0	5.10	13.0	11.9	7.14	10.5	66.0	391	7.34	62.2	990	13,234
9	13.7	5.17	13.2	12.1	7.48	10.6	68.0	403	7.70	62.6	1016	13,255
10	14.2	5.22	13.4	12.2	7.77	10.6	69.3	411	7.99	62.9	1040	–

Table 2.2 Condition numbers of the BPX-preconditioned Laplacian on the analytic arc seen on the right hand side

Level	p=1	p=2	p=3	p=4
3	5.04 (21.8)	12.4 (8.64)	31.8 (31.8)	184 (184)
4	11.1 (90.2)	16.3 (34.3)	54.7 (32.9)	291 (173)
5	25.3 (368)	19.0 (139)	70.1 (98.9)	376 (171)
6	31.9 (1492)	21.4 (560)	79.2 (401)	436 (322)
7	37.4 (6015)	23.1 (2255)	84.4 (1620)	471 (1297)
8	42.1 (241,721)	24.3 (9062)	87.3 (6506)	490 (5217)
9	45.7 (969,301)	25.2 (36,353)	89.0 (26,121)	500 (20,945)
10	48.8 (388,690)	25.9 (145,774)	90.1 (104,745)	505 (83,975)



The bracketed numbers are the related condition numbers without preconditioning

Table 2.3 Condition numbers of the BPX-preconditioned Laplacian for cubic B-splines on different geometries in case of using a BPX-SSOR preconditioning on each level

Level	Square	Analytic arc	$\mathscr{C}^0$-map of the L-shape	Singular $\mathscr{C}^1$-map of the L-shape
3	3.61	3.65	3.67	3.80
4	6.58	6.97	7.01	7.05
5	8.47	10.2	10.2	14.8
6	9.73	13.1	13.2	32.2
7	10.5	14.9	15.2	77.7
8	11.0	15.9	16.3	180
9	11.2	16.5	17.0	411
10	11.4	16.9	17.7	933

In doing so, the condition numbers can be improved impressively. In Table 2.3, we list the ℓ_2-condition numbers for the BPX-preconditioned Laplacian in case of cubic B-splines in two spatial dimensions. By comparing the numbers with those found in Tables 2.1 and 2.2 one can infer that the related condition numbers are all reduced by a factor about five. Note that the setup, storage and application of the operator defined in (2.33) is still of optimal linear complexity.

Finally, we provide numerical results in order to demonstrate the preconditioning and to specify the dependence on the spatial dimension n and the spline degree p. We consider an approximation of the homogeneous Dirichlet problem for the Poisson equation on the n-dimensional unit cube $\hat{\Omega} = (0,1)^n$ for $n = 1, 2, 3$. The mesh on level j is obtained by subdividing the cube j-times dyadically into 2^n

subcubes of mesh size $h_j = 2^{-j}$. On this subdivision, we consider the B-splines of degree $p = 1, 2, 3, 4$ as defined in Sect. 2.2.1. The ℓ_2-condition numbers of the related stiffness matrices, preconditioned by the BPX-preconditioner (2.32), are shown in Table 2.1. The condition numbers seem to be independent of the level j, but they depend on the spline degree p and the space dimension n for $n > 1$. For fourth order problems on the sphere, corresponding results for the bi-Laplacian with and without BPX preconditioning were presented in [60].

We study next the dependence of the condition numbers on the parametric mapping $\mathbf{F}$. We consider the case $n = 2$ in case of a smooth mapping (see the plot on the right hand side of Table 2.2 for an illustration of the mapping). As one can see from Table 2.2, the condition numbers are at most about a factor of five higher than the related values in Table 2.1. Nearly the same observation holds if we replace the parametric mapping by a $\mathscr{C}^0$-parametrization which maps the unit square onto an L-shaped domain, see [10].

If we consider a singular map $\mathbf{F}$, that is, a mapping that does not satisfy (2.11), the condition numbers grow considerably as expected, see [10]. But even in this case, the BPX-preconditioner with SSOR acceleration (2.33) is able to drastically reduce the condition numbers of the system matrix in all examples, see Table 2.3.

For further remarks concerning multiplicative multilevel preconditioners as the so-called multigrid methods in the context of isogeometric analysis together with references, one may consult [55].

2.3 Problem Classes

The variational problems to be investigated further will first be formulated in the following abstract form.

2.3.1 An Abstract Operator Equation

Let $\mathscr{H}$ be a Hilbert space with norm $\|\cdot\|_{\mathscr{H}}$ and let $\mathscr{H}'$ be the normed dual of $\mathscr{H}$ endowed with the norm

$$\|w\|_{\mathscr{H}'} := \sup_{v\in\mathscr{H}} \frac{\langle v, w\rangle}{\|v\|_{\mathscr{H}}} \tag{2.34}$$

where $\langle\cdot,\cdot\rangle$ denotes the dual pairing between $\mathscr{H}$ and $\mathscr{H}'$.

Given $F \in \mathscr{H}'$, we seek a solution to the operator equation

$$\mathscr{L} U = F \tag{2.35}$$

where $\mathscr{L} : \mathscr{H} \to \mathscr{H}'$ is a linear operator which is assumed to be a bounded bijection, that is,

$$\|\mathscr{L} V\|_{\mathscr{H}'} \sim \|V\|_{\mathscr{H}}, \qquad V \in \mathscr{H}. \tag{2.36}$$

We call the operator equation *well-posed* since (2.35) implies for any given data $F \in \mathscr{H}'$ the existence and uniqueness of the solution $U \in \mathscr{H}$ which depends continuously on the data.

In the following subsections, we describe some problem classes which can be placed into this framework. In particular, these examples will have the format that $\mathscr{H}$ is a product space

$$\mathscr{H} := H_{1,0} \times \cdots \times H_{M,0} \tag{2.37}$$

where each of the $H_{i,0} \subseteq H_i$ is a Hilbert space (or a closed subspace of a Hilbert space H_i determined, e.g., by homogeneous boundary conditions). The spaces H_i will be Sobolev spaces living on a domain $\Omega \subset \mathbb{R}^n$ or on (part of) its boundary. According to the definition of $\mathscr{H}$, the elements $V \in \mathscr{H}$ will consist of M components $V = (v_1, \ldots, v_M)^T$, and we define $\|V\|^2_{\mathscr{H}} := \sum_{i=1}^M \|v_i\|^2_{H_i}$. The dual space $\mathscr{H}'$ is then endowed with the norm

$$\|W\|_{\mathscr{H}'} := \sup_{V \in \mathscr{H}} \frac{\langle V, W\rangle}{\|V\|_{\mathscr{H}}} \tag{2.38}$$

where $\langle V, W\rangle := \sum_{i=1}^M \langle v_i, w_i\rangle_i$ in terms of the dual pairing $\langle \cdot, \cdot\rangle_i$ between H_i and H_i'.

We next formulate four classes which fit into this format. The first two concern are elliptic boundary value problems with included essential boundary conditions, and elliptic boundary value problems formulated as saddle point problem with boundary conditions treated by means of Lagrange Multipliers. For an introduction into elliptic boundary value problems and saddle point problems together with the functional analytic background one can, e.g., resort to [7]. Based on these formulations, we afterwards introduce certain control problems. A recurring theme in the derivation of the system of operator equation is the minimization of a quadratic functional subject to linear constraints.

2.3.2 Elliptic Boundary Value Problems

Let $\Omega \subset \mathbb{R}^n$ be a bounded domain with piecewise smooth boundary $\partial\Omega := \Gamma \cup \Gamma_N$. We consider the scalar second order boundary value problem

$$\begin{aligned} -\nabla \cdot (\mathbf{a}\nabla y) + cy &= f \quad \text{in } \Omega,\\ y &= g \quad \text{on } \Gamma,\\ (\mathbf{a}\nabla y)\cdot \mathbf{n} &= 0 \quad \text{on } \Gamma_N, \end{aligned} \tag{2.39}$$

where $\mathbf{n} = \mathbf{n}(\mathbf{x})$ is the outward normal at $\mathbf{x} \in \Gamma$, $\mathbf{a} = \mathbf{a}(\mathbf{x}) \in \mathbb{R}^{n\times n}$ is uniformly positive definite and bounded on Ω and $c \in L_\infty(\Omega)$. Moreover, f and g are some given right hand side and boundary data. With the usual definition of the bilinear form

$$a(v,w) := \int_\Omega (\mathbf{a}\nabla v \cdot \nabla w + cvw)\, d\mathbf{x}, \tag{2.40}$$

the weak formulation of (2.39) requires in the case $g \equiv 0$ to find $y \in \mathscr{H}$ where

$$\mathscr{H} := H^1_{0,\Gamma}(\Omega) := \{v \in H^1(\Omega) :\ v|_\Gamma = 0\}, \tag{2.41}$$

or

$$\mathscr{H} := \{v \in H^1(\Omega) : \int_\Omega v(\mathbf{x})\, d\mathbf{x} = 0\} \quad \text{when } \Gamma = \emptyset, \tag{2.42}$$

such that

$$a(y,v) = \langle v, f\rangle, \quad v \in \mathscr{H}. \tag{2.43}$$

The Neumann-type boundary conditions on Γ_N are implicitly satisfied in the weak formulation (2.43), therefore called *natural boundary conditions*. In contrast, the Dirichlet boundary conditions on Γ have to be posed explicitly, for this reason called *essential boundary conditions*. The easiest way to achieve this for homogeneous Dirichlet boundary conditions when $g \equiv 0$ is to include them into the solution space as above in (2.41). In the nonhomogeneous case $g \not\equiv 0$ on Γ in (2.39) and $\Gamma \neq \emptyset$, one can reduce the problem to a problem with homogeneous boundary conditions by *homogenization* as follows. Let $w \in H^1(\Omega)$ be such that $w = g$ on Γ. Then $\tilde{y} := y - w$ satisfies $a(\tilde{y}, v) = a(y,v) - a(w,v) = \langle v, f\rangle - a(w,v) =: \langle v, \tilde{f}\rangle$ for all $v \in \mathscr{H}$ defined in (2.41), and on Γ one has $\tilde{y} = g - w \equiv 0$, that is, $\tilde{y} \in \mathscr{H}$. Thus, it suffices to consider the weak form (2.43) with eventually modified right hand side. (A second possibility which allows to treat inhomogeneous boundary conditions explicitly in the context of saddle point problems will be discussed below in Sect. 2.3.3.)

The crucial property is that the bilinear form defined in (2.40) is continuous and elliptic on $\mathscr{H}$,

$$a(v, v) \sim \|v\|_{\mathscr{H}}^2 \qquad \text{for any } v \in \mathscr{H}, \tag{2.44}$$

see, e.g., [7].

By Riesz' representation theorem, the bilinear form defines a linear operator $A : \mathscr{H} \to \mathscr{H}'$ by

$$\langle w, Av \rangle := a(v, w), \qquad v, w \in \mathscr{H}, \tag{2.45}$$

which is under the above assumptions a bounded linear bijection, that is,

$$c_A \|v\|_{\mathscr{H}} \;\leq\; \|Av\|_{\mathscr{H}'} \;\leq\; C_A \|v\|_{\mathscr{H}} \qquad \text{for any } v \in \mathscr{H}. \tag{2.46}$$

Here we only consider the case where A is symmetric. With corresponding alterations, the material in the subsequent sections can also be derived for the nonsymmetric case with corresponding changes with respect to the employed algorithms.

The relation (2.46) entails that given any $f \in \mathscr{H}'$, there exists a unique $y \in \mathscr{H}$ which solves the linear system

$$Ay = f \qquad \text{in } \mathscr{H}' \tag{2.47}$$

derived from (2.43). This linear operator equation where the operator defines a bounded bijection in the sense of (2.46) is the simplest case of a well-posed variational problem (2.35). Adhering to the notation in Sect. 2.3.1, we have here $M = 1$ and $\mathscr{L} = A$.

2.3.3 Saddle Point Problems Involving Boundary Conditions

A collection of saddle point problems or, more general, multiple field formulations including first order system formulations of the elliptic boundary value problem (2.39) and the three field formulation of the Stokes problem with inhomogeneous boundary conditions have been rephrased as well-posed variational problems in the above sense in [35], see also further references cited therein.

Here a particular saddle point problem derived from (2.39) shall be considered which will be recycled later in the context of control problems. In fact, this formulation is particularly appropriate to handle essential Dirichlet boundary conditions.

Recall from, e.g., [7], that the solution $y \in \mathscr{H}$ of (2.43) is also the unique minimizer of the minimization problem

$$\inf_{v \in \mathscr{H}} \mathscr{J}(v), \qquad \mathscr{J}(v) := \frac{1}{2} a(v, v) - \langle v, f \rangle. \tag{2.48}$$

This means that y is a zero for its first order variational derivative of $\mathscr{J}$, that is, $\delta \mathscr{J}(y; v) = 0$. We denote here and in the following by $\delta^M \mathscr{J}(v; w_1, \ldots, w_M)$ the M-th variation of $\mathscr{J}$ at v in directions $w_1, \ldots, w_M$, see e.g., [72]. In particular, for $M = 1$

$$\delta \mathscr{J}(v; w) := \lim_{\varepsilon \to 0} \frac{\mathscr{J}(v + \varepsilon w) - \mathscr{J}(v)}{\varepsilon} \tag{2.49}$$

is the (Gateaux) derivative of $\mathscr{J}$ at v in direction w.

In order to generalize (2.48) to the case of nonhomogeneous Dirichlet boundary conditions g, we formulate this as minimizing J over $v \in H^1(\Omega)$ subject to constraints in form of the essential boundary conditions $v = g$ on Γ. Using techniques from nonlinear optimization theory, one can employ a *Lagrange multiplier* p to append the constraints to the optimization functional J defined in (2.48). Satisfying the constraint is guaranteed by taking the supremum over all such Lagrange multipliers before taking the infimum. Thus, minimization subject to a constraint leads to the problem of finding a *saddle point* (y, p) of the *saddle point problem*

$$\inf_{v \in H^1(\Omega)} \sup_{q \in (H^{1/2}(\Gamma))'} \mathscr{J}(v) + \langle v - g, q \rangle_\Gamma. \tag{2.50}$$

Some comments on the choice of the Lagrange multiplier space and the dual form $\langle \cdot, \cdot \rangle_\Gamma$ in (2.50) are in order. The boundary expression $v = g$ actually means taking the *trace* of $v \in H^1(\Omega)$ to $\Gamma \subseteq \partial\Omega$ which we explicitly write from now on $\gamma v := v|_\Gamma$. Classical trace theorems which may be found in [43] state that for any $v \in H^1(\Omega)$ one looses '$\frac{1}{2}$ order of smoothness' when taking traces so that one ends up with $\gamma v \in H^{1/2}(\Gamma)$. Thus, when the data g is also such that $g \in H^{1/2}(\Gamma)$, the expression in (2.50) involving the dual form $\langle \cdot, \cdot \rangle_\Gamma := \langle \cdot, \cdot \rangle_{H^{1/2}(\Gamma) \times (H^{1/2}(\Gamma))'}$ is well-defined, and so is the selection of the multiplier space $(H^{1/2}(\Gamma))'$. In case of Dirichlet boundary conditions on the whole boundary of Ω, i.e., the case $\Gamma \equiv \partial\Omega$, one can identify $(H^{1/2}(\Gamma))' = H^{-1/2}(\Gamma)$.

The above formulation (2.50) was first investigated in [2]. Another standard technique from optimization to handle minimization problems under constraints is to append the constraints to $J(v)$ by means of a *penalty parameter* ε as follows, cf. [3]. For the case of homogeneous Dirichlet boundary conditions, one could introduce the functional $J(v) + (2\varepsilon)^{-1} \|\gamma v\|^2_{H^{1/2}(\Gamma)}$. (The original formulation in [3] uses the term $\|\gamma v\|^2_{L_2(\Gamma)}$.) Although the linear system derived from this formulation is still elliptic—the bilinear form is of the type $a(v, v) + \varepsilon^{-1}(\gamma v, \gamma v)_{H^{1/2}(\Gamma)}$—the spectral condition number of the corresponding operator A_ε depends on ε. The choice of ε is typically attached to the discretization of an underlying grid with grid spacing h for Ω of the form $\varepsilon \sim h^\alpha$ when $h \to 0$ for some exponent $\alpha > 0$ chosen such that one retains the optimal approximation order of the underlying scheme. Thus, the spectral condition number of the operators in such systems depends polynomially on (at least) $h^{-\alpha}$. Consequently, iterative solution schemes such as the conjugate gradient method converge as slow as without preconditioning for A, and so far no optimal preconditioners for this situation are known.

It should also be mentioned that the way of treating essential boundary conditions by Lagrange multipliers can be extended to *fictitious domain methods* which may be used for problems with changing boundaries such as shape optimization problems [46, 49]. There one embeds the domain Ω into a larger, simple domain $\Box$, and formulates (2.50) with respect to $H^1(\Box)$ and dual form on the changing boundary Γ [52]. One should note, however, that for Γ a proper subset of $\partial\Omega$, there may occur some ambiguity in the relation between the fictitious domain formulation and the corresponding strong form (2.39).

In order to bring out the role of the trace operator, we define in addition to (2.40) a second bilinear form on $H^1(\Omega) \times (H^{1/2}(\Gamma))'$ by

$$b(v, q) := \int_\Gamma (\gamma v)(s)\, q(s)\, ds \tag{2.51}$$

so that the saddle point problem (2.50) may be rewritten as

$$\inf_{v \in H^1(\Omega)} \sup_{q \in (H^{1/2}(\Gamma))'} \mathscr{J}(v, q), \qquad \mathscr{J}(v, q) := J(v) + b(v, q) - \langle g, q\rangle_\Gamma. \tag{2.52}$$

Computing zeroes of the first order variations of $\mathscr{J}$, now with respect to both v and q, yields the system of equations that a saddle point (y, p) has to satisfy

$$\begin{aligned} a(y, v) + b(v, p) &= \langle v, f\rangle, \qquad & v &\in H^1(\Omega), \\ b(y, q) &= \langle g, q\rangle_\Gamma, \qquad & q &\in (H^{1/2}(\Gamma))'. \end{aligned} \tag{2.53}$$

Defining the linear operator $B : H^1(\Omega) \to H^{1/2}(\Gamma)$ and its adjoint $B' : (H^{1/2}(\Gamma))' \to (H^1(\Omega))'$ by $\langle Bv, q\rangle_\Gamma = \langle v, B'q\rangle_\Gamma := b(v, q)$, this can be rewritten as the linear operator equation from $\mathscr{H} := H^1(\Omega) \times (H^{1/2}(\Gamma))'$ to $\mathscr{H}'$ as follows: Given $(f, g) \in \mathscr{H}'$, find $(y, p) \in \mathscr{H}$ that solves

$$\begin{pmatrix} A & B' \\ B & 0 \end{pmatrix} \begin{pmatrix} y \\ p \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}. \tag{2.54}$$

It can be shown that the Lagrange multiplier is given by $p = -\mathbf{n} \cdot \mathbf{a}\nabla y$ and can here be interpreted as a *stress force* on the boundary [2].

Let us briefly investigate the properties of B representing the trace operator. Classical trace theorems from, e.g., [43], state that for any $f \in H^s(\Omega)$, $1/2 < s < 3/2$, one has

$$\|f|_\Gamma\|_{H^{s-1/2}(\Gamma)} \lesssim \|f\|_{H^s(\Omega)}. \tag{2.55}$$

Conversely, for every $g \in H^{s-1/2}(\Gamma)$, there exists some $f \in H^s(\Omega)$ such that $f|_\Gamma = g$ and

$$\|f\|_{H^s(\Omega)} \lesssim \|g\|_{H^{s-1/2}(\Gamma)}. \tag{2.56}$$

Note that the range of s extends accordingly if Γ is more regular. Estimate (2.55) immediately entails for $s = 1$ that $B : H^1(\Omega) \to H^{1/2}(\Gamma)$ is continuous. Moreover, the second property (2.56) means B is surjective, i.e., $rangeB = H^{1/2}(\Gamma)$ and $\ker B' = \{0\}$, which yields that the *inf–sup condition*

$$\inf_{q\in(H^{1/2}(\Gamma))'} \sup_{v\in H^1(\Omega)} \frac{\langle Bv, q\rangle_\Gamma}{\|v\|_{H^1(\Omega)}\,\|q\|_{(H^{1/2}(\Gamma))'}} \gtrsim 1 \tag{2.57}$$

is satisfied.

At this point it will be more convenient to consider (2.54) as a saddle point problem in abstract form on $\mathscr{H} = Y \times Q$. Thus, we identify $Y = H^1(\Omega)$ and $Q = (H^{1/2}(\Gamma))'$ and linear operators $A : Y \to Y'$ and $B : Y \to Q'$.

The abstract theory of saddle point problems states that existence and uniqueness of a solution pair $(y, p) \in \mathscr{H}$ holds if A and B are continuous, A is invertible on $\ker B \subseteq Y$ and the range of B is closed in Q', see, e.g., [7, 9, 42]. The properties for B and the continuity for A have been assured above. In addition, we will always deal here with operators A which are invertible on $\ker B$, which cover the standard cases of the Laplacian ($\mathbf{a} = I$ and $c \equiv 0$) and the Helmholtz operator ($\mathbf{a} = I$ and $c = 1$).

Consequently,

$$\mathscr{L} := \begin{pmatrix} A & B' \\ B & 0 \end{pmatrix} : \mathscr{H} \to \mathscr{H}' \tag{2.58}$$

is linear bijection, and one has the mapping property

$$\left\| \mathscr{L}\begin{pmatrix} v \\ q \end{pmatrix} \right\|_{\mathscr{H}'} \sim \left\| \begin{pmatrix} v \\ q \end{pmatrix} \right\|_{\mathscr{H}} \tag{2.59}$$

for any $(v, q) \in \mathscr{H}$ with constants depending on upper and lower bounds for A, B. Thus, the operator equation (2.54) is established to be a well-posed variational problem in the sense of Sect. 2.3.1: for given $(f, g) \in \mathscr{H}'$, there exists a unique solution $(y, p) \in \mathscr{H} = Y \times Q$ which continuously depends on the data.

2.3.4 Parabolic Boundary Value Problems

More recently, weak full space-time formulation for one linear parabolic equation became popular which allow us to consider time just as another space variable as follows.

Let again $\Omega \subset \mathbb{R}^n$ be a bounded Lipschitz domain with boundary $\partial\Omega$, and denote by $\Omega_T := I \times \Omega$ with time interval $I := (0, T)$ the time-space cylinder for functions $f = f(t, x)$ depending on time t and space x. The parameter $T < \infty$ will always denote a fixed final time. Let Y be a dense subspace of $H := L_2(\Omega)$ which is continuously embedded in $L_2(\Omega)$ and denote by Y' its topological dual. The associated dual form is denoted by $\langle\cdot,\cdot\rangle_{Y'\times Y}$ or, shortly $\langle\cdot,\cdot\rangle$. Later we will use $\langle\cdot,\cdot\rangle$ also for time-space duality with the precise meaning clear from the context. Norms will be indexed by the corresponding spaces. Following [59], Chapter III, p. 100, let for a.e. $t \in I$ there be bilinear forms $a(t;\cdot,\cdot) : Y \times Y \to \mathbb{R}$ so that $t \mapsto a(t;\cdot,\cdot)$ is measurable on I and that $a(t;\cdot,\cdot)$ is continuous and elliptic on Y, i.e., there exists constants $0 < \alpha_1 \leq \alpha_2 < \infty$ independent of t such that a.e. $t \in I$

$$\begin{aligned} a(t; v, w) &\leq \alpha_2 \|v\|_Y \|w\|_Y, && v, w \in Y, \\ a(t; v, v) &\geq \alpha_1 \|v\|_Y^2, && v \in Y. \end{aligned} \tag{2.60}$$

Define accordingly a linear operator $A = A(t) : Y \to Y'$ by

$$\langle A(t)v, w\rangle := a(t; v, w), \qquad v, w \in Y. \tag{2.61}$$

Denoting by $\mathscr{L}(V, W)$ the set of all bounded linear functions from V to W, we have by (2.60) $A(t) \in \mathscr{L}(Y, Y')$ for a.e. $t \in I$. Typically, $A(t)$ will be a scalar linear elliptic differential operator of order two on Ω and $Y = H_0^1(\Omega)$. We denote by $L_2(I; Z)$ the space of all functions $v = v(t, x)$ for which for a.e. $t \in I$ one has $v(t,\cdot) \in Z$. Instead of $L_2(I; Z)$, we will write this space as the tensor product of the two separable Hilbert spaces, $L_2(I) \otimes Z$, which, by Theorem 12.6.1 in [1], can be identified. This fact will be frequently employed also in the sequel.

The standard semi-weak form a linear evolution equation is the following, see e.g. [40]. Given an initial condition $y_0 \in H$ and right hand side $f \in L_2(I; Y')$, find y in some function space on Ω_T such that

$$\begin{aligned} \langle \tfrac{\partial y(t,\cdot)}{\partial t}, v\rangle + \langle A(t)\, y(t,\cdot), v\rangle &= \langle f(t,\cdot), v\rangle \text{ for all } v \in Y \text{ and a.e. } t \in (0, T), \\ \langle y(0,\cdot), v\rangle &= \langle y_0, v\rangle \qquad \text{for all } v \in H. \end{aligned} \tag{2.62}$$

For $Y = H_0^1(\Omega)$, the weak formulation of the first equation includes homogeneous Dirichlet conditions $y(t,\cdot)|_{\partial\Omega} = 0$ for a.e. $t \in I$.

The *space-time variational formulation* for (2.62) will be based on the *solution space*

$$\begin{aligned}\mathscr{Y} &:= L_2(I;Y)\cap H^1(I;Y') = (L_2(I)\otimes Y)\cap\Big(H^1(I)\otimes Y'\Big)\\ &= \{w\in L_2(I;Y) : \tfrac{\partial w(t,\cdot)}{\partial t}\in L_2(I;Y')\}\end{aligned}\tag{2.63}$$

equipped with the graph norm

$$\|w\|^2_{\mathscr{Y}} := \|w\|^2_{L_2(I;Y)} + \|\tfrac{\partial w(t,\cdot)}{\partial t}\|^2_{L_2(I;Y')}\tag{2.64}$$

and the Cartesian product *space of test functions*

$$\mathscr{V} := L_2(I;Y)\times H = (L_2(I)\otimes Y)\times H\tag{2.65}$$

equipped for $v=(v_1,v_2)\in\mathscr{V}$ with the norm

$$\|v\|^2_{\mathscr{V}} := \|v_1\|^2_{L_2(I;Y)} + \|v_2\|^2_H\tag{2.66}$$

Note that $v_1 = v_1(t,x)$ and $v_2 = v_2(x)$.

Integration of (2.62) over $t\in I$ leads to the variational problem to find for given $f\in\mathscr{V}'$ a function $y\in\mathscr{Y}$

$$b(y,v) = \langle f,v\rangle \qquad \text{for all } v=(v_1,v_2)\in\mathscr{V},\tag{2.67}$$

where the bilinear form $b(\cdot,\cdot):\mathscr{Y}\times\mathscr{V}\to\mathbb{R}$ is defined by

$$b(w,(v_1,v_2)) := \int_I\Big(\langle\tfrac{\partial w(t,\cdot)}{\partial t},v_1(t,\cdot)\rangle + \langle A(t)\,w(t,\cdot),v_1(t,\cdot)\rangle\Big)\,dt + \langle w(0,\cdot),v_2\rangle\tag{2.68}$$

and the right hand side $\langle f,\cdot\rangle:\mathscr{V}\to\mathbb{R}$ by

$$\langle f,v\rangle := \int_I\langle f(t,\cdot),v_1(t,\cdot)\rangle\,dt + \langle y_0,v_2\rangle\tag{2.69}$$

for $v=(v_1,v_2)\in\mathscr{V}$. It was proven in [37, Chapter XVIII, §3] that the operator defined by the bilinear form $b(\cdot,\cdot)$ is an isomorphism with respect to the spaces $\mathscr{Y}$ and $\mathscr{V}$. An alternative, shorter proof given in [66] is based on a characterization of bounded invertibility of linear operators between Hilbert spaces and provides detailed bounds on the norms of the operator and its inverse as follows.

Theorem 4 *The operator $B \in \mathscr{L}(\mathscr{Y}, \mathscr{V}')$ defined by $\langle Bw, v\rangle := b(w, v)$ for $w \in \mathscr{Y}$ and $v \in \mathscr{V}$ with $b(\cdot,\cdot)$ from (2.68) and spaces $\mathscr{Y}$, $\mathscr{V}$ defined in (2.63), (2.65) is boundedly invertible: There exist constants $0 < \beta_1 \leq \beta_2 < \infty$ such that*

$$\|B\|_{\mathscr{Y}\to\mathscr{V}'} \leq \beta_2 \quad \textit{and} \quad \|B^{-1}\|_{\mathscr{V}'\to\mathscr{Y}} \leq \frac{1}{\beta_1}. \tag{2.70}$$

As proved in [66], the continuity constant β_2 and the inf–sup condition constant β_1 for $b(\cdot,\cdot)$ satisfy

$$\beta_1 \geq \frac{\min(\alpha_1\alpha_2^{-2}, \alpha_1)}{\sqrt{2\max(\alpha_1^{-2}, 1) + \varrho^2}}, \qquad \beta_2 \leq \sqrt{2\max(1, \alpha_2^2) + \varrho^2}, \tag{2.71}$$

where α_1, α_2 are the constants from (2.60) bounding $A(t)$, and ϱ is defined as

$$\varrho := \sup_{0\not\equiv w\in\mathscr{Y}} \frac{\|w(0,\cdot)\|_H}{\|w\|_{\mathscr{Y}}}.$$

We like to recall from [37, 40] that $\mathscr{Y}$ is continuously embedded in $\mathscr{C}^0(I; H)$ so that the pointwise in time initial condition in (2.62) is well-defined. From this it follows that the constant ρ is bounded uniformly in the choice of $\mathscr{Y} \hookrightarrow H$.

For the sequel, it will be useful to explicitly identify the dual operator $B^* : \mathscr{V} \to \mathscr{Y}'$ of B which is defined by

$$\langle Bw, v\rangle =: \langle w, B^*v\rangle. \tag{2.72}$$

In fact, it follows from the definition of the bilinear form (2.68) on $\mathscr{Y} \times \mathscr{V}$ by integration by parts for the first term with respect to time, and using the dual $A(t)^*$ w.r.t. space that

$$\begin{aligned} b(w, (v_1, v_2)) &= \int_I \Big(\langle w(t,\cdot), \tfrac{\partial v_1(t,\cdot)}{\partial t}\rangle + \langle w(t,\cdot), A(t)^* v_1(t,\cdot)\rangle\Big)\, dt \\ &\quad + \langle w(0,\cdot), v_2\rangle + \langle w(t,\cdot), v_2\rangle|_0^T \\ &= \int_I \Big(\langle w(t,\cdot), \tfrac{\partial v_1(t,\cdot)}{\partial t}\rangle + \langle w(t,\cdot), A(t)^* v_1(t,\cdot)\rangle\Big)\, dt \\ &\quad + \langle w(T,\cdot), v_2\rangle \\ &=: \langle w, B^*v\rangle. \end{aligned} \tag{2.73}$$

Note that the first term of the right hand side defining B^* which involves $\frac{\partial}{\partial t} v_1(t,\cdot)$ is still well-defined with respect to t as an element of $\mathscr{Y}'$ on account of $w \in \mathscr{Y}$.

2.3.5 PDE-Constrained Control Problems: Distributed Control

A class of problems where the numerical solution of systems (2.47) is required repeatedly are certain control problems with PDE-constraints described next. Adhering to the notation from Sect. 2.3.2, consider as a guiding model for the subsequent discussion the objective to minimize a quadratic functional of the form

$$\mathscr{J}(y,u) = \frac{1}{2}\|y - y_*\|_{\mathscr{Z}}^2 + \frac{\omega}{2}\|u\|_{\mathscr{U}}^2, \tag{2.74}$$

subject to linear constraints

$$Ay = f + u \qquad \text{in } H' \tag{2.75}$$

where $A : H \to H'$ is defined as above in (2.61) satisfying (2.46) and $f \in H$ is given. Reserving the symbol $\mathscr{H}$ for the resulting product space in view of the notation in Sect. 2.3.1, the space H is in this subsection defined as in (2.41) or in (2.42). In order for a solution y of (2.75), the *state* of the system, to be well-defined, the problem formulation has to ensure that the unknown *control* u appearing on the right hand side is at least in H'. This can be achieved by choosing the *control space* $\mathscr{U}$ whose norm appears in (2.74) such that it is as least as smooth as H'. The second ingredient in the functional (2.74) is a data fidelity term which tries to match the system state y to some prescribed target state y_*, measured in some norm which is typically weaker than $\|\cdot\|_H$. Thus, we require that the *observation space* $\mathscr{Z}$ and the control space $\mathscr{U}$ are such that the continuous embeddings

$$\|v\|_{H'} \lesssim \|v\|_{\mathscr{U}}, \quad v \in \mathscr{U}, \qquad \|v\|_{\mathscr{Z}} \lesssim \|v\|_H, \quad v \in H, \tag{2.76}$$

hold. Mostly one has investigated the simplest cases of norms which occur for $\mathscr{U} = \mathscr{Z} = L_2(\Omega)$ and which are covered by these assumptions [59]. The parameter $\omega \geq 0$ balances the norms in (2.74).

Since the control appears in all of the right hand side of (2.75), such control problems are termed problems with *distributed* control. Although their practical value is of a rather limited nature, distributed control problems help to bring out the basic mechanisms. Note that when the observed data are *compatible* in the sense that $y_* \equiv A^{-1}f$, the control problem has the trivial solution $u \equiv 0$ which yields $\mathscr{J}(y,u) \equiv 0$.

Solution schemes for the control problem (2.74) subject to the constraints (2.75) can be based on the system of operator equations derived next by the same variational principles as employed in the previous section, using a Lagrange multiplier p to enforce the constraints. Defining the Lagrangian functional

$$\mathrm{Lagr}(y,p,u) := \mathscr{J}(y,u) + \langle p, Ay - f - u\rangle \tag{2.77}$$

on $H \times H \times H'$, the first order necessary conditions or *Karush-Kuhn-Tucker (KKT) conditions* $\delta\,\mathrm{Lagr}(x) = 0$ for $x = p, y, u$ can be derived as

$$\begin{aligned} Ay &= f + u \\ A'p &= -S(y - y_*) \\ \omega R u &= p. \end{aligned} \tag{2.78}$$

Here the linear operators S and R can be interpreted as Riesz operators defined by the inner products $(\cdot,\cdot)_{\mathscr{Z}}$ and $(\cdot,\cdot)_{\mathscr{U}}$. The system (2.78) may be written in saddle point form as

$$\mathscr{L}V := \begin{pmatrix} \mathscr{A} & \mathscr{B}' \\ \mathscr{B} & 0 \end{pmatrix} V := \begin{pmatrix} S & 0 & A' \\ 0 & \omega R & -I \\ A & -I & 0 \end{pmatrix} \begin{pmatrix} y \\ u \\ p \end{pmatrix} = \begin{pmatrix} Sy_* \\ 0 \\ f \end{pmatrix} =: F \tag{2.79}$$

on $\mathscr{H} := H \times H \times H'$.

Remark 2 We can also allow for $\mathscr{Z}$ in (2.74) to be a *trace space* on part of the boundary $\partial\Omega$ as long as the corresponding condition (2.76) is satisfied [53].

The class of control problems where the control is exerted through Neumann boundary conditions can also be written in this form since in this case the control still appears on the right hand side of a single operator equation of a form like (2.75), see [29].

Well-posedness of the system (2.79) can now be established by applying the conditions for saddle point problems stated in Sect. 2.3.3. For the control problems here and below we will, however, follow a different route which better supports efficient numerical solution schemes. The idea is as follows. While the PDE constraints (2.75) that govern the system are fixed, there is in many applications some ambiguity with respect to the choice of the spaces $\mathscr{Z}$ and $\mathscr{U}$. L_2 norms are easily realized in finite element discretizations, although in some applications like glass cooling smoother norms for the observation $\|\cdot\|_{\mathscr{Z}}$ are desirable [63]. Once $\mathscr{Z}$ and $\mathscr{U}$ are fixed, there is only a single parameter ω to balance the two norms in (2.74). *Modelling* the objective functional is therefore an issue where more flexibility may be advantageous. Specifically in a multiscale setting, one may want to weight contributions on different scales by multiple parameters.

The wavelet setting which we describe below allows for this flexibility. It is based on formulating the objective functional in terms of weighted wavelet coefficient sequences which are equivalent to $\mathscr{Z}$, $\mathscr{U}$ and which, in addition, support an efficient numerical implementation. Once wavelet discretizations are introduced, we formulate below control problems with such objective functionals.

2.3.6 PDE-Constrained Control Problems: Dirichlet Boundary Control

Even more involved as the control problems with distributed control encountered in the previous section are those problems with Dirichlet boundary control which are, however, practically much more relevant.

An illustrative guiding model for this case is the problem to minimize for some given data y_* the quadratic functional

$$\mathscr{J}(y,u) = \frac{1}{2}\|y - y_*\|_{\mathscr{Z}}^2 + \frac{\omega}{2}\|u\|_{\mathscr{U}}^2, \tag{2.80}$$

where, adhering to the notation in Sect. 2.3.2 the state y and the control u are coupled through the linear second order elliptic boundary value problem

$$\begin{aligned} -\nabla\cdot(\mathbf{a}\nabla y) + ky &= f && \text{in } \Omega,\\ y &= u && \text{on } \Gamma,\\ (\mathbf{a}\nabla y)\cdot\mathbf{n} &= 0 && \text{on } \Gamma_N. \end{aligned} \tag{2.81}$$

The appearance of the control u as a Dirichlet boundary condition in (2.81) is referred to as a *Dirichlet boundary control*. In view of the treatment of essential Dirichlet boundary conditions in the context of saddle point problems derived in Sect. 2.3.3, we write the PDE constraints (2.81) in the operator form (2.54) on $Y \times Q$ where $Y = H^1(\Omega)$ and $Q = (H^{1/2}(\Gamma))'$. The model control problem with Dirichlet boundary control then reads as follows: Minimize for given data $y_* \in \mathscr{Z}$ and $f \in Y'$ the quadratic functional

$$\mathscr{J}(y,u) = \frac{1}{2}\|y - y_*\|_{\mathscr{Z}}^2 + \frac{\omega}{2}\|u\|_{\mathscr{U}}^2 \tag{2.82}$$

subject to

$$\begin{pmatrix} A & B' \\ B & 0 \end{pmatrix}\begin{pmatrix} y \\ p \end{pmatrix} = \begin{pmatrix} f \\ u \end{pmatrix}. \tag{2.83}$$

In view of the problem formulation in Sect. 2.3.5 and the discussion of the choice of the observation space $\mathscr{Z}$ and the control space, here we require analogously that $\mathscr{Z}$ and $\mathscr{U}$ are such that the continuous embeddings

$$\|v\|_{Q'} \lesssim \|v\|_{\mathscr{U}}, \quad v \in \mathscr{U}, \qquad \|v\|_{\mathscr{Z}} \lesssim \|v\|_Y, \quad v \in Y, \tag{2.84}$$

hold. In view of Remark 2, also the case of observations on part of the boundary $\partial\Omega$ can be taken into account [54]. Part of the numerical results are for such a situation shown in Fig. 2.4.

Remark 3 It should be mentioned that the simple choice $\mathscr{U} = L_2(\Gamma)$ which is used in many applications of Dirichlet control problems is *not* covered here. There may arise the problem of well-posedness in this case which we briefly discuss. Note that the constraints (2.81) or, in weak form (2.54), guarantee a unique weak solution $y \in Y = H^1(\Omega)$ provided that the boundary term u satisfies $u \in Q' = H^{1/2}(\Gamma)$. In the framework of control problems, this smoothness of u therefore has to be required either by the choice of $\mathscr{U}$ or by the choice of $\mathscr{Z}$ (such as $\mathscr{Z} = H^1(\Omega)$) which would assure $By \in Q'$. In the latter case, we could relax condition (2.84) on $\mathscr{U}$.

In the context of flow control problems, an H^1 norm on the boundary for the control has been used in [45].

Similarly as stated at the end of Sect. 2.3.5, we can derive now by variational principles the first order necessary conditions for a coupled *system* of saddle point problems. Well-posedness of this system can then again be established by applying the conditions for saddle point problems from Sect. 2.3.3 where the inf-sup condition for the saddle point problem (2.54) yields an inf-sup condition for the exterior saddle point problem of interior saddle point problems [51]. However, also in this case, we follow the ideas mentioned at the end of Sect. 2.3.6 and pose a corresponding control problem in terms of wavelet coefficients.

2.3.7 PDE-Constrained Control Problems: Parabolic PDEs

Finally, we consider the following tracking-type control problem constrained by an evolution PDE as formulated in Sect. 2.3.4.

We wish to minimize for some given target state y_* and fixed end time $T > 0$ the quadratic functional

$$J(y,u) := \tfrac{\omega_1}{2}\|y - y_*\|^2_{L_2(I;Z)} + \tfrac{\omega_2}{2}\|y(T,\cdot) - y_*(T,\cdot)\|^2_Z + \tfrac{\omega_3}{2}\|u\|^2_{L_2(I;U)} \tag{2.85}$$

over the state $y = y(t,x)$ and the control $u = u(t,x)$ subject to

$$By = Eu + f \qquad \text{in } \mathscr{V}' \tag{2.86}$$

where B is defined by Theorem 4 and $f \in \mathscr{V}'$ is given by (2.69). The real weight parameters $\omega_1, \omega_2 \geq 0$ are such that $\omega_1 + \omega_2 > 0$ and $\omega_3 > 0$. The space Z by which the integral over Ω in the first two terms in (2.85) is indexed is to satisfy $Z \supseteq Y$ with continuous embedding. Although there is in the wavelet framework great flexibility in choosing even fractional Sobolev spaces for Z, for transparency, we pick here $Z = Y$. A more general choice only results in multiplications of vectors in wavelet coordinate with diagonal matrices of the form (2.96) below, see [29]. Moreover, we suppose that the operator E is a linear operator $E : U \to \mathscr{V}'$ extending $\int_I \langle u(t,\cdot), v_1(t,\cdot)\rangle\, dt$ trivially, that is, $E \equiv (I,0)^T$. In order to generate a well-posed problem, the space U in (2.85) must be chosen to enforce that Eu is at least in $\mathscr{V}'$. We pick here the natural case $U = Y'$ which is also the weakest possible one. More general cases for both situations which result again in multiplication with diagonal matrices for wavelet coordinate vectors are discussed in [29].

2.4 Wavelets

The numerical solution of the classes of problems introduced above hinges on the availability of appropriate wavelet bases for the function spaces under consideration which are all particular Hilbert spaces. first introduce the three basic properties that we require our wavelet bases to satisfy.

Afterwards, construction principles for wavelets based on multiresolution analysis of function spaces on bounded domains will be given.

2.4.1 Basic Properties

In view of the problem classes considered above, we need to have a wavelet basis for each occurring function space at our disposal. A *wavelet basis* for a Hilbert space H is here understood as a collection of functions

$$\Psi_H := \{\psi_{H,\lambda} : \lambda \in \mathit{I\!I}_H\} \subset H \tag{2.87}$$

which are indexed by elements λ from an infinite index set $\in \mathit{I\!I}_H$. Each of the λ comprises different information $\lambda = (j, \mathbf{k}, \mathbf{e})$ such as the *refinement scale* or *level of resolution* j and a spatial location $\mathbf{k} = \mathbf{k}(\lambda) \in \mathbb{Z}^n$. In more than one space dimensions, the basis functions are built from taking tensor products of certain univariate functions, and in this case the third index $\mathbf{e}$ contains information on the *type* of wavelet. We will frequently use the symbol $|\lambda| := j$ to have access to the resolution level j. In the univariate case on all of $\mathbb{R}$, $\psi_{H,\lambda}$ is typically generated by means of shifts and dilates of a single function ψ, i.e., $\psi_\lambda = \psi_{j,k} = 2^{j/2}\psi(2^j \cdot -k)$, $j, k \in \mathbb{Z}$, normalized with respect to $\|\cdot\|_{L_2}$. On bounded domains, the structure of the functions is essentially the same up to modifications near the boundary.

The three crucial properties that we will assume the wavelet basis to have for the sequel are the following.

Riesz Basis Property (R) Every $v \in H$ has a unique expansion in terms of Ψ_H,

$$v = \sum_{\lambda \in \mathit{I\!I}_H} v_\lambda \, \psi_{H,\lambda} =: \mathbf{v}^T \, \Psi_H, \quad \mathbf{v} := (v_\lambda)_{\lambda \in \mathit{I\!I}_H}, \tag{2.88}$$

and its expansion coefficients satisfy a *norm equivalence*, that is, for any $\mathbf{v} = \{v_\lambda : \lambda \in I\!I_H\}$ one has

$$c_H \, \|\mathbf{v}\|_{\ell_2(I\!I_H)} \;\leq\; \|\mathbf{v}^T \Psi_H\|_H \;\leq\; C_H \, \|\mathbf{v}\|_{\ell_2(I\!I_H)}, \quad \mathbf{v} \in \ell_2(I\!I_H), \tag{2.89}$$

where $0 < c_H \leq C_H < \infty$. This means that wavelet expansions induce *isomorphisms* between certain function spaces and sequence spaces. It will be convenient in the following to abbreviate ℓ_2 norms without subscripts as $\|\cdot\| := \|\cdot\|_{\ell_2(I\!I_H)}$ when the index set is clear from the context. If the precise format of the constants does not matter, we write the norm equivalence (2.89) shortly as

$$\|\mathbf{v}\| \;\sim\; \|\mathbf{v}^T \Psi_H\|_H, \quad \mathbf{v} \in \ell_2(I\!I_H). \tag{2.90}$$

Locality (L) The functions $\psi_{H,\lambda}$ are have compact support which decreases with increasing level $j = |\lambda|$, i.e.,

$$diam\,(supp\psi_{H,\lambda}) \;\sim\; 2^{-|\lambda|}. \tag{2.91}$$

Cancellation Property (CP) There exists an integer $\tilde{d} = \tilde{d}_H$ such that

$$\langle v, \psi_{H,\lambda}\rangle \;\lesssim\; 2^{-|\lambda|(n/2-n/p+\tilde{d})} |v|_{W_p^{\tilde{d}}(supp\,\psi_{H,\lambda})}. \tag{2.92}$$

Thus, integrating against a wavelet has the effect of taking an $\tilde{d}$th order difference which annihilates the smooth part of v. This property is for wavelets defined on Euclidean domains typically realized by constructing Ψ_H in such a way that it possesses a *dual* or *biorthogonal* basis $\tilde{\Psi}_H \subset H'$ such that the multiresolution spaces $\tilde{S}_j := span\{\tilde{\psi}_{H,\lambda} : |\lambda| < j\}$ contain all polynomials of order $\tilde{d}$. Here *dual basis* means that $\langle \psi_{H,\lambda}, \tilde{\psi}_{H,\nu}\rangle = \delta_{\lambda,\nu}$, $\lambda, \nu \in I\!I_H$.

A few remarks on these properties are in order. In (R), the norm equivalence (2.90) is crucial since it means complete control over a function measured in $\|\cdot\|_H$ from above and below by its expansion coefficients: small changes in the coefficients only causes small changes in the function which, together with the locality (L), also means that local changes stay local. This stability is an important feature which is used for deriving optimal preconditioners and driving adaptive approximations where, again, the locality is crucial. Finally, the cancellation property (CP) entails that smooth functions have small wavelet coefficients which, on account of (2.89) may be neglected in a controllable way. Moreover, (CP) can be used to derive quasi-sparse representations of a wide class of operators.

By duality arguments one can show that (2.89) is equivalent to the existence of a biorthogonal collection which is *dual* or *biorthogonal* to Ψ_H,

$$\tilde{\Psi}_H := \{\tilde{\psi}_{H,\lambda} : \lambda \in I\!I_H\} \subset H', \quad \langle \psi_{H,\lambda}, \tilde{\psi}_{H,\mu}\rangle = \delta_{\lambda,\mu}, \qquad \lambda, \mu \in I\!I_H, \tag{2.93}$$

which is a Riesz basis for H', that is, for any $\tilde{v} = \tilde{\mathbf{v}}^T\, \tilde{\Psi}_H \in H'$ one has

$$C_H^{-1}\, \|\tilde{\mathbf{v}}\| \;\le\; \|\tilde{\mathbf{v}}^T \tilde{\Psi}_H\|_{H'} \;\le\; c_H^{-1}\, \|\tilde{\mathbf{v}}\|, \tag{2.94}$$

see [23, 25, 51]. Here and in the sequel the tilde expresses that the collection $\tilde{\Psi}_H$ is a dual basis to a primal one for the space identified by the subscript, so that $\tilde{\Psi}_H = \Psi_{H'}$.

Above in (2.89), we have already introduced the following shorthand notation which simplifies the presentation of many terms. We will view Ψ_H both as in (2.87) as a *collection* of functions as well as a (possibly infinite) column *vector* containing all functions always assembled in some fixed unspecified order. For a countable collection of functions Θ and some single function σ, the term $\langle\Theta, \sigma\rangle$ is to be understood as the column vector with entries $\langle\theta, \sigma\rangle$, $\theta \in \Theta$, and correspondingly $\langle\sigma, \Theta\rangle$ the row vector. For two collections Θ, Σ, the quantity $\langle\Theta, \Sigma\rangle$ is then a (possibly infinite) matrix with entries $(\langle\theta, \sigma\rangle)_{\theta\in\Theta,\, \sigma\in\Sigma}$ for which $\langle\Theta, \Sigma\rangle = \langle\Sigma, \Theta\rangle^T$. This also implies for a (possibly infinite) matrix $\mathbf{C}$ that $\langle\mathbf{C}\Theta, \Sigma\rangle = \mathbf{C}\langle\Theta, \Sigma\rangle$ and $\langle\Theta, \mathbf{C}\Sigma\rangle = \langle\Theta, \Sigma\rangle\mathbf{C}^T$.

In this notation, the *biorthogonality* or *duality conditions* (2.93) can be reexpressed as

$$\langle\Psi, \tilde{\Psi}\rangle = \mathbf{I} \tag{2.95}$$

with the infinite identity matrix $\mathbf{I}$.

Wavelets with the above properties can actually obtained in the following way. This concerns, in particular, a scaling depending on the regularity of the space under consideration. In our case, H will always be a Sobolev space $H^s = H^s(\Omega)$ or a closed subspace of $H^s(\Omega)$ determined by homogeneous boundary conditions, or its dual. For $s < 0$, H^s is interpreted as above as the dual of H^{-s}. One typically obtains the wavelet basis Ψ_H for H from an *anchor basis* $\Psi = \{\psi_\lambda : \lambda \in I\!I = I\!I_H\}$ which is a Riesz basis for $L_2(\Omega)$, meaning that Ψ is scaled such that $\|\psi_\lambda\|_{L_2(\Omega)} \sim 1$. Moreover, its dual basis $\tilde{\Psi}$ is also a Riesz basis for $L_2(\Omega)$. Ψ and $\tilde{\Psi}$ are constructed in such a way that rescaled versions of *both bases* Ψ, $\tilde{\Psi}$ form Riesz bases for a whole range of (closed subspaces of) Sobolev spaces H^s, for $0 < s < \gamma, \tilde{\gamma}$, respectively. Consequently, one can derive that for each $s \in (-\tilde{\gamma}, \gamma)$ the collection

$$\Psi_s := \{2^{-s|\lambda|}\psi_\lambda : \lambda \in I\!I\} =: \mathbf{D}^{-s}\Psi \tag{2.96}$$

is a Riesz basis for H^s [23]. This means that there exist positive finite constants c_s, C_s such that

$$c_s\, \|\mathbf{v}\| \;\le\; \|\mathbf{v}^T \Psi_s\|_{H^s} \;\le\; C_s\, \|\mathbf{v}\| \quad \mathbf{v} \in \ell_2(I\!I), \tag{2.97}$$

holds for each $s \in (-\tilde{\gamma}, \gamma)$. Such a scaling represented by a diagonal matrix $\mathbf{D}^s$ introduced in (2.96) will play an important role later on. The analogous expression in terms of the dual basis reads

$$\tilde{\Psi}_s := \{2^{s|\lambda|}\,\tilde{\psi}_\lambda : \lambda \in I\!I\} = \mathbf{D}^s\,\tilde{\Psi}, \tag{2.98}$$

where $\tilde{\Psi}_s$ forms a Riesz basis of H^s for $s \in (-\gamma, \tilde{\gamma})$. This entails the following fact. For $\tau \in (-\tilde{\gamma}, \gamma)$ the mapping

$$D^\tau : v = \mathbf{v}^T\Psi \mapsto (\mathbf{D}^\tau\mathbf{v})^T\Psi = \mathbf{v}^T\mathbf{D}^\tau\Psi = \sum_{\lambda\in I\!I} v_\lambda\, 2^{\tau|\lambda|}\psi_\lambda \tag{2.99}$$

acts as a shift operator between Sobolev scales which means that

$$\|D^\tau v\|_{H^s} \sim \|v\|_{H^{s+\tau}} \sim \|\mathbf{D}^{s+\tau}\mathbf{v}\|, \text{ if } s,\ s+\tau \in (-\tilde{\gamma}, \gamma). \tag{2.100}$$

Concrete constructions of wavelet bases with the above properties for parameters $\gamma, \tilde{\gamma} \le 3/2$ on a bounded Lipschitz domain Ω can be found in [33, 34]. This suffices for the above mentioned examples where the relevant Sobolev regularity indices range between -1 and 1.

2.4.2 Norm Equivalences and Riesz Maps

As we have seen, the scaling provided by $\mathbf{D}^{-s}$ is an important feature to establish norm equivalences (2.97) for the range $s \in (-\tilde{\gamma}, \gamma)$ of Sobolev spaces H^s. However, there are several other norms which are *equivalent* to $\|\cdot\|_{H^s}$ which may later be used in the objective functional (2.74) in the context of control problems. This issue addresses the *mathematical model* which we briefly discuss now.

We first consider norm equivalences for the L_2 norm. Let as before Ψ be the anchor wavelet basis for L_2 for which the *Riesz operator* $\mathbf{R} = \mathbf{R}_{L_2}$ is the (infinite) Gramian matrix with respect to the inner product $(\cdot,\cdot)_{L_2}$ defined as

$$\mathbf{R} := (\Psi, \Psi)_{L_2} = \langle\Psi, \Psi\rangle. \tag{2.101}$$

Expanding Ψ in terms of $\tilde{\Psi}$ and recalling the duality (2.95), this entails

$$\mathbf{I} = \langle\Psi, \tilde{\Psi}\rangle = \left\langle \langle\Psi, \Psi\rangle\tilde{\Psi}, \tilde{\Psi}\right\rangle = \mathbf{R}\langle\tilde{\Psi}, \tilde{\Psi}\rangle \quad \text{ or } \quad \mathbf{R}^{-1} = \langle\tilde{\Psi}, \tilde{\Psi}\rangle. \tag{2.102}$$

$\mathbf{R}$ may be interpreted as the transformation matrix for the change of basis from $\tilde{\Psi}$ to Ψ, that is, $\Psi = \mathbf{R}\tilde{\Psi}$.

For any $w = \mathbf{w}^T\Psi \in L_2$, we now obtain the identities

$$\|w\|_{L_2}^2 = (\mathbf{w}^T\Psi, \mathbf{w}^T\Psi)_{L_2} = \mathbf{w}^T\langle\Psi, \Psi\rangle\,\mathbf{w} = \mathbf{w}^T\mathbf{R}\mathbf{w} = \|\mathbf{R}^{1/2}\mathbf{w}\|^2 =: \|\hat{\mathbf{w}}\|^2. \tag{2.103}$$

Expanding w with respect to the basis $\hat{\Psi} := \mathbf{R}^{-1/2}\Psi = \mathbf{R}^{1/2}\tilde{\Psi}$, that is, $w = \hat{\mathbf{w}}^T\hat{\Psi}$, yields $\|w\|_{L_2} = \|\hat{\mathbf{w}}\|$. On the other hand, we get from (2.97) with $s = 0$

$$c_0^2\,\|\mathbf{w}\|^2 \;\le\; \|w\|_{L_2}^2 \;\le\; C_0^2\,\|\mathbf{w}\|^2. \tag{2.104}$$

From this we can derive the *condition number* $\kappa(\Psi)$ of the wavelet basis in terms of the extreme eigenvalues of $\mathbf{R}$ by defining

$$\kappa(\Psi) := \left(\frac{C_0}{c_0}\right)^2 = \frac{\lambda_{\max}(\mathbf{R})}{\lambda_{\min}(\mathbf{R})} = \kappa(\mathbf{R}) \sim 1, \tag{2.105}$$

where $\kappa(\mathbf{R})$ also denotes the spectral condition number of $\mathbf{R}$ and where the last relation is assured by the asymptotic estimate (2.104). However, the absolute constants will have an impact on numerical results in specific cases.

For a Hilbert space H denote by Ψ_H a wavelet basis for H satisfying (R), (L), (CP) with a corresponding dual basis $\tilde{\Psi}_H$. The (infinite) Gramian matrix with respect to the inner product $(\cdot,\cdot)_H$ inducing $\|\cdot\|_H$ which is defined by

$$\mathbf{R}_H := (\Psi_H, \Psi_H)_H \tag{2.106}$$

will be also called *Riesz operator*. The space L_2 is covered trivially by $\mathbf{R}_0 = \mathbf{R}$. For any function $v := \mathbf{v}^T\Psi_H \in H$ we have then the identity

$$\begin{aligned}\|v\|_H^2 = (v,v)_H = (\mathbf{v}^T\Psi_H, \mathbf{v}^T\Psi_H)_H = \mathbf{v}^T(\Psi_H,\Psi_H)_H\,\mathbf{v}\\ = \mathbf{v}^T\mathbf{R}_H\mathbf{v} = \|\mathbf{R}_H^{1/2}\mathbf{v}\|^2.\end{aligned} \tag{2.107}$$

Note that in general $\mathbf{R}_H$ may not be explicitly computable, in particular, when H is a fractional Sobolev space.

Again referring to (2.97), we obtain as in (2.105) for the more general case

$$\kappa(\Psi_s) := \left(\frac{C_s}{c_s}\right)^2 = \frac{\lambda_{\max}(\mathbf{R}_{H^s})}{\lambda_{\min}(\mathbf{R}_{H^s})} = \kappa(\mathbf{R}_{H^s}) \sim 1 \quad \text{for each } s \in (-\tilde{\gamma}, \gamma). \tag{2.108}$$

Thus, all Riesz operators on the applicable scale of Sobolev spaces are spectrally equivalent. Moreover, comparing (2.108) with (2.105), we get

$$\frac{c_s}{C_0}\,\|\mathbf{R}^{1/2}\mathbf{v}\| \;\le\; \|\mathbf{R}_{H^s}^{1/2}\mathbf{v}\| \;\le\; \frac{C_s}{c_0}\,\|\mathbf{R}^{1/2}\mathbf{v}\|. \tag{2.109}$$

Of course, in practice, the constants appearing in this equation may be much sharper, as the bases for Sobolev spaces with different exponents are only obtained by a diagonal scaling which preserves much of the structure of the original basis for L_2.

We summarize these results for further reference.

Proposition 1 *In the above notation, we have for any $v = \mathbf{v}^T \Psi_s \in H^s$ the norm equivalences*

$$\|v\|_{H^s} = \|\mathbf{R}_{H^s}^{1/2}\mathbf{v}\| \sim \|\mathbf{R}^{1/2}\mathbf{v}\| \sim \|\mathbf{v}\| \qquad \textit{for each } s \in (-\tilde{\gamma}, \gamma). \tag{2.110}$$

2.4.3 Representation of Operators

A final ingredient concerns the *wavelet representation* of linear operators in terms of wavelets. Let H, V be Hilbert spaces with wavelet bases Ψ_H, Ψ_V and corresponding duals $\tilde{\Psi}_H$, $\tilde{\Psi}_V$, and suppose that $\mathscr{L} : H \to V$ is a linear operator with dual $\mathscr{L}' : V' \to H'$ defined by $\langle v, \mathscr{L}'w\rangle := \langle \mathscr{L}v, w\rangle$ for all $v \in H, w \in V$.

We shall make frequent use of this representation and its properties.

Remark 4 The wavelet representation of $\mathscr{L} : H \to V$ with respect to the bases $\Psi_H, \tilde{\Psi}_V$ of H, V', respectively, is given by

$$\mathbf{L} := \langle \tilde{\Psi}_V, \mathscr{L}\Psi_H\rangle, \quad \mathscr{L}v = (\mathbf{L}\mathbf{v})^T \Psi_V. \tag{2.111}$$

Thus, the expansion coefficients of $\mathscr{L}v$ in the basis that spans the range space of $\mathscr{L}$ are obtained by applying the *infinite* matrix $\mathbf{L} = \langle \tilde{\Psi}_V, \mathscr{L}\Psi_H\rangle$ to the coefficient vector of v. Moreover, boundedness of $\mathscr{L}$ implies boundedness of $\mathbf{L}$ in ℓ_2, i.e.,

$$\|\mathscr{L}v\|_V \lesssim \|v\|_H, \quad v \in H, \quad \text{implies} \quad \|\mathbf{L}\| := \sup_{\|\mathbf{v}\|_{\ell_2(\mathit{I\!I}_H)} \le 1} \|\mathbf{L}\mathbf{v}\|_{\ell_2(\mathit{I\!I}_V)} \lesssim 1. \tag{2.112}$$

Proof Any image $\mathscr{L}v \in V$ can naturally be expanded with respect to Ψ_V as $\mathscr{L}v = \langle \mathscr{L}v, \tilde{\Psi}_V\rangle\Psi_V$. Expanding in addition v in the basis Ψ_H, $v = \mathbf{v}^T\Psi_H$ yields

$$\mathscr{L}v = \mathbf{v}^T\langle \mathscr{L}\Psi_H, \tilde{\Psi}_V\rangle\Psi_V = (\langle \mathscr{L}\Psi_H, \tilde{\Psi}_V\rangle^T\mathbf{v})^T\Psi_V = (\langle \tilde{\Psi}_V, \mathscr{L}\Psi_H\rangle\mathbf{v})^T\Psi_V. \tag{2.113}$$

As for (2.112), we can infer from (2.89) and (2.111) that

$$\|\mathbf{L}\mathbf{v}\|_{\ell_2(\mathit{I\!I}_V)} \sim \|(\mathbf{L}\mathbf{v})^T\Psi_V\|_V = \|Lv\|_V \lesssim \|v\|_H \sim \|\mathbf{v}\|_{\ell_2(\mathit{I\!I}_H)},$$

which confirms the claim. □

2.4.4 Multiscale Decomposition of Function Spaces

In this section, the basic construction principles of the biorthogonal wavelets with properties (R), (L) and (CP) are summarized, see, e.g., [24]. Their cornerstones are *multiresolution analyses* of the function spaces under consideration and the concept of *stable completions*. These concepts are free of Fourier techniques and can therefore be applied to derive constructions of wavelets on domains or manifolds which are subsets of $\mathbb{R}^n$.

Multiresolution of L_2 Practical constructions of wavelets typically start out with multiresolution analyses of function spaces. Consider a *multiresolution* $\mathscr{S}$ of L_2 which consists of closed subspaces S_j of L_2, called *trial spaces*, such that they are nested and their union is dense in L_2,

$$S_{j_0} \subset S_{j_0+1} \subset \ldots \subset S_j \subset S_{j+1} \subset \ldots L_2, \qquad \mathrm{clos}_{L_2}\Big(\bigcup_{j=j_0}^{\infty} S_j\Big) = L_2. \tag{2.114}$$

The index j is the refinement level which appeared already in the elements of the index set $I\!I$ in (2.87), starting with some coarsest level $j_0 \in \mathbb{N}_0$. We abbreviate for a finite subset $\Theta \subset L_2$ the linear span of Θ as

$$S(\Theta) = span\{\Theta\}.$$

Typically the multiresolution spaces S_j have the form

$$S_j = S(\Phi_j), \qquad \Phi_j = \{\phi_{j,k} : k \in \Delta_j\}, \tag{2.115}$$

for some finite index set Δ_j, where the set $\{\Phi_j\}_{j=j_0}^{\infty}$ is *uniformly stable* in the sense that

$$\|\mathbf{c}\|_{\ell_2(\Delta_j)} \sim \|\mathbf{c}^T \Phi_j\|_{L_2}, \qquad \mathbf{c} = \{c_k\}_{k\in\Delta_j} \in \ell_2(\Delta_j), \tag{2.116}$$

holds uniformly in j. Here we have used again the shorthand notation

$$\mathbf{c}^T \Phi_j = \sum_{k\in\Delta_j} c_k \phi_{j,k}$$

and Φ_j denotes both the (column) vector containing the functions $\phi_{j,k}$ as well as the set of functions (2.115).

The collection Φ_j is called *single scale basis* since all its elements live only on one scale j. In the present context of multiresolution analysis, Φ_j is also called *generator basis* or shortly *generators* of the multiresolution. We assume that the $\phi_{j,k}$ are compactly supported with

$$diam(supp\phi_{j,k}) \sim 2^{-j}. \tag{2.117}$$

It follows from (2.116) that they are scaled such that

$$\|\phi_{j,k}\|_{L_2} \sim 1 \tag{2.118}$$

holds. It is known that nestedness (2.114) together with stability (2.116) implies the existence of matrices $\mathbf{M}_{j,0} = (m^j_{r,k})_{r\in\Delta_{j+1},k\in\Delta_j}$ such that the two-scale relation

$$\phi_{j,k} = \sum_{r\in\Delta_{j+1}} m^j_{r,k}\phi_{j+1,r}, \quad k \in \Delta_j, \tag{2.119}$$

is satisfied. We can essentially simplify the subsequent presentation of the material by viewing (2.119) as a matrix-vector equation which then attains the compact form

$$\Phi_j = \mathbf{M}^T_{j,0}\Phi_{j+1}. \tag{2.120}$$

Any set of functions satisfying an equation of this form, the *refinement* or *two-scale relation*, will be called *refinable*.

Denoting by $[X, Y]$ the space of bounded linear operators from a normed linear space X into the normed linear space Y, one has that

$$\mathbf{M}_{j,0} \in [\ell_2(\Delta_j), \ell_2(\Delta_{j+1})]$$

is *uniformly sparse* which means that the number of entries in each row or column is uniformly bounded. Furthermore, one infers from (2.116) that

$$\|\mathbf{M}_{j,0}\| = \mathcal{O}(1), \quad j \geq j_0, \tag{2.121}$$

where the corresponding operator norm is defined as

$$\|\mathbf{M}_{j,0}\| := \sup_{\mathbf{c}\in\ell_2(\Delta_j),\ \|\mathbf{c}\|_{\ell_2(\Delta_j)}=1} \|\mathbf{M}_{j,0}\mathbf{c}\|_{\ell_2(\Delta_{j+1})}.$$

Since the union of $\mathscr{S}$ is dense in L_2, a basis for L_2 can be assembled from functions which span any complement between two successive spaces S_j and S_{j+1}, i.e.,

$$S(\Phi_{j+1}) = S(\Phi_j) \oplus S(\Psi_j) \tag{2.122}$$

where

$$\Psi_j = \{\psi_{j,k} : k \in \nabla_j\}, \qquad \nabla_j := \Delta_{j+1} \setminus \Delta_j. \tag{2.123}$$

The functions Ψ_j are called *wavelet functions* or shortly *wavelets* if, among other conditions detailed below, the union $\{\Phi_j \cup \Psi_j\}$ is still uniformly stable in the sense of (2.116). Since (2.122) implies $S(\Psi_j) \subset S(\Phi_{j+1})$, the functions in Ψ_j must also satisfy a matrix-vector relation of the form

$$\Psi_j = \mathbf{M}_{j,1}^T \Phi_{j+1} \tag{2.124}$$

with a matrix $\mathbf{M}_{j,1}$ of size $(\#\Delta_{j+1}) \times (\#\nabla_j)$. Furthermore, (2.122) is equivalent to the fact that the linear operator composed of $\mathbf{M}_{j,0}$ and $\mathbf{M}_{j,1}$,

$$\mathbf{M}_j = (\mathbf{M}_{j,0}, \mathbf{M}_{j,1}), \tag{2.125}$$

is *invertible* as a mapping from $\ell_2(\Delta_j \cup \nabla_j)$ onto $\ell_2(\Delta_{j+1})$. One can also show that the set $\{\Phi_j \cup \Psi_j\}$ is uniformly stable if and only if

$$\|\mathbf{M}_j\|, \|\mathbf{M}_j^{-1}\| = \mathscr{O}(1), \quad j \to \infty. \tag{2.126}$$

The particular cases that will be important for practical purposes are when not only $\mathbf{M}_{j,0}$ and $\mathbf{M}_{j,1}$ are uniformly sparse but also the inverse of $\mathbf{M}_j$. We denote this inverse by $\mathbf{G}_j$ and assume that it is split into

$$\mathbf{G}_j = \mathbf{M}_j^{-1} = \begin{pmatrix} \mathbf{G}_{j,0} \\ \mathbf{G}_{j,1} \end{pmatrix}. \tag{2.127}$$

A special situation occurs when

$$\mathbf{G}_j = \mathbf{M}_j^{-1} = \mathbf{M}_j^T$$

which corresponds to the case of L_2 *orthogonal wavelets* [36]. A systematic construction of more general $\mathbf{M}_j$, $\mathbf{G}_j$ for spline-wavelets can be found in [34], see also [24] for more examples, including the hierarchical basis.

Thus, the identification of the functions Ψ_j which span the complement of $S(\Phi_j)$ in $S(\Phi_{j+1})$ is equivalent to completing a given refinement matrix $\mathbf{M}_{j,0}$ to an invertible matrix $\mathbf{M}_j$ in such a way that (2.126) is satisfied. Any such completion $\mathbf{M}_{j,1}$ is called *stable completion* of $\mathbf{M}_{j,0}$. In other words, the problem of the construction of compactly supported wavelets can equivalently be formulated as an algebraic problem of finding the (uniformly) sparse completion of a (uniformly) sparse matrix $\mathbf{M}_{j,0}$ in such a way that its inverse is also (uniformly) sparse. The fact that inverses of sparse matrices are usually dense elucidates the difficulties in the constructions.

The concept of stable completions has been introduced in [14] for which a special case is known as the *lifting scheme* [69]. Of course, constructions that yield compactly supported wavelets are particularly suited for computations in numerical analysis.

Combining the two-scale relations (2.120) and (2.124), one can see that $\mathbf{M}_j$ performs a change of bases in the space S_{j+1},

$$\begin{pmatrix} \Phi_j \\ \Psi_j \end{pmatrix} = \begin{pmatrix} \mathbf{M}_{j,0}^T \\ \mathbf{M}_{j,1}^T \end{pmatrix} \Phi_{j+1} = \mathbf{M}_j^T \Phi_{j+1}. \tag{2.128}$$

Conversely, applying the inverse of $\mathbf{M}_j$ to both sides of (2.128) results in the *reconstruction identity*

$$\Phi_{j+1} = \mathbf{G}_j^T \begin{pmatrix} \Phi_j \\ \Psi_j \end{pmatrix} = \mathbf{G}_{j,0}^T \Phi_j + \mathbf{G}_{j,1}^T \Psi_j. \tag{2.129}$$

Fixing a *finest resolution level* J, one can repeat the decomposition (2.122) so that $S_J = S(\Phi_J)$ can be written in terms of the functions from the coarsest space supplied with the complement functions from all intermediate levels,

$$S(\Phi_J) = S(\Phi_{j_0}) \oplus \bigoplus_{j=j_0}^{J-1} S(\Psi_j). \tag{2.130}$$

Thus, every function $v \in S(\Phi_J)$ can be written in its *single-scale representation*

$$v = (\mathbf{c}_J)^T \Phi_J = \sum_{k \in \Delta_J} c_{J,k} \phi_{J,k} \tag{2.131}$$

as well as in its *multi-scale form*

$$v = (\mathbf{c}_{j_0})^T \Phi_{j_0} + (\mathbf{d}_{j_0})^T \Psi_{j_0} + \cdots + (\mathbf{d}_{J-1})^T \Psi_{J-1} \tag{2.132}$$

with respect to the *multiscale* or *wavelet basis*

$$\Psi^J := \Phi_{j_0} \cup \bigcup_{j=j_0}^{J-1} \Psi_j =: \bigcup_{j=j_0-1}^{J-1} \Psi_j \tag{2.133}$$

Often the single-scale representation of a function may be easier to compute and evaluate while the multi-scale representation allows one to separate features of the underlying function characterized by different length scales. Since therefore both representations are advantageous, it is useful to determine the transformation between the two representations, commonly referred to as the *Wavelet Transform*,

$$\mathbf{T}_J : \ell_2(\Delta_J) \to \ell_2(\Delta_J), \qquad \mathbf{d}^J \mapsto \mathbf{c}_J, \tag{2.134}$$

where

$$\mathbf{d}^J := (\mathbf{c}_{j_0}, \mathbf{d}_{j_0}, \ldots, \mathbf{d}_{J-1})^T.$$

The previous relations (2.128) and (2.129) indicate that this will involve the matrices $\mathbf{M}_j$ and $\mathbf{G}_j$. In fact, $\mathbf{T}_J$ has the representation

$$\mathbf{T}_J = \mathbf{T}_{J,J-1} \cdots \mathbf{T}_{J,j_0}, \tag{2.135}$$

where each factor has the form

$$\mathbf{T}_{J,j} := \begin{pmatrix} \mathbf{M}_j & \mathbf{0} \\ \mathbf{0} & \mathbf{I}^{(\#\Delta_J - \#\Delta_{j+1})} \end{pmatrix} \in \mathbb{R}^{(\#\Delta_J)\times(\#\Delta_J)}. \tag{2.136}$$

Schematically $\mathbf{T}_J$ can be visualized as a pyramid scheme

$$\begin{tikzcd}
\mathbf{c}_{j_0} \arrow[r, "\mathbf{M}_{j_0,0}"] & \mathbf{c}_{j_0+1} \arrow[r, "\mathbf{M}_{j_0+1,0}"] & \mathbf{c}_{j_0+2} \arrow[r] & \cdots\ \mathbf{c}_{J-1} \arrow[r, "\mathbf{M}_{J-1,0}"] & \mathbf{c}_J \\
\mathbf{d}_{j_0} \arrow[ur, "\mathbf{M}_{j_0,1}"] & \mathbf{d}_{j_0+1} \arrow[ur, "\mathbf{M}_{j_0+1,1}"] & \mathbf{d}_{j_0+2} \arrow[ur] \cdots & \mathbf{d}_{J-1} \arrow[ur, "\mathbf{M}_{J-1,1}"] &
\end{tikzcd} \tag{2.137}$$

Accordingly, the inverse transform $\mathbf{T}_J^{-1}$ can be written also in product structure (2.135) in reverse order involving the matrices $\mathbf{G}_j$ as follows:

$$\mathbf{T}_J^{-1} = \mathbf{T}_{J,j_0}^{-1} \cdots \mathbf{T}_{J,J-1}^{-1}, \tag{2.138}$$

where each factor has the form

$$\mathbf{T}_{J,j}^{-1} := \begin{pmatrix} \mathbf{G}_j & \mathbf{0} \\ \mathbf{0} & \mathbf{I}^{(\#\Delta_J - \#\Delta_{j+1})} \end{pmatrix} \in \mathbb{R}^{(\#\Delta_J)\times(\#\Delta_J)}. \tag{2.139}$$

The corresponding pyramid scheme is then

$$\begin{tikzcd}
\mathbf{c}_J \arrow[r, "\mathbf{G}_{J-1,0}"] \arrow[dr, "\mathbf{G}_{J-1,1}"] & \mathbf{c}_{J-1} \arrow[r, "\mathbf{G}_{J-2,0}"] \arrow[dr, "\mathbf{G}_{J-2,1}"] & \mathbf{c}_{J-2} \arrow[r] \arrow[dr] & \cdots \arrow[r, "\mathbf{G}_{j_0,0}"] \arrow[dr, "\mathbf{G}_{j_0,1}"] & \mathbf{c}_{j_0} \\
 & \mathbf{d}_{J-1} & \mathbf{d}_{J-2} & \mathbf{d}_{J-1} & \mathbf{d}_{j_0}
\end{tikzcd} \tag{2.140}$$

Remark 5 Property (2.126) and the fact that $\mathbf{M}_j$ and $\mathbf{G}_j$ can be applied in $(\#\Delta_{j+1})$ operations uniformly in j entails that the complexity of applying $\mathbf{T}_J$ or $\mathbf{T}_J^{-1}$ using the pyramid scheme is of order $\mathscr{O}(\#\Delta_J) = \mathscr{O}(\dim S_J)$ uniformly in J. For this reason, $\mathbf{T}_J$ is called the *Fast Wavelet Transform* (FWT). Note that there is no need to explicitly assemble $\mathbf{T}_J$ or $\mathbf{T}_J^{-1}$.

In Table 2.4 spectral condition numbers for the Fast Wavelet Transform (FWT) for different constructions of biorthogonal wavelets on the interval computed in [62] are displayed.

Since $\cup_{j \geq j_0} S_j$ is dense in L_2, a basis for the whole space L_2 is obtained when letting $J \to \infty$ in (2.133),

$$\begin{aligned} \Psi &:= \bigcup_{j=j_0-1}^{\infty} \Psi_j = \{\psi_{j,k} : (j,k) \in I\!I\}, \qquad \Psi_{j_0-1} := \Phi_{j_0} \\ I\!I &:= \left\{\{j_0\} \times \Delta_{j_0}\right\} \cup \bigcup_{j=j_0}^{\infty} \left\{\{j\} \times \nabla_j\right\}. \end{aligned} \tag{2.141}$$

The next theorem from [23] illustrates the relation between Ψ and $\mathbf{T}_J$.

Theorem 5 *The multiscale transformations $\mathbf{T}_J$ are well-conditioned in the sense*

$$\|\mathbf{T}_J\|, \|\mathbf{T}_J^{-1}\| = \mathscr{O}(1), \quad J \geq j_0, \tag{2.142}$$

if and only if the collection Ψ defined by (2.141) is a Riesz basis for L_2, i.e., every $v \in L_2$ has unique expansions

$$v = \sum_{j=j_0-1}^{\infty} \langle v, \tilde{\Psi}_j \rangle \Psi_j = \sum_{j=j_0-1}^{\infty} \langle v, \Psi_j \rangle \tilde{\Psi}_j, \tag{2.143}$$

where $\tilde{\Psi}$ defined analogously as in (2.141) is also a Riesz basis for L_2 which is biorthogonal or dual to Ψ,

$$\langle \Psi, \tilde{\Psi} \rangle = \mathbf{I} \tag{2.144}$$

such that

$$\|v\|_{L_2} \sim \|\langle \tilde{\Psi}, v \rangle\|_{\ell_2(I\!I)} \sim \|\langle \Psi, v \rangle\|_{\ell_2(I\!I)}. \tag{2.145}$$

We briefly explain next how the functions in $\tilde{\Psi}$, denoted as *wavelets dual to* Ψ, or *dual wavelets*, can be determined. Assume that there is a second multiresolution $\tilde{\mathscr{S}}$ of L_2 satisfying (2.114) where

$$\tilde{S}_j = S(\tilde{\Phi}_j), \qquad \tilde{\Phi}_j = \{\tilde{\phi}_{j,k} : k \in \Delta_j\} \tag{2.146}$$

and $\{\tilde{\Phi}_j\}_{j=j_0}^{\infty}$ is uniformly stable in j in the sense of (2.116). Let the functions in $\tilde{\Phi}_j$ also have compact support satisfying (2.117). Furthermore, suppose that the biorthogonality conditions

$$\langle \Phi_j, \tilde{\Phi}_j \rangle = \mathbf{I} \tag{2.147}$$

hold. We will often refer to Φ_j as the *primal* and to $\tilde{\Phi}_j$ as the *dual generators*. The nestedness of the $\tilde{S}_j$ and the stability again implies that $\tilde{\Phi}_j$ is refinable with some matrix $\tilde{\mathbf{M}}_{j,0}$, similar to (2.120),

$$\tilde{\Phi}_j = \tilde{\mathbf{M}}_{j,0}^T \tilde{\Phi}_{j+1}. \tag{2.148}$$

The problem of determining biorthogonal wavelets now consists in finding bases $\Psi_j, \tilde{\Psi}_j$ for the complements of $S(\Phi_j)$ in $S(\Phi_{j+1})$, and of $S(\tilde{\Phi}_j)$ in $S(\tilde{\Phi}_{j+1})$, such that

$$S(\Phi_j) \perp S(\tilde{\Psi}_j), \qquad S(\tilde{\Phi}_j) \perp S(\Psi_j) \tag{2.149}$$

and

$$S(\Psi_j) \perp S(\tilde{\Psi}_r), \quad j \neq r, \tag{2.150}$$

holds. The connection between the concept of stable completions and the dual generators and wavelets is made by the following result which is a special case from [14].

Proposition 2 *Suppose that the biorthogonal collections* $\{\Phi_j\}_{j=j_0}^{\infty}$ *and* $\{\tilde{\Phi}_j\}_{j=j_0}^{\infty}$ *are both uniformly stable and refinable with refinement matrices* $\mathbf{M}_{j,0}$, $\tilde{\mathbf{M}}_{j,0}$*, i.e.,*

$$\Phi_j = \mathbf{M}_{j,0}^T \Phi_{j+1}, \qquad \tilde{\Phi}_j = \tilde{\mathbf{M}}_{j,0}^T \tilde{\Phi}_{j+1}, \tag{2.151}$$

and satisfy the duality condition (2.147). Assume that $\check{\mathbf{M}}_{j,1}$ *is any stable completion of* $\mathbf{M}_{j,0}$ *such that*

$$\check{\mathbf{M}}_j := (\mathbf{M}_{j,0}, \check{\mathbf{M}}_{j,1}) = \check{\mathbf{G}}_j^{-1} \tag{2.152}$$

satisfies (2.126).

Then

$$\mathbf{M}_{j,1} := (\mathbf{I} - \mathbf{M}_{j,0} \tilde{\mathbf{M}}_{j,0}^T) \check{\mathbf{M}}_{j,1} \tag{2.153}$$

is also a stable completion of $\mathbf{M}_{j,0}$*, and* $\mathbf{G}_j = \mathbf{M}_j^{-1} = (\mathbf{M}_{j,0}, \mathbf{M}_{j,1})^{-1}$ *has the form*

$$\mathbf{G}_j = \begin{pmatrix} \tilde{\mathbf{M}}_{j,0}^T \\ \check{\mathbf{G}}_{j,1} \end{pmatrix}. \tag{2.154}$$

Moreover, the collections of functions

$$\Psi_j := \mathbf{M}_{j,1}^T \Phi_{j+1}, \qquad \tilde{\Psi}_j := \check{\mathbf{G}}_{j,1} \tilde{\Phi}_{j+1} \tag{2.155}$$

form biorthogonal systems,

$$\langle \Psi_j, \tilde{\Psi}_j \rangle = \mathbf{I}, \qquad \langle \Psi_j, \tilde{\Phi}_j \rangle = \langle \Phi_j, \tilde{\Psi}_j \rangle = \mathbf{0}, \tag{2.156}$$

so that

$$S(\Psi_j) \perp S(\tilde{\Psi}_r), \quad j \neq r, \qquad S(\Phi_j) \perp S(\tilde{\Psi}_j), \quad S(\tilde{\Phi}_j) \perp S(\Psi_j). \tag{2.157}$$

In particular, the relations (2.147), (2.156) imply that the collections

$$\Psi = \bigcup_{j=j_0-1}^{\infty} \Psi_j, \qquad \tilde{\Psi} := \bigcup_{j=j_0-1}^{\infty} \tilde{\Psi}_j := \tilde{\Phi}_{j_0} \cup \bigcup_{j=j_0}^{\infty} \tilde{\Psi}_j \tag{2.158}$$

are biorthogonal,

$$\langle \Psi, \tilde{\Psi} \rangle = \mathbf{I}. \tag{2.159}$$

Remark 6 It is important to note that the properties needed in addition to (2.159) in order to ensure (2.145) are neither properties of the complements nor of their bases $\Psi, \tilde{\Psi}$ but of the multiresolution sequences $\mathscr{S}$ and $\tilde{\mathscr{S}}$. These can be phrased as approximation and regularity properties and appear in Theorem 6.

We briefly recall yet another useful point of view. The operators

$$\begin{aligned} P_j v &:= \langle v, \tilde{\Phi}_j \rangle \Phi_j = \langle v, \tilde{\Psi}^j \rangle \Psi^j = \langle v, \tilde{\Phi}_{j_0} \rangle \Phi_{j_0} + \sum_{r=j_0}^{j-1} \langle v, \tilde{\Psi}_r \rangle \Psi_r \\ P_j' v &:= \langle v, \Phi_j \rangle \tilde{\Phi}_j = \langle v, \Psi^j \rangle \tilde{\Psi}^j = \langle v, \Phi_{j_0} \rangle \tilde{\Phi}_{j_0} + \sum_{r=j_0}^{j-1} \langle v, \Psi_r \rangle \tilde{\Psi}_r \end{aligned} \tag{2.160}$$

are projectors onto

$$S(\Phi_j) = S(\Psi^j) \qquad \text{and} \qquad S(\tilde{\Phi}_j) = S(\tilde{\Psi}^j) \tag{2.161}$$

respectively, which satisfy

$$P_r P_j = P_r, \quad P'_r P'_j = P'_r, \qquad r \leq j. \tag{2.162}$$

Remark 7 Let $\{\Phi_j\}_{j=j_0}^{\infty}$ be uniformly stable. The P_j defined by (2.160) are uniformly bounded if and only if $\{\tilde{\Phi}_j\}_{j=j_0}^{\infty}$ is also uniformly stable. Moreover, the P_j satisfy (2.162) if and only if the $\tilde{\Phi}_j$ are refinable as well. Note that then (2.147) implies

$$\mathbf{M}_{j,0}^T \tilde{\mathbf{M}}_{j,0} = \mathbf{I}. \tag{2.163}$$

In terms of the projectors, the uniform stability of the complement bases Ψ_j, $\tilde{\Psi}_j$ means that

$$\|(P_{j+1} - P_j)v\|_{L_2} \sim \|\langle \tilde{\Psi}_j, v\rangle\|_{\ell_2(\nabla_j)}, \quad \|(P'_{j+1} - P'_j)v\|_{L_2} \sim \|\langle \Psi_j, v\rangle\|_{\ell_2(\nabla_j)}, \tag{2.164}$$

so that the L_2 norm equivalence (2.145) is equivalent to

$$\|v\|_{L_2}^2 \sim \sum_{j=j_0}^{\infty} \|(P_j - P_{j-1})v\|_{L_2}^2 \sim \sum_{j=j_0}^{\infty} \|(P'_j - P'_{j-1})v\|_{L_2}^2 \tag{2.165}$$

for any $v \in L_2$, where $P_{j_0-1} = P'_{j_0-1} := 0$.

The whole concept derived so far lives from both Φ_j *and* $\tilde{\Phi}_j$. It should be pointed out that in the algorithms one actually does not need $\tilde{\Phi}_j$ explicitly for computations.

We recall next results that guarantee norm equivalences of the type (2.89) for Sobolev spaces.

Multiresolution of Sobolev Spaces Let now $\mathscr{S}$ be a multiresolution sequence consisting of closed subspaces of H^s with the property (2.114) whose union is dense in H^s. The following result from [23] ensures under which conditions norm equivalences hold for the H^s-norm.

Theorem 6 *Let $\{\Phi_j\}_{j=j_0}^{\infty}$ and $\{\tilde{\Phi}_j\}_{j=j_0}^{\infty}$ be uniformly stable, refinable, biorthogonal collections and let the $P_j : H^s \to S(\Phi_j)$ be defined by (2.160). If the* Jackson-type estimate

$$\inf_{v_j \in S_j} \|v - v_j\|_{L_2} \lesssim 2^{-sj} \|v\|_{H^s}, \quad v \in H^s, \; 0 < s \leq \bar{d}, \tag{2.166}$$

and the Bernstein inequality

$$\|v_j\|_{H^s} \lesssim 2^{sj} \|v_j\|_{L_2}, \quad v_j \in S_j, \; s < \bar{t}, \tag{2.167}$$

hold for

$$S_j = \left\{ \begin{matrix} S(\Phi_j) \\ S(\tilde{\Phi}_j) \end{matrix} \right\} \; \textit{with} \text{ order } \bar{d} = \left\{ \begin{matrix} d \\ \tilde{d} \end{matrix} \right\} \; \textit{and } \bar{t} = \left\{ \begin{matrix} t \\ \tilde{t} \end{matrix} \right\}, \tag{2.168}$$

then for

$$0 < \sigma := \min\{d, t\}, \qquad 0 < \tilde{\sigma} := \min\{\tilde{d}, \tilde{t}\}, \tag{2.169}$$

one has

$$\|v\|_{H^s}^2 \sim \sum_{j=j_0}^{\infty} 2^{2sj} \|(P_j - P_{j-1})v\|_{L_2}^2, \quad s \in (-\tilde{\sigma}, \sigma). \tag{2.170}$$

Recall that we always write $H^s = (H^{-s})'$ for $s < 0$.
The regularity of $\mathscr{S}$ and $\tilde{\mathscr{S}}$ is characterized by

$$t := \sup\{s : S(\Phi_j) \subset H^s, \; j \geq j_0\}, \qquad \tilde{t} := \sup\{s : S(\tilde{\Phi}_j) \subset H^s, \; j \geq j_0\} \tag{2.171}$$

Recalling the representation (2.164), we can immediately derive the following fact.

Corollary 2 *Suppose that the assumptions in Theorem 6 hold. Then we have the norm equivalence*

$$\|v\|_{H^s}^2 \sim \sum_{j=j_0-1}^{\infty} 2^{2sj} \|\langle \tilde{\Psi}_j, v \rangle\|_{\ell_2(\nabla_j)}^2, \quad s \in (-\tilde{\sigma}, \sigma). \tag{2.172}$$

In particular for $s = 0$ the Riesz basis property of the Ψ, $\tilde{\Psi}$ relative to L_2(2.145) is recovered. For many applications it suffices to have (2.170) or (2.172) only for certain $s > 0$ for which one only needs to require (2.166) and (2.167) for $\{\Phi_j\}_{j=j_0}^{\infty}$. The Jackson estimates (2.166) of order $\tilde{d}$ for $S(\tilde{\Phi}_j)$ imply the cancellation properties (CP) (2.92), see, e.g., [26].

Remark 8 When the wavelets live on $\Omega \subset \mathbb{R}^n$, (2.166) means that all polynomials up to order $\tilde{d}$ are contained in $S(\tilde{\Phi}_j)$. One also says that $S(\tilde{\Phi}_j)$ is *exact* of order $\tilde{d}$. On account of (2.144), this implies that the wavelets $\psi_{j,k}$ are orthogonal to polynomials up to order $\tilde{d}$ or have $\tilde{d}$th order *vanishing moments*. By Taylor expansion, this in turn yields (2.92).

We will later use the following generalization of the discrete norms (2.165). Let for $s \in \mathbb{R}$

$$|||v|||_s := \left(\sum_{j=j_0}^{\infty} 2^{2sj} \|(P_j - P_{j-1})v\|_{L_2}^2 \right)^{1/2} \tag{2.173}$$

which by the relations (2.164) is also equivalent to

$$\left| v \right|_s := \left(\sum_{j=j_0-1}^{\infty} 2^{2sj} \|\langle \tilde{\Psi}_j, v \rangle\|_{\ell_2(\nabla_j)}^2 \right)^{1/2} . \tag{2.174}$$

In this notation, (2.170) and (2.172) read

$$\|v\|_{H^s} \sim |||v|||_s \sim \left| v \right|_s . \tag{2.175}$$

In terms of such discrete norms, Jackson and Bernstein estimates hold with constants equal to one [51], which turns out to be useful later in Sect. 2.5.2.

Lemma 1 *Let $\{\Phi_j\}_{j=j_0}^{\infty}$ and $\{\tilde{\Phi}_j\}_{j=j_0}^{\infty}$ be uniformly stable, refinable, biorthogonal collections and let the P_j be defined by (2.160). Then the estimates*

$$\left| v - P_j v \right|_{s'} \leq 2^{-(j+1)(s-s')} \left| v \right|_s, \qquad v \in H^s, \;\; s' \leq s \leq d, \tag{2.176}$$

and

$$\left| v_j \right|_s \leq 2^{j(s-s')} \left| v_j \right|_{s'}, \qquad v_j \in S(\Phi_j), \;\; s' \leq s \leq d, \tag{2.177}$$

are valid, and correspondingly for the dual side.

The same results hold for the norm $|||\cdot|||$ defined in (2.173).

Reverse Cauchy–Schwarz Inequalities The biorthogonality condition (2.147) implies together with direct and inverse estimates the following reverse Cauchy–Schwarz inequalities for finite-dimensional spaces [28]. It will be one essential ingredient for the discussion of the LBB condition in Sect. 2.5.2.

Lemma 2 *Let the assumptions in Theorem 6 be valid such that the norm equivalence (2.170) holds for $(-\tilde{\sigma}, \sigma)$ with $\sigma, \tilde{\sigma}$ defined in (2.169). Then for any $v \in S(\Phi_j)$ there exists some $\tilde{v}^* = \tilde{v}^*(v) \in S(\tilde{\Phi}_j)$ such that*

$$\|v\|_{H^s} \|\tilde{v}^*\|_{H^{-s}} \lesssim \langle v, \tilde{v}^* \rangle \tag{2.178}$$

for any $0 \leq s < \min(\sigma, \tilde{\sigma})$.

The proof of this result given in [28] for $s = 1/2$ in terms of the projectors P_j defined in (2.160) and corresponding duals P_j' immediately carries over to more general s. Recalling the representation (2.161) in terms of wavelets, the reverse Cauchy inequality (2.178) attains the following sharp form.

Lemma 3 ([51]) *Let the assumptions of Lemma 1 hold. Then for every $v \in S(\Phi_j)$ there exists some $\tilde{v}^* = \tilde{v}^*(v) \in S(\tilde{\Phi}_j)$ such that*

$$\|v\|_s \, \|\tilde{v}^*\|_{-s} = \langle v, \tilde{v}^* \rangle \tag{2.179}$$

for any $0 \leq s \leq \min(\sigma, \tilde{\sigma})$.

Proof Every $v \in S(\Phi_j)$ can be written as

$$v = \sum_{r=j_0-1}^{j-1} 2^{sr} \sum_{k \in \nabla_r} v_{r,k} \psi_{r,k}.$$

Setting now

$$\tilde{v}^* := \sum_{r=j_0-1}^{j-1} 2^{-sr} \sum_{k \in \nabla_r} v_{r,k} \tilde{\psi}_{r,k}$$

with the same coefficients $v_{j,k}$, the definition of $\|\cdot\|_s$ yields by biorthogonality (2.159)

$$\|v\|_s \, \|\tilde{v}^*\|_{-s} = \sum_{r=j_0-1}^{j-1} \sum_{k \in \nabla_r} |v_{j,k}|^2.$$

Combining this with the observation

$$\langle v, \tilde{v}^* \rangle = \sum_{r=j_0-1}^{j-1} \sum_{k \in \nabla_r} |v_{j,k}|^2$$

confirms (2.179). □

Remark 9 The previous proof reveals that the identity (2.179) is also true for elements from infinite-dimensional spaces H^s and $(H^s)'$ for which Ψ and $\tilde{\Psi}$ are Riesz bases.

Biorthogonal Wavelets on $\mathbb{R}$ The construction of biorthogonal spline-wavelets on $\mathbb{R}$ from [18] for $L_2 = L_2(\mathbb{R})$ employs the multiresolution framework introduced at the beginning of this section. There the $\phi_{j,k}$ are generated through the dilates and translates of a single function $\phi \in L_2$,

$$\phi_{j,k} = 2^{j/2}\phi(2^j \cdot -k). \tag{2.180}$$

This corresponds to the idea of a *uniform* virtual underlying grid, explaining the terminology *uniform refinements*. B-Splines on uniform grids are known to satisfy refinement relations (2.119) in addition to being compactly supported and having L_2-stable integer translates. For computations, they have the additional advantage that they can be expressed as piecewise polynomials. In the context of variational formulations for second order boundary value problems, a well-used example are the nodal finite elements $\phi_{j,k}$ generated by the cardinal B-Spline of order two, i.e., the piecewise linear continuous function commonly called the 'hat function'. For cardinal B-Splines as generators, a whole class of dual generators $\tilde{\phi}_{j,k}$ (of arbitrary smoothness at the expense of larger supports) can be constructed which are also generated by one single function $\tilde{\phi}$ through translates and dilates. By Fourier techniques, one can construct from $\phi, \tilde{\phi}$ then a pair of biorthogonal wavelets $\psi, \tilde{\psi}$ whose dilates and translates built as in (2.180) constitute Riesz bases for $L_2(\mathbb{R})$.

By taking tensor products of these functions, one can generate biorthogonal wavelet bases for $L_2(\mathbb{R}^n)$.

Biorthogonal Wavelets on Domains Some constructions that exist by now have as a core ingredient tensor products of one-dimensional wavelets on an *interval* derived from the biorthogonal wavelets from [18] on $\mathbb{R}$. On finite intervals in $\mathbb{R}$, the corresponding constructions are usually based on keeping the elements of $\Phi_j, \tilde{\Phi}_j$ supported *inside* the interval while modifying those translates overlapping the end points of the interval so as to preserve a desired degree of polynomial exactness. A general detailed construction satisfying all these requirements has been proposed in [34]. Here just the main ideas for constructing a biorthogonal pair $\Phi_j, \tilde{\Phi}_j$ and corresponding wavelets satisfying the above requirements are sketched, where we apply the techniques derived at the beginning of this section.

We start out with those functions from two collections of biorthogonal generators $\Phi_j^{\mathbb{R}}, \tilde{\Phi}_j^{\mathbb{R}}$ for some fixed $j \geq j_0$ living on the whole real line whose support has nonempty intersection with the interval $(0, 1)$. In order to treat the boundary effects separately, we assumed that the coarsest resolution level j_0 is large enough so that, in view of (2.117), functions overlapping one end of the interval vanish at the other. One then leaves as many functions from the collection $\Phi_j^{\mathbb{R}}, \tilde{\Phi}_j^{\mathbb{R}}$ living in the interior of the interval untouched and modifies only those near the interval ends. Note that keeping just the restrictions to the interval of those translates overlapping the end points would destroy stability (and also the cardinality of the primal and dual basis functions living on $(0, 1)$ since their supports do not have the same size). Therefore, modifications at the end points are necessary; also, just discarding them from the collections (2.115), (2.146) would produce an error near the end points. The basic idea is essentially the same for all constructions of orthogonal and biorthogonal wavelets on $\mathbb{R}$ adapted to an interval. Namely, one takes *fixed* linear combinations of all functions in $\Phi_j^{\mathbb{R}}, \tilde{\Phi}_j^{\mathbb{R}}$ living near the ends of the interval in such a way that monomials up to the exactness order are reproduced there and such that the generator bases have the same cardinality. Because of the boundary modifications, the collections of generators are there no longer biorthogonal. However, one can show in the case of cardinal B-Splines as primal generators (which is a widely used class for numerical analysis) that biorthogonalization is indeed possible. This yields collections denoted by $\Phi_j^{(0,1)}, \tilde{\Phi}_j^{(0,1)}$ which then satisfy (2.147) on (0, 1) and all assumptions required in Proposition 2.

For the construction of corresponding wavelets, first an *initial* stable completion $\check{\mathbf{M}}_{j,1}$ is computed by applying Gaussian eliminations to factor $\mathbf{M}_{j,0}$ and then to find a uniformly stable inverse of $\check{\mathbf{M}}_j$. Here we exploit that for cardinal B-Splines as generators the refinement matrices $\mathbf{M}_{j,0}$ are totally positive. Thus, they can be stably decomposed by Gaussian elimination without pivoting. Application of Proposition 2 then gives the corresponding biorthogonal wavelets $\Psi_j^{(0,1)}, \tilde{\Psi}_j^{(0,1)}$ on (0, 1) which satisfy the requirements in Corollary 2. It turns out that these wavelets coincide in the interior of the interval again with those on all of $\mathbb{R}$ from [18]. An example of the primal wavelets for $d = 2$ generated by piecewise linear continuous functions is displayed in Fig. 2.2 on the left. After constructing these basic versions, one can then perform local transformations near the ends of the interval in order to improve the condition or L_2 stability constants, see [11, 62] for corresponding results and numerical examples.

We display spectral condition numbers for the FWT for two different constructions of biorthogonal wavelets on the interval computed in [62] in Table 2.4. The first column denotes the finest level on which the spectral condition numbers of the FWT are computed. The next column contains the numbers for the construction of biorthogonal spline-wavelets on the interval from [34] for the case $d = 2, \tilde{d} = 4$ while the last column displays the numbers for a scaled version derived in [11]. We will see later in Sect. 2.5.1 how the transformation $\mathbf{T}_J$ is used for preconditioning.

Along these lines, also biorthogonal generators and wavelets with homogeneous (Dirichlet) boundary conditions can be constructed. Since the $\Phi_j^{(0,1)}$ are locally near the boundary monomials which all vanish at 0, 1 except for one, removing the one from $\Phi_j^{(0,1)}$ which corresponds to the constant function produces a collection of generators with homogeneous boundary conditions at 0, 1. In order for the

Table 2.4 Computed spectral condition numbers [62] for the Fast Wavelet Transform for different constructions of biorthogonal wavelets on the interval [11, 34]

j	$\kappa_2(\mathbf{T}_{\mathrm{DKU}})$	$\kappa_2(\mathbf{T}_{\mathrm{B}})$
4	4.743e+00	4.640e+00
5	6.221e+00	6.024e+00
6	8.154e+00	6.860e+00
7	9.473e+00	7.396e+00
8	1.023e+01	7.707e+00
9	1.064e+01	7.876e+00
10	1.086e+01	7.965e+00
11	1.097e+01	8.011e+00
12	1.103e+01	8.034e+00
13	1.106e+01	8.046e+00
14	1.107e+01	8.051e+00
15	1.108e+01	8.054e+00
16	1.108e+01	8.056e+00

moment conditions (2.92) still to hold for the Ψ_j, the dual generators have to have *complementary* boundary conditions. A corresponding construction has been carried out in [30] and implemented in [11]. Homogeneous boundary conditions of higher order can be generated accordingly.

By taking tensor products of the wavelets on $(0, 1)$, in this manner biorthogonal wavelets for Sobolev spaces on $(0, 1)^n$ with or without homogeneous boundary conditions are obtained. This construction can be further extended to any other domain or manifold which is the image of a regular parametric mapping of the unit cube. Some results on the construction of wavelets on manifolds are summarized in [25]. There are essentially two approaches. The first idea is based on domain decomposition and consists in 'gluing' generators across interelement boundaries, see, e.g., [13, 31]. These approaches all have in common that the norm equivalences (2.172) for $H^s = H^s(\Gamma)$ can be shown to hold only for the range $-1/2 < s < 3/2$, due to the fact that duality arguments apply only for this range because of the nature of a modified inner product to which biorthogonality refers. The other approach which overcomes the above limitations on the ranges for which the norm equivalences hold has been developed in [32] based on previous characterizations of function spaces as Cartesian products from [15]. The construction in [32] has been optimized and implemented to construct wavelet bases on the sphere in [56, 64], see Fig. 2.2.

Of course, there are also different attempts to construct wavelet bases with the above properties without using tensor products. A construction of biorthogonal spline-wavelets on triangles introduced by [68] has been implemented in two spatial dimensions with an application to the numerical solution of a linear elliptic boundary value problem in [48].

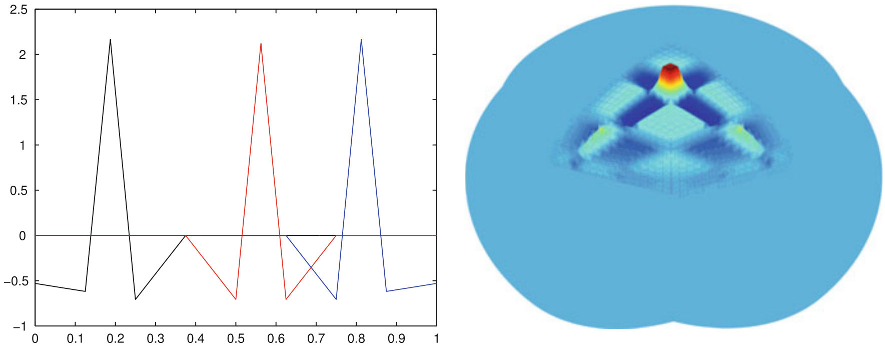


Fig. 2.2 Primal wavelets for $d = 2$ on $[0, 1]$ (left) and on a sphere (right) from [64]

2.5 Problems in Wavelet Coordinates

2.5.1 Elliptic Boundary Value Problems

We now consider the wavelet representation of the elliptic boundary value problem from Sect. 2.3.2. Let for $\mathscr{H}$ given by (2.41) or (2.42) $\Psi_{\mathscr{H}}$ be a wavelet basis with corresponding dual $\tilde{\Psi}_{\mathscr{H}}$ which satisfies the properties (R), (L) and (CP) from Sect. 2.4.1. Following the recipe from Sect. 2.4.3, expanding $y = \mathbf{y}^T \Psi_{\mathscr{H}}$, $f = \mathbf{f}^T \tilde{\Psi}_{\mathscr{H}}$ and recalling (2.45), the wavelet representation of the elliptic boundary value problem (2.47) is given by

$$\mathbf{A}\mathbf{y} = \mathbf{f} \tag{2.181}$$

where

$$\mathbf{A} := a(\Psi_{\mathscr{H}}, \Psi_{\mathscr{H}}), \qquad \mathbf{f} := \langle \Psi_{\mathscr{H}}, f \rangle. \tag{2.182}$$

Then the mapping property (2.46) and the Riesz basis property (R) yield the following fact.

Proposition 3 *The infinite matrix* $\mathbf{A}$ *is a boundedly invertible mapping from* $\ell_2 = \ell_2(I\!I_{\mathscr{H}})$ *into itself, and there exists finite positive constants* $c_{\mathbf{A}} \leq C_{\mathbf{A}}$ *such that*

$$c_{\mathbf{A}} \|\mathbf{v}\| \leq \|\mathbf{A}\mathbf{v}\| \leq C_{\mathbf{A}} \|\mathbf{v}\|, \qquad \mathbf{v} \in \ell_2(I\!I_{\mathscr{H}}). \tag{2.183}$$

Proof For any $v \in \mathscr{H}$ with coefficient vector $\mathbf{v} \in \ell_2$, we have by the lower estimates in (2.89), (2.46) and the upper inequality in (2.94), respectively,

$$\|\mathbf{v}\| \leq c_{\mathscr{H}}^{-1} \|v\|_{\mathscr{H}} \leq c_{\mathscr{H}}^{-1} c_A^{-1} \|Av\|_{\mathscr{H}'} = c_{\mathscr{H}}^{-1} c_A^{-1} \|(\mathbf{A}\mathbf{v})^T \tilde{\Psi}_{\mathscr{H}}\|_{\mathscr{H}'} \leq c_{\mathscr{H}}^{-2} c_A^{-1} \|\mathbf{A}\mathbf{v}\|$$

where we have used the wavelet representation (2.111) for A. Likewise, the converse estimate

$$\|\mathbf{A}\mathbf{v}\| \leq C_{\mathscr{H}} \|Av\|_{\mathscr{H}'} \leq C_{\mathscr{H}} C_A \|v\|_{\mathscr{H}} \leq C_{\mathscr{H}}^2 C_A \|\mathbf{v}\|$$

follows by the lower inequality in (2.94) and the upper estimates in (2.46) and (2.89). The constants appearing in (2.183) are therefore identified as $c_{\mathbf{A}} := c_{\mathscr{H}}^2 c_A$ and $C_{\mathbf{A}} := c_{\mathscr{H}}^2 C_A$. □

In the present situation where $\mathbf{A}$ is defined via the elliptic bilinear form $a(\cdot, \cdot)$, Proposition 3 entails the following result with respect to *preconditioning*. Let for $I\!I = I\!I_{\mathscr{H}}$ the symbol Λ denote *any* finite subset of the index set $I\!I$. For the corresponding set of wavelets $\Psi_\Lambda := \{\psi_\lambda : \lambda \in \Lambda\}$ denote by $S_\Lambda := span \Psi_\Lambda$ the respective finite-dimensional subspace of $\mathscr{H}$. For the wavelet representation of

A in terms of Ψ_Λ,

$$\mathbf{A}_\Lambda := a(\Psi_\Lambda, \Psi_\Lambda), \tag{2.184}$$

we obtain the following result.

Proposition 4 *If $a(\cdot,\cdot)$ is $\mathscr{H}$-elliptic according to (2.44), the finite matrix $\mathbf{A}_\Lambda$ is symmetric positive definite and its spectral condition number is bounded uniformly in Λ, i.e.,*

$$\kappa_2(\mathbf{A}_\Lambda) \ \leq \frac{C_\mathbf{A}}{c_\mathbf{A}}, \tag{2.185}$$

where $c_\mathbf{A}$, $C_\mathbf{A}$ are the constants from (2.183).

Proof Clearly, since $\mathbf{A}_\Lambda$ is just a finite section of $\mathbf{A}$, we have $\|\mathbf{A}_\Lambda\| \ \leq \ \|\mathbf{A}\|$. On the other hand, by assumption, $a(\cdot,\cdot)$ is $\mathscr{H}$-elliptic which entails that $a(\cdot,\cdot)$ is also elliptic on every finite subspace $S_\Lambda \subset \mathscr{H}$. Thus, we infer $\|\mathbf{A}_\Lambda^{-1}\| \leq \|\mathbf{A}^{-1}\|$, and we have

$$c_\mathbf{A}\|\mathbf{v}_\Lambda\| \leq \|\mathbf{A}_\Lambda\mathbf{v}_\Lambda\| \leq C_\mathbf{A}\|\mathbf{v}_\Lambda\|, \qquad \mathbf{v}_\Lambda \in S_\Lambda. \tag{2.186}$$

Together with the definition $\kappa_2(\mathbf{A}_\Lambda) \ := \ \|\mathbf{A}_\Lambda\|\,\|\mathbf{A}_\Lambda^{-1}\|$ we obtain the claimed estimate. □

In other words, representations of A with respect to properly scaled wavelet bases for $\mathscr{H}$ entail well-conditioned system matrices $\mathbf{A}_\Lambda$ independent of Λ. This in turn means that the convergence speed of an iterative solver applied to the corresponding finite system

$$\mathbf{A}_\Lambda\mathbf{y}_\Lambda = \mathbf{f}_\Lambda \tag{2.187}$$

does not deteriorate as $\Lambda \to \infty$.

In summary, ellipticity implies stability of the Galerkin discretizations for any set $\Lambda \subset I\!I$. This is not the case for finite versions of the saddle point problems discussed in Sect. 2.5.2.

Fast Wavelet Transform Let us briefly summarize how in the situation of uniform refinements, i.e., when $S(\Phi_J) = S(\Psi^J)$, the Fast Wavelet Transformation (FWT) $\mathbf{T}_J$ can be used for preconditioning linear elliptic operators, together with a diagonal scaling induced by the norm equivalence (2.172) [27]. Here we recall the notation from Sect. 2.4.4 where the wavelet basis is in fact the (unscaled) anchor basis from Sect. 2.4.1. Thus, the norm equivalence (2.89) using the scaled wavelet basis Ψ_H is the same as (2.172) in the anchor basis. Recall that the norm equivalence (2.172) implies that every $v \in H^s$ can be expanded uniquely in terms of the Ψ and its expansion coefficients $\mathbf{v}$ satisfy

$$\|v\|_{H^s} \ \sim \ \|\mathbf{D}^s\mathbf{v}\|_{\ell_2}$$

where $\mathbf{D}^s$ is a diagonal matrix with entries $\mathbf{D}^s_{(j,k),(j',k')} = 2^{sj}\delta_{j,j'}\delta_{k,k'}$. For $\mathscr{H} \subset H^1(\Omega)$, the case $s = 1$ is relevant.

In a stable Galerkin scheme for (2.43) with respect to $S(\Psi^J) = S(\Psi_\Lambda)$, we have therefore already identified the diagonal (scaling) matrix $\mathbf{D}_J$ consisting of the finite portion of the matrix $\mathbf{D} = \mathbf{D}^1$ for which $j_0 - 1 \le j \le J - 1$. The representation of A with respect to the (unscaled) wavelet basis Ψ^J can be expressed in terms of the Fast Wavelet Transform $\mathbf{T}_J$, that is,

$$\langle \Psi^J, A\Psi^J \rangle \;=\; \mathbf{T}_J^T \,\langle \Phi_J, A\Phi_J \rangle \,\mathbf{T}_J, \tag{2.188}$$

where Φ_J is the single-scale basis for $S(\Psi^J)$. Thus, we first set up the operator equation as in Finite Element settings in terms of the single-scale basis Φ_J. Applying the Fast Wavelet Transform $\mathbf{T}_J$ together with $\mathbf{D}_J$ yields that the operator

$$\mathbf{A}_J \;:=\; \mathbf{D}_J^{-1}\,\mathbf{T}_J^T\,\langle \Phi_J, A\Phi_J \rangle\,\mathbf{T}_J\,\mathbf{D}_J^{-1} \tag{2.189}$$

has uniformly bounded condition numbers independent of J. This can be seen by combining the properties of A according to (2.46) with the norm equivalences (2.89) and (2.94).

It is known that the boundary adaptations of the generators and wavelets aggravate the absolute values of the condition numbers. Nevertheless, these constants can be greatly reduced by sophisticated biorthogonalizations of the boundary adapted functions [11]. Numerical tests confirm that the absolute constants can further be improved by taking instead of $\mathbf{D}_J^{-1}$ the inverse of the diagonal of $\langle \Psi^J, A\Psi^J \rangle$ for the scaling in (2.189) [11, 17, 62]. Table 2.5 displays the condition numbers for discretizations of an operator in two spatial dimensions for boundary adapted biorthogonal spline-wavelets in the case $d = 2$, $\tilde{d} = 4$ with such a scaling.

2.5.2 Saddle Point Problems Involving Boundary Conditions

As in the previous situation, we first derive an infinite wavelet representation of the saddle point problem introduced in Sect. 2.3.3.

Let for $\mathscr{H} = Y \times Q$ with $Y = H^1(\Omega)$, $Q = (H^{1/2}(\Gamma))'$ two collections of wavelet bases Ψ_Y, Ψ_Q be available, each satisfying (R), (L) and (CP), with respective duals $\tilde{\Psi}_Y$, $\tilde{\Psi}_Q$. Similar to the previous case, we expand $y = \mathbf{y}^T\Psi_Y$ and $p = \mathbf{p}^T\Psi_Q$ and test with the elements from Ψ_Y, Ψ_Q. Then (2.54) attains the form

$$\mathbf{L}\begin{pmatrix}\mathbf{y}\\ \mathbf{p}\end{pmatrix} := \begin{pmatrix}\mathbf{A} & \mathbf{B}^T\\ \mathbf{B} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{y}\\ \mathbf{p}\end{pmatrix} = \begin{pmatrix}\mathbf{f}\\ \mathbf{g}\end{pmatrix}, \tag{2.190}$$

where

$$\begin{aligned} \mathbf{A} &:= \langle \Psi_Y, A\Psi_Y \rangle \quad \mathbf{f} := \langle \Psi_Y, f \rangle, \\ \mathbf{B} &:= \langle \Psi_Q, B\Psi_Y \rangle, \ \mathbf{g} := \langle \Psi_Q, g \rangle. \end{aligned} \tag{2.191}$$

In view of the above assertions, the operator $\mathbf{L}$ is an ℓ_2-automorphism, i.e., for every $(\mathbf{v}, \mathbf{q}) \in \ell_2(\mathit{I\!I}) = \ell_2(\mathit{I\!I}_Y \times \mathit{I\!I}_Q)$ we have

$$c_{\mathbf{L}} \left\| \begin{pmatrix} \mathbf{v} \\ \mathbf{q} \end{pmatrix} \right\| \ \le \ \left\| \mathbf{L} \begin{pmatrix} \mathbf{v} \\ \mathbf{q} \end{pmatrix} \right\| \ \le \ C_{\mathbf{L}} \left\| \begin{pmatrix} \mathbf{v} \\ \mathbf{q} \end{pmatrix} \right\| \tag{2.192}$$

with constants $c_{\mathbf{L}}, C_{\mathbf{L}}$ only depending on $c_{\mathscr{L}}, C_{\mathscr{L}}$ from (2.59) and the constants in the norm equivalences (2.89) and (2.94).

For saddle point problems with an operator $\mathbf{L}$ satisfying (2.192), finite sections are in general not uniformly stable in the sense of (2.186). In fact, for discretizations on uniform grids, the validity of the corresponding mapping property relies on a suitable stability condition, see e.g. [9, 42]. The relevant facts derived in [28] are as follows.

The bilinear form $a(\cdot, \cdot)$ defined in (2.40) is for $c > 0$ elliptic on all of $Y = H^1(\Omega)$ and, hence, also on any finite-dimensional subspace of Y. Let there be two multiresolution analyses $\mathscr{Y}$ of $H^1(\Omega)$ and $\mathscr{Q}$ of Q where the discrete spaces are $Y_j \subset H^1(\Omega)$ and $Q_\Lambda =: Q_\ell \subset (H^{1/2}(\Gamma))'$. With the notation from Sect. 2.4.4 and in addition superscripts referring to the domain on which the functions live, these spaces are represented by

$$\begin{aligned} Y_j &= S(\Phi_j^\Omega) = S(\Psi^{j,\Omega}), \ \tilde{Y}_j = S(\tilde{\Phi}_j^\Omega) = S(\tilde{\Psi}^{j,\Omega}), \\ Q_\ell &= S(\Phi_\ell^\Gamma) = S(\Psi^{\ell,\Gamma}), \ \tilde{Q}_\ell = S(\tilde{\Phi}_\ell^\Gamma) = S(\tilde{\Psi}^{\ell,\Gamma}). \end{aligned} \tag{2.193}$$

Here the indices j and ℓ refer to mesh sizes on the domain and the boundary,

$$h_\Omega \sim 2^{-j} \qquad \text{and} \qquad h_\Gamma \sim 2^{-\ell}.$$

The discrete inf–sup condition, the *LBB condition*, for the pair Y_j, Q_ℓ requires that there exists a constant $\beta_1 > 0$ *independent* of j and ℓ such that

$$\inf_{q \in Q_\ell} \ \sup_{v \in Y_j} \frac{b(v, q)}{\|v\|_{H^1(\Omega)} \, \|q\|_{(H^{1/2}(\Gamma))'}} \ \ge \ \beta_1 > 0 \tag{2.194}$$

holds. We have investigated in [28] the general case in arbitrary spatial dimensions where the Q_ℓ are *not* trace spaces of Y_j. Employing the reverse Cauchy-Schwarz inequalities from Sect. 2.4.4, one can show that (2.194) is satisfied provided that $h_\Gamma (h_\Omega)^{-1} = 2^{j-\ell} \ge c_\Omega > 1$, similar to a condition which was known for bivariate polygons and particular finite elements [2, 41].

Table 2.5 Spectral condition numbers of the operators **A** and **L** for different constructions of biorthogonal wavelets on the interval [62]

j	$\kappa_2(\mathbf{A}_{\mathrm{DKU}})$	$\kappa_2(\mathbf{A}_{\mathrm{B}})$	$\kappa_2(\mathbf{L}_{\mathrm{DKU}})$	$\kappa_2(\mathbf{L}_{\mathrm{DKU}})$
3	5.195e+02	1.898e+01	1.581e+02	4.147e+01
4	6.271e+02	1.066e+02	1.903e+02	1.050e+02
5	6.522e+02	1.423e+02	1.997e+02	1.399e+02
6	6.830e+02	1.820e+02	2.112e+02	1.806e+02
7	7.037e+02	2.162e+02	2.318e+02	2.145e+02
8	7.205e+02	2.457e+02	2.530e+02	2.431e+02
9	7.336e+02	2.679e+02	2.706e+02	2.652e+02

It should be mentioned that the obstructions caused by the LBB condition can be avoided by means of stabilization techniques proposed, e.g., in [67] where, however, the location of the boundary of Ω relative to the mesh is somewhat constrained. Another stabilization strategy based on wavelets has been investigated in [6]. A related approach which systematically avoids restrictions of the LBB type is based on least squares techniques [35].

It is particularly interesting that adaptive schemes based on wavelets like the one in Sect. 2.6.2 can be designed in such a way that the LBB condition is *automatically* enforced which was first observed in [22]. More on this subject can be found in [26].

In order to get an impression of the value of the constants for the condition numbers for $\mathbf{A}_\Lambda$ in (2.185) and the corresponding ones for the saddle point operator on uniform grids (2.192), we mention an example investigated and implemented in [62]. In this example, $\Omega = (0,1)^2$ and Γ is one face of its boundary. In Table 2.5 from [62], the spectral condition numbers of **A** and **L** with respect to two different constructions of wavelets for the case $d = 2$ and $\tilde{d} = 4$ are displayed. We see next to the first column in which the refinement level j is listed the spectral condition numbers of **A** with the wavelet construction from [34] denoted by $\mathbf{A}_{\mathrm{DKU}}$ and with the modification introduced in [11] and a further transformation [62] denoted by $\mathbf{A}_{\mathrm{B}}$. The last columns contain the respective numbers for the saddle point matrix **L** where $\kappa_2(\mathbf{L}) := \sqrt{\kappa(\mathbf{L}^T\mathbf{L})}$.

2.5.3 Control Problems: Distributed Control

We now discuss appropriate wavelet formulations for PDE-constrained control problems with distributed control as introduced in Sect. 2.3.5. Let for any space $\mathcal{V} \in \{H, \mathcal{Z}, \mathcal{U}\}$ $\Psi_{\mathcal{V}}$ denote a wavelet basis with the properties (R), (L), (CP) for $\mathcal{V}$ with dual basis $\tilde{\Psi}_{\mathcal{V}}$.

Let $\mathcal{Z}, \mathcal{U}$ satisfy the embedding (2.76). In terms of wavelet bases, the corresponding canonical injections correspond in view of (2.96) to a multiplication by a diagonal matrix. That is, let $\mathbf{D}_{\mathcal{Z}}, \mathbf{D}_H$ be such that

$$\Psi_{\mathcal{Z}} = \mathbf{D}_{\mathcal{Z}} \Psi_H, \quad \tilde{\Psi}_H = \mathbf{D}_H \Psi_{\mathcal{U}}. \tag{2.195}$$

Since $\mathscr{Z}$ possibly induces a weaker and $\mathscr{U}$ a stronger topology, the diagonal matrices $\mathbf{D}_{\mathscr{Z}}, \mathbf{D}_H$ are such that their entries are nondecreasing in scale, and there is a finite constant C such that

$$\|\mathbf{D}_{\mathscr{Z}}^{-1}\|, \|\mathbf{D}_H^{-1}\| \leq C. \tag{2.196}$$

For instance, for $H = H^\alpha$, $\mathscr{Z} = H^\beta$, or for $H' = H^{-\alpha}$, $\mathscr{U} = H^{-\beta}$, $0 \leq \beta \leq \alpha$, $\mathbf{D}_{\mathscr{Z}}, \mathbf{D}_H$ have entries $(\mathbf{D}_{\mathscr{Z}})_{\lambda,\lambda} = (\mathbf{D}_H)_{\lambda,\lambda} = (\mathbf{D}^{\alpha-\beta})_{\lambda,\lambda} = 2^{(\alpha-\beta)|\lambda|}$.

We expand y in Ψ_H and u in a wavelet basis $\Psi_{\mathscr{U}}$ for $\mathscr{U} \subset H'$,

$$u = \mathbf{u}^T \Psi_U = (\mathbf{D}_H^{-1}\mathbf{u})^T \Psi_{H'}. \tag{2.197}$$

Following the derivation in Sect. 2.5.1, the linear constraints (2.75) attain the form

$$\mathbf{Ay} = \mathbf{f} + \mathbf{D}_H^{-1}\mathbf{u} \tag{2.198}$$

where

$$\mathbf{A} := a(\Psi_H, \Psi_H), \qquad \mathbf{f} := \langle \Psi_H, f \rangle. \tag{2.199}$$

Recall that $\mathbf{A}$ has been assumed to be symmetric. The objective functional (2.80) is stated in terms of the norms $\|\cdot\|_{\mathscr{Z}}$ and $\|\cdot\|_{\mathscr{U}}$. For an exact representation of these norms, corresponding Riesz operators $\mathbf{R}_{\mathscr{Z}}$ and $\mathbf{R}_{\mathscr{U}}$ defined analogously to (2.106) would come into play which may not be explicitly computable if $\mathscr{Z}, \mathscr{U}$ are fractional Sobolev spaces. On the other hand, as mentioned before, such a cost functional in many cases serves the purpose of yielding unique solutions while there is some ambiguity in its exact formulation. Hence, in search for a formulation which best supports numerical realizations, it is often sufficient to employ norms which are *equivalent* to $\|\cdot\|_{\mathscr{Z}}$ and $\|\cdot\|_{\mathscr{U}}$. In view of the discussion in Sect. 2.4.2, we can work for the norms $\|\cdot\|_{\mathscr{Z}}$, $\|\cdot\|_{\mathscr{U}}$ only with the diagonal scaling matrices $\mathbf{D}^s$ induced by the regularity of $\mathscr{Z}, \mathscr{U}$, or we can in addition include the Riesz map $\mathbf{R}$ defined in (2.101). In the numerical studies in [11], a somewhat better quality of the solution is observed when $\mathbf{R}$ is included. In order to keep track of the appearance of the Riesz maps in the linear systems derived below, we choose here the latter variant.

Moreover, we expand the given observation function $y_* \in \mathscr{Z}$ as

$$y_* = \langle y_*, \tilde{\Psi}_{\mathscr{Z}} \rangle \Psi_{\mathscr{Z}} =: (\mathbf{D}_{\mathscr{Z}}^{-1}\mathbf{y}_*)^T \Psi_{\mathscr{Z}} = \mathbf{y}_*^T \Psi_H. \tag{2.200}$$

The way the vector $\mathbf{y}_*$ is defined here for notational convenience may by itself actually have infinite norm in ℓ_2. However, its occurrence will always include premultiplication by $\mathbf{D}_{\mathscr{Z}}^{-1}$ which is therefore always well-defined. In view of (2.110), we obtain the relations

$$\|y - y_*\|_{\mathscr{Z}} \sim \|\mathbf{R}^{1/2}\mathbf{D}_{\mathscr{Z}}^{-1}(\mathbf{y} - \mathbf{y}_*)\| \sim \|\mathbf{D}_{\mathscr{Z}}^{-1}(\mathbf{y} - \mathbf{y}_*\|. \tag{2.201}$$

Note that here $\mathbf{R} = \langle \Psi, \Psi \rangle$ (and not $\mathbf{R}^{-1}$) comes into play since y, y_* have been expanded in a scaled version of the primal wavelet basis Ψ. Hence, equivalent norms for $\| \cdot \|_{\mathscr{Z}}$ may involve $\mathbf{R}$. As for describing equivalent norms for $\| \cdot \|_{\mathscr{U}}$, recall that u is expanded in the basis Ψ_U for $U \subset H'$. Consequently, $\mathbf{R}^{-1}$ is the natural matrix to take into account when considering equivalent norms, i.e., we choose here

$$\|u\|_{\mathscr{U}} \sim \|\mathbf{R}^{-1/2}\mathbf{u}\|. \tag{2.202}$$

Finally, we formulate the following control problem in (infinite) wavelet coordinates.

(DCP) *For given data* $\mathbf{D}_{\mathscr{Z}}^{-1}\mathbf{y}_* \in \ell_2(\mathit{II}_{\mathscr{Z}})$, $\mathbf{f} \in \ell_2(\mathit{II}_H)$, *and weight parameter* $\omega > 0$, *minimize the quadratic functional*

$$\check{\mathbf{J}}(\mathbf{y}, \mathbf{u}) := \tfrac{1}{2} \|\mathbf{R}^{1/2}\mathbf{D}_{\mathscr{Z}}^{-1}(\mathbf{y} - \mathbf{y}_*)\|^2 + \tfrac{\omega}{2} \|\mathbf{R}^{-1/2}\mathbf{u}\|^2 \tag{2.203}$$

over $(\mathbf{y}, \mathbf{u}) \in \ell_2(\mathit{II}_H) \times \ell_2(\mathit{II}_H)$ *subject to the linear constraints*

$$\mathbf{A}\mathbf{y} = \mathbf{f} + \mathbf{D}_H^{-1}\mathbf{u}. \tag{2.204}$$

Remark 10 Problem (DCP) can be viewed as (discretized yet still infinite-dimensional) *representation* of the linear-quadratic control problem (2.74) together with (2.75) in wavelet coordinates in the following sense. The functional $\check{\mathbf{J}}(\mathbf{y}, \mathbf{u})$ defined in (2.203) is equivalent to the functional $J(y, u)$ from (2.74) in the sense that there exist constants $0 < c_J \le C_J < \infty$ such that

$$c_J\, \check{\mathbf{J}}(\mathbf{y}, \mathbf{u}) \ \le\ J(y, u) \ \le\ C_J\, \check{\mathbf{J}}(\mathbf{y}, \mathbf{u}) \tag{2.205}$$

holds for any $y = \mathbf{y}^T \Psi_H \in H$, given $y_* = (\mathbf{D}_{\mathscr{Z}}^{-1}\mathbf{y}_*)^T \Psi_{\mathscr{Z}} \in \mathscr{Z}$ and any $u = \mathbf{u}^T \Psi_{\mathscr{U}} \in \mathscr{U}$. Moreover, in the case of compatible data $y_* = A^{-1} f$ yielding $J(y, u) \equiv 0$, the respective minimizers coincide, and $\mathbf{y}_* = \mathbf{A}^{-1}\mathbf{f}$ yields $\check{\mathbf{J}}(\mathbf{y}, \mathbf{u}) \equiv \mathbf{0}$. In this sense the new functional (2.203) captures the essential features of the model minimization functional.

Once problem (DCP) is posed, we can apply variational principles to derive necessary and sufficient conditions for a unique solution. All control problems considered here are in fact simple in this regard, as we have to minimize a quadratic functional subject to linear constraints, for which the necessary conditions are also sufficient. In principle, there are two ways to derive the optimality conditions for (DCP). We have encountered in Sect. 2.3.5 already the technique via the Lagrangian.

We define for (DCP) the *Lagrangian* introducing the *Lagrange multiplier*, *adjoint variable* or *adjoint state* $\mathbf{p}$ as

$$\mathbf{Lagr}(\mathbf{y}, \mathbf{p}, \mathbf{u}) := \check{\mathbf{J}}(\mathbf{y}, \mathbf{u}) + \langle \mathbf{p}, \mathbf{A}\mathbf{y} - \mathbf{f} - \mathbf{D}_H^{-1}\mathbf{u} \rangle. \tag{2.206}$$

Then the KKT conditions $\delta\,\mathbf{Lagr}(\mathbf{w}) = \mathbf{0}$ for $\mathbf{w} = \mathbf{p}, \mathbf{y}, \mathbf{u}$ are, respectively,

$$\mathbf{A}\mathbf{y} = \quad \mathbf{f} + \mathbf{D}_H^{-1}\mathbf{u}, \tag{2.207a}$$

$$\mathbf{A}^T\mathbf{p} = -\mathbf{D}_{\mathscr{Z}}^{-1}\mathbf{R}\mathbf{D}_{\mathscr{Z}}^{-1}(\mathbf{y} - \mathbf{y}_*) \tag{2.207b}$$

$$\omega\mathbf{R}^{-1}\mathbf{u} = \quad \mathbf{D}_H^{-1}\mathbf{p}. \tag{2.207c}$$

The first system resulting from the variation with respect to the Lagrange multiplier always recovers the original constraints (2.204) and will be referred to as the *primal system* or the *state equation*. Accordingly, we call (2.207b) the *adjoint* or *dual system*, or the *costate equation*. The third Eq. (2.207c) is sometimes denoted as the *design equation*. Although $\mathbf{A}$ is symmetric, we continue to write $\mathbf{A}^T$ for the operator of the adjoint system to distinguish it from the primal system.

The coupled system (2.207) later is to be solved. However, in order to derive convergent iterations and deduce complexity estimates, a different formulation will be advantageous. It is based on the fact that $\mathbf{A}$ is according to Proposition 3 a boundedly invertible mapping on ℓ_2. Thus, we can formally invert (2.198) to obtain $\mathbf{y} = \mathbf{A}^{-1}\mathbf{f} + \mathbf{A}^{-1}\mathbf{D}_H^{-1}\mathbf{u}$. Substitution into (2.203) yields a functional depending only on $\mathbf{u}$,

$$\mathbf{J}(\mathbf{u}) := \tfrac{1}{2}\,\|\mathbf{R}^{1/2}\mathbf{D}_{\mathscr{Z}}^{-1}\left(\mathbf{A}^{-1}\mathbf{D}_H^{-1}\mathbf{u} - (\mathbf{y}_* - \mathbf{A}^{-1}\mathbf{f})\right)\|^2 + \tfrac{\omega}{2}\,\|\mathbf{R}^{-1/2}\mathbf{u}\|^2. \tag{2.208}$$

Employing the abbreviations

$$\mathbf{Z} := \quad \mathbf{R}^{1/2}\mathbf{D}_{\mathscr{Z}}^{-1}\mathbf{A}^{-1}\mathbf{D}_H^{-1}, \tag{2.209a}$$

$$\mathbf{G} := -\mathbf{R}^{1/2}\mathbf{D}_{\mathscr{Z}}^{-1}(\mathbf{A}^{-1}\mathbf{f} - \mathbf{y}_*), \tag{2.209b}$$

the functional simplifies to

$$\mathbf{J}(\mathbf{u}) = \tfrac{1}{2}\|\mathbf{Z}\mathbf{u} - \mathbf{G}\|^2 + \tfrac{\omega}{2}\,\|\mathbf{R}^{-1/2}\mathbf{u}\|^2. \tag{2.210}$$

Proposition 5 ([53]) *The functional* $\mathbf{J}$ *is twice differentiable with first and second variation*

$$\delta\mathbf{J}(\mathbf{u}) = (\mathbf{Z}^T\mathbf{Z} + \omega\mathbf{R}^{-1})\mathbf{u} - \mathbf{Z}^T\mathbf{G}, \qquad \delta^2\mathbf{J}(\mathbf{u}) = \mathbf{Z}^T\mathbf{Z} + \omega\mathbf{R}^{-1}. \tag{2.211}$$

In particular, $\mathbf{J}$ *is convex so that a unique minimizer exists.*

Setting

$$\mathbf{Q} := \mathbf{Z}^T\mathbf{Z} + \omega\mathbf{R}^{-1}, \qquad \mathbf{g} := \mathbf{Z}^T\mathbf{G}, \tag{2.212}$$

the unique minimizer $\mathbf{u}$ of (2.210) is given by solving

$$\delta \mathbf{J}(\mathbf{u}) = \mathbf{0} \tag{2.213}$$

or, equivalently, the system

$$\mathbf{Q}\mathbf{u} = \mathbf{g}. \tag{2.214}$$

By definition (2.212), $\mathbf{Q}$ is a symmetric positive definite (infinite) matrix. Hence, finite versions of (2.214) could be solved by gradient or conjugate gradient iterative schemes. As the convergence speed of any such iteration depends on the spectral condition number of $\mathbf{Q}$, it is important to note that the following result.

Proposition 6 *The (infinite) matrix* $\mathbf{Q}$ *is uniformly bounded on* ℓ_2*, i.e., there exist constants* $0 < c_{\mathbf{Q}} \le C_{\mathbf{Q}} < \infty$ *such that*

$$c_{\mathbf{Q}}\,\|\mathbf{v}\| \le \|\mathbf{Q}\mathbf{v}\| \le C_{\mathbf{Q}}\,\|\mathbf{v}\|, \qquad \mathbf{v} \in \ell_2. \tag{2.215}$$

The proof follows from (2.46) and (2.196) [29]. Of course, in order to make such iterative schemes for (2.214) practically feasible, the explicit inversion of $\mathbf{A}$ in the definition of $\mathbf{Q}$ has to be avoided and replaced by an iterative solver in turn. This is where the system (2.207) will come into play. In particular, the third equation (2.207c) has the following interpretation which will turn out to be very useful later.

Proposition 7 *If we solve for a given control vector* $\mathbf{u}$ *successively (2.204) for* $\mathbf{y}$ *and (2.207b) for* $\mathbf{p}$*, then the* residual *for (2.214) attains the form*

$$\mathbf{Q}\mathbf{u} - \mathbf{g} = \omega \mathbf{R}^{-1}\mathbf{u} - \mathbf{D}_U^{-1}\mathbf{p}. \tag{2.216}$$

Proof Solving consecutively (2.204) and (2.207b) and recalling the definitions of $\mathbf{Z}$, $\mathbf{g}$ (2.209a), (2.212) we obtain

$$\begin{aligned}
\mathbf{D}_H^{-1}\mathbf{p} &= -\mathbf{D}_H^{-1}(\mathbf{A}^{-T}\mathbf{D}_{\mathscr{Z}}^{-1}\mathbf{R}\mathbf{D}_{\mathscr{Z}}^{-1}(\mathbf{y} - \mathbf{y}_*)) \\
&= -\mathbf{Z}^T\mathbf{R}^{1/2}\mathbf{D}_{\mathscr{Z}}^{-1}(\mathbf{A}^{-1}\mathbf{f} + \mathbf{A}^{-1}\mathbf{D}_H^{-1}\mathbf{u} - \mathbf{y}_*) \\
&= \ \ \mathbf{Z}^T\mathbf{G} - \mathbf{Z}^T\mathbf{R}^{1/2}\mathbf{D}_{\mathscr{Z}}^{-1}\mathbf{A}^{-1}\mathbf{D}_H^{-1}\mathbf{u} \\
&= \ \ \mathbf{g} - \mathbf{Z}^T\mathbf{Z}\mathbf{u}.
\end{aligned}$$

Hence, the residual $\mathbf{Q}\mathbf{u} - \mathbf{g}$ attains the form

$$\mathbf{Q}\mathbf{u} - \mathbf{g} = (\mathbf{Z}^T\mathbf{Z} + \omega\mathbf{R}^{-1})\mathbf{u} - \mathbf{g} = \omega\mathbf{R}^{-1}\mathbf{u} - \mathbf{D}_H^{-1}\mathbf{p},$$

where we have used the definition of $\mathbf{Q}$ from (2.212). □

Having derived the optimality conditions (2.207), the next issue is their efficient numerical solution. In view of the fact that the system (2.207) still involves infinite matrices and vectors, this also raises the question how to derive computable finite versions. By now we have investigated two scenarios.

The first version with respect to *uniform discretizations* is based on choosing finite-dimensional subspaces of the function spaces under consideration. The second version which deals with *adaptive discretizations* is actually based on the infinite system (2.207). In both scenarios, a fully iterative numerical scheme for the solution of (2.207) is designed along the following lines. The basic iteration scheme is a *gradient* or *conjugate gradient iteration* for (2.214) as an *outer iteration* where each application of $\mathbf{Q}$ is in turn realized by solving the primal and the dual system (2.204) and (2.207b) also by a gradient or conjugate gradient method as *inner iterations*.

For *uniform* discretizations for which we wanted to test numerically the role of equivalent norms and the influence of Riesz maps in the cost functional (2.203), we have used in [12] as central iterative scheme the conjugate gradient (CG) method. Since the interior systems are only solved up to discretization error accuracy, the whole procedure may therefore be viewed as an *inexact conjugate gradient (CG) method*. We stress already at this point that the iteration numbers of such a method do *not* depend on the discretization level as finite versions of all involved operators are also uniformly well-conditioned in the sense of (2.215). In each step of the outer iteration, the error will be reduced by a fixed factor ρ. Combined with a *nested iteration strategy*, it will be shown that this yields an asymptotically optimal method in the amount of arithmetic operations.

Starting from the infinite coupled system (2.207), we have investigated in [29] *adaptive schemes* which, given any prescribed accuracy $\varepsilon > 0$, solve (2.207) such that the error for $\mathbf{y}$, $\mathbf{u}$, $\mathbf{p}$ is controlled by ε. Here we have used a *gradient scheme* as basic iterative scheme since it somehow simplifies the analysis, see Sect. 2.6.2.

2.5.4 Control Problems: Dirichlet Boundary Control

Having derived a representation in wavelet coordinates for both the saddle point problem from Sect. 2.3.3 and the PDE-constrained control problem in the previous section, it is straightforward to find also an appropriate representation of the control problem with Dirichlet boundary control introduced in Sect. 2.3.6. In order not to be overburdened with notation, we specifically choose the control space on the boundary as $\mathscr{U} := Q(= (H^{1/2}(\Gamma))')$. For the more general situation covered by (2.84), a diagonal matrix with nondecreasing entries like in (2.195) would come into play to switch between $\mathscr{U}$ and Q. Thus, the exact wavelet representation of the constraints (2.83) is given by the system (2.190), where we exchange the given Dirichlet boundary term $\mathbf{g}$ by $\mathbf{u}$ in the present situation to express the dependence on the control in the right hand side, i.e.,

$$\mathbf{L}\begin{pmatrix}\mathbf{y}\\ \mathbf{p}\end{pmatrix} := \begin{pmatrix}\mathbf{A} & \mathbf{B}^T\\ \mathbf{B} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{y}\\ \mathbf{p}\end{pmatrix} = \begin{pmatrix}\mathbf{f}\\ \mathbf{u}\end{pmatrix}. \tag{2.217}$$

The derivation of a representer of the initial objective functional (2.82) is under the embedding condition (2.84) $\|v\|_{\mathscr{Z}} \lesssim \|v\|_Y$ for $v \in Y$ now the same as in the previous section, where all reference to the space H is to be exchanged by reference to Y. We end up with the following minimization problem in wavelet coordinates for the case of Dirichlet boundary control. **(DCP)** *For given data* $\mathbf{D}_{\mathscr{Z}}^{-1}\mathbf{y}_* \in \ell_2(I\!I_{\mathscr{Z}})$, $\mathbf{f} \in \ell_2(I\!I_Y)$*, and weight parameter* $\omega > 0$*, minimize the quadratic functional*

$$\check{\mathbf{J}}(\mathbf{y},\mathbf{u}) := \tfrac{1}{2}\,\|\mathbf{R}^{1/2}\mathbf{D}_{\mathscr{Z}}^{-1}(\mathbf{y}-\mathbf{y}_*)\|^2 + \tfrac{\omega}{2}\,\|\mathbf{R}^{-1/2}\mathbf{u}\|^2 \tag{2.218}$$

over $(\mathbf{y},\mathbf{u}) \in \ell_2(I\!I_Y) \times \ell_2(I\!I_Y)$ *subject to the linear constraints (2.217),*

$$\mathbf{L}\begin{pmatrix}\mathbf{y}\\ \mathbf{p}\end{pmatrix} = \begin{pmatrix}\mathbf{f}\\ \mathbf{u}\end{pmatrix}.$$

The corresponding Karush-Kuhn-Tucker conditions can be derived by the same variational principles as in the previous section by defining a Lagrangian in terms of the functional $\check{\mathbf{J}}(\mathbf{y},\mathbf{u})$ and appending the constraints (2.198) with the help of additional Lagrange multipliers $(\mathbf{z},\boldsymbol{\mu})^T$, see [53]. We obtain in this case a system of coupled saddle point problems

$$\mathbf{L}\begin{pmatrix}\mathbf{y}\\ \mathbf{p}\end{pmatrix} = \begin{pmatrix}\mathbf{f}\\ \mathbf{u}\end{pmatrix} \tag{2.219a}$$

$$\mathbf{L}^T\begin{pmatrix}\mathbf{z}\\ \boldsymbol{\mu}\end{pmatrix} = \begin{pmatrix}-\omega\mathbf{D}_{\mathscr{Z}}^{-1}\mathbf{R}\mathbf{D}_{\mathscr{Z}}^{-1}(\mathbf{y}-\mathbf{y}_*)\\ \mathbf{0}\end{pmatrix} \tag{2.219b}$$

$$\mathbf{u} = \boldsymbol{\mu}. \tag{2.219c}$$

Again, the first system appearing here, the *primal system*, are just the constraints (2.198) while (2.95) will be referred to as the *dual* or *adjoint system*. The specific form of the right hand side of the dual system emerges from the particular formulation of the minimization functional (2.218). The (here trivial) equation (2.219c) stems from measuring $\mathbf{u}$ just in ℓ_2, representing measuring the control in its natural trace norm. Instead of replacing $\boldsymbol{\mu}$ by $\mathbf{u}$ in (2.95) and trying to solve the resulting equations, (2.219c) will be essential to devise an inexact gradient scheme. In fact, since $\mathbf{L}$ in (2.198) is an invertible operator, we can rewrite $\check{\mathbf{J}}(\mathbf{y},\mathbf{u})$ by formally inverting (2.198) as a functional of $\mathbf{u}$, that is, $\mathbf{J}(\mathbf{u}) := \check{\mathbf{J}}(\mathbf{y}(\mathbf{u}),\mathbf{u})$ as above. The following result will be very useful for the design of the outer–inner iterative solvers

Proposition 8 *The first variation of* $\mathbf{J}$ *satisfies*

$$\delta \mathbf{J}(\mathbf{u}) \; = \; \mathbf{u} - \boldsymbol{\mu}, \tag{2.220}$$

where $(\mathbf{u}, \boldsymbol{\mu})$ *are part of the solution of (2.219). Moreover,* $\mathbf{J}$ *is convex so that a unique minimizer exists.*

Hence, Eq. (2.219c) is just $\delta \mathbf{J}(\mathbf{u}) \; = \; \mathbf{0}$. For a unified treatment below of both control problems considered in these notes, it will be useful to rewrite (2.219c) like in (2.214) as a condensed equation for the control $\mathbf{u}$ alone. We formally invert (2.217) and (2.219b) and obtain

$$\mathbf{Q}\mathbf{u} = \mathbf{g} \tag{2.221}$$

with the abbreviations

$$\mathbf{Q} := \mathbf{Z}^T\mathbf{Z} + \omega \mathbf{I}, \quad \mathbf{g} := \mathbf{Z}^T(\mathbf{y}_* - \mathbf{T}_\square \mathbf{L}^{-1} \mathbf{I}_\square \mathbf{f}) \tag{2.222}$$

and

$$\mathbf{Z} := \mathbf{T}_\square \mathbf{L}^{-1} \mathbf{I}_\square, \qquad \mathbf{I}_\square := \begin{pmatrix} \mathbf{0} \\ \mathbf{I} \end{pmatrix}, \qquad \mathbf{T}_\square := (\mathbf{T} \;\; \mathbf{0}). \tag{2.223}$$

Proposition 9 *The vector* $\mathbf{u}$ *as part of the solution vector* $(\mathbf{y}, \mathbf{p}, \mathbf{z}, \boldsymbol{\mu}, \mathbf{u})$ *of (2.219) coincides with the unique solution* $\mathbf{u}$ *of the condensed equations (2.221).*

2.6 Iterative Solution

Each of the four problem classes discussed above lead to the problem to finally solve a system

$$\delta \mathbf{J}(\mathbf{q}) \; = \; \mathbf{0} \tag{2.224}$$

or, equivalently, a linear system

$$\mathbf{M}\mathbf{q} = \mathbf{b}, \tag{2.225}$$

where $\mathbf{M} : \ell_2 \to \ell_2$ is a (possibly infinite) symmetric positive definite matrix satisfying

$$c_{\mathbf{M}} \|\mathbf{v}\| \le \|\mathbf{M}\mathbf{v}\| \le C_{\mathbf{M}} \|\mathbf{v}\|, \quad \mathbf{v} \in \ell_2, \tag{2.226}$$

for some constants $0 < c_{\mathbf{M}} \leq C_{\mathbf{M}} < \infty$ and where $\mathbf{b} \in \ell_2$ is some given right hand side.

A simple *gradient method* for solving (2.224) is

$$\mathbf{q}_{k+1} := \mathbf{q}_k - \alpha\, \delta\mathbf{J}(\mathbf{q}_k), \qquad k = 0, 1, 2, \ldots \tag{2.227}$$

with some initial guess $\mathbf{q}_0$. In all of the previously considered situations, it has been asserted that there exists a fixed parameter α, depending on bounds for the second variation of $\mathbf{J}$, such that (2.227) converges and reduces the error in each step by at least a fixed factor $\rho < 1$, i.e.,

$$\|\mathbf{q} - \mathbf{q}_{k+1}\| \leq \rho\|\mathbf{q} - \mathbf{q}_k\|, \quad k = 0, 1, 2, \ldots, \tag{2.228}$$

where ρ is determined by

$$\rho := \|\mathbf{I} - \alpha\mathbf{M}\| < 1.$$

Hence, the scheme (2.227) is a convergent iteration for the possibly infinite system (2.225). Next we will need to discuss how to reduce the infinite systems to computable finite versions.

2.6.1 Finite Systems on Uniform Grids

Let us first consider finite-dimensional trial spaces with respect to uniform discretizations. For each of the Hilbert spaces H, this means in the wavelet setting to pick the index set of all indices up to some *highest refinement level* J, i.e.,

$$I\!I_{J,H} := \{\lambda \in I\!I_H : |\lambda| \leq J\} \subset I\!I_H$$

satisfying $N_{J,H} := \# I\!I_{J,H} < \infty$. The representation of operators is then built as in Sect. 2.4.3 with respect to this truncated index set which corresponds to deleting all rows and columns that refer to indices λ such that $|\lambda| > J$, and correspondingly for functions. There is by construction also a *coarsest level* of resolution denoted by j_0.

Computationally the representation of operators according to (2.111) is in the case of uniform grids always realized as follows. First, the operator is set up in terms of the *generator basis* on the finest level J. This generator basis simply consists of tensor products of B-Splines, or linear combinations of these near the boundaries. The representation of an operator in the *wavelet basis* is then achieved by applying the Fast Wavelet Transform (FWT) which needs $\mathscr{O}(N_{J,H})$ arithmetic operations and is therefore asymptotically optimal, see, e.g., [24, 34, 51] and Sect. 2.4.4.

In order not to overburden the notation, let in this subsection the resulting system for $N = N_{J,H}$ unknowns again be denoted by

$$\mathbf{M}\mathbf{q} = \mathbf{b}, \tag{2.229}$$

where now $\mathbf{M} : \mathbb{R}^N \to \mathbb{R}^N$ is a symmetric positive definite matrix satisfying (2.226) on $\mathbb{R}^N$. It will be convenient to abbreviate the residual using an approximation $\tilde{\mathbf{q}}$ to $\mathbf{q}$ for (2.229) as

$$\textsc{Resd}(\tilde{\mathbf{q}}) := \mathbf{M}\tilde{\mathbf{q}} - \mathbf{b}. \tag{2.230}$$

We will employ a basic conjugate gradient method that iteratively computes an approximate solution $\mathbf{q}_K$ to (2.229) with given initial vector $\mathbf{q}_0$ and given tolerance $\varepsilon > 0$ such that

$$\|\mathbf{M}\mathbf{q}_K - \mathbf{b}\| = \|\textsc{Resd}(\mathbf{q}_K)\| \leq \varepsilon, \tag{2.231}$$

where K denotes the number of iterations used. Later we specify ε depending on the discretization for which (2.229) is set up. The following scheme CG contains a routine $\textsc{App}(\eta_k, \mathbf{M}, \mathbf{d}_k)$ which in view of the problem classes discussed above is to have the property that it approximately computes the product $\mathbf{M}\mathbf{d}_k$ up to a tolerance $\eta_k = \eta_k(\varepsilon)$ depending on ε, i.e., the output $\mathbf{m}_k$ of $\textsc{App}(\eta_k, \mathbf{M}, \mathbf{d}_k)$ satisfies

$$\|\mathbf{m}_k - \mathbf{M}\mathbf{d}_k\| \leq \eta_k. \tag{2.232}$$

For the cases where $\mathbf{M} = \mathbf{A}$, this is simply the matrix-vector multiplication $\mathbf{M}\mathbf{d}_k$. For the situations where $\mathbf{M}$ may involve the solution of an additional system, this multiplication will be only approximative. The routine is as follows.

```
CG [$\varepsilon, \mathbf{q}_0, \mathbf{M}, \mathbf{b}$] → $\mathbf{q}_K$
(I) SET $\mathbf{d}_0 := \mathbf{b} - \mathbf{M}\mathbf{q}_0$ AND $\mathbf{r}_0 := -\mathbf{d}_0$. LET $k = 0$.
(II) WHILE $\|\mathbf{r}_k\| > \varepsilon$
        $\mathbf{m}_k := \textsc{App}(\eta_k(\varepsilon), \mathbf{M}, \mathbf{d}_k)$
        $\alpha_k := \frac{(\mathbf{r}_k)^T \mathbf{r}_k}{(\mathbf{d}_k)^T \mathbf{m}_k}$                $\mathbf{q}_{k+1} := \mathbf{q}_k + \alpha_k \mathbf{d}_k$
        $\mathbf{r}_{k+1} := \mathbf{r}_k + \alpha_k \mathbf{m}_k$                $\beta_k := \frac{(\mathbf{r}_{k+1})^T \mathbf{r}_{k+1}}{(\mathbf{r}_k)^T \mathbf{r}_k}$        (2.233)
        $\mathbf{d}_{k+1} := -\mathbf{r}_{k+1} + \beta_k \mathbf{d}_k$
        $k := k + 1$
(III) SET $K := k - 1$.
```

Let us briefly discuss in the case $\mathbf{M} = \mathbf{A}$ that the final iterate $\mathbf{q}_K$ indeed satisfies (2.231). From the newly computed iterate $\mathbf{q}_{k+1} = \mathbf{q}_k + \alpha_k \mathbf{d}_k$ it follows by applying $\mathbf{M}$ on both sides that $\mathbf{M}\mathbf{q}_{k+1} - \mathbf{b} = \mathbf{M}\mathbf{q}_k - \mathbf{b} + \alpha_k \mathbf{M}\mathbf{d}_k$ which is the same as $\textsc{Resd}(\mathbf{q}_{k+1}) = \textsc{Resd}(\mathbf{q}_k) + \alpha_k \mathbf{M}\mathbf{d}_k$. By the initialization for $\mathbf{r}_k$ used above, this in turn is the updating term for $\mathbf{r}_k$, hence, $\mathbf{r}_k = \textsc{Resd}(\mathbf{q}_k)$. After the stopping criterion based on $\mathbf{r}_k$ is met, the final iterate $\mathbf{q}_K$ observes (2.231).

The routine CG computes the *residual* up to the stopping criterion ε. From the residual, we can in view of (2.226) estimate the *error* in the solution as

$$\|\mathbf{q} - \mathbf{q}_K\| = \|\mathbf{M}^{-1}(\mathbf{b} - \mathbf{M}\mathbf{q}_K)\| \le \|\mathbf{M}^{-1}\| \, \|\textsc{Resd}(\mathbf{q}_K)\| \le \frac{\varepsilon}{c_{\mathbf{M}}}, \tag{2.234}$$

that is, it may deviate from the norm of the residual from a factor proportional to the smallest eigenvalue of $\mathbf{M}$.

Distributed Control Let us now apply the solution scheme to the situation from Sect. 2.5.3 where $\mathbf{Q}$ now involves the inversion of finite-dimensional systems (2.207a) and (2.207b). The material in the remainder of this subsection is essentially contained in [12].

We begin with a specification of the approximate computation of the right hand side $\mathbf{b}$ which also contains applications of $\mathbf{A}^{-1}$.

RHS $[\zeta, \mathbf{A}, \mathbf{f}, \mathbf{y}_*] \to \mathbf{b}_\zeta$

(I) CG $[\frac{c_{\mathbf{A}}}{2C} \frac{c_{\mathbf{A}}}{C^2 C_0^2} \zeta, \mathbf{0}, \mathbf{A}, \mathbf{f}] \to \mathbf{b}_1$

(II) CG $[\frac{c_{\mathbf{A}}}{2C} \zeta, \mathbf{0}, \mathbf{A}^T, -\mathbf{D}_{\mathscr{Z}}^{-1} \mathbf{R} \mathbf{D}_{\mathscr{Z}}^{-1} (\mathbf{b}_1 - \mathbf{y}_*)] \to \mathbf{b}_2$

(III) $\mathbf{b}_\zeta := \mathbf{D}_H^{-1} \mathbf{b}_2$.

The tolerances used within the two conjugate gradient methods depend on the constants $c_{\mathbf{A}}, C, C_0$ from (2.46), (2.196) and (2.104), respectively. Since the additional factor $c_{\mathbf{A}}(CC_0)^{-2}$ in the stopping criterion in step (I) in comparison to step (II) is in general smaller than one, this means that the primal system needs to be solved more accurately than the adjoint system in step (II).

Proposition 10 *The result* $\mathbf{b}_\zeta$ *of* RHS $[\zeta, \mathbf{A}, \mathbf{f}, \mathbf{y}_*]$ *satisfies*

$$\|\mathbf{b}_\zeta - \mathbf{b}\| \le \zeta. \tag{2.235}$$

Proof Recalling the definition (2.212) of $\mathbf{b}$, step (III) and step (II) yield

$$\begin{aligned}
\|\mathbf{b}_\zeta - \mathbf{b}\| &\le \|\mathbf{D}_H^{-1}\| \, \|\mathbf{b}_2 - \mathbf{D}_H \mathbf{b}\| \\
&\le C \|\mathbf{A}^{-T}\| \, \|\mathbf{A}^T \mathbf{b}_2 - \mathbf{D}_{\mathscr{Z}}^{-1} \mathbf{R} \mathbf{D}_{\mathscr{Z}}^{-1} (\mathbf{A}^{-1}\mathbf{f} - \mathbf{b}_1 + \mathbf{b}_1 - \mathbf{y}_*)\| \\
&\le \frac{C}{c_{\mathbf{A}}} \left(\frac{c_{\mathbf{A}}}{2C} \zeta + \|\mathbf{D}_{\mathscr{Z}}^{-1} \mathbf{R} \mathbf{D}_{\mathscr{Z}}^{-1} (\mathbf{A}^{-1}\mathbf{f} - \mathbf{b}_1)\| \right).
\end{aligned} \tag{2.236}$$

Employing the upper bounds for $\mathbf{D}_{\mathscr{Z}}^{-1}$ and $\mathbf{R}$, we arrive at

$$\begin{aligned}\|\mathbf{b}_\zeta - \mathbf{b}\| &\leq \frac{C}{c_{\mathbf{A}}}\Big(\frac{c_{\mathbf{A}}}{2C}\zeta + C^2C_0^2\,\|\mathbf{A}^{-1}\|\,\|\mathbf{f} - \mathbf{A}\mathbf{b}_1\|\Big)\\ &\leq \frac{C}{c_{\mathbf{A}}}\left(\frac{c_{\mathbf{A}}}{2C}\zeta + \frac{C^2C_0^2}{c_{\mathbf{A}}}\frac{c_{\mathbf{A}}}{2C}\frac{c_{\mathbf{A}}}{C^2C_0^2}\zeta\right) \;= \zeta.\end{aligned} \tag{2.237}$$

□

Accordingly, an approximation $\mathbf{m}_\eta$ to the matrix-vector product $\mathbf{Qd}$ is the output of the following routine App.
App $[\eta, \mathbf{Q}, \mathbf{d}] \to \mathbf{m}_\eta$

(I) CG $[\frac{c_{\mathbf{A}}}{3C}\frac{c_{\mathbf{A}}}{C^2C_0^2}\eta, \mathbf{0}, \mathbf{A}, \mathbf{f} + \mathbf{D}_H^{-1}\mathbf{d}] \to \mathbf{y}_\eta$
(II) CG $[\frac{c_{\mathbf{A}}}{3C}\eta, \mathbf{0}, \mathbf{A}^T, -\mathbf{D}_{\mathscr{Z}}^{-1}\mathbf{R}\mathbf{D}_Z^{-1}(\mathbf{y}_\eta - \mathbf{y}_*)] \to \mathbf{p}_\eta$
(III) $\mathbf{m}_\eta := \mathbf{g}_{\eta/3} + \omega\mathbf{R}^{-1}\mathbf{d} - \mathbf{D}_H^{-1}\mathbf{p}_\eta$.

The choice of the tolerances for the interior application of CG in steps (i) and (ii) will become clear from the following result.

Proposition 11 *The result* $\mathbf{m}_\eta$ *of* App$[\eta, \mathbf{Q}, \mathbf{d}]$ *satisfies*

$$\|\mathbf{m}_\eta - \mathbf{Qd}\| \leq \eta. \tag{2.238}$$

Proof Denote by $\mathbf{y}_\mathbf{d}$ the exact solution of (2.207a) with $\mathbf{d}$ in place of $\mathbf{u}$ on the right hand side, and by $\mathbf{p}_\mathbf{d}$ the exact solution of (2.207b) with $\mathbf{y}_\mathbf{d}$ on the right hand side. Then we deduce from step (iii) and (2.216) combined with (2.104) and (2.196)

$$\begin{aligned}\|\mathbf{m}_\eta - \mathbf{Qd}\| &= \|\mathbf{g}_{\eta/3} - \mathbf{g} + \omega\mathbf{R}^{-1}\mathbf{d} - \mathbf{D}_U^{-1}\mathbf{p}_\eta \;- (\mathbf{Qd} - \mathbf{g})\|\\ &\leq \frac{1}{3}\eta + \|\omega\mathbf{R}^{-1}\mathbf{d} - \mathbf{D}_U^{-1}\mathbf{p}_\eta - (\omega\mathbf{R}^{-1}\mathbf{d} - \mathbf{D}_U^{-1}\mathbf{p}_\mathbf{d})\|\\ &\leq \frac{1}{3}\eta + C\|\mathbf{p}_\mathbf{d} - \mathbf{p}_\eta\|.\end{aligned} \tag{2.239}$$

Denote by $\hat{\mathbf{p}}$ the exact solution of (2.207b) with $\mathbf{y}_\eta$ on the right hand side. Then we have $\mathbf{p}_\mathbf{d} - \hat{\mathbf{p}} = -\mathbf{A}^{-T}\mathbf{D}_Z^{-1}\mathbf{R}\mathbf{D}_Z^{-1}(\mathbf{y}_\mathbf{d} - \mathbf{y}_\eta)$. It follows by (2.46), (2.104) and (2.196) that

$$\|\mathbf{p}_\mathbf{d} - \hat{\mathbf{p}}\| \leq \frac{C^2C_0^2}{c_{\mathbf{A}}}\|\mathbf{y}_\mathbf{d} - \mathbf{y}_\eta\| \leq \frac{1}{3C}\eta, \tag{2.240}$$

where the last estimate follows by the choice of the threshold in step (i). Finally, the combination (2.239) and (2.240) together with (2.235) and the stopping criterion in step (ii) readily confirms that

$$\begin{aligned}\|\mathbf{m}_\eta - \mathbf{Qd}\| &\le \frac{1}{3}\eta + C\left(\|\mathbf{p_d} - \hat{\mathbf{p}}\| + \|\hat{\mathbf{p}} - \mathbf{p}_\eta\|\right)\\ &\le \frac{1}{3}\eta + C\left(\frac{1}{3C}\eta + \frac{1}{3C}\eta\right) = \eta. \qquad \square\end{aligned}$$

The effect of perturbed applications of $\mathbf{M}$ in CG and more general Krylov subspace schemes with respect to convergence has been investigated in a numerical linear algebra context for a given linear system (2.229) in several papers, see, e.g., [70]. Here we have chosen the η_i to be proportional to the outer accuracy ε incorporating a safety factor accounting for the values of β_i and $\|\mathbf{r}_i\|$.

Finally, we can formulate a full nested iteration strategy for finite systems (2.207) on uniform grids which employs outer and inner CG routines as follows. The scheme starts at the coarsest level of resolution j_0 with some initial guess $\mathbf{u}_0^{j_0}$ and successively solves (2.214) with respect to each level j until the norm of the current residual is below the discretization error on that level.

In wavelet coordinates, $\|\cdot\|$ corresponds to the energy norm. If we employ as in [12] on the primal side for approximation linear combinations of B-splines of order d (degree $d-1$, see Sect. 2.2.1), the discretization error is for smooth solutions expected to be proportional to $2^{-(d-1)j}$ (compare (2.15)). Then the refinement level is successively increased until on the finest level J a prescribed tolerance proportional to the discretization error $2^{-(d-1)J}$ is met. In the following, superscripts on vectors denote the refinement level on which this term is computed. The given data $\mathbf{y}_*^j$, $\mathbf{f}^j$ are supposed to be accessible on all levels. On the coarsest level, the solution of (2.214) is computed exactly up to double precision by QR decomposition. Subsequently, the results from level j are prolongated onto the next higher level $j+1$. Using wavelets, this is accomplished by simply adding zeros: wavelet coordinates have the character of differences, this prolongation corresponds to the exact representation in higher resolution wavelet coordinates. The resulting *Nested-Iteration-Incomplete-Conjugate-Gradient* Algorithm is the following.

```
NEICG [$J$] → $\mathbf{u}^J$

 (I) INITIALIZATION FOR COARSEST LEVEL $j := j_0$
     (1) COMPUTE RIGHT HAND SIDE $\mathbf{g}^{j_0} = (\mathbf{Z}^T\mathbf{G})^{j_0}$ BY QR DECOMPOSITION
         USING (2.209).
     (2) COMPUTE SOLUTION $\mathbf{u}^{j_0}$ OF (2.214) BY QR DECOMPOSITION.
(II) WHILE $j < J$
     (1) PROLONGATE $\mathbf{u}^j \to \mathbf{u}_0^{j+1}$ BY ADDING ZEROS, SET $j := j+1$.
     (2) COMPUTE RIGHT HAND SIDE USING RHS $[2^{-(d-1)j}, \mathbf{A}, \mathbf{f}^j, \mathbf{y}_*^j] \to \mathbf{g}^j$.
     (3) COMPUTE SOLUTION OF (2.214) USING CG $[2^{-(d-1)j}, \mathbf{u}_0^j, \mathbf{Q}, \mathbf{g}^j] \to \mathbf{u}^j$.
```

Recall that step (II.3) requires multiple calls of APP$[\eta, \mathbf{Q}, \mathbf{d}]$, which in turn invokes both CG$[\ldots, \mathbf{A}, \ldots]$ as well as CG$[\ldots, \mathbf{A}^T, \ldots]$ in each application.

On account of (2.46) and (2.215), finite versions of the system matrices $\mathbf{A}$ and $\mathbf{Q}$ have uniformly bounded condition numbers, entailing that each CG routine employed in the process reduces the error by a fixed rate $\rho < 1$ in each iteration step. Let $N_J \sim 2^{nJ}$ be the total number of unknowns (for $\mathbf{y}^J$, $\mathbf{u}^J$ and $\mathbf{p}^J$) on the highest level J. Employing the CG method only on the highest level, one needs $\mathcal{O}(J) = \mathcal{O}(\log \varepsilon)$ iterations to achieve the prescribed discretization error accuracy $\varepsilon_J = 2^{-(d-1)J}$. As each application of $\mathbf{A}$ and $\mathbf{Q}$ requires $\mathcal{O}(N_J)$ operations, the solution of (2.214) by CG only on the finest level requires $\mathcal{O}(J\,N_J)$ arithmetic operations.

Theorem 7 ([12]) *If the residual (2.216) is computed up to discretization error proportional to* $2^{-(d-1)j}$ *on each level* j *and the corresponding solutions are taken as initial guesses for the next higher level,* NEICG *is an asymptotically optimal method in the sense that it provides the solution* $\mathbf{u}^J$ *up to discretization error on level* J *in an overall amount of* $\mathcal{O}(N_J)$ *arithmetic operations.*

Proof In the above notation, nested iteration allows one to get rid of the factor J in the total amount of operations. Starting with the exact solution on the coarsest level j_0, in view of the uniformly bounded condition numbers of $\mathbf{A}$ and $\mathbf{Q}$, one needs only a fixed amount of iterations to reduce the error up to discretization error accuracy $\varepsilon_j = 2^{-(d-1)j}$ on each subsequent level j, taking the solution from the previous level as initial guess. Thus, on each level, one needs $\mathcal{O}(N_j)$ operations to realize discretization error accuracy. Since the spaces are nested and the number of unknowns on each level grows like $N_j \sim 2^{nj}$, by a geometric series argument the total number of arithmetic operations stays proportional to $\mathcal{O}(N_J)$. □

Numerical Examples As an illustration of the ingredients for a distributed control problem, we consider the following example taken from [12] with the Helmholtz operator in (2.39) ($\mathbf{a} = I$, $c = 1$) and homogeneous Dirichlet boundary condition. A non-constant right hand side $f(x) := 1 + 2.3\exp(-15|x - \frac{1}{2}|)$ is chosen, and the target state is set to a constant $y_* \equiv 1$. We first investigate the role the different norms $\|\cdot\|_{\mathcal{Z}}$ and $\|\cdot\|_{\mathcal{U}}$ in (2.74), encoded in the diagonal matrices $\mathbf{D}_{\mathcal{Z}}, \mathbf{D}_H$ from (2.195), have on the solution. We see in Fig. 2.3 for the choice $\mathcal{U} = L_2$ and $\mathcal{Z} = H^s(0, 1)$ for different values of s varying between 0 and 1 the solution y (left) and the corresponding control u (right) for fixed weight $\omega = 1$. As s is increased, a stronger tendency of y towards the prescribed state $y_* \equiv 1$ can be observed which is, however, deterred from reaching this state by the homogeneous boundary conditions. Extensive studies of this type can be found in [11, 12].

As an example displaying the performance of the proposed fully iterative scheme NEICG in two spatial dimensions, Table 2.6 from [12] is included. This is an example of a control problem for the Helmholtz operator with Neumann boundary conditions. The stopping criterion for the outer iteration (relative to $\|\cdot\|$ which corresponds to the energy norm) on level j is chosen to be proportional to 2^{-j}. The second column displays the final value of the residual of the outer CG scheme

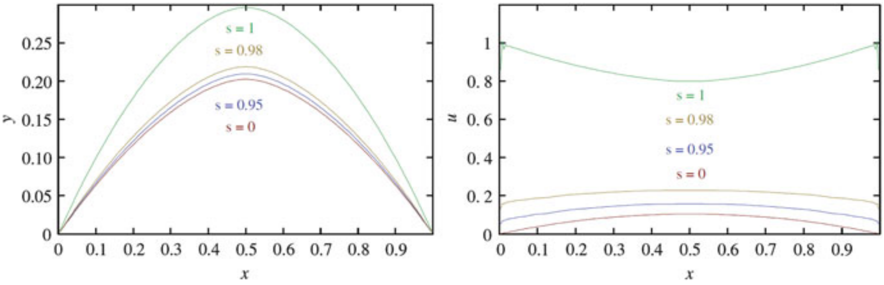


Fig. 2.3 Distributed control problem for elliptic problem with Dirichlet boundary conditions, a peak as right hand side f, $y_* \equiv 1$, $\omega = 0$, $\mathscr{U} = L_2$ and varying $\mathscr{X} = H^s(0, 1)$

Table 2.6 Iteration history for a two-dimensional distributed control problem with Neumann boundary conditions, $\omega = 1$, $\mathscr{X} = H^1(\Omega)$, $\mathscr{U} = (H^{0.5}(\Omega))'$

j	$\|\mathbf{r}_K^j\|$	#O	#E	#A	#R	$\|R(\mathbf{y}^J)-\mathbf{y}^j\|$	$\|\mathbf{y}^J - P(\mathbf{y}^j)\|$	$\|R(\mathbf{u}^J)-\mathbf{u}^j\|$	$\|\mathbf{u}^J - P(\mathbf{u}^j)\|$
3						6.86e−03	1.48e−02	1.27e−04	4.38e−04
4	1.79e−05	5	12	5	8	2.29e−03	7.84e−03	4.77e−05	3.55e−04
5	1.98e−05	5	14	6	9	6.59e−04	3.94e−03	1.03e−05	2.68e−04
6	4.92e−06	7	13	5	9	1.74e−04	1.96e−03	2.86e−06	1.94e−04
7	3.35e−06	7	12	5	9	4.55e−05	9.73e−04	9.65e−07	1.35e−04
8	2.42e−06	7	11	5	10	1.25e−05	4.74e−04	7.59e−07	8.88e−05
9	1.20e−06	8	11	5	10	4.55e−06	2.12e−04	4.33e−07	5.14e−05
10	4.68e−07	9	10	5	9	3.02e−06	3.02e−06	2.91e−07	2.91e−07

on this level, i.e., $\|\mathbf{r}_K^j\| = \|\mathrm{R}\textsc{esd}(\mathbf{u}_K^j)\|$. The next three columns show the number of outer CG iterations (#O) for $\mathbf{Q}$ according to the APP scheme followed by the maximum number of inner iterations for the primal system (#E), the adjoint system (#A) and the design equation (#R). We see very well the effect of the uniformly bounded condition numbers of the involved operators. The last columns display different versions of the actual error in the state $\mathbf{y}$ and the control $\mathbf{u}$ when compared to the fine grid solution (R denotes restriction of the fine grid solution to the actual grid, and P prolongation). Here we can see the effect of the constants appearing in (2.234), that is, the error is very well controlled via the residual. More results for up to three spatial dimensions can be found in [11, 12].

Dirichlet Boundary Control For the system of saddle point problems (2.219) arising from the control problem with Dirichlet boundary control in Sect. 2.3.6, also a fully iterative algorithm NEICG can be designed along the above lines. Again the design equation (2.219c) for $\mathbf{u}$ serves as the equation for which a basic iterative scheme (2.227) can be posed. Of course, the CG method for $\mathbf{A}$ then has to be replaced by a convergent iterative scheme for saddle point operators $\mathbf{L}$ like Uzawa's algorithm. Also the discretization has to be chosen such that the LBB condition is satisfied, see Sect. 2.5.2. Details can be found in [53]. Alternatively, since $\mathbf{L}$ has a uniformly bounded condition number, the CG scheme can, in principle, also be

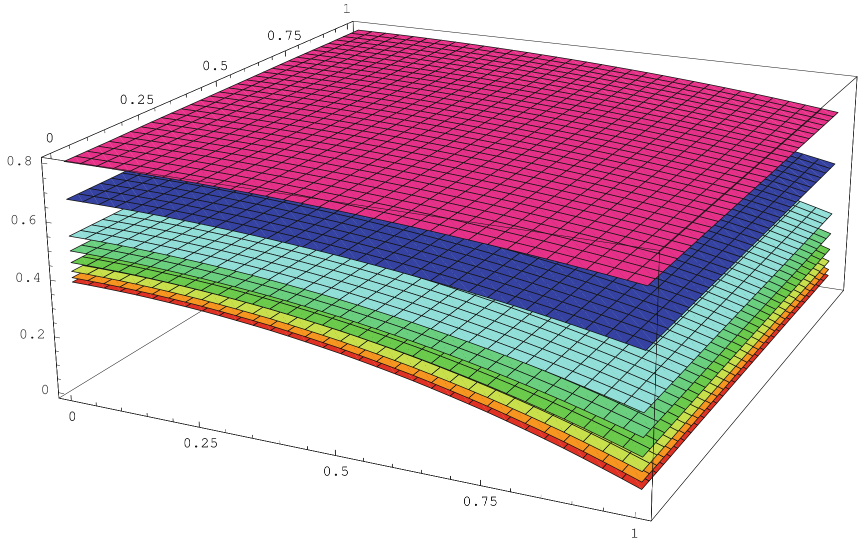


Fig. 2.4 State y of the Dirichlet boundary control problem using the objective functional $J(y,u) = \frac{1}{2}\|y - y_*\|^2_{H^s(\Gamma_y)} + \frac{1}{2}\|u\|^2_{H^{1/2}(\Gamma)}$ for $s = 0.1, 0.2, 0.3, 0.4, 0.5, 0.7, 0.9$ (from bottom to top) on resolution level $J = 5$

applied to $\mathbf{L}^T\mathbf{L}$. The performance of wavelet schemes on uniform grids for such systems of saddle point problems arising from optimal control is currently under investigation [62].

Numerical Example For illustration of the choice of different norms for the Dirichlet boundary control problem, consider the following example taken from [62]. Here we actually have the situation of controlling the system through the control boundary Γ on the right hand side of Fig. 2.4 while a prescribed state $y_* \equiv 1$ on the observation boundary Γ_y opposite the control boundary is to be achieved. The right hand side is chosen as constant $f \equiv 1$, and $\omega = 1$. Each layer in Fig. 2.4 corresponds to the state y for different values of s when the observation term is measured in $H^s(\Gamma_y)$, that is, the objective functional (2.82) contains a term $\|y - y_*\|^2_{H^s(\Gamma_y)}$ for $s = 1/10, 2/10, 3/10, 4/10, 5/10, 7/10, 9/10$ from bottom to top. We see that as the smoothness index s for the observation increases, the state moves towards the target state at the observation boundary.

2.6.2 *Adaptive Schemes*

In case of the appearance of singularities caused by the data or the domain, a prescribed accuracy may require discretizations with respect to uniform grids to spend a large amount of degrees of freedom in areas where the solution is actually smooth. Hence, although the above numerical scheme NEICG is of optimal linear complexity, the degrees of freedom are not implanted in an optimal way. In these situations, one expects adaptive schemes to work favourably which judiciously place degrees of freedom where singularities occur. Thus, the guiding line for adaptive schemes is to reduce the total amount of degrees of freedom when compared to discretizations on a uniform grid. This does not mean that the previous investigations with respect to uniform discretizations are dispensable. In fact, the above results on conditioning carry over to the adaptive case, the solvers are still linear in the amount of arithmetic operations and, in particular, one expects to recover the uniform situation when the solutions are smooth. Much on adaptivity for variational problems and the relation to nonlinear approximation can be found in [26].

The starting point for adaptive wavelet schemes systematically derived for variational problems in [19–21] is the infinite formulation in wavelet coordinates as derived for the different problem classes in Sect. 2.5. These algorithms have been proven to be optimal in the sense that they match the optimal work/ accuracy rate of the wavelet-best N-term approximation, a concept which has been introduced in [19]. The schemes start out with formulating algorithmic ingredients which are then step by step reduced to computable quantities. We follow in this section the material for the distributed control problem from [29]. An extension to Dirichlet control problem involving saddle point problems can be found in [54]. It should be pointed out that the theory is neither confined to symmetric $\mathbf{A}$ nor to the positive definite case.

Algorithmic Ingredients We start out again with a very simple iterative scheme for the design equation. In view of (2.215) and the fact that $\mathbf{Q}$ is positive definite, there exists a fixed positive parameter α such that in the *Richardson iteration* (which is a special case of a gradient method)

$$\mathbf{u}^{k+1} = \mathbf{u}^k + \alpha(\mathbf{g} - \mathbf{Q}\mathbf{u}^k) \tag{2.241}$$

the error is reduced in each step by at least a factor

$$\rho := \|\mathbf{I} - \alpha\mathbf{Q}\| < 1, \tag{2.242}$$

$$\|\mathbf{u} - \mathbf{u}^{k+1}\| \leq \rho\, \|\mathbf{u} - \mathbf{u}^k\|, \quad k = 0, 1, 2, \ldots, \tag{2.243}$$

where $\mathbf{u}$ is the exact solution of (2.214). As the involved system is still infinite, we aim at carrying out this iteration approximately with dynamically updated accuracy tolerances.

The central idea of the wavelet-based adaptive schemes is to start from the infinite system in wavelet coordinates (2.207) and step by step reduce the routines to computable versions of applying the infinite matrix $\mathbf{Q}$ and the evaluation of the right hand side $\mathbf{g}$ of (2.214) involving the inversion of $\mathbf{A}$. The main conceptual tools from [19–21] are the following.

We first assume that we have a routine at our disposal with the following property. Later it will be shown how to realize this routine in the concrete case.

RES $[\eta, \mathbf{Q}, \mathbf{g}, \mathbf{v}] \rightarrow \mathbf{r}_\eta$ DETERMINES FOR A GIVEN TOLERANCE $\eta > 0$ A FINITELY SUPPORTED SEQUENCE $\mathbf{r}_\eta$ SATISFYING

$$\|\mathbf{g} - \mathbf{Q}\mathbf{v} - \mathbf{r}_\eta\| \leq \eta. \tag{2.244}$$

The schemes considered below will also contain the following routine.

COARSE $[\eta, \mathbf{w}] \rightarrow \mathbf{w}_\eta$ DETERMINES FOR ANY FINITELY SUPPORTED INPUT VECTOR $\mathbf{w}$ A VECTOR $\mathbf{w}_\eta$ WITH SMALLEST POSSIBLE SUPPORT SUCH THAT

$$\|\mathbf{w} - \mathbf{w}_\eta\| \leq \eta. \tag{2.245}$$

This ingredient will eventually play a crucial role in controlling the complexity of the scheme although its role is not yet apparent at this stage. A detailed description of COARSE can be found in [19]. The basic idea is to first sort the entries of $\mathbf{w}$ by size. Then one subtracts squares of their moduli until the sum reaches η^2, starting from the smallest entry. A quasi-sorting based on binary binning can be shown to avoid the logarithmic term in the sorting procedure at the expense of the resulting support size being at most a fixed constant of the minimal size, see [4].

Next a *perturbed iteration* is designed which converges in the following sense: for every target accuracy ε, the scheme produces after finitely many steps a finitely supported approximate solution with accuracy ε. To obtain a correctly balanced interplay between the routines RES and COARSE, we need the following control parameter. Given (an estimate of) the reduction rate ρ and the step size parameter α from (2.242), let K denote the minimal integer ℓ for which $\rho^{\ell-1}(\alpha\ell + \rho) \leq \frac{1}{10}$.

Denoting in the following always by $\mathbf{u}$ the exact solution of (2.214), a perturbed version of (2.241) for a fixed target accuracy $\varepsilon > 0$ is the following.

SOLVE $[\varepsilon, \mathbf{Q}, \mathbf{g}, \overline{\mathbf{q}}^0, \varepsilon_0] \rightarrow \overline{\mathbf{q}}_\varepsilon$

```
(I) GIVEN AN INITIAL GUESS $\overline{\mathbf{q}}^0$ AND AN ERROR BOUND $\|\mathbf{q} - \overline{\mathbf{q}}^0\| \leq \varepsilon_0$; SET $j = 0$.
(II) IF $\varepsilon_j \leq \varepsilon$, STOP AND SET $\overline{\mathbf{q}}_\varepsilon := \overline{\mathbf{q}}^j$. OTHERWISE SET $\mathbf{v}^0 := \overline{\mathbf{q}}^j$.
    (II.1) FOR $k = 0, \ldots, K-1$ COMPUTE RES $[\rho^k\varepsilon_j, \mathbf{Q}, \mathbf{g}, \mathbf{v}^k] \rightarrow \mathbf{r}^k$ AND
                $\mathbf{v}^{k+1} := \mathbf{v}^k + \alpha\mathbf{r}^k$.        (2.246)
    (II.2) APPLY COARSE $[\frac{2}{5}\varepsilon_j, \mathbf{v}^K] \rightarrow \overline{\mathbf{q}}^{j+1}$; SET $\varepsilon_{j+1} := \frac{1}{2}\varepsilon_j$, $j + 1 \rightarrow j$ AND GO TO (II).
```

In the case that no particular initial guess is known, we initialize $\overline{\mathbf{q}}^0 = \mathbf{0}$, set $\varepsilon_0 := c_{\mathbf{Q}}^{-1}\|\mathbf{g}\|$ and briefly write then SOLVE $[\varepsilon, \mathbf{Q}, \mathbf{g}] \rightarrow \overline{\mathbf{q}}_\varepsilon$.

In a straightforward manner, perturbation arguments yield the convergence of this algorithm [20, 21].

Proposition 12 *The iterates* $\overline{\mathbf{q}}^j$ *generated by* SOLVE $[\varepsilon, \mathbf{Q}, \mathbf{g}]$ *satisfy*

$$\|\mathbf{q} - \overline{\mathbf{q}}^j\| \leq \varepsilon_j \qquad \textit{for any} \quad j \geq 0, \tag{2.247}$$

where $\varepsilon_j = 2^{-j}\varepsilon_0$.

In order to derive appropriate numerical realizations of SOLVE, recall that (2.214) is equivalent to the KKT conditions (2.207). Although the matrix $\mathbf{A}$ is always assumed to be symmetric here, the distinction between the system matrices for the primal and the dual system, $\mathbf{A}$ and $\mathbf{A}^T$, may be helpful.

The strategy for approximating in each step the residual $\mathbf{g} - \mathbf{Q}\mathbf{u}^k$, that is, realization of the routine RES for the problem (2.214), is based upon the result stated in Proposition 7. In turn, this requires solving the two auxiliary systems in (2.207). Since the residual only has to be approximated, these systems will have to be solved only approximately. These approximate solutions, in turn, will be provided again by employing SOLVE but this time with respect to suitable residual schemes tailored to the systems in (2.207). In our special case, the matrix $\mathbf{A}$ is symmetric positive definite, and the choice of wavelet bases ensures the validity of (2.46). Thus, (2.242) holds for $\mathbf{A}$ and $\mathbf{A}^T$ so that the scheme SOLVE can indeed be invoked. Although we conceptually use the fact that a gradient iteration for the reduced problem (2.214) reduces the error for $\mathbf{u}$ in each step by a fixed amount, employing (2.207) for the evaluation of the residuals will generate as byproducts approximate solutions to the exact solution triple $(\mathbf{y}, \mathbf{p}, \mathbf{u})$ of (2.207). Under this hypothesis, we formulate next the ingredients for suitable versions SOLVE$_{\text{PRM}}$ and SOLVE$_{\text{ADJ}}$ of SOLVE for the systems in (2.207). Specifically, this requires identifying residual routines RES$_{\text{PRM}}$ and RES$_{\text{ADJ}}$ for the systems SOLVE$_{\text{PRM}}$ and SOLVE$_{\text{ADJ}}$. The main task in both cases is to apply the operators $\mathbf{A}$, $\mathbf{A}^T$, $\mathbf{D}_H^{-1}$ and $\mathbf{R}^{1/2}\mathbf{D}_{\mathscr{Z}}^{-1}$. Again we assume for the moment that routines for the application of these operators are available, i.e., that for any $\mathbf{L} \in \{\mathbf{A}, \mathbf{A}^T, \mathbf{D}_H^{-1}, \mathbf{R}^{1/2}\mathbf{D}_{\mathscr{Z}}^{-1}\}$ we have a scheme at our disposal with the following property.

APPLY $[\eta, \mathbf{L}, \mathbf{v}] \rightarrow \mathbf{w}_\eta$ DETERMINES FOR ANY FINITELY SUPPORTED INPUT VECTOR $\mathbf{v}$ AND ANY TOLERANCE $\eta > 0$ A FINITELY SUPPORTED OUTPUT $\mathbf{w}_\eta$ WHICH SATISFIES

$$\|\mathbf{L}\mathbf{v} - \mathbf{w}_\eta\| \leq \eta. \tag{2.248}$$

The scheme SOLVE$_{\text{PRM}}$ for the first system in (2.207) is then defined by

$$\text{SOLVE}_{\text{PRM}}\,[\eta, \mathbf{A}, \mathbf{D}_H^{-1}, \mathbf{f}, \mathbf{v}, \overline{\mathbf{y}}^0, \varepsilon_0] := \text{SOLVE}\,[\eta, \mathbf{A}, \mathbf{f} + \mathbf{D}_H^{-1}\mathbf{v}, \overline{\mathbf{y}}^0, \varepsilon_0],$$

where $\overline{\mathbf{y}}^0$ is an initial guess for the solution $\mathbf{y}$ of $\mathbf{A}\mathbf{y} = \mathbf{f} + \mathbf{D}_H^{-1}\mathbf{v}$ with accuracy ε_0. The scheme RES for Step (II) in SOLVE is in this case realized by a new routine RES$_{\text{PRM}}$ defined as follows.

RES$_{\text{PRM}}$ $[\eta, \mathbf{A}, \mathbf{D}_H^{-1}, \mathbf{f}, \mathbf{v}, \overline{\mathbf{y}}] \rightarrow \mathbf{r}_\eta$ DETERMINES FOR ANY POSITIVE TOLERANCE η, A GIVEN FINITELY SUPPORTED $\mathbf{v}$ AND ANY FINITELY SUPPORTED INPUT $\overline{\mathbf{y}}$ A

FINITELY SUPPORTED APPROXIMATE RESIDUAL $\mathbf{r}_\eta$ SATISFYING (2.244),

$$\|\mathbf{f} + \mathbf{D}_H^{-1}\mathbf{v} - \mathbf{A}\overline{\mathbf{y}} - \mathbf{r}_\eta\| \leq \eta, \tag{2.249}$$

AS FOLLOWS:

(I) APPLY $[\frac{1}{3}\eta, \mathbf{A}, \overline{\mathbf{y}}] \to \mathbf{w}_\eta$;
(II) COARSE $[\frac{1}{3}\eta, \mathbf{f}] \to \mathbf{f}_\eta$;
(III) APPLY $[\frac{1}{3}\eta, \mathbf{D}_H^{-1}, \mathbf{v}] \to \mathbf{z}_\eta$;
(IV) set $\mathbf{r}_\eta := \mathbf{f}_\eta + \mathbf{z}_\eta - \mathbf{w}_\eta$.

By triangle inequality, one can for RES$_{\text{PRM}}$ and the subsequent variants of RES show that indeed (2.249) or (2.244) holds.

Similarly, one needs a version of SOLVE for the approximate solution of the second system (2.207b), $\mathbf{A}^T\mathbf{p} = -\mathbf{D}_{\mathscr{Z}}^{-1}\mathbf{R}\mathbf{D}_{\mathscr{Z}}^{-1}(\mathbf{y} - \mathbf{y}_*)$, which depends on an approximate solution $\overline{\mathbf{y}}$ of the primal system and possibly on some initial guess $\overline{\mathbf{p}}^0$ with accuracy ε_0. Here we set

$$\text{SOLVE}_{\text{ADJ}}\,[\eta, \mathbf{A}, \mathbf{D}_{\mathscr{Z}}^{-1}, \mathbf{y}_*, \overline{\mathbf{y}}, \overline{\mathbf{p}}^0, \varepsilon_0] := \text{SOLVE}\,[\eta, \mathbf{A}^T, \mathbf{D}_{\mathscr{Z}}^{-1}\mathbf{R}\mathbf{D}_{\mathscr{Z}}^{-1}(\mathbf{y} - \overline{\mathbf{y}}), \overline{\mathbf{p}}^0, \varepsilon_0].$$

As usual we assume that the data $\mathbf{f}$, $\mathbf{y}_*$ are approximated in a preprocessing step with sufficient accuracy. A suitable residual approximation scheme RES$_{\text{ADJ}}$ for Step (II) of this version of SOLVE is the following where the main issue is the approximate evaluation of the right hand side.

RES$_{\text{ADJ}}$ $[\eta, \mathbf{A}, \mathbf{D}_{\mathscr{Z}}^{-1}, \mathbf{y}_*, \overline{\mathbf{y}}, \overline{\mathbf{p}}] \to \mathbf{r}_\eta$ DETERMINES FOR ANY POSITIVE TOLERANCE η, GIVEN FINITELY SUPPORTED DATA $\overline{\mathbf{y}}, \mathbf{y}_*$ AND ANY FINITELY SUPPORTED INPUT $\overline{\mathbf{p}}$ AN APPROXIMATE RESIDUAL $\mathbf{r}_\eta$ SATISFYING (2.244), I.E.,

$$\| -\mathbf{D}_{\mathscr{Z}}^{-1}\mathbf{R}\mathbf{D}_{\mathscr{Z}}^{-1}(\overline{\mathbf{y}} - \mathbf{y}_*) - \mathbf{A}^T\overline{\mathbf{p}} - \mathbf{r}_\eta\| \leq \eta, \tag{2.250}$$

AS FOLLOWS:

(i) APPLY $[\frac{1}{3}\eta, \mathbf{A}^T, \overline{\mathbf{p}}] \to \mathbf{w}_\eta$;
(ii) APPLY $[\frac{1}{6}\,\eta, \mathbf{D}_{\mathscr{Z}}^{-1}, \overline{\mathbf{y}}] \to \mathbf{z}_\eta$; COARSE $[\frac{1}{6}\,\eta, \mathbf{y}_*] \to (\mathbf{y}_*)_\eta$;
SET $\mathbf{d}_\eta := (\mathbf{y}_Z)_\eta - \mathbf{z}_\eta$;
APPLY $[\frac{1}{6}\eta, \mathbf{D}_{\mathscr{Z}}^{-1}, \mathbf{d}_\eta] \to \hat{\mathbf{v}}_\eta$; APPLY $[\frac{1}{6}\eta, \mathbf{R}, \hat{\mathbf{v}}_\eta] \to \mathbf{v}_\eta$;
(iii) SET $\mathbf{r}_\eta := \mathbf{v}_\eta - \mathbf{w}_\eta$.

Finally, we can define the residual scheme for the version of SOLVE applied to (2.214). We shall refer to this specification as SOLVE$_{\text{DCP}}$ with corresponding residual scheme is RES$_{\text{DCP}}$. Since the scheme is based on Proposition 7, it will involve several parameters stemming from the auxiliary systems (2.207).

RES$_{\text{DCP}}$ $[\eta, \mathbf{Q}, \mathbf{g}, \tilde{\mathbf{y}}, \delta_y, \tilde{\mathbf{p}}, \delta_p, \mathbf{v}, \delta_v] \to (\mathbf{r}_\eta, \tilde{\mathbf{y}}, \delta_y, \tilde{\mathbf{p}}, \delta_p)$ DETERMINES FOR ANY APPROXIMATE SOLUTION TRIPLE $(\tilde{\mathbf{y}}, \tilde{\mathbf{p}}, \mathbf{v})$ OF THE SYSTEM (2.207) SATISFYING

$$\|\mathbf{y} - \tilde{\mathbf{y}}\| \leq \delta_y, \quad \|\mathbf{p} - \tilde{\mathbf{p}}\| \leq \delta_p, \quad \|\mathbf{u} - \mathbf{v}\| \leq \delta_v, \tag{2.251}$$

an approximate residual $\mathbf{r}_\eta$ such that

$$\|\mathbf{g} - \mathbf{Q}\mathbf{v} - \mathbf{r}_\eta\| \le \eta. \tag{2.252}$$

Moreover, the initial approximations $\tilde{\mathbf{y}}, \tilde{\mathbf{p}}$ are overwritten by new approximations $\tilde{\mathbf{y}}, \tilde{\mathbf{p}}$ satisfying (2.251) with new bounds δ_y and δ_p defined in (2.253) below, as follows:

```
(i)   $\textsc{Solve}_{\mathrm{PRM}}\,[\frac{1}{3}c_{\mathbf{A}}\,\eta, \mathbf{A}, \mathbf{D}_H^{-1}, \mathbf{f}, \mathbf{v}, \tilde{\mathbf{y}}, \delta_y] \to \mathbf{y}_\eta$;
(ii)  $\textsc{Solve}_{\mathrm{ADJ}}\,[\frac{1}{3}\eta, \mathbf{A}, \mathbf{D}_{\mathscr{Z}}^{-1}, \mathbf{y}_*, \mathbf{y}_\eta, \tilde{\mathbf{p}}, \delta_p] \to \mathbf{p}_\eta$;
(iii) Apply $[\frac{1}{3}\eta, \mathbf{D}_H^{-1}, \mathbf{p}_\eta] \to \mathbf{q}_\eta$; set $\mathbf{r}_\eta := \mathbf{q}_\eta - \omega\mathbf{v}$;
(iv)  set $\xi_y := c_{\mathbf{A}}^{-1}\,\delta_v + \frac{1}{3}c_{\mathbf{A}}\eta,\ \ \xi_p := c_{\mathbf{A}}^{-2}\,\delta_v + \frac{2}{3}\eta$; replace $\tilde{\mathbf{y}}, \delta_y$ and $\tilde{\mathbf{p}}, \delta_p$ by
```

$$\begin{aligned} \tilde{\mathbf{y}} &:= \textsc{Coarse}[4\xi_y, \mathbf{y}_\eta], \quad & \delta_y &:= 5\,\xi_y, \\ \tilde{\mathbf{p}} &:= \textsc{Coarse}[4\xi_p, \mathbf{p}_\eta], \quad & \delta_p &:= 5\,\xi_p. \end{aligned} \tag{2.253}$$

Step (iv) already indicates the conditions on the tolerance η and the accuracy bound δ_v under which the new error bounds in (2.253) are actually tighter. The precise relation between η and δ_v in the context of $\textsc{Solve}_{\mathrm{DCP}}$ is not apparent yet and emerges as well as the claimed estimates (2.252) and (2.253) from the complexity analysis in [29].

Finally, the scheme $\textsc{Solve}_{\mathrm{DCP}}$ attains the following form with the error reduction factor ρ from (2.242) and α from (2.241).

$\textsc{Solve}_{\mathrm{DCP}}\ [\varepsilon, \mathbf{Q}, \mathbf{g}] \to \overline{\mathbf{u}}_\varepsilon$

```
(i)  Let $\overline{\mathbf{q}}^0 := \mathbf{0}$ and $\varepsilon_0 := c_{\mathbf{A}}^{-1}(\|\mathbf{y}_Z\| + c_{\mathbf{A}}^{-1}\|\mathbf{f}\|)$.
     Let $\tilde{\mathbf{y}} := \mathbf{0}, \tilde{\mathbf{p}} := \mathbf{0}$ and set $j = 0$.
     Define $\delta_y := \delta_{y,0} := c_{\mathbf{A}}^{-1}(\|\mathbf{f}\| + \varepsilon_0)$ and $\delta_p := \delta_{p,0} := c_{\mathbf{A}}^{-1}(\delta_{y,0} + \|\mathbf{y}_Z\|)$.
(ii) If $\varepsilon_j \le \varepsilon$, stop and set $\overline{\mathbf{u}}_\varepsilon := \overline{\mathbf{u}}^j, \overline{\mathbf{y}}_\varepsilon = \tilde{\mathbf{y}}, \overline{\mathbf{p}}_\varepsilon = \tilde{\mathbf{p}}$.
     Otherwise set $\mathbf{v}^0 := \overline{\mathbf{u}}^j$.

     (ii.1) For $k = 0, \dots, K-1$, compute
            $\textsc{Res}_{\mathrm{DCP}}\ [\rho^k\varepsilon_j, \mathbf{Q}, \mathbf{g}, \tilde{\mathbf{y}}, \delta_y, \tilde{\mathbf{p}}, \delta_p, \mathbf{v}^k, \delta_k] \to (\mathbf{r}^k, \tilde{\mathbf{y}}, \delta_y, \tilde{\mathbf{p}}, \delta_p)$,
            where $\delta_0 := \varepsilon_j$ and $\delta_k := \rho^{k-1}(\alpha k + \rho)\varepsilon_j$;
            set
```

$$\mathbf{v}^{k+1} := \mathbf{v}^k + \alpha\mathbf{r}^k. \tag{2.254}$$

```
     (ii.2) Coarse $[\frac{2}{5}\varepsilon_j, \mathbf{v}^K] \to \overline{\mathbf{u}}^{j+1}$; set $\varepsilon_{j+1} := \frac{1}{2}\varepsilon_j$, $j+1 \to j$ and go to (ii).
```

By overwriting $\tilde{\mathbf{y}}$, $\tilde{\mathbf{p}}$ at the last stage prior to the termination of $\textsc{Solve}_{\mathrm{DCP}}$ one has $\delta_v \le \varepsilon$, $\eta \le \varepsilon$, so that the following fact is an immediate consequence of (2.253).

Proposition 13 *The outputs $\overline{\mathbf{y}}_\varepsilon$ and $\overline{\mathbf{p}}_\varepsilon$ produced by* $\textsc{Solve}_{\mathrm{DCP}}$ *in addition to $\mathbf{u}_\varepsilon$ are approximations to the exact solutions $\mathbf{y}, \mathbf{p}$ of (2.207) satisfying*

$$\|\mathbf{y} - \overline{\mathbf{y}}_\varepsilon\| \le 5\varepsilon\,(c_{\mathbf{A}}^{-1} + \tfrac{1}{3}c_{\mathbf{A}}), \qquad \|\mathbf{p} - \overline{\mathbf{p}}_\varepsilon\| \le 5\varepsilon\,(c_{\mathbf{A}}^{-2} + \tfrac{2}{3}).$$

Complexity Analysis Proposition 12 states that the routine SOLVE converges for an arbitrary given accuracy provided that there is a routine RES satisfying the property (2.244). Then we have broken down step by step the necessary ingredients to derive computable versions which satisfy these requirements. What we finally want to show is that the routines are *optimal* in the sense that they provide the optimal work/accuracy rate in terms of best N-term approximation. The complexity analysis given next also reveals the role of the routine COARSE within the algorithms and the particular choices of the thresholds in Step (IV) of RES$_{\text{DCP}}$.

In order to be able to assess the quality of the adaptive algorithm, the notion of *optimality* has to be clarified first in the present context.

Definition 1 The scheme SOLVE has an *optimal work/accuracy rate s* if the following holds: Whenever the error of *best N-term approximation* satisfies

$$\|\mathbf{q}-\mathbf{q}_N\| := \min_{\# supp \mathbf{v} \le N} \|\mathbf{q}-\mathbf{v}\| \lesssim N^{-s},$$

then the solution $\overline{\mathbf{q}}_\varepsilon$ is generated by SOLVE at an expense that also stays proportional to $\varepsilon^{-1/s}$ and in that sense matches the best N-term approximation rate.

Note that this implies that $\#\text{supp}\,\overline{\mathbf{q}}_\varepsilon$ also stays proportional to $\varepsilon^{-1/s}$. Thus, our benchmark is that whenever the solution of (2.214) can be approximated by N terms at rate s, SOLVE recovers that rate asymptotically. If $\mathbf{q}$ is known, the wavelet-best N-term approximation $\mathbf{q}_N$ of $\mathbf{q}$ is given by picking the N largest terms in modulus from $\mathbf{q}$, of course. However, when $\mathbf{q}$ is the (unknown) solution of (2.214) this information is certainly not available.

Since we are here in the framework of sequence spaces ℓ_2, the formulation of appropriate criteria for complexity will be based on a characterization of sequences which are *sparse* in the following sense. We consider sequences $\mathbf{v}$ for which the best N-term approximation error decays at a particular rate (*Lorentz spaces*). That is, for any given threshold $0 < \eta \le 1$, the number of terms exceeding that threshold is controlled by some function of this threshold. In particular, set for some $0 < \tau < 2$

$$\ell_\tau^w := \{\mathbf{v} \in \ell_2 : \#\{\lambda \in I\!I : |v_\lambda| > \eta\} \le C_\mathbf{v}\, \eta^{-\tau}, \text{ for all } 0 < \eta \le 1\}. \tag{2.255}$$

This determines a strict subspace of ℓ_2 only when $\tau < 2$. Smaller τ's indicate sparser sequences. Let $C_\mathbf{v}$ for a given $\mathbf{v} \in \ell_\tau^w$ be the smallest constant for which (2.255) holds. Then one has $|\mathbf{v}|_{\ell_\tau^w} := \sup_{n\in\mathbb{N}} n^{1/\tau} v_n^* = C_\mathbf{v}^{1/\tau}$,where $\mathbf{v}^* = (v_n^*)_{n\in\mathbb{N}}$ is a non-decreasing rearrangement of $\mathbf{v}$. Furthermore, $\|\mathbf{v}\|_{\ell_\tau^w} := \|\mathbf{v}\| + |\mathbf{v}|_{\ell_\tau^w}$ is a quasi-norm for ℓ_τ^w. Since the continuous embeddings $\ell_\tau \hookrightarrow \ell_\tau^w \hookrightarrow \ell_{\tau+\varepsilon} \hookrightarrow \ell_2$ hold for $\tau < \tau+\varepsilon < 2$, ℓ_τ^w is 'close' to ℓ_τ and is therefore called *weak* ℓ_τ. The following crucial result connects sequences in ℓ_τ^w to best N-term approximation [19].

Proposition 14 *Let positive real numbers s and τ be related by*

$$\frac{1}{\tau} = s + \frac{1}{2}. \tag{2.256}$$

Then $\mathbf{v} \in \ell_\tau^w$ *if and only if* $\|\mathbf{v} - \mathbf{v}_N\| \lesssim N^{-s} \|\mathbf{v}\|_{\ell_\tau^w}$.

The property that an array of wavelet coefficients $\mathbf{v}$ belongs to ℓ_τ is equivalent to the fact that the expansion $\mathbf{v}^T \Psi_H$ in terms of a wavelet basis Ψ_H for a Hilbert space H belongs to a certain *Besov space* which describes a much weaker regularity measure than a Sobolev space of corresponding order, see, e.g., [16, 39]. Thus, Proposition 14 expresses how much loss of regularity can be compensated by judiciously placing the degrees of freedom in a nonlinear way in order to retain a certain optimal order of error decay.

A key criterion for a scheme SOLVE to exhibit an optimal work/accuracy rate can be formulated through the following property of the respective residual approximation. The routine RES is called τ^*-*sparse* for some $0 < \tau^* < 2$ if the following holds: Whenever the solution $\mathbf{q}$ of (2.214) belongs to ℓ_τ^w for some $\tau^* < \tau < 2$, then for any $\mathbf{v}$ with finite support the output $\mathbf{r}_\eta$ of RES $[\eta, \mathbf{Q}, \mathbf{g}, \mathbf{v}]$ satisfies

$$\|\mathbf{r}_\eta\|_{\ell_\tau^w} \lesssim \max\{\|\mathbf{v}\|_{\ell_\tau^w}, \|\mathbf{q}\|_{\ell_\tau^w}\}$$

and

$$\#supp\mathbf{r}_\eta \lesssim \eta^{-1/s} \max\{\|\mathbf{v}\|_{\ell_\tau^w}^{1/s}, \|\mathbf{q}\|_{\ell_\tau^w}^{1/s}\}$$

where s and τ are related by (2.256), and the number of floating point operations needed to compute $\mathbf{r}_\eta$ stays proportional to $\#supp\mathbf{r}_\eta$.

The analysis in [20] then yields the following result.

Theorem 8 *If* RES *is* τ^**-sparse and if the exact solution* $\mathbf{q}$ *of (2.214) belongs to* ℓ_τ^w *for some* $\tau > \tau^*$*, then for every* $\varepsilon > 0$ *algorithm* SOLVE $[\varepsilon, \mathbf{Q}, \mathbf{g}]$ *produces after finitely many steps an output* $\overline{\mathbf{q}}_\varepsilon$ *(which, according to Proposition 12, always satisfies* $\|\mathbf{q} - \overline{\mathbf{q}}_\varepsilon\| < \varepsilon$*) with the following properties: For s and* τ *related by (2.256), one has*

$$\#supp\overline{\mathbf{q}}_\varepsilon \lesssim \varepsilon^{-1/s} \|\mathbf{q}\|_{\ell_\tau^w}^{1/s}, \qquad \|\overline{\mathbf{q}}_\varepsilon\|_{\ell_\tau^w} \lesssim \|\mathbf{q}\|_{\ell_\tau^w}, \tag{2.257}$$

and the number of floating point operations needed to compute $\overline{\mathbf{q}}_\varepsilon$ *remains proportional to* $\#supp\overline{\mathbf{q}}_\varepsilon$.

Hence, τ^*-sparsity of the routine RES implies for SOLVE asymptotically optimal work/accuracy rates for a certain range of decay rates given by τ^*. We stress that the algorithm itself does *not* require any a-priori knowledge about the solution such as its actual best N-term approximation rate. Theorem 8 also states that controlling the ℓ_τ^w-norm of the quantities generated in the computations is crucial. This finally explains the role of COARSE in Step (II.2) of SOLVE in terms of the following result [19].

Lemma 4 *Let* $\mathbf{v} \in \ell_\tau^w$ *and let* $\mathbf{w}$ *be any finitely supported approximation such that* $\|\mathbf{v} - \mathbf{w}\| \le \frac{1}{5}\eta$*. Then the output* $\mathbf{w}_\eta$ *of* COARSE $[\frac{4}{5}\eta, \mathbf{w}]$ *satisfies*

$$\#supp\mathbf{w}_\eta \ \lesssim\ \|\mathbf{v}\|_{\ell_\tau^w}^{1/\tau}\, \eta^{-1/s}, \quad \|\mathbf{v} - \mathbf{w}_\eta\| \ \lesssim\ \eta, \quad and \quad \|\mathbf{w}_\eta\|_{\ell_\tau^w} \ \lesssim\ \|\mathbf{v}\|_{\ell_\tau^w}. \tag{2.258}$$

This can be interpreted as follows. If an error bound for a given finitely supported approximation $\mathbf{w}$ is known, a certain coarsening using only knowledge about $\mathbf{w}$ produces a new approximation to (the possibly unknown) $\mathbf{v}$ which gives rise to a slightly larger error but realizes the optimal relation between support and accuracy up to a uniform constant. In the scheme SOLVE, this means that by the coarsening step the ℓ_τ^w-norms of the iterates $\mathbf{v}^K$ are controlled.

It remains to establish that for SOLVE$_{\text{DCP}}$ the corresponding routine RES$_{\text{DCP}}$ is τ^*-sparse. The following results from [29] reduce this question to the efficiency of APPLY. We say that APPLY $[\cdot, \mathbf{L}, \cdot]$ is τ^*-*efficient* for some $0 < \tau^* < 2$ if for any finitely supported $\mathbf{v} \in \ell_\tau^w$, for $0 < \tau^* < \tau < 2$, the output $\mathbf{w}_\eta$ of APPLY $[\eta, \mathbf{L}, \mathbf{v}]$ satisfies $\|\mathbf{w}_\eta\|_{\ell_\tau^w} \lesssim \|\mathbf{v}\|_{\ell_\tau^w}$ and $\#\text{supp}\,\mathbf{w}_\eta \lesssim \eta^{-1/s}\|\mathbf{v}\|_{\ell_\tau^w}^{1/s}$ for $\eta \to 0$. Here the constants depend only on τ as $\tau \to \tau^*$ and s, τ satisfy (2.256). Moreover, the number of floating point operations needed to compute $\mathbf{w}_\eta$ is to remain proportional to $\#\text{supp}\,\mathbf{w}_\eta$.

Proposition 15 *If the* APPLY *schemes in* RES$_{\text{PRM}}$ *and* RES$_{\text{ADJ}}$ *are* τ^**-efficient for some* $\tau^* < 2$*, then* RES$_{\text{DCP}}$ *is* τ^**-sparse whenever there exists a constant* C *such that* $C\eta \ge \max\{\delta_v, \delta_p\}$ *and*

$$\max\{\|\tilde{\mathbf{p}}\|_{\ell_\tau^w}, \|\tilde{\mathbf{y}}\|_{\ell_\tau^w}, \|\mathbf{v}\|_{\ell_\tau^w}\} \le C\left(\|\mathbf{y}\|_{\ell_\tau^w} + \|\mathbf{p}\|_{\ell_\tau^w} + \|\mathbf{u}\|_{\ell_\tau^w}\right),$$

where $\mathbf{v}$ *is the current finitely supported input and* $\tilde{\mathbf{y}}, \tilde{\mathbf{p}}$ *are the initial guesses for the exact solution components* $(\mathbf{y}, \mathbf{p})$*.*

Theorem 9 *If the* APPLY *schemes appearing in* RES$_{\text{PRM}}$ *and* RES$_{\text{ADJ}}$ *are* τ^**-efficient for some* $\tau^* < 2$ *and the components of the solution* $(\mathbf{y}, \mathbf{p}, \mathbf{u})$ *of (2.207) all belong to the respective space* ℓ_τ^w *for some* $\tau > \tau^*$*, then the approximate solutions* $\mathbf{y}_\varepsilon, \mathbf{p}_\varepsilon, \mathbf{u}_\varepsilon$*, produced by* SOLVE$_{\text{DCP}}$ *for any target accuracy* ε*, satisfy*

$$\|\mathbf{y}_\varepsilon\|_{\ell_\tau^w} + \|\mathbf{p}_\varepsilon\|_{\ell_\tau^w} + \|\mathbf{u}_\varepsilon\|_{\ell_\tau^w} \ \lesssim\ \|\mathbf{y}\|_{\ell_\tau^w} + \|\mathbf{p}\|_{\ell_\tau^w} + \|\mathbf{u}\|_{\ell_\tau^w}, \tag{2.259}$$

and

$$(\#\text{supp}\,\mathbf{y}_\varepsilon) + (\#\text{supp}\,\mathbf{p}_\varepsilon) + (\#\text{supp}\,\mathbf{u}_\varepsilon) \ \lesssim\ \left(\|\mathbf{y}\|_{\ell_\tau^w}^{1/s} + \|\mathbf{p}\|_{\ell_\tau^w}^{1/s} + \|\mathbf{u}\|_{\ell_\tau^w}^{1/s}\right)\varepsilon^{-1/s}, \tag{2.260}$$

where the constants only depend on τ when τ approaches τ^*. Moreover, the number of floating point operations required during the execution of Solve$_{\text{DCP}}$ *remains proportional to the right hand side of (2.260).*

Thus, the practical realization of Solve$_{\text{DCP}}$ providing optimal work/accuracy rates for a possibly large range of decay rates of the error of best N-term approximation hinges on the availability of τ^*-efficient schemes Apply with possibly small τ^* for the involved operators.

For the approximate application of wavelet representations of a wide class of operators, including differential operators, one can indeed devise efficient schemes which is a consequence of the cancellation properties (CP) together with the norm equivalences (2.89) for the relevant function spaces. For the example considered above, the τ^*-efficiency of $\mathbf{A}$ defined in (2.198) can be shown whenever $\mathbf{A}$ is s^*-compressible where τ^* and s^* are related by (2.256). One knows that s^* is the larger the higher the 'regularity' of the operator and the order of cancellation properties of the wavelets are. Estimates for s^* in terms of these quantities for spline wavelets and the above differential operator A can be found in [5]. These were refined and extended to trace operators in [62]. Hence, Theorem 9 guarantees asymptotically optimal complexity bounds for $\tau > \tau^*$. This means that the scheme Solve$_{\text{DCP}}$ recovers rates of the error of best N-term approximation of order N^{-s} for $s < s^*$.

When describing the control problem, it has been pointed out that the wavelet framework allows for a flexible choice of norms in the control functional which is reflected by the diagonal matrices $\mathbf{D}_{\mathscr{Z}}$ and $\mathbf{D}_H$ in (DCP), (2.203) together with (2.204). The following result states that multiplication by either $\mathbf{D}_{\mathscr{Z}}^{-1}$ or $\mathbf{D}_H^{-1}$ makes a sequence more compressible, that is, they produce a shift in weak ℓ_τ spaces [29].

Proposition 16 *For $\beta > 0$, $\mathbf{p} \in \ell_\tau^w$ implies $\mathbf{D}^{-\beta}\mathbf{p} \in \ell_{\tau'}^w$, where $\frac{1}{\tau'} := \frac{1}{\tau} + \frac{\beta}{d}$.*

We can conclude the following. Whatever the sparsity class of the adjoint variable $\mathbf{p}$ is, the control $\mathbf{u}$ is in view of (2.207c) even sparser. This means also that although the control $\mathbf{u}$ may be accurately recovered with relatively few degrees of freedom, the overall solution complexity is in the above case bounded from below by the less sparse auxiliary variable $\mathbf{p}$.

The application of these techniques to control problems constrained by parabolic PDEs can be found in [44]. For an extension of these techniques to control problems involving PDEs with possibly infinite stochastic coefficients which introduce a substantial difficulty, one may consult [57, 58].

References

1. J.P. Aubin, *Applied Functional Analysis*, 2nd edn. (Wiley, Hoboken, 2000)
2. I. Babuška, The finite element method with Lagrange multipliers. Numer. Math. **20**, 179–192 (1973)
3. I. Babuška, The finite element method with penalty. Math. Comput. **27**, 221–228 (1973)

4. A. Barinka, Fast evaluation tools for adaptive wavelet schemes, Ph.D. dissertation, RWTH Aachen, 2004
5. A. Barinka, T. Barsch, Ph. Charton, A. Cohen, S. Dahlke, W. Dahmen, K. Urban, Adaptive wavelet schemes for elliptic problems — implementation and numerical experiments. SIAM J. Sci. Comput. **23**, 910–939 (2001)
6. S. Bertoluzza, Wavelet stabilization of the Lagrange multiplier method. Numer. Math. **86**, 1–28 (2000)
7. D. Braess, *Finite Elements: Theory, Fast Solvers and Applications in Solid Mechanics*, 2nd edn. (Cambridge University Press, Cambridge, 2001)
8. J.H. Bramble, J.E. Pasciak, J. Xu, Parallel multilevel preconditioners. Math. Comput. **55**, 1–22 (1990)
9. F. Brezzi, M. Fortin, *Mixed and Hybrid Finite Element Methods* (Springer, New York, 1991)
10. A. Buffa, H. Harbrecht, A. Kunoth, G. Sangalli, Multilevel preconditioning for isogeometric analysis, Comput. Methods Appl. Mech. Eng. **265**, 63–70 (2013)
11. C. Burstedde, Wavelets methods for linear-quadratic, elliptic optimal control problems, Ph.D. dissertation, University of Bonn, Bonn, 2005
12. C. Burstedde, A. Kunoth, Fast iterative solution of elliptic control problems in wavelet discretizations. J. Comput. Appl. Math. **196**(1), 299–319 (2006)
13. C. Canuto, A. Tabacco, K. Urban, The wavelet element method, part I: construction and analysis. Appl. Comput. Harmon. Anal. **6**, 1–52 (1999)
14. J.M. Carnicer, W. Dahmen, J.M. Peña, Local decomposition of refinable spaces. Appl. Comput. Harmon. Anal. **3**, 127–153 (1996)
15. Z. Ciesielski, T. Figiel, Spline bases in classical function spaces on compact C^{∞} manifolds: part I and II. Stud. Math. **76**, 1–58 and 95–136 (1983)
16. A. Cohen, *Numerical Analysis of Wavelet Methods*. Studies in Mathematics and Its Applications, vol. 32 (Elsevier, New York, 2003)
17. A. Cohen, R. Masson, Adaptive wavelet methods for second order elliptic problems, preconditioning and adaptivity. SIAM J. Sci. Comput. **21**, 1006–1026 (1999)
18. A. Cohen, I. Daubechies, J.-C. Feauveau, Biorthogonal bases of compactly supported wavelets, Commun. Pure Appl. Math. **45**, 485–560 (1992)
19. A. Cohen, W. Dahmen, R. DeVore, Adaptive wavelet methods for elliptic operator equations – convergence rates. Math. Comput. **70**, 27–75 (2001)
20. A. Cohen, W. Dahmen, R. DeVore, Adaptive wavelet methods II – beyond the elliptic case. Found. Comput. Math. **2**, 203–245 (2002)
21. A. Cohen, W. Dahmen, R. DeVore, Adaptive wavelet schemes for nonlinear variational problems. SIAM J. Numer. Anal. **41**(5), 1785–1823 (2003)
22. S. Dahlke, W. Dahmen, K. Urban, Adaptive wavelet methods for saddle point problems — optimal convergence rates. SIAM J. Numer. Anal. **40**, 1230–1262 (2002)
23. W. Dahmen, Stability of multiscale transformations. J. Fourier Anal. Appl. **2**, 341–361 (1996)
24. W. Dahmen, Wavelet and multiscale methods for operator equations. Acta Numer. **6**, 55–228 (1997)
25. W. Dahmen, Wavelet methods for PDEs – some recent developments. J. Comput. Appl. Math. **128**, 133–185 (2001)
26. W. Dahmen, Multiscale and wavelet methods for operator equations, in *Multiscale Problems and Methods in Numerical Simulation*, ed. by C. Canuto. C.I.M.E. Lecture Notes in Mathematics, vol. 1825 (Springer, Heidelberg, 2003), pp. 31–96
27. W. Dahmen, A. Kunoth, Multilevel preconditioning. Numer. Math. **63**, 315–344 (1992)
28. W. Dahmen, A. Kunoth, Appending boundary conditions by Lagrange multipliers: analysis of the LBB condition. Numer. Math. **88**, 9–42 (2001)
29. W. Dahmen, A. Kunoth, Adaptive wavelet methods for linear-quadratic elliptic control problems: convergence rates. SIAM J. Control. Optim. **43**(5), 1640–1675 (2005)
30. W. Dahmen, R. Schneider, Wavelets with complementary boundary conditions — function spaces on the cube. Results Math. **34**, 255–293 (1998)

31. W. Dahmen, R. Schneider, Composite wavelet bases for operator equations. Math. Comput. **68**, 1533–1567 (1999)
32. W. Dahmen, R. Schneider, Wavelets on manifolds I: construction and domain decomposition. SIAM J. Math. Anal. **31**, 184–230 (1999)
33. W. Dahmen, R. Stevenson, Element-by-element construction of wavelets satisfying stability and moment conditions. SIAM J. Numer. Anal. **37**, 319–325 (1999)
34. W. Dahmen, A. Kunoth, K. Urban, Biorthogonal spline wavelets on the interval – stability and moment conditions. Appl. Comput. Harmon. Anal. **6**, 132–196 (1999)
35. W. Dahmen, A. Kunoth, R. Schneider, Wavelet least squares methods for boundary value problems. SIAM J. Numer. Anal. **39**, 1985–2013 (2002)
36. I. Daubechies, Orthonormal bases of compactly supported wavelets. Commun. Pure Appl. Math. **41**, 909–996 (1988)
37. R. Dautray, J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*. Evolution Problems, vol. 5 (Springer, Berlin, 2000)
38. C. de Boor, *A Practical Guide to Splines*, revised edn. (Springer, New York, 2011)
39. R.A. DeVore, Nonlinear approximation. Acta Numer. **7**, 51–150 (1998)
40. L.C. Evans, *Partial Differential Equations* (AMS, Providence, 1998)
41. V. Girault, R. Glowinski, Error analysis of a fictitious domain method applied to a Dirichlet problem. Jpn. J. Ind. Appl. Math. **12**, 487–514 (1995)
42. V. Girault, P.-A. Raviart, *Finite Element Methods for Navier–Stokes Equations* (Springer, Berlin, 1986)
43. P. Grisvard, *Elliptic Problems in Nonsmooth Domains* (Pitman, New York, 1985)
44. M.D. Gunzburger, A. Kunoth, Space-time adaptive wavelet methods for optimal control problems constrained by parabolic evolution equations. SIAM J. Control. Optim. **49**(3), 1150–1170 (2011)
45. M.D. Gunzburger, H.C. Lee, Analysis, approximation, and computation of a coupled solid/fluid temperature control problem. Comput. Methods Appl. Mech. Eng. **118**, 133–152 (1994)
46. J. Haslinger, R.A.E. Mäkinen, *Introduction to Shape Optimization: Theory, Approximation, and Computation* (SIAM, Philadelphia, 2003)
47. S. Jaffard, Wavelet methods for fast resolution of elliptic problems. SIAM J. Numer. Anal. **29**, 965–986 (1992)
48. J. Krumsdorf, Finite element wavelets for the numerical solution of elliptic partial differential equations on polygonal domains, Diploma thesis (in English), Universität Bonn, Bonn, January 2004
49. K. Kunisch, G. Peichl, Shape optimization for mixed boundary value problems based on an embedding domain method. Dyn. Contin. Discret. Impuls. Syst. **4**, 439–478 (1998)
50. A. Kunoth, *Multilevel Preconditioning* (Verlag Shaker, Aachen, 1994)
51. A. Kunoth, Wavelet methods — elliptic boundary value problems and control problems, in *Advances in Numerical Mathematics* (Teubner, Leipzig, 2001)
52. A. Kunoth, Wavelet techniques for the fictitious domain—Lagrange multiplier approach. Numer. Algorithm **27**, 291–316 (2001)
53. A. Kunoth, Fast iterative solution of saddle point problems in optimal control based on wavelets. Comput. Optim. Appl. **22**, 225–259 (2002)
54. A. Kunoth, Adaptive wavelet methods for an elliptic control problem with Dirichlet boundary control. Numer. Algorithm **39**(1–3), 199–220 (2005)
55. A. Kunoth, Multilevel preconditioning for variational problems, in *Isogeometric Analysis and Applications*, ed. by B. Jüttler, B. Simeon. Lecture Notes in Computational Sciences and Engineering (Springer, 2014), pp. 247–281
56. A. Kunoth, J. Sahner, Wavelets on manifolds: an optimized construction. Math. Comput. **75**, 1319–1349 (2006)
57. A. Kunoth, Chr. Schwab, Analytic regularity and GPC approximation for control problems constrained by parametric elliptic and parabolic PDEs. SIAM J. Control. Optim. **51**(3), 2442–2471 (2013)

58. A. Kunoth, Chr. Schwab, Sparse adaptive tensor Galerkin approximations of stochastic pde-constrained control problems. SIAM/ASA J. Uncertain. Quantif. **4** (2016), 1034–1059
59. J.L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations* (Springer, Berlin, 1971)
60. J. Maes, A. Kunoth, A. Bultheel, BPX-type preconditioners for 2nd and 4th order elliptic problems on the sphere. SIAM J. Numer. Anal. **45**(1), 206–222(2007)
61. P. Oswald, On discrete norm estimates related to multilevel preconditioners in the finite element method, in *Constructive Theory of Functions*, ed. by K.G. Ivanov, P. Petrushev, B. Sendov. Proceedings of International Conference Varna 1991 (Bulgarian Academy of Sciences, Sofia, 1992), pp. 203–214
62. R. Pabel, *Wavelet Methods for PDE Constrained Control Problems with Dirichlet Boundary Control* (Shaker, Maastricht, 2007). https://doi.org/10.2370/236_232
63. R. Pinnau, G. Thömmes, Optimal boundary control of glass cooling processes. Math. Methods Appl. Sci. **27**, 1261–1281 (2004)
64. J. Sahner, On the optimized construction of wavelets on manifolds, Diploma thesis (in English), Universität Bonn, Bonn, September 2003
65. L.L. Schumaker, *Spline Functions: Basic Theory*, 3rd edn. (Cambridge Mathematical Library, Cambridge University Press, Cambridge, 2007)
66. Chr. Schwab, R. Stevenson, Space-time adaptive wavelet methods for parabolic evolution equations. Math. Comput. **78**, 1293–1318 (2009)
67. R. Stenberg, On some techniques for approximating boundary conditions in the finite element method. J. Comput. Appl. Math. **63**, 139–148 (1995)
68. R. Stevenson, Locally supported, piecewise polynomial biorthogonal wavelets on non-uniform meshes. Constr. Approx. **19**, 477–508(2003)
69. W. Sweldens, The lifting scheme: a construction of second generation wavelets. SIAM J. Math. Anal. **29**, 511–546 (1998)
70. J. van den Eshof, G.L.G. Sleijpen, Inexact Krylov subspace methods for linear systems. SIAM J. Matrix Anal. Appl. **26**, 125–153 (2004)
71. H. Yserentant, On the multilevel splitting of finite element spaces. Numer. Math. **49**, 379–412 (1986)
72. E. Zeidler, *Nonlinear Functional Analysis and its Applications; III: Variational Methods and Optimization* (Springer, New York, 1985)