A UNIFIED FRAMEWORK FOR ADAPTIVE BDDC*

CLEMENS PECHSTEIN† AND CLARK R. DOHRMANN‡

Abstract. In this theoretical study, we explore how to automate the selection of weights and primal constraints in BDDC methods for general SPD problems. In particular, we address the three-dimensional case and non-diagonal weight matrices such as the deluxe scaling. We provide an overview of existing approaches, show connections between them, and present new theoretical results: A localization of the global BDDC estimate leads to a reliable condition number bound and to a local generalized eigenproblem on each glob, i.e., each subdomain face, edge, and possibly vertex. We discuss how the eigenvectors corresponding to the smallest eigenvalues can be turned into generalized primal constraints. These can be either treated as they are or (which is much simpler to implement) be enforced by (possibly stronger) classical primal constraints. We show that the second option is the better one. Furthermore, we discuss equivalent versions of the face and edge eigenproblem which match with previous works and show an optimality property of the deluxe scaling. Lastly, we give a localized algorithm which guarantees the definiteness of the matrix $\bar{S}$ underlying the BDDC preconditioner under mild assumptions on the subdomain matrices.

Key words. preconditioning, domain decomposition, iterative substructuring, BDDC, FETI-DP, primal constraints, adaptive coarse space, deluxe scaling, generalized eigenvalue problems, parallel sum

AMS subject classifications. 65F08, 65N30, 65N35, 65N55

1. Introduction. The method of balancing domain decomposition by constraints (BDDC) [20] (see [19, 34] for closely related methods) is, together with the dual-primal finite element tearing and interconnecting (FETI-DP) method [33], among the most-advanced non-overlapping domain decomposition methods for partial differential equations. The two methods can be considered as dual to each other, and for symmetric positive definite (SPD) problems, the corresponding preconditioned operators have identical spectrum (up to values of 1 and 0) [12, 72, 75, 77].

For a variety of PDEs discretized by the finite element method, the condition number of the preconditioned system can be bounded by $C(1 + \log(H/h))^2$, where $H/h$ is the maximal ratio of the subdomain diameter and the element size. Covered cases are scalar diffusion problems [59, 63, 74, 81], linear elasticity [62] as well as positive definite problems in $H(\text{curl})$ [14, 18, 24, 125] and $H(\text{div})$ [88, 89]. Beyond the SPD case, algorithms and theory have been extended to certain saddle point problems such as Stokes flow [48, 71], almost incompressible elasticity [36, 60, 91], Reissner-Mindlin plates [69], and porous media flow [104, 107, 114, 117]. The same kind of bound has been obtained for spectral elements [90], boundary elements [92, 93], mortar methods [43, 44], discontinuous Galerkin [16, 18, 28, 29, 103], and isogeometric analysis [7, 8, 9, 40, 64]. Without giving a list of further references, we note that BDDC and FETI-DP were successfully applied to many more problems, mostly of a mechanical type. Preconditioners based on a Schur complement approximation similar to BDDC were recently proposed by Kraus et al. [65] and Schöberl [102].

The constant $C$ in the bound is usually independent of the subdomain diameters and mesh sizes and thus also of the number of subdomains, which is necessary for scalability.

—Received June 23, 2016. Accepted May 23, 2017. Published online on July 24, 2017. Recommended by S. C. Brenner.

†CST–Computer Simulation Technology AG, a Dassault Systèmes company, Bad Nauheimer Str. 19, 64289 Darmstadt, Germany (formerly: Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenberger Str. 69, 4040 Linz, Austria), (clemens.pechstein@3ds.com).

Ideally, $C$ is also independent of problem parameters, typically coefficient jumps [36, 63, 96], coefficient ratios [24], or geometry details [14, 59]. As shown in [63, 75, 77], at least for SPD problems, part of the analysis is problem-independent, and the condition number estimate reduces to a single norm estimate of a projection operator ($P_D$). For a given decomposition into subdomains, this estimate is influenced by two sources: (i) the weights/scalings and (ii) the primal constraints.

(i) Several scalings have been used in the literature. The multiplicity scaling is not robust for coefficient jumps. A coefficient-dependent scaling, sometimes called $\rho$-scaling, based on constant values per vertex/edge/face leads to robustness for coefficient jumps between subdomains. The stiffness scaling takes more information into account and may look promising but can lead to very high condition numbers in the case of irregular meshes [59] or mildly varying coefficients [93, 98]. A trade off “between” the latter two for jumps along interfaces has been proposed in [93, 98]; see also [94]. All the scalings above involve diagonal weight matrices. The deluxe scaling introduced in [23] (for early experiments see also [22]) breaks with this rule as the weights are dense matrices per subdomain edge/vertex. For subdomain faces, it was observed several times that the deluxe scaling can lead to very good results [8, 18, 24, 54, 69, 88]. Computationally economic versions are discussed in [24, 53].

(ii) The selection of good primal constraints is not an easy task either. On the one hand, choosing too few constraints leads to poor performance of the preconditioner [113, Algorithm A]. On the other hand, choosing too many constraints results in a large coarse problem, which leads to a computationally inefficient method. Although large coarse problems can be alleviated using multiple levels [78, 108, 115, 116], it is better to keep the coarse problem size at a necessary minimum. For scalar diffusion and linear elasticity with coefficients that are constant in each subdomain, good selection algorithms are available; see [113] as well as [105] and the references therein. For hard problems with varying coefficients or coefficient jumps along subdomain interfaces, these recipes may happen to work but can also easily lead to poor performance [30, 57, 67, 96] (see [94, 97, 98] for the classical FETI method). This has led to problem-adapted algorithms for choosing primal constraints, called adaptive BDDC/FETI-DP, which we discuss in the following. Although the adaptive choice means more computational work, this can pay off in highly parallel regimes, where local operations are expected to be comparably cheap [51, 123, 124].

Mandel and Sousedík [76] were the first to investigate, for general diagonal scalings, the influence of primal constraints under quite general assumptions on SPD problems and in an algebraic framework. They came up with a condition number indicator which is based on a local estimate per closed face $F$, reading

$$\sum_{i \in N_F} |\Xi_{iF}(P_D w_i)|_{S_i}^2 \leq \omega_F \sum_{i \in N_F} |w_i|_{S_i}^2.$$

Here, $N_F$ is the set of subdomains shared by face $F$, $\Xi_{iF}$ extracts the degrees of freedom on $F$, the projection $P_D$ will be defined below, $| \cdot |_{S_i}$ is the subdomain (semi)norm, and the estimate must hold for all functions $w$ in the broken space $W$ vanishing on all but the subdomains in $N_F$ and satisfying all primal constraints between these subdomains. The best constant $\omega_F$ is the maximal eigenvalue of an associated generalized eigenproblem and as such computable. The maximum of all indicators $\omega_F$ turned out to be quite reliable for some practical applications. The eigenvectors corresponding to the largest eigenvalues can also be used to create new, adaptive constraints in order to enhance the condition number. Together with Šístek, this approach was extended to three-dimensional problems [79, 108].

The idea of replacing difficult local estimates by local generalized eigenproblems has been used before, e.g., in smoothed aggregation multigrid [13], balancing Neumann-Neumann methods [11], or spectral AMGe [17]. More recently, this technique has been used in over-
lapping Schwarz methods [26, 31, 32, 35, 85, 109, 120]; see also the recent monograph [25]. Spillane and Rixen [110] have employed it for the classical FETI method; see also [38]. Kraus, Lymbery, and Margenov [65] use a similar idea in the context of the additive Schur complement approximation. Other works on BDDC and FETI-DP will be mentioned below.

There are four limitations of the method in [76, 79, 108]:

(a) The theory considers only diagonal scaling matrices.
(b) In the original works, the local bounds are only indicators and were not (yet) proven to be reliable.
(c) Primal constraints in BDDC and FETI-DP are usually linear conditions between functions on two different subdomains involving the degrees of freedom of a glob, i.e., a vertex, an open edge, or an open face. The eigenvectors corresponding to the largest eigenvalues of the generalized eigenproblem associated with \( F \) above, however, typically involve the degrees of freedom on the closed face \( F \). One possibility is to split the eigenvectors and create new, adaptive constraints on the corresponding open face \( F \) and the edges forming its boundary. This can possibly lead to unnecessary constraints. Another possibility (actually the one suggested in [79]) is to disregard the conditions on the face boundary, but this is not supported theoretically.
(d) It is assumed that the initial set of primal constraints already controls the kernel of the underlying PDE such as the rigid body modes of elasticity; this is needed to realize the (formal) matrix inverse \( \hat{S}^{-1} \) in the BDDC preconditioner. It would be good if the local eigenproblems could even detect these kernels and guarantee that \( \hat{S} \) is definite.

Issue (b) has only been resolved quite recently. In [54], Klawonn, Radtke, and Rheinbach show that for two-dimensional problems, where all vertices are chosen primal, the maximum of all indicators \( \omega_\tau \) serves as a reliable condition number bound up to a benign factor. In that work, more general scaling matrices are also considered. In their recent article [49], Klawonn, Kühn, and Rheinbach show a reliable condition number bound for general three-dimensional problems, where all vertices are chosen primal, using a diagonal scaling matrix. Up to a benign factor, the bound is the maximum over all the indicators \( \omega_\tau \) and some additional indicators associated with those subdomain edges that share four or more subdomains. To guarantee the reliability, the obtained face constraints are split into face and edge constraints as described above. The authors also provide some recipes on how the additional work for the edge indicators can be minimized.

A suggestion to resolve issue (d) for the Poisson equation and linear elasticity was recently presented in [6] involving perturbed operators in the BDDC preconditioner which are guaranteed to be invertible while not degrading the condition number too much.

In our article, we briefly review the new approach in [49] and show that it can be equally obtained from a pair-based localization of the \( P_D \) estimate. In the main part of our work, however, we take a different path and provide a similar framework as in [79] but using a glob-based localization. Here, a glob is an open subdomain face, edge, or possibly vertex. On each glob \( G \), we define an indicator \( \omega_G \) associated with the local estimate

$$
\sum_{i \in \mathcal{N}_G} |\Xi_G(P_D w_i)|^2_{S_i} \leq \omega_G \sum_{i \in \mathcal{N}_G} |w_i|^2_{S_i}.
$$

Here, \( \mathcal{N}_G \) is the set of subdomains shared by \( G \), \( \Xi_G \) extracts the degrees of freedom on \( G \), and the estimate must hold for all functions \( w \) in the broken space \( W \) vanishing on all but the subdomains \( \mathcal{N}_G \) and with all primal constraints enforced between \( w_i \), \( w_j \), for \( i, j \in \mathcal{N}_G \). The best local indicator \( \omega_G \) can again be obtained by a generalized eigenproblem, and the corresponding eigenvectors associated with the smallest eigenvalues can be used to create adaptive constraints. Solutions are given to all of the above issues:
(a) We allow general scaling matrices that only need to be block-diagonal with respect to the partitioning into globs.

(b) Up to a benign factor, the maximum over all indicators $\omega_G$ serves as a reliable computable condition number bound.

(c) The constraints on open faces need not be split and can be used as they are. The eigenvectors obtained on subdomain edges, however, are not in the usual form of primal constraints. We show that we can use them as they are (thereby generalizing the notion of primal constraints), or convert them to classical primal constraints, which is more efficient and fully supported by theory.

(d) The local eigenproblems stay well-defined even if the set of initial primal constraints is empty. Under mild assumptions on the subdomain matrices, we can show that using essentially the eigenvectors corresponding to zero eigenvalues as primal constraints guarantees that the inverse $\tilde{S}^{-1}$ appearing in the BDDC preconditioner exists. Our approach is different from [6] and more general.

In the following, we would like to comment on other approaches to this problem. A first group of papers considers two-dimensional problems, where all vertices are a priori chosen as primal. On subdomain faces (there called edges), generalized eigenvalue problems (and sometimes analytic estimates) are used to adaptively choose additional primal constraints.

- Klawonn, Radtke, and Rheinbach [50, 53] consider scalar diffusion and compressible elasticity with discontinuous coefficients discretized by $P^1$ finite elements. They propose to use three generalized eigenproblems per face,

$$
S_{iF}^* v = \lambda M_{iF} v, \quad S_{jF}^* v = \lambda M_{jF} v, \quad S_{iF}^* v = \lambda \hat{\rho}_i S_{iF}^* v,
$$

where $\hat{\rho}_k$ is the maximal coefficient on the subdomain $k$ and $M_{kF}$ is a scaled mass matrix. The discrete Sobolev inequality $|v|_{S_{kF}}^2 \leq C_1 |v|_{S_{kF}}^2 + C_2 |v|_{M_F}^2$ completes the theory and leads to a reliable method for scalar diffusion and linear elasticity with varying coefficients. The authors use a coefficient-dependent scaling based on the values $\hat{\rho}_k$, similar to the $\rho$-scaling.

- Chung and Kim [45] have worked out a fully algebraic approach (though limited to two-dimensional problems). They propose to use two eigenproblems per face,

$$
(S_{iF} + S_{jF}) v = \lambda (S_{iF}^* + S_{jF}^*) v, \quad S_{iF}^* v = \lambda S_{iF}^* v.
$$

General scalings are allowed, but the condition number bound depends on the norm of the scaling matrices. For the multiplicity and the deluxe scaling, this norm is bounded by 1.

In both approaches, in contrast to [76, 79], several (simpler) eigenproblems/estimates are combined. Moreover, the influence of the primal constraints on the neighboring vertices (on $\partial F$) are not included in the local eigenproblems. These two issues raise the question whether the obtained primal constraints are really necessary, or in other words, whether the local bound is efficient; see also [53, 119].

In our approach, we follow Mandel and Sousedík [76] and use a natural eigenproblem that directly follows from the localization (1.1) of the global $P_D$ estimate. This eigenproblem
involve unknowns on all subdomains shared by the glob, i.e., for a face about twice as many as for the eigenproblems (1.2)–(1.3). Here, the (good) influence of a priori chosen primal does on neighboring globs can (but need not) be included. Disregarding them leads to a much simpler implementation, but including them can reduce the number of primal constraints needed for a desired condition number bound. Besides that, we have collected a number of abstract tools for modifying/simplifying generalized eigenproblems.

Intermediate steps of our work are documented in the form of slides [22, 95]. In [95], we show that for the deluxe scaling, on each subdomain face $F$ shared by the subdomains $i$ and $j$, one can alternatively use the generalized eigenproblem
\begin{equation}
(S^*_i : S^*_j) v = \lambda(S_{iF} : S_{jF}) v \tag{1.4}
\end{equation}
where the colon $:$ denotes the parallel sum of matrices introduced by Anderson and Duffin [3]. This idea has recently been discussed in a publication by Klawonn, Radtke, and Rheinbach [54], comparing three different methods for the two-dimensional case: the method by Mandel and Sousedík [76], their own approach [53], and our intermediate approach [95] involving the parallel sum, for which they propose a variant for general scalings,
\begin{equation}
(S^*_i : S^*_j) v = \lambda(D_{iF}^T S_{iF} D_{jF} + D_{jF}^T S_{jF} D_{iF}) v, \tag{1.5}
\end{equation}
where $D_{iF}$ are the face scaling matrices. A sound theory for all three cases is given but limited to the two-dimensional case. Moreover, economic variants are proposed, where $S_{iF}, S^*_i, S^*_j$ etc. are replaced by matrices where not all subdomain degrees of freedom are eliminated but only those at a certain distance from the face $F$. Kim, Chung, and Wang [46, 47] have also compared the method by Chung and Kim [45] with (1.5). Zampini [122, 123, 124] as well as Calvo and Widlund [15, 119] have experimented with (1.4) too and give suggestions for the three-dimensional case.

In our current paper, we show a new theoretical link: if one disregards the influence of neighboring globs, then the natural generalized eigenproblem corresponding to (1.1) on face $G = F$ shared by the subdomains $i$ and $j$ is equivalent to (1.5). In case of the deluxe scaling, (1.5) is identical to (1.4). Moreover, we show that the deluxe scaling minimizes the matrix trace of the left-hand side matrix in (1.5), which is in favor of making the eigenvalues larger. Whereas in [95], we have used the parallel sum as an auxiliary tool, our new minimizing result shows that it is really encoded into BDDC.

The three-dimensional case including subdomain edges has turned out to be a particularly hard problem. For simplicity, consider an edge $E$ shared by three subdomains $i, j, k$. Calvo and Widlund [15, 119] suggest to use
\begin{equation}
(S^*_i : S^*_j : S^*_k) v = \lambda(T_{iE} + T_{jE} + T_{kE}) v \tag{1.6}
\end{equation}
in the context of deluxe scaling, where $T_{iE} = S_{iE} : (S_{jE} + S_{kE})$. Kim, Chung, and Wang [46, 47] give a choice for general scalings:
\begin{equation}
(S^*_i : S^*_j : S^*_k) v = \lambda(A_{iE} + A_{jE} + A_{kE}) v, \tag{1.7}
\end{equation}
where $A_{iE} = D_{jE}^T S_{iE} D_{jE} + D_{kE}^T S_{iE} D_{kE}$. We provide two alternatives. Firstly, one can use the natural edge eigenproblem, optionally simplified by discarding the primal constraints on neighboring globs. We then show how to use the eigenvectors obtained as constraints in the BDDC algorithm. Secondly, we show that with further simplifications, the natural eigenproblem can be decoupled into $n-1$ independent eigenproblems where $n$ is the number of subdomains shared by the edge. When recombining the decoupled problems, one obtains (1.7) in general and (1.6) in case of the deluxe scaling.
Let us note that Stefano Zampini has experimented with
\[
(S_{IE}^* : S_{JE}^* : S_{KE}^*)v = \lambda (S_{IE} : S_{JE} : S_{KE})v,
\]
which behaves robustly for some \( H(\text{curl}) \) problems [125], but a theoretical validation is yet missing (and we do not show any).

Apparently, the eigenproblems (1.6) and (1.7) are simpler than the natural one corresponding to (1.1), but the primal constraints resulting from (1.6), (1.7) may be unnecessary. Vice versa, the natural eigenproblem corresponding to (1.1) will lead to efficient constraints but is more complicated to compute. Our decoupled choice is in between.

Note that for all the eigenproblems considered, we show how initially chosen primal constraints on the respective glob \((G, F, E)\) can be built in. Essentially, the eigenproblems have to be projected onto the space where the initial constraints hold.

We hope that our theoretical study will contribute to a better understanding of the proposed methods and the links between them and to help identify a good trade-off between (a) the more efficient but also more complicated “natural” eigenproblems and (b) simpler eigenproblems that potentially lead to unnecessary constraints but are easier to compute.

The remainder of this paper is organized as follows: In Section 2 we discuss the problem setting, the BDDC preconditioner, an abstract theory for the condition number, and primal constraints on globs. Section 3 provides a localization of the global \( P_D \) estimate under mild assumptions on the weight/scaling matrices. Moreover, we localize the condition for \( S \) to be definite. The local estimate is turned into an eigenproblem, which is discussed in detail in Section 4. Section 5 is devoted to the choice of the adaptive constraints for both the face and edge eigenproblems. Section 6 discusses the deluxe scaling and its optimality property. In Section 7 we combine the local definiteness condition from Section 3 and some abstract results from Section 4 to show how in practice and under mild assumptions on the subdomain matrices, the global definiteness of \( S \) can be guaranteed. An appendix contains auxiliary, technical results.

Our paper is meant to be comprehensive and self-contained. To get an overview, we recommend to skip the sections marked with an asterisk (\(*\)) for the first time. Experienced BDDC readers may initially skip Section 2 as well.

Some Notation: \( X^* \) denotes the algebraic dual of the finite-dimensional (real) vector space \( X \). We always identify the bi-dual \( X^{**} \) with \( X \). For an Euclidean space \( X = \mathbb{R}^n \), we even identify \( X^* \) with \( X \). For a linear operator \( A : X \rightarrow Y \), the transpose \( A^T : Y^* \rightarrow X^* \) is given by \( \langle y^* , Ax \rangle = \langle A^T y^* , x \rangle \) for \( x \in X \), \( y^* \in Y^* \), where \( \langle \cdot , \cdot \rangle \) are the dual pairings. A linear operator \( A : X \rightarrow X^* \) (with \( X \) finite-dimensional) is said to be symmetric if \( \langle Ax , y \rangle = \langle A^T y , x \rangle \) for all \( x , y \in X \), positive semi-definite if \( \langle Ax , x \rangle \geq 0 \) for all \( x \in X \), and positive definite if \( \langle Ax , x \rangle > 0 \) for all \( x \in X \setminus \{0\} \). Symmetric and positive semi-definite (SPSD) operators \( A, B : X \rightarrow X^* \) have the following properties, which we will use frequently:

(i) \( \langle Ax , x \rangle = 0 \iff x \in \ker(A) \),
(ii) \( \ker(A + B) = \ker(A) \cap \ker(B) \),
(iii) \( \text{range}(A + B) = \text{range}(A) + \text{range}(B) \),
(iv) \( \|x\|_A := \langle Ax , x \rangle^{1/2} \) is a semi-norm on \( X \).

If \( P : X \rightarrow X \) is a projection \( (P^2 = P) \), then \( X = \ker(P) \oplus \text{range}(P) \), where \( \oplus \) denotes the direct sum. Moreover, \( (I - P) \) is a projection too, and the identities \( \ker(I - P) = \text{range}(P) \) and \( \text{range}(I - P) = \ker(P) \) hold. Product spaces are denoted by \( V_1 \times \cdots \times V_N \) or \( \otimes_{i=1}^N V_i \).

## 2. BDDC in an algebraic setting

In this section, we summarize the main components of the BDDC method and fix the relevant notation. For the related FETI-DP method, see Appendix B. We give abstract definitions of globs (equivalence classes of degrees of freedom), classical primal constraints, and generalized primal constraints.
2.1. Problem setting. We essentially follow the approach and notation in [75] and require quite minimal assumptions. The problem to be solved is the system of linear equations

$$\text{find } \hat{u} \in U : \quad R^T S R \hat{u} = R^T g,$$

where

$$S = \begin{bmatrix} S_1 & 0 \\ \vdots & \ddots \\ 0 & S_N \end{bmatrix}, \quad g = \begin{bmatrix} g_1 \\ \vdots \\ g_N \end{bmatrix}, \quad R = \begin{bmatrix} R_1 \\ \vdots \\ R_N \end{bmatrix},$$

with SPSD matrices $S_i$. The assembled system matrix $S$ is assumed to be definite such that (2.1) has a unique solution. Let $W_i$ be the (real) Euclidean space of subdomain (interface) degrees of freedom (dofs) and $U$ the Euclidean space of global (interface) dofs such that

$$R_i : U \rightarrow W_i, \quad R : U \rightarrow W := W_1 \times \ldots \times W_N,$$

$$S_i : W_i \rightarrow W_i, \quad S : W \rightarrow W.$$

For an illustration see also Figure 2.3 (left). We simply call the indices $i = 1, \ldots, N$ subdomains. Each matrix $R_i$ corresponds to a local-to-global mapping

$$g_i : \{1, \ldots, \dim(W_i)\} \rightarrow \{1, \ldots, \dim(U)\}$$

and $(R_i)_{\ell k} = 1$ if and only if $k = g_i(\ell)$ (the local dof $\ell$ on the subdomain $i$ corresponds to the global dof $k$), and 0 otherwise. We assume that each mapping $g_i$ is injective. Therefore, $R_i$ has full row rank, and we conclude that

$$R_i R_i^T = I, \quad R_i^T R_i = \text{diag}(\mu_k^{(i)}), \quad \mu_k^{(i)} \in \{0, 1\}.$$

Moreover, $R^T R = \text{diag}(\mu_k^{\text{dim}(U)})$ with $\mu_k = \sum_{i=1}^N \mu_k^{(i)}$ being the multiplicity of the dof $k$. We assume throughout that $\mu_k \geq 2$ for all $k$, which implies in particular that $R$ has full column rank and the subspace

$$\widetilde{W} := \text{range}(R)$$

is isomorphic to $U$.

**Remark 2.1.** Let us note that the assumptions we made so far are fulfilled not only for the standard discretizations of the Poisson equation or compressible linear elasticity. Merely, all our definitions, assumptions (with the exception of Section 7), and the derived theoretical results hold for a series of SPD problems in $H^p(\text{curl})$ [14, 24, 125] and $H^p(\text{div})$ [88, 89], spectral elements [90], boundary elements [92, 93], mortar discretizations [43, 44], discontinuous Galerkin discretizations [16, 18, 28, 29, 40, 103], or isogeometric analysis [7, 8, 9, 40, 64].

**Remark 2.2.** Typically, the matrices $S_i$ are constructed from (larger) subdomain finite element stiffness matrices $A_i$ based on a non-overlapping domain decomposition (e.g., using a graph partitioner) by the (formal) static condensation of non-shared dofs. For the corresponding BDDC preconditioner for the non-condensed system, see, e.g., [79]. We stress that the

---

1Note that $R^T$ in (2.1) actually maps $W^*$ to $U^*$ and assembles local contributions to the global residual (i.e., a functional), whereas $R_i^T$ in (2.2) plays a different role as it extends a function in $W_i$ to $U$ by putting all dofs to 0 that are not shared by subdomain $i$. 
matrices $A_i$ themselves need not be SPSD but only their Schur complements $S_i$, e.g., as in [71, 91]. Otherwise, the theory in this paper (and most of the available, relevant literature) is based heavily on energy arguments, and in this sense, the SPSD condition on $S_i$ can (so far) not be relaxed.

**Remark 2.3.** The assumption that each dof is at least shared by two subdomains is purely to simplify our presentation. All our results can be generalized to the case $n_k \geq 1$, which is, e.g., convenient for BETI [68]. Moreover, we could allow that $R_i$ is rank-deficient and assume that $R_j R_i^T$ is diagonal with ones and zeros. Then, however, some formulas would require adaptations. Such “phantom dofs” appear in the TFETI method [27]. See also [93] for both cases.

**2.2. The BDDC preconditioner.** There are two main ingredients for the BDDC preconditioner. The first one is the averaging operator

$$E_D : W \rightarrow U, \quad E_D w := \sum_{i=1}^N R_i^T D_i w_i,$$

where $D_i : W_i \rightarrow W_i$ are weight matrices that need not be SPSD but fulfill the following condition.

**Condition 2.4 (partition of unity).**

$$\sum_{i=1}^N R_i^T D_i R_i = I \quad \text{(or equivalently } E_D R = I).$$

**Proposition 2.5.** Under Condition 2.4, $\text{range}(E_D) = U$, and $R E_D : W \rightarrow W$ is a projection onto $\tilde{W}$.

**Proof.** We have

$$U \supset \text{range}(E_D) \supset \text{range}(E_D R) = U, \quad \text{and} \quad (R E_D)^2 = R E_D R E_D = R E_D.$$ 

Finally, $\text{range}(R E_D) = R(\text{range}(E_D)) = \text{range}(R) = \tilde{W}$. □

The simplest weights are given by the multiplicity scaling, $D_i = \text{diag}(1/\mu_{g_i}(\ell))_{\ell=1}^{\dim(W_i)}$, where $g_i(\ell)$ is the global dof corresponding to the local dof $\ell$ on the subdomain $i$. In some papers ([75, p. 180], [72, 79]), the weight matrices $D_i$ are assumed to be diagonal with positive entries. In the current paper, we allow more general weights (see Condition 3.4 below). A special choice, the deluxe scaling, is discussed in Section 6.

The second ingredient is an intermediate subspace $\tilde{W}$ that fulfills the following conditions.

**Condition 2.6.**

$$\tilde{W} \subset \tilde{W} \subset W.$$ 

**Condition 2.7.** $S$ is definite on $\tilde{W}$ (ker($S$) $\cap$ $\tilde{W}$ = $\{0\}$).

The construction of $\tilde{W}$ is further described in Section 2.5.2 below. Condition 2.7 is needed for both the practical application of the BDDC preconditioner and its analysis, and it will be further discussed in Section 3.1 as well as in Section 7. Let

$$\tilde{I} : \tilde{W} \rightarrow W$$

denote the natural embedding operator and define the restricted operator

$$\tilde{S} := \tilde{I}^T S \tilde{I} : \tilde{W} \rightarrow \tilde{W}^*.$$
Due to Condition 2.7, $\tilde{S}$ is definite and thus has a well-defined inverse. The BDDC preconditioner for problem (2.1) is defined by

$$M_{BDDC}^{-1} := E_D (\tilde{I} \tilde{S}^{-1} \tilde{I}^T) E_D^T : U \rightarrow U.$$ 

If we explicitly choose a basis for $\tilde{W}$, then $\tilde{I}$ and $\tilde{S}$ have matrix representations and $\tilde{S}^{-1}$ can be constructed via a block Cholesky factorization (see e.g. [62, 72]). Depending on the structure of the space $\tilde{W}$, this can cause a loss of sparsity, which leads to inefficient local solvers when using, e.g., nested dissection. The original BDDC method [20] is based on primal dofs (explained in Section 2.5), and it provides an efficient algorithm (Appendix C) to realize $\tilde{I} \tilde{S}^{-1} \tilde{I}^T$ using a change of basis only implicitly and preserving sparsity. A more general construction of the space $\tilde{W}$ (cf. [79]) has certain importance for our work as well and will be investigated in Section 2.6, Section 5.4, and Appendix C.3.

2.3. Abstract analysis. Theorem 2.9 below has been shown several times in the literature (see, e.g., [75, 77]). For its statement we need the projection operator

$$P_D := I - R E_D : W \rightarrow W.$$ 

The following properties can be derived from Proposition 2.5.

**Proposition 2.8.** Under Condition 2.4 and Condition 2.6,

(i) $P_D^2 = P_D$,

(ii) $P_D w \neq 0 \iff w \in \text{range}(R) = \tilde{W}$,

(iii) $P_D w \in \tilde{W} \iff w \in W$, in particular $P_D(\tilde{W}) \subset \tilde{W}$, $\text{range}(P_D) \cap \tilde{W} = P_D(\tilde{W})$.

**Theorem 2.9 ([75, Theorem 5]).** Let the assumptions from Section 2.1 hold, and let Condition 2.4 (partition of unity), Condition 2.6 ($\tilde{W} \subset W$), and Condition 2.7 ($S$ is definite on $\tilde{W}$) be fulfilled. Then

$$\lambda_{\min}(M_{BDDC}^{-1}(\tilde{S})) \geq 1.$$ 

Moreover, the three estimates

(2.4) $|R E_D w|_S^2 \leq \omega |w|_S^2 \quad \forall w \in \tilde{W}$,

(2.5) $|P_D w|_S^2 \leq \omega |w|_S^2 \quad \forall w \in \tilde{W}$,

$$\lambda_{\max}(M_{BDDC}^{-1}(\tilde{S})) \leq \omega$$

are equivalent. Summarizing, (2.5) implies $\kappa(M_{BDDC}^{-1}(\tilde{S})) \leq \omega$.

A proof based on the fictitious space lemma is provided in Appendix A; see also [65].

**Remark 2.10.** In general, the definiteness of $\tilde{S}$ does not follow from (2.4) or (2.5). Consider one global dof ($U = \mathbb{R}$) shared by two subdomains with $S_1 = D_1 = 1$, $S_2 = D_2 = 0$, and $\tilde{W} = W = \mathbb{R}^2$. Then $\tilde{S}$ is singular, but $|P_D w|_S^2 = 0$ and $|R E_D w|_S^2 = |w|_S^2$.

**Remark 2.11.** For a fixed problem matrix $S$ and weight matrices $\tilde{D}_i$, consider two BDDC preconditioners based on some spaces $\tilde{W}^{(1)} \subset \tilde{W}^{(2)}$ (typically meaning that $\tilde{W}^{(1)}$ has more primal constraints than $\tilde{W}^{(2)}$), and let $\lambda_{\max}^{(1)}$, $\lambda_{\max}^{(2)}$ denote the corresponding maximal eigenvalues. Then $\lambda_{\max}^{(1)} \leq \lambda_{\max}^{(2)}$. Since in practice, $\lambda_{\min}$ is close or even equal to 1 [12, 72, 76, 77], we can expect the smaller space (with the larger set of primal constraints) to lead to a smaller condition number.
2.4. Globs. In BDDC and FETI-DP the intermediate space $\tilde{W}$ is described using primal dofs or coarse dofs. In this particular paper, we restrict ourselves to primal dofs that are associated with globs.

**Definition 2.12 (globs).** For each global dof $k \in \{1, \ldots, \dim(U)\}$, we define the set

$$\mathcal{N}_k := \{i = 1, \ldots, N : \mu_k^{(i)} = 1\}$$

of sharing subdomains. The set $\{1, \ldots, \dim(U)\}$ of global dofs is partitioned into equivalence classes, called globs, with respect to the equivalence relation $k \sim k' \iff \mathcal{N}_k = \mathcal{N}_{k'}$. We denote by $\mathcal{G}$ the set of all globs and by $\mathcal{N}_G$ the set of subdomains shared by glob $G$. Finally, we define the set

$$\mathcal{G}_i := \{G \in \mathcal{G} : i \in \mathcal{N}_G\}$$

of globs for the subdomain $i$. If $|\mathcal{N}_G| = 2$, we call $G$ a face, and we denote the set of all faces (of the subdomain $i$) by $\mathcal{F}$ ($\mathcal{F}_i$, respectively).

**Definition 2.13 (glob relationships).** A glob $G_1$ is called an ancestor of $G_2$ if $\mathcal{N}_{G_1} \supseteq \mathcal{N}_{G_2}$, and $G_1$ is called a parent of $G_2$ if $G_1$ is an ancestor of $G_2$ and there is no other glob $G_3$ with $\mathcal{N}_{G_1} \supseteq \mathcal{N}_{G_3} \supseteq \mathcal{N}_{G_2}$. Certainly, a glob can have several parents. If a glob has no parents, we call it a base glob. Two globs $G_1 \neq G_2$ are called neighbors if $|\mathcal{N}_{G_1} \cap \mathcal{N}_{G_2}| \geq 2$, i.e., if they share at least two common subdomains.

Figure 2.1 illustrates these definitions (assuming a relatively fine mesh and a finite element space with node-based dofs such that sets of nodes appear as geometrical sets).

**Remark 2.14.** For general partitions of 3D finite element meshes, e.g., obtained from a graph partitioner, it can be hard to classify globs geometrically, in particular, to distinguish between vertices and edges. For some rules/heuristics, see [55, Section 2], [62, Section 3], [24, Section 5]. For our purposes, such a classification is not needed. The above definition also

\[\text{Note that many different definitions of globs are used in the literature: sometimes globs are geometrical sets}\ [82, 93] \text{and sometimes the set of globs excludes vertices [76].}\]
Fig. 2.2. Illustration of dofs and globs for left: conforming standard $P^1$ finite element discretization, middle: $C^2$-continuous B-spline discretization (grid lines displayed), right: standard $P^1$ discontinuous Galerkin discretization. White rectangles belong to the subdomain vertex, solid dots belong to subdomain edges; a dashed line marks the dofs of a subdomain.

resembles the fat faces/edges/vertices of isogeometric analysis (cf. Figure 2.2 (middle) and [7]), and it is also applicable to discontinuous Galerkin discretizations (cf. Figure 2.2 (right) and [29]). Moreover, the setting is not only limited to two- and three-dimensional problems. Lastly, note that our theory holds for any alternative definition of globs that refines Definition 2.12 in the sense that each glob of Definition 2.12 is a union of the refined globs. For instance, one may want to split a glob if it is not connected geometrically; see also [62, 93, 123, 124].

Definition 2.15. Let $U_G$ denote the Euclidean space of dofs on $G$ (with a fixed numbering). For any $i \in \mathcal{N}_G$, let $R_{iG} : W_i \rightarrow U_G$ be the (zero-one) restriction matrix (of full rank) extracting these dofs such that $R_{iG} R_{iG}^T = I$.

Since $U_G$ has a fixed dof numbering, we conclude that there exists a matrix $\hat{R}_G : U \rightarrow U_G$ such that

$$R_{iG} R_{iG} = \hat{R}_G \quad \forall i \in \mathcal{N}_G, \quad \hat{R}_G \hat{R}_G^T = I \quad \forall G \in \mathcal{G};$$

see also Figure 2.3 (right). Since the globs are disjoint to each other,

$$R_{iG_1} R_{iG_2}^T = \begin{cases} I & \text{if } G_1 = G_2 \in \mathcal{G}_i, \\ 0 & \text{otherwise}. \end{cases}$$

We define the cut-off/filter matrices

$$\Xi_{iG} := R_{iG}^T R_{iG}, \quad \Xi_G := \text{diag}(\Xi_{iG})_{i=1}^N, \quad \hat{\Xi}_G := \hat{R}_G^T \hat{R}_G,$$

which are diagonal matrices with entry 1 if the corresponding dof is on $G$ and 0 otherwise. From the previous definitions and properties we conclude that

$$\Xi_{iG} R_i = R_i \hat{\Xi}_G, \quad \Xi_G R = R \hat{\Xi}_G, \quad \Xi^2_G = \Xi_G, \quad \hat{\Xi}^2_G = \hat{\Xi}_G.$$

By construction, we have the following partitions of unity on $W_i$ and $U$,

$$\sum_{G \in \mathcal{G}_i} \Xi_{iG} = I, \quad \sum_{G \in \mathcal{G}} \hat{\Xi}_G = I,$$

as well as the following characterization of the “continuous” space (cf. [75])

$$\hat{\mathcal{W}} := \text{range}(R) = \{w \in W : \forall G \in \mathcal{G} \forall i, j \in \mathcal{N}_G : R_{iG} w_i - R_{jG} w_j = 0\}.$$
2.5. Primal dofs and the space $\tilde{W}$. Various definitions of primal dofs have been used for FETI-DP \cite{33, 62, 63, 70, 93, 113} and BDDC \cite{20, 75, 76} in the literature. Here, we require that a primal dof must be associated with a glob and is nothing but a linear combination of regular dofs within that glob. In Section 2.5.3 below, we discuss a more general definition of primal dofs and the space $\tilde{W}$ based on closed globs, which we, however, do not use in the main part of our theory.

2.5.1. Classical primal dofs. The following definition is more common in BDDC methods, which is why we term it “classical”; see Section 2.5.3 for a more general definition.

**Definition 2.16.** Classical primal dofs on the open glob $G$ are described by a matrix $Q_J^G : U_G \to U_{\Pi G} := \mathbb{R}^{n_{\Pi G}}$,

where $n_{\Pi G} \geq 0$ is the number of primal dofs associated with a glob $G$. The subspace of $U_G$ where the primal dofs vanish is

$$U_{G \Delta} := \{y \in U_G : Q_J^G y = 0\}.$$ 

We set

$$W_{\Pi i} := \bigotimes_{G \in G_i} U_{\Pi G}, \quad W_{\Pi} := \bigotimes_{i=1}^N W_{\Pi i}, \quad \text{and} \quad U_{\Pi} := \bigotimes_{G \in G} U_{\Pi G} \cong \mathbb{R}^{n_{\Pi}},$$

with $n_{\Pi} = \sum_{G \in G} n_{\Pi G}$ the total number of primal dofs. Analogously to Section 2.1, we can find zero-one matrices

$$R_{\Pi i} : U_{\Pi} \to W_{\Pi i}, \quad R_{\Pi} : U_{\Pi} \to W_{\Pi}, \quad \text{and} \quad R_{\Pi G} : W_{\Pi i} \to U_{\Pi G},$$

and a matrix $\hat{R}_{\Pi G} : U_{\Pi} \to U_{\Pi G}$ such that $R_{\Pi G} R_{\Pi i} = \hat{R}_{\Pi G}$ independent of $i \in N_G$. Let

$$C_i : W_i \to W_{\Pi i}, \quad C_i := \sum_{G \in G_i} R_{\Pi G} Q_J^G R_G,$$

be the matrix evaluating all primal dofs associated with the subdomain $i$, and define the **dual subspaces** \cite{75, 113}

$$W_{i \Delta} := \ker(C_i) = \{w_i \in W_i : \forall G \in G_i : Q_J^G R_G w_i = 0\}, \quad W_{\Delta} := \bigotimes_{i=1}^N W_{i \Delta}.$$ 

**Remark 2.17.** The operators/spaces $R_{\Pi i}, U_{\Pi i}, W_{\Pi i}$ correspond to $R_c, U_c, X$, respectively, from \cite[Section 2.3]{76}. The operator $Q_P$ from \cite{75, 76} reads

$$Q_P^T = \sum_{G \in G} \hat{R}_{\Pi G}^T Q_J^G R_G : U \to U_{\Pi}$$
in our notation. So Definition 2.16 is equivalent to saying that \( Q_G^T \) is block-diagonal with respect to the partitions of (primal) dofs into globs.

The next condition states that the primal dofs on \( G \) are linearly independent. This can always be achieved by a (modified) Gram-Schmidt orthonormalization or, more generally, by a QR factorization [37, Section 5.2].

**Condition 2.18 (linearly independent primal dofs).** For each glob \( G \in \mathcal{G} \), the columns of the matrix \( Q_G \) are linearly independent.

The following condition is needed later on:

**Condition 2.19 (\( C_i \) surjective).**

\[
\ker(C_i^T) = \{0\} \quad \text{for all } i = 1, \ldots, N.
\]

**Proposition 2.20.** Let \( \{Q_G^T\}_{G \in \mathcal{G}} \) be primal dofs in the sense of Definition 2.16. Then Condition 2.18 is equivalent to Condition 2.19.

**Proof.** Recall that \( C_i^T = \sum_{G \in \mathcal{G}_i} R_{iG}^T Q_G R_{iG} \), i.e., \( C_i^T \) is block-diagonal with respect to the partition of \( W_i \) into globs and to the partition of \( W_{i\Pi} \) into \( \{U_{iG}\}_{G \in \mathcal{G}} \). Hence \( C_i^T \) is injective if and only if all the matrices \( \{Q_G\}_{G \in \mathcal{G}} \) are injective. \( \square \)

Some special primal dofs control all dofs on a glob (in applications, these are typically subdomain vertices):

**Definition 2.21 (totally primal glob).** We call a glob \( G \) totally primal if \( Q_G^T \) is injective (typically the identity). The set of globs (for the subdomain \( i \)) which are not totally primal is denoted by \( G^* \) (\( G^*_i \) respectively).

### 2.5.2. The space \( \widehat{W} \).

Following [20, 72, 75, 76], we define the “partially continuous space” \( \widehat{W} \) based on primal dofs.

**Definition 2.22.** For given primal dofs \( \{Q_G^T\}_{G \in \mathcal{G}} \) in the sense of Definition 2.16, we set

\[
\widehat{W} := \{ w \in W : \forall G \in \mathcal{G} \forall i, j \in N_G: Q_G^T (R_{iG} w_i - R_{jG} w_j) = 0 \}.
\]

Obviously, the space above fulfills Condition 2.6, i.e., \( \widehat{W} \subset \widetilde{W} \subset W \). The following characterization can be shown using the properties of the restriction matrices \( R_{i\Pi} \); cf. [75, 76, Section 2.3].

**Proposition 2.23.** If the primal dofs are linearly independent (Condition 2.18), then

\[
\widetilde{W} = \{ w \in W : \exists u_{i\Pi} \in U_{i\Pi} \forall i = 1, \ldots, N: C_i w_i = R_{i\Pi} u_{i\Pi} \}.
\]
The side conditions in (2.14) are called \textit{primal constraints}, and they fulfill two purposes: First, we need enough constraints such that Condition 2.7 holds ($\mathcal{S}$ is invertible). Second, additional constraints may be needed to get a good constant in the bound (2.5) (recall Remark 2.11: the smaller the space $\mathcal{W}$, the (potentially) smaller the constant $\omega$). In particular this is important for 3D problems or parameter-dependent problems. The first purpose is treated in Section 3.1 and in Section 7. The rest of the paper is mainly devoted to the second purpose. Here, one has to take into account that, although a smaller space leads to a better condition number, the amount of coupling within $\mathcal{W}$ should be kept at a minimum, otherwise the algorithm is not efficient. For example, if $\mathcal{W} = \mathcal{W}_1$, then $\mathcal{S}$ (which should actually be cheaper to invert) is the same as the original problem matrix.

Before proceeding, we provide two basic results on the space $\mathcal{W}$. The first one clarifies its dimension.

**Proposition 2.24.** If the primal dofs are linearly independent (Condition 2.18), then $\dim(\mathcal{W}) = n_{\Pi} + \sum_{i=1}^{N} \dim(W_{i\Delta})$.

The second result allows us to write $\mathcal{W}$ as a direct sum of a 	extit{continuous} and a 	extit{discontinuous} space; see also [75, Section 5], [113, Section 6.4].

**Proposition 2.25.** If the primal dofs are linearly independent (Condition 2.18), then

$$\mathcal{W} = \mathcal{W}_\Pi \oplus W_\Delta,$$

where $\mathcal{W}_\Pi = \text{range}(\hat{\Phi}) \subset \mathcal{W}$ is given by the full-rank matrix

$$\hat{\Phi} : U_\Pi \to \mathcal{W}, \quad \hat{\Phi} := RQ_P = R \sum_{G \in \mathcal{G}} \hat{R}_G Q_G \hat{R}_{\Pi G}.$$

Moreover, $\hat{\Phi}_i = C_i^T R_{\Pi_i} = (\sum_{G \in \mathcal{G}_i} \hat{R}_G Q_G R_{\Pi \Pi_i}) R_{\Pi_i}$, so the basis has local support.

**Remark 2.26.** If the primal dofs are orthogonal, i.e., for all $G \in \mathcal{G}$: $Q_G^T Q_G = I$, then $C_i \hat{\Phi}_i = I$. Otherwise, one can redefine $\hat{\Phi}$ to fulfill the latter property; cf. [75, Lemma 9].

### 2.5.3. Primal dofs on closed globs

In some references and implementations, primal dofs are defined on the \textit{closure of globs}; cf. [52, 79, 113].

**Definition 2.27.** The closure $\mathcal{G}$ of a glob $G$ is given by $G$ and all its ancestors, i.e.,

$$\mathcal{G} := \bigcup_{G' \in \mathcal{G}_G^\tau} G', \text{ where } \mathcal{G}_G^\tau := \{G' \in \mathcal{G} : N_{G'} \supseteq N_G\}.$$

Let $U_{\mathcal{G}}$ denote the space of dofs on $\mathcal{G}$ (with a fixed numbering). Analogously to the above, we can find zero-one matrices $R_{\mathcal{G}} : W_i \to U_{\mathcal{G}}$ and $\hat{R}_{\mathcal{G}} : U \to U_{\mathcal{G}}$ extracting these dofs such that $R_{\mathcal{G}} R_i = \hat{R}_{\mathcal{G}}$ independent of $i \in N_G$.

**Definition 2.28.** Primal dofs on the closed glob $\mathcal{G}$ are described by a matrix

$$Q_{\mathcal{G}}^T : U_{\mathcal{G}} \to U_{\Pi G} := \mathbb{R}^{n_{\Pi G}}.$$

The analogous definitions of $C_i : W_i \to W_{\Pi i}$ and $Q_i^T : U \to U_{\Pi i}$ are

$$C_i = \sum_{G \in \mathcal{G}_i} \hat{R}_{\Pi \Pi i} Q_G^T R_{\Pi i}, \quad Q_i^T = \sum_{G \in \mathcal{G}_i} \hat{R}_{\Pi \Pi i} Q_G^L \hat{R}_{\mathcal{G}},$$

and the space $\mathcal{W}$ can now be defined as in (2.15).
Recall that for the classical primal dofs (on “open” globs), the proof of Proposition 2.20 is very simple. For the closed case, an analogue is not presently known. Yet, the following is easily verified:

**Proposition 2.29.** Let \( R_{|G|} := \tilde{R}_G \tilde{R}_G^T \) denote the restriction matrix from the dofs on \( G \) to the (fewer) dofs on the open glob \( G \). If for each \( G \in \mathcal{G} \), the matrix \( R_{\mathcal{G} \mathcal{G}} Q_{\mathcal{G} \mathcal{G}} \) has full column rank, then also \( Q_{\mathcal{G} \mathcal{G}} \) has full column rank (analogous to Condition 2.18).

If \( R_{\mathcal{G} \mathcal{G}} \) has linearly dependent columns, then we can split each primal dof on the closed glob \( G \) into primal dofs on all the open globs \( G' \in \mathcal{G}' \), orthonormalize them together with the existing ones, and finally obtain linearly independent primal dofs on open globs (Condition 2.18). However, to our best knowledge, no algorithm exists to date which gets Condition 2.18 to hold by modifying \( Q_{\mathcal{G} \mathcal{G}} \) without increasing the overall number of primal dofs. See also [79, p. 1819]. This is one of the reasons why we use Definition 2.16.

### 2.6. Generalized primal constraints

Mandel, Sousedík, and Šístek [79] use a more general definition of the space \( \tilde{W} \), which is of central importance to our own work:

\[
\tilde{W} = \{ w \in W : Lw = 0 \},
\]

where \( L : W \to X := \mathbb{R}^M \) is a matrix with \( M \) linearly independent rows. One easily shows that \( \tilde{W} \subset \tilde{W} \subset W \) (Condition 2.6) holds if and only if \( LR = 0 \), or equivalently,

\[
Lw = 0 \quad \forall w \in \tilde{W}.
\]

Apparently, Definition 2.22 (based on the classical primal dofs) is a special case of (2.16) but not vice versa. For the general form (2.16), the application \( y = \tilde{S}^{-1} \tilde{T} \psi \) for \( \psi \in W \) is equivalent to solving the global saddle point problem

\[
\begin{bmatrix}
S & L^T \\
L & 0
\end{bmatrix}
\begin{bmatrix}
y \\
z
\end{bmatrix} = \begin{bmatrix}
\psi \\
0
\end{bmatrix}.
\]

For the special case of \( L \) discussed below, a more viable option is given in Appendix C.

**Remark 2.30.** Actually, for any space \( \tilde{W} \) with \( \tilde{W} \subset \tilde{W} \subset W \) (Condition 2.6), there is a matrix \( L \) such that (2.16)–(2.17) holds. In a FETI-DP framework (see Appendix B), the property (2.17) implies that \( L = LB \) for some \( L \), and thus, such constraints can be implemented by deflation [41, 46, 47, 53, 58]. The balancing Neumann-Neumann method [73] can be interpreted as a BDDC method with (2.16), however, the constraints \( L \) are global; cf. [93, p. 110].

In [79], Mandel et al. require that each constraint (each row of \( L \)) is local to a glob, i.e., for each glob \( G \in \mathcal{G} \), there exist matrices \( L_{jG} : U_G \to X_G, j \in \mathcal{N}_G \), such that

\[
Lw = \sum_{G \in \mathcal{G}} R_{XG} \sum_{j \in \mathcal{N}_G} L_{jG} R_{jG} w_j,
\]

where \( X \) is isomorphic to \( \otimes_{G \in \mathcal{G}} X_G \) and \( R_{XG} : X \to X_G \) is the zero-one matrix extracting the “\( G \)” component. If \( L \) is of form (2.19) then

(i) \( L \) has linearly independent rows if and only if the block row matrix \( \cdots \{L_{jG}\} \cdots \}_{j \in \mathcal{N}_G} \) has linearly independent rows for all \( G \in \mathcal{G} \).

(ii) \( Lw = 0 \) holds if and only if

\[
\sum_{j \in \mathcal{N}_G} L_{jG} R_{jG} w_j = 0 \quad \forall G \in \mathcal{G},
\]
(iii) \( L_w = 0 \) for all \( w \in \widehat{W} \) (Condition (2.17)) if and only if

\[
\sum_{j \in N_G} L_{jG} = 0 \quad \forall G \in \mathcal{G}.
\]

The above form of constraints is important to our study because our localized bounds (implying the central bound (2.5) of the \( P_D \) operator) hold (and are sharp) if constraints of the form (2.19), (2.21) are imposed (in addition to previously fixed primal constraints). In particular, they pop out of local generalized eigenproblems associated with globs that share more than two subdomains and that involve more than just a few dofs such as subdomain edges.

Mandel, Sousedík, and Šístek provide an algorithm for the efficient solution of the global saddle point problem (2.18) based on the multifrontal massively parallel sparse direct solver MUMPS [1]. In Appendix C.3, we give an extension of the algorithm proposed in [20] which realizes \( I \hat{S}^{-1} I^\top \) by solving local saddle point problems and one global (coarse) SPD problem. Under the perspective of the extended algorithm, BDDC with generalized (but still glob-based) primal constraints becomes amenable for multiple levels [56, 78, 108, 116]. This is because the coarse problem is again an SPD problem that can be subassembled from SPSD subdomain contributions. Thus, the subdomains of the second level can be defined as agglomerates of the first level subdomains. Multiple levels can be a rather attractive option for problems with high contrast coefficients [18, 53, 96, 97, 98] and/or a detailed underlying geometry [21, 24, 76]. Nevertheless, as we will show in Section 5.4 below, rather than using the generalized primal constraints, it is much more favorable to use potentially stronger classical primal constraints and the conventional algorithm from Appendix C.1–C.2 (which is naturally amenable to multiple levels). Although our result holds for the general case, we will describe it later in Section 5, when needed.

3. Localization. In this section, we provide a local condition implying the global definiteness of \( \hat{S} \) (Section 3.1, Condition 2.7). After introducing our mild assumptions on the weight/scaling matrices \( D_i \) and showing some technical results in Section 3.2, we provide local estimates implying the global estimate (2.5) of the \( P_D \)-operator in Section 3.3. We also review a similar approach by Klawonn, Kühn, and Rheinbach [49] (Section 3.4). Throughout this section, we assume a space \( \widehat{W} \) based on classical primal dofs (Definition 2.16 and Definition 2.22).

3.1. A local, glob-based condition for the definiteness of \( \hat{S} \). The problem of how to guarantee definiteness of \( \hat{S} \) already arose in the original FETI-DP method [33]. Suitable choices of primal constraints are known for scalar diffusion and linear elasticity problems ([70, 113]). For the general SPD case, however, an all-purpose recipe is yet missing (to our best knowledge). As one can see easily, the definiteness of \( S \) on \( \widehat{W} \) (Condition 2.7) implies the necessary local condition

\[
S_i \text{ is definite on } W_i \Delta \quad \forall i = 1, \ldots, N,
\]

which is, however, not sufficient (see Figure 3.1 for a counterexample).

Condition 3.1 below is local and sufficient (Lemma 3.2) although not necessary. In Section 7, we provide an algorithm that computes a set of primal constraints such that Condition 3.1 can be guaranteed under mild assumptions on the problem. For each glob \( G \in \mathcal{G} \), we define the (Euclidean) space

\[
W_{\mathcal{N}_G} := \{ w = [w_i]_{i \in \mathcal{N}_G} : w_i \in W_i \},
\]
where \([w_i]_{i \in \mathcal{N}_G}\) simply designates a block vector. We denote by \(w_{\mathcal{N}_G} \in W_{\mathcal{N}_G}\) the restriction of \(w \in W\) to the subdomains in \(\mathcal{N}_G\) and define the subspace

\[
\tilde{W}_{\mathcal{N}_G} := \{w \in W_{\mathcal{N}_G} : \exists z \in \widetilde{W}: w_i = z, \forall i \in \mathcal{N}_G\}
\]

\[
= \{w \in W_{\mathcal{N}_G} : \forall i \neq j \in \mathcal{N}_G \forall G', \{i, j\} \subset \mathcal{N}_{G'}: Q_G^\top (R_{iG'}w_i - R_{jG'}w_j) = 0\},
\]

i.e., the space of functions living “around” \(G\), where (previously fixed) primal constraints are enforced on all the neighboring globs of \(G\); cf. Definition 2.13. See Figure 3.2 for a two-dimensional example where \(G\) is a vertex. If \(G\) is a typical edge in a three-dimensional problem, then in addition to the previously fixed constraints on the edge \(G\), also previously fixed constraints on the neighboring vertices, edges, and faces are enforced.

**CONDITION 3.1** (local kernel condition). For each glob \(G \in \mathcal{G}^*\) (i.e., not totally primal), assume that

\[
\forall w \in \tilde{W}_{\mathcal{N}_G}: \quad (\forall i \in \mathcal{N}_G: S_i w_i = 0) \implies (\forall i, j \in \mathcal{N}_G: R_{iG}w_i = R_{jG}w_j).
\]

**LEMMA 3.2.** **Condition 3.1 implies Condition 2.7** \((S\) is definite on \(\tilde{W}\)).

**Proof.** Let Condition 3.1 hold and let \(w \in \ker(S) \cap \tilde{W}\) be arbitrary but fixed. Then \(S_i w_i = 0\) for all \(i = 1, \ldots, N\). Due to Condition 3.1 for all not totally primal globs \(G\),

\[
\forall i, j \in \mathcal{N}_G: R_{iG}w_i = R_{jG}w_j.
\]

On the remaining totally primal globs, we get the same condition from Definition 2.22 and Definition 2.21. So, all dofs are continuous across all globs, and with (2.11), \(w \in \tilde{W}\). Since \(\ker(S) \cap \tilde{W} = \{0\}\) (cf. Section 2.1), \(w = 0\). Summarizing, \(\ker(S) \cap \tilde{W} = \{0\}\). \(\square\)

**REMARK 3.3.** Condition 3.1 is similar to but substantially different from [76, Assumption 8]. The latter reads as follows. For all faces \(F \in \mathcal{F}\),

\[
\forall w \in \tilde{W}_{\mathcal{N}_F}: \quad (S_i w_i = 0, S_j w_j = 0) \implies (R_{iF}w_i = R_{jF}w_j),
\]

where \([w_i]_{i \in \mathcal{N}_F}\) simply designates a block vector. We denote by \(w_{\mathcal{N}_F} \in W_{\mathcal{N}_F}\) the restriction of \(w \in W\) to the subdomains in \(\mathcal{N}_F\) and define the subspace

\[
\tilde{W}_{\mathcal{N}_F} := \{w \in W_{\mathcal{N}_F} : \exists z \in \tilde{W}: w_i = z, \forall i \in \mathcal{N}_F\}
\]

\[
= \{w \in W_{\mathcal{N}_F} : \forall i \neq j \in \mathcal{N}_F \forall G', \{i, j\} \subset \mathcal{N}_{G'}: Q_G^\top (R_{iG'}w_i - R_{jG'}w_j) = 0\},
\]

i.e., the space of functions living “around” \(F\), where (previously fixed) primal constraints are enforced on all the neighboring globs of \(F\); cf. Definition 2.13. See Figure 3.2 for a two-dimensional example where \(F\) is a vertex. If \(F\) is a typical edge in a three-dimensional problem, then in addition to the previously fixed constraints on the edge \(F\), also previously fixed constraints on the neighboring vertices, edges, and faces are enforced.

**CONDITION 3.1** (local kernel condition). For each glob \(F \in \mathcal{G}^*\) (i.e., not totally primal), assume that

\[
\forall w \in \tilde{W}_{\mathcal{N}_F}: \quad (\forall i \in \mathcal{N}_F: S_i w_i = 0) \implies (\forall i, j \in \mathcal{N}_F: R_{iF}w_i = R_{jF}w_j).
\]

**LEMMA 3.2.** **Condition 3.1 implies Condition 2.7** \((S\) is definite on \(\tilde{W}\)).

**Proof.** Let Condition 3.1 hold and let \(w \in \ker(S) \cap \tilde{W}\) be arbitrary but fixed. Then \(S_i w_i = 0\) for all \(i = 1, \ldots, N\). Due to Condition 3.1 for all not totally primal globs \(F\),

\[
\forall i, j \in \mathcal{N}_F: R_{iF}w_i = R_{jF}w_j.
\]

On the remaining totally primal globs, we get the same condition from Definition 2.22 and Definition 2.21. So, all dofs are continuous across all globs, and with (2.11), \(w \in \tilde{W}\). Since \(\ker(S) \cap \tilde{W} = \{0\}\) (cf. Section 2.1), \(w = 0\). Summarizing, \(\ker(S) \cap \tilde{W} = \{0\}\). \(\square\)

**REMARK 3.3.** Condition 3.1 is similar to but substantially different from [76, Assumption 8]. The latter reads as follows. For all faces \(F \in \mathcal{F}\),

\[
\forall w \in \tilde{W}_{\mathcal{N}_F}: \quad (S_i w_i = 0, S_j w_j = 0) \implies (R_{iF}w_i = R_{jF}w_j),
\]
where \( \{i, j\} = N_F \) and \( \overline{F} \) is the closed face (Definition 2.27). Under the additional assumption that for each glob \( G \in \mathcal{G} \setminus F \), one can connect each pair \( \{i, j\} \subset N_G \) via a path through faces (which is fulfilled for usual domain decompositions), one can show that Condition 2.7 holds. Neither Condition 3.1 nor (3.4) are necessary for Condition 2.7 to hold.

### 3.2. Assumption on the weight matrices.

In our subsequent theory, we need the following, mild assumption on the scaling matrices \( D_i \):

**Assumption 3.4 (\( D_i \) block diagonal).** Each scaling matrix \( D_i \) is block diagonal with respect to the glob partition, i.e., there exist matrices \( D_{iG} \):

\[
D_{iG}^T D_{iG} R_{iG} = R_{iG} D_{iG} R_{iG}.
\]

The condition below is a glob-wise partition of unity, and the proposition thereafter is easily verified.

**Condition 3.5 (glob-wise partition of unity).** For each glob \( G \in \mathcal{G} \), there holds

\[
\sum_{j \in N_G} D_{jG} = I.
\]

**Proposition 3.6.** Let Assumption 3.4 hold. Then for all \( G \in \mathcal{G} \) and \( i \in N_G \):

\[
(\ref{eq:3.5}) \quad \Xi_{iG} D_i = D_i \Xi_{iG}, \quad \Xi_{iG} D = D \Xi_{iG}
\]

(where \( D = \text{diag}(D_i) \)). and

\[
(\ref{eq:3.6}) \quad E_D w = \sum_{G \in \mathcal{G}} \sum_{i \in N_G} D_{iG} R_{iG} w_i.
\]

Moreover, Condition 2.4 (partition of unity) is equivalent to Condition 3.5.

**Proof.** Firstly, we show (3.5) and (3.6):

\[
\Xi_{iG} D_i \overset{\text{Ass. 3.4}}{=} \sum_{G \in \mathcal{G}_i} \Xi_{iG} R_{iG}^T D_{iG} R_{iG} \overset{\text{Ass. 3.4}}{=} D_i \Xi_{iG},
\]

\[
E_D w = \sum_{i=1}^N R_i^T D_i w_i \overset{\text{Ass. 3.4}}{=} \sum_{i=1}^N \sum_{G \in \mathcal{G}_i} R_i^T R_{iG}^T D_{iG} R_{iG} w_i = \sum_{G \in \mathcal{G}} \sum_{i \in N_G} D_{iG} R_{iG} w_i.
\]

Secondly, (3.6) implies \( E_D R = \sum_{G \in \mathcal{G}} \sum_{i \in N_G} D_{iG} \hat{R}_{iG} \).

If Condition 2.4 holds, then the left-hand side evaluates to \( I \), and we obtain Condition 3.5 by multiplying from the left by \( \hat{R}_G \) and from the right by \( \hat{R}_G^T \) (for an arbitrary \( G \in \mathcal{G} \) and using (2.6)). Conversely, if Condition 3.5 holds, then the right-hand side evaluates to \( I \) due to (2.10), thus Condition 2.4 is fulfilled.

The following two results will be helpful for Section 3.3.

**Lemma 3.7.** Let Assumption 3.4 (\( D_i \) block diagonal) and Condition 3.5 (glob-wise partition of unity) hold. Then

\[
(i) \quad \Xi_G E_D = E_D \Xi_G, \quad (R E_D \Xi_G)^2 = R E_D \Xi_G, \quad \Xi_G P_D = P_D \Xi_G, \quad (P_D \Xi_G)^2 = P_D \Xi_G.
\]
Applying \( \Xi \) the other assertions follow immediately from (2.9), the fact that (3.7) Lemma 3.7(ii) and the space Definition 2.22. Since (iii) If \( G \) is totally primal (\( G \neq G^* \), cf. Section 2.5.1), then
\[
\Xi_G P_D w = 0 \quad \Longleftrightarrow \quad (\forall i, j \in N_G : R_{iG}w_i = R_{jG}w_j).
\]

Proof. (i) By definition, \( E_D = R^T D \) with \( D = \text{diag}(D_i)_{i=1}^N \). From (2.9), (3.5) we get
\[
\hat{\Xi}_G E_D = \hat{\Xi}_G R^T D = R^T \Xi_G D = R^T D \Xi_G = E_D \Xi_G.
\]
The other assertions follow immediately from (2.9), the fact that \( (RE_D)^2 = RE_D \) (Proposition 3.6 and Proposition 2.5), and the definition of \( P_D \).

(ii) From the definitions of \( E_D \) and \( P_D \) we get
\[
R_{iG}(P_D w)_i = R_{iG}w_i - \hat{R}_G E_D w \quad \forall i \in N_G.
\]

Applying \( \hat{R}_G \) to formula (3.6), we find that
\[
\hat{R}_G E_D w \overset{(2.7)}{=} \sum_{G' \in \mathcal{G}} \hat{R}_G \hat{\Xi}_G \sum_{j \in N_{G'}} D_{jG'} R_{jG}w_j = \sum_{j \in N_G} D_{jG} R_{jG}w_j.
\]

Substituting the latter result into (3.7) yields
\[
R_{iG}^{\top} R_{iG}(P_D w)_i = R_{iG}^{\top} \left( R_{iG}w_i - \sum_{j \in N_G} D_{jG} R_{jG}w_j \right).
\]
The definition of \( \Xi_{iG} \) and Condition 3.5 yield the desired formula.

(iii) If \( G \neq G^* \) and \( w \in \hat{W} \), then \( Q^G_G(R_{iG}w_i - R_{jG}w_j) = 0 \) for all \( i, j \in N_G \); cf. Definition 2.22. Since \( Q^G_G \) is non-singular, \( R_{iG}w_i = R_{jG}w_j \), and Lemma 3.7(ii) implies that \( \Xi_G P_D w = 0 \).

### 3.3. A glob-based localization of the \( P_D \) estimate (2.5).
Recall the formula stated in Lemma 3.7(ii) and the space \( W_{\mathcal{N}_G} \) from (3.2). We define
\[
P_{D,G} : W_{\mathcal{N}_G} \to W_{\mathcal{N}_G} : \quad (P_{D,G} w)_i := R_{iG}^{\top} \sum_{j \in N_G \setminus \{i\}} D_{jG}(R_{iG}w_i - R_{jG}w_j).
\]

**Lemma 3.8.** Let Assumption 3.4 (\( D_i \) block diagonal) and Condition 3.5 (glob-wise partition of unity) hold. Then
\( i \) : \( \Xi_G(P_D w)_i = (P_{D,G} w)_{\mathcal{N}_G}, \forall w \in W \forall i \in N_G \), where \( w_{\mathcal{N}_G} = [w_j]_{j \in N_G} \),
\( ii \) : \( P_{D,G} = P_{D,G} \),
\( iii \) : \( \ker(P_{D,G}) = \{ w \in W_{\mathcal{N}_G} : \forall i, j \in N_G : R_{iG}w_i = R_{jG}w_j \}, \)
\( iv \) : \( P_{D,G} w \in \hat{W}_{\mathcal{N}_G} \Longleftrightarrow w \in \hat{W}_{\mathcal{N}_G} \) with the space \( \hat{W}_{\mathcal{N}_G} \) from (3.3), in particular, \( P_{D,G} \hat{W}_{\mathcal{N}_G} \subset W_{\mathcal{N}_G} \) and \( \text{range}(P_{D,G}) \cap \hat{W}_{\mathcal{N}_G} = P_{D,G} \hat{W}_{\mathcal{N}_G} \),
\( v \) : there exists a projection operator \( \hat{P}_{D,G} : \hat{W}_{\mathcal{N}_G} \to \hat{W}_{\mathcal{N}_G} \) such that \( P_{D,G} I_{\mathcal{N}_G} = I_{\mathcal{N}_G} \hat{P}_{D,G} \), where \( \hat{I}_{\mathcal{N}_G} : \hat{W}_{\mathcal{N}_G} \to W_{\mathcal{N}_G} \) is the natural embedding.
Proof. Part (i) follows from Lemma 3.7(ii) and the definition of \( P_{D,G} \), Part (ii) from Lemma 3.7(i). Part (iii) can be derived using Lemma 3.7(ii). Part (iv): for \( y \in W_{NG} \) and \( w = P_{D,G}y \) one easily shows that
\[
Q_G^T R_i G w_i = Q_G^T R_i G (\Xi_i G y_i - R_{ij G} \bar{y}_G), \quad \text{where} \quad \bar{y}_G = \sum_{j \in NG} D_{ij} R_{ij G} w_j ,
\]
\[
Q_G^T (R_{ij G} w_i - R_{ij G} y_j) = Q_G^T (R_{ij G} y_i - R_{ij G} y_j).
\]
Finally, Part (v) follows from Parts (ii) and (iii). □

Remark 3.9. If \( \bar{W} \) does not originate from primal dofs on open globs (Definition 2.16), then Parts (iii) and (v) do not necessarily hold.

As the next theorem shows, the global bound (2.5) can be established from local bounds associated with individual globs (with \( \bar{W}_{NG} \), \( P_{D,G} \) defined as in (3.3), (3.8), respectively):

Local glob estimate:
\[
(3.9) \quad \sum_{i \in NG} |(P_{D,G} w_i)_{i} |^2_{S_i} \leq \omega_G \sum_{i \in NG} |w_i|^2_{S_i} \quad \forall w \in \bar{W}_{NG}.
\]

Theorem 3.10. Let Assumption 3.4 (\( D_i \) block diagonal) and Condition 3.5 (glob-wise partition of unity) be fulfilled, and let \( \bar{W} \) be defined by classical primal dofs (Definition 2.22). For each glob \( G \in G^* \) (that is not totally primal), assume that the local estimate (3.9) holds with some constant \( \omega_G < \infty \). Then the global \( P_D \)-estimate (2.5) holds with
\[
\omega = \left( \max_{i=1,...,N} |G_i^*| \right) \left( \max_{G \in G^*} \omega_G \right),
\]
where \( |G_i^*| \) denotes the cardinality of the set \( G_i^* \). In particular, if, in addition, \( \bar{S} \) is definite (Condition 2.7), then Theorem 2.9 implies \( \kappa_{BBDC} \leq \omega \).

Proof. Firstly, we use (2.10), Lemma 3.7(iii), and Lemma 3.8(i) to obtain
\[
(P_D w)_i = \sum_{G \in \bar{G}_i} \Xi_{iG}(P_D w)_i = \sum_{G \in \bar{G}_i^*} \Xi_{iG}(P_D w)_i = \sum_{G \in \bar{G}_i^*} (P_{D,G} w_{NG})_i.
\]
Secondly, the Cauchy-Bunyakovsky-Schwarz inequality and the local bounds (3.9) imply
\[
\sum_{i=1}^N |(P_D w)_i |^2_{S_i} \leq \sum_{i=1}^N |G_i^*| \sum_{G \in \bar{G}_i^*} |(P_{D,G} w_{NG})_i |^2_{S_i}
\]
\[
\leq \left( \max_{i=1,...,N} |G_i^*| \right) \sum_{j \in NG} \sum_{G \in \bar{G}_i^*} |(P_{D,G} w)_j |^2_{S_j}
\]
\[
\leq \left( \max_{i=1,...,N} |G_i^*| \right) \sum_{j \in NG} \sum_{G \in \bar{G}_i^*} \omega_G |w_j|^2_{S_j}
\]
\[
\leq \left( \max_{i=1,...,N} |G_i^*| \right) \sum_{i=1}^N \sum_{G \in \bar{G}_i^*} \omega_G |w_i|^2_{S_i}.
\]
Finally, \( \sum_{G \in \bar{G}_i^*} \omega_G \leq \left( \max_{i=1,...,n} |G_i^*| \right) \left( \max_{G \in G^*} \omega_G \right). \) □

The arguments in the proof above are not new and are used in all the known theoretical condition number bounds of FETI, FETI-DP, and BDDC for specific PDEs and discretizations;
see, e.g., [8, 23, 61, 63, 80, 81, 113]. The more recent works [18, 53] make implicitly use of Theorem 3.10, and a similar result for the two-dimensional case can be found in [54, Theorem 5.1].

**Remark 3.11.** If Assumption 3.4 did not hold, i.e., if the matrices $D_i$ were not block-diagonal with respect to the globs, then we would need an estimate of the form

$$
\sum_{i \in N_G} |\Xi_{iG}(P_D w)_i|^2_{S_i} \leq \omega_G \sum_{j \in N_G^+} |w_j|^2_{S_j} \quad \forall w \in \tilde{W},
$$

where $N_G^+$ are the subdomains of $N_G$ and all their next neighbors.

**Remark 3.12.** Certainly, if the local glob estimate (3.9) holds on a larger space than $\tilde{W_{N_G}}$, then we get a similar result (possibly with a pessimistic bound $\omega_G$). A possible choice for such a space is

$$
\tilde{W}_{N_G}^G := \{ w = [w_i]_{i \in N_G} : \forall i \neq j \in N_G : Q_{iG}^T R_i^T w_i = Q_{jG}^T R_j^T w_j \},
$$

i.e., the space of functions living “around” $G$, where only the primal constraints associated with $G$ are enforced. We shall make use of this later in Section 4.2, Strategy 4.

**Remark 3.13.** Whereas the local estimate (3.9) is glob-based, other localizations used in the literature are subdomain-based. For example, translating the suggestion of Kraus et al. [65, Section 5] to our framework leads to the estimate

$$
|(P_D w)_i|^2_{S_i} \leq \omega_i \sum_{j \in N_i} |w_j|^2_{S_j} \quad \forall w \in \tilde{W}_{N_i},
$$

where $N_i$ are the neighboring subdomains of $i$ and $\tilde{W}_{N_i}$ is the restriction of $\tilde{W}$ to these. Another option, related to the work by Spillane and Rixen [110], is

$$
\sum_{j \in N_i} (P_D w)_j|^2_{S_j} \leq \omega_i |w_i|^2_{S_i},
$$

for all $w \in W$ that vanish in all but the $i$-th subdomain.

### 3.4. A review of a pair-based localization

The local estimate (3.9) was first proposed in [76] (see also [79, 106, 108]), however, there in slightly different form on every closed face $F$,

$$
(\Xi_{iF}(P_D w)_i|^2_{S_i} + |\Xi_{jF}(P_D w)_j|^2_{S_j}) \leq \omega_F (|w_i|^2_{S_i} + |w_j|^2_{S_j}),
$$

where $N_F = \{i, j\}$, $\Xi_{iF} := \sum_{G \in F} \Xi_{iG}$ is the filter matrix corresponding to $F$, and (3.10) must hold for all $w \in W$ with $w_k = 0$ for $k \notin \{i, j\}$ and with all primal constraints enforced between $w_i$ and $w_j$. Under Assumption 3.4 ($D_i$ block diagonal) and Condition 3.5 (glob-wise partition of unity), the estimate can be expressed using a space $\tilde{W}_F$ and an operator $P_{D,F}$ defined analogously to $\tilde{W}_{N_G}$ and $P_{D,G}$, respectively. In [76, 79], the local bounds are used to define the condition number indicator

$$
\tilde{\omega} := \max_{F \in \mathcal{F}} \omega_F.
$$

If every glob $G$ is either totally primal or $|N_G| = 2$ (typical for two-dimensional problems), then it does not matter whether one uses the open or closed face, and (2.5) holds with $\omega = \tilde{\omega}$. Thus, $\tilde{\omega}$ is indeed a reliable bound for the condition number; see also [54, Theorem 5.1].

For the three-dimensional case, the reliability of (3.11) was open for quite a long time. In their recent preprint [49], Klawonn, Kühn, and Rheinbach show that in general, (3.11) is
reliable if (i) all vertices are totally primal and (ii) one includes some estimates associated with those subdomain edges that share more than three subdomains. In the following, we present this latest theory under a slightly different perspective.

If Assumption 3.4 (glob-wise partition of unity) holds then

\[(P_D w)_i = \sum_{j \in N_i} \sum_{G \in \mathcal{G}^*_{i,j}} R^T_{iG} D_{jG} (R_{iG} w_i - R_{jG} w_j),\]

where \(N_i := \bigcup_{G \in \mathcal{G}_{i,N}} N_G\) is the set of neighboring subdomains of the subdomain \(i\). This formula motivates a neighbor-based viewpoint and the following definition.

**Definition 3.14** (generalized facet). For each pair \(\{i, j\}, i \neq j\), we define the generalized facet

\[\Gamma_{ij} := \bigcup_{G \in \mathcal{G}^*_{i,j}} G,\]

e.g., the set of dofs shared by the subdomains \(i\) and \(j\), excluding totally primal dofs. Note that \(\Gamma_{ji} = \Gamma_{ij}\). The set of non-trivial generalized facets is given by

\[\mathcal{Y}^* := \{\Gamma_{ij} : i, j = 1, \ldots, N, i \neq j, \Gamma_{ij} \neq \emptyset\}.\]

**Remark 3.15.** Most of these generalized facets are closed faces. Assume that every vertex is chosen totally primal, then in two dimensions, all generalized facets are actually closed faces. In three dimensions, if we have a regular subdomain edge \(E\) shared by four or more subdomains, then for each pair \(i \neq j\) with \(\{i, j\} \in \mathcal{N}_E\) where no face \(F\) exists such that \(\{i, j\} \in \mathcal{N}_F\), we get a generalized facet \(\Gamma_{ij}\). According to [49, 99], for decompositions generated from a graph partitioner, most of the subdomain edges share only three subdomains.

We fix an ordering of the dofs for each set \(\Gamma_{ij}\) and denote by \(R_{i\Gamma_{ij}} : W_i \rightarrow U_{\Gamma_{ij}}\) the corresponding zero-one restriction matrix. For each sub-glob \(G \subset \Gamma_{ij}\), we denote by \(R_{G\Gamma_{ij}} : U_{\Gamma_{ij}} \rightarrow U_G\) the zero-one restriction matrix such that \(R_{iG} = R_{G\Gamma_{ij}} R_{i\Gamma_{ij}}\).

Moreover, for each pair \((i, j)\) with \(\Gamma_{ij} \in \mathcal{Y}^*\), we denote by \(W_i, \tilde{W}_i, W_j, \tilde{W}_j\) the restriction of \(W, \tilde{W}\), respectively, to the two components \(i, j\). The restriction of a vector \(w \in \tilde{W}\) or \(W\) is denoted by \(w_{ij}\). With this notation, we deduce from (3.12) that

\[(P_D w)_i = \sum_{j : \Gamma_{ij} \in \mathcal{Y}^*} R^T_{i\Gamma_{ij}} \left( \sum_{G \subset \Gamma_{ij}} R^T_{G\Gamma_{ij}} D_{jG} R_{G\Gamma_{ij}} \right) (R_{iG} w_i - R_{jG} w_j)
= \sum_{j : \Gamma_{ij} \in \mathcal{Y}^*} R^T_{i\Gamma_{ij}} D_{j\Gamma_{ij}} (R_{i\Gamma_{ij}} w_i - R_{j\Gamma_{ij}} w_j)
= (P_{D,\Gamma_{ij}} w_{ij})_i\]

where \(P_{D,\Gamma_{ij}} : W_{ij} \rightarrow W_{ij}\). The following result was first shown in [49, Lemma 6.1] with essentially the same constant.

**Lemma 3.16.** Let Assumption 3.4 (glob-wise partition of unity) be fulfilled. If for every \(\Gamma_{ij} \in \mathcal{Y}^*\), the inequality

\[|\langle P_{D,\Gamma_{ij}} w_{ij} \rangle_{S_j}^2 + |\langle P_{D,\Gamma_{ij}} w_{ij} \rangle_{S_j}^2 | \leq \omega_{ij} \left( |w_i|^2_{S_i} + |w_j|^2_{S_j} \right) \quad \forall w_{ij} \in \tilde{W}_{ij}\]

holds, then

\[|P_D w|_{S_i}^2 \leq \omega |w|_{S_i}^2 \quad \forall w \in \tilde{W},\]

with \(\omega = \left( \max_{i=1,\ldots,N} n_i^2 \right) \left( \max_{\Gamma_{ij} \in \mathcal{Y}^*} \omega_{ij} \right), \) where \(n_i := |\{j : \Gamma_{ij} \in \mathcal{Y}^*\}|.\)
Proof. The Cauchy-Bunyakovsky-Schwarz inequality implies
\[
\sum_{i=1}^{N} |(P_D w)|_S^2_i \leq \sum_{i=1}^{N} n_i \sum_{\Gamma_{ij} \in \Gamma} |(P_{D,\Gamma_{ij}} w_{ij})|_S^2_i
\]
\[
\leq \left( \max_{i=1,\ldots,N} n_i \right) \sum_{\Gamma_{ij} \in \Gamma} \left( |(P_{D,\Gamma_{ij}} w_{ij})|_S^2_i + |(P_{D,\Gamma_{ij}} w_{ij})|_S^2_j \right).
\]
Employing the local estimate and using Cauchy-Bunyakovsky-Schwarz another time yields
\[
\sum_{i=1}^{N} |(P_D w)|_S^2_i \leq \left( \max_{i=1,\ldots,N} n_i \right) \sum_{\Gamma_{ij} \in \Gamma^*} \omega_{ij} \left( |w_i|_S^2_i + |w_j|_S^2_j \right)
\]
\[
= \left( \max_{i=1,\ldots,N} n_i \right) \sum_{i=1}^{N} \sum_{\Gamma_{ij} \in \Gamma^*} \omega_{ij} |w_i|_S^2_i
\]
\[
\leq \left( \max_{i=1,\ldots,N} n_i^2 \right) \left( \max_{\Gamma_{ij} \in \Gamma^*} \omega_{ij} \right) \sum_{i=1}^{N} |w_i|_S^2_i.
\]

Unlike the glob-based operator \( P_{D,G} \), the pair-based operator \( P_{D,\Gamma_{ij}} \) fails to be a projection. For this reason and the fact that adaptive constraints on the generalized facets \( \Gamma_{ij} \) would have to be specially treated (e.g., split) in order to ensure that the constraints associated with each subdomain are linearly independent, we do not pursue the pair-based localization further. Note, however, that parts (not all) of our theory could be transferred to the pair-based localization.

4. The glob eigenproblem for general scalings. The local glob estimate (3.9) is directly related to a generalized eigenproblem \( A = \lambda B \), where \( A, B \) correspond to the right- and left-hand side of the estimate, respectively, and the best constant is the inverse of the minimal eigenvalue. We show this relation in detail (Section 4.1), allowing both \( A, B \) to be singular (in this, our presentation differs from [77, 79]). Next, we show how to reduce generalized eigenproblems by using Schur complements and how to modify them, obtaining the same or related estimates. In Section 4.2, we discuss the eigenproblem associated with estimate (3.9) and provide some strategies on how it could be computed in practice.

4.1. Technical tools for generalized eigenproblems*. The following definition and lemma are common knowledge but stated and proved for the sake of completeness; see also [37, Section 7.7.1], [79, Lemma 2], [109, Definition 2.10, Lemma 2.11], and [31, 53] for similar results.

**Definition 4.1.** Let \( V \) be a finite-dimensional (real) Hilbert space and \( A, B : V \rightarrow V^* \) linear operators. We call \((\lambda, y)\) a (real) generalized eigenpair of \((A, B)\) if either

(a) \( \lambda \in \mathbb{R} \) and \( y \in V \setminus \{0\} \) fulfill \( Ay = \lambda By \), or

(b) \( \lambda = \infty \) and \( y \in \ker(B) \setminus \{0\} \).

We will not need complex eigenvalues in the sequel. In this text, we say that \( \lambda \) is a genuine eigenvalue of \((A, B)\) if there is an associated eigenvector in \( V \setminus (\ker(A) \cap \ker(B)) \).

Apparently, \( \lambda \) is a generalized eigenvalue of \((A, B)\) if and only if \( 1/\lambda \) is a generalized eigenvalue of \((B, A)\), where \( 1/0 := \infty \) and \( 1/\infty := 0 \). The eigenspaces corresponding to \( \lambda = 0 \) and \( \lambda = \infty \) are \( \ker(A) \) and \( \ker(B) \), respectively. If \( \ker(A) \cap \ker(B) \) is non-trivial, then every \((\lambda, y)\) with \( \lambda \in \mathbb{R} \cup \{\infty\} \) and \( y \in \ker(A) \cap \ker(B) \) is a generalized eigenpair. If an eigenvalue \( \lambda \) has only eigenvectors in \( \ker(A) \cap \ker(B) \), then we call it ambiguous in
the sequel. If $B$ is non-singular, then the generalized eigenvalues of $(A, B)$ are the same as the regular eigenvalues of $B^{-1/2}AB^{-1/2}$, and if $B$ is SPD, then they are the same as those of $B^{-1/2}AB^{-1/2}$, where $B^{1/2}$ is the SPD matrix square root. The next lemma treats the general SPD case, and its proof is given on page 297.

**Lemma 4.2.** Let $V$ be a finite-dimensional (real) Hilbert space and $A, B : V \to V^*$ linear operators that are SPD. Then there exist at least $n = \dim(V) - \dim(\ker(B))$ genuine generalized eigenvalues

$$0 \leq \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n < \infty$$

and a basis $\{y_k\}_{k=1}^{\dim(V)}$ of $V$ such that $(\lambda_k, y_k)_{k=1}^n$ and $(\infty, y_k)_{k=n+1}^{\dim(V)}$ are generalized eigenpairs of $(A, B)$ and

$$\langle By_k, y_\ell \rangle = \delta_{k\ell}, \quad \langle Ay_k, y_\ell \rangle = \lambda_k \delta_{k\ell} \quad \forall k, \ell = 1, \ldots, n,$n

and $\ker(B) = \text{span}\{y_k\}_{k=n+1}^{\dim(V)}$. Furthermore, for any $k \in \{0, \ldots, n-1\}$ with $\lambda_{k+1} > 0$,

$$\langle Bz, z \rangle \leq \frac{1}{\lambda_{k+1}} \langle Az, z \rangle \quad \forall z \in V, \quad \langle By_\ell, z \rangle = 0, \quad \ell = 1, \ldots, k.$$

The constant in this bound cannot be improved.

The next result is interesting in itself; cf. [109, Lemma 2.11].

**Corollary 4.3.** Let $k \in \{0, \ldots, n-1\}$ with $\lambda_{k+1} > 0$ as in the previous lemma and let $\Pi_k : V \to V$ be the projection defined by $\Pi_kv := \sum_{\ell=1}^k \langle Bv, y_\ell \rangle y_\ell$. Then

$$|\Pi_kv|_A \leq |v|_A, \quad |(I - \Pi_k)v|_A \leq |v|_A, \quad \Pi_kv|_B \leq |v|_B, \quad |(I - \Pi_k)v|_B \leq |v|_B.$$

Moreover,

$$|(I - \Pi_k)v|_B^2 \leq \frac{1}{\lambda_{k+1}} |(I - \Pi_k)v|_A^2 \leq \frac{1}{\lambda_{k+1}} |v|_A^2 \quad \forall v \in V.$$

**Proof.** All the estimates can be easily verified by expanding $v \in V$ with respect to the eigenvectors, i.e., $v = \sum_{k=1}^{\dim(V)} \beta_k y_k$, and by using the results of Lemma 4.2. \Box

For the proof of Lemma 4.2, we need an auxiliary result.

**Principle 4.4** ("Schur principle": reduction of infinite eigenvalues by Schur complement). Let $V$ be a finite-dimensional (real) Hilbert space and $A, B : V \to V^*$ two linear and self-adjoint operators. Let $V_2 \subset \ker(B)$ be a subspace and $V_1$ some complementary space such that $V = V_1 \oplus V_2$ (direct sum, not necessarily orthogonal). In that situation, we may identify $V$ with $V_1 \times V_2$ and write

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_{11} & 0 \\ 0 & 0 \end{bmatrix}.$$

Assume that $\ker(A_{22}) \subset \ker(A_{12})$ (cf. Lemma D.3), and let $S_1 := A_{11} - A_{12}A_{22}^*A_{21}$ be a generalized Schur complement (cf. Appendix D). Then the following holds:

(i) $(\lambda, y)$ is a generalized eigenpair of $(A, B)$ if and only if either $\lambda = \infty$ and $y \in \ker(B)$, or

$$y = \begin{bmatrix} y_1 \\ -A_{22}^*A_{21}y_1 + v^K_2 \end{bmatrix} \text{ for some } v^K_2 \in \ker(A_{22}),$$

and $(\lambda, y_1)$ is a generalized eigenpair of $(S_1, B_{11})$. 


The constant cannot be improved due to the Courant-Fisher minimax principle [37, Theo-

rem 8.1.2]. Let \( y_k \) be a basis of \( \ker(B) \) and set

\[
\lambda_k = \tilde{\lambda}_k, \quad y_k = \begin{bmatrix} \tilde{y}_k \\ -A_{22}^T A_{21} \tilde{y}_k \end{bmatrix}
\]

for \( k = 1, \ldots, n \).

Now all the statements follow from Principle 4.4. \( \square \)
The “Schur” Principle 4.4 is not only valuable for the proof of Lemma 4.2 but will be quite useful in our subsequent theory and method as it provides a way to reduce an eigenproblem by keeping all the finite eigenvalues. Conversely, Principle 4.4 can be used to unroll a Schur complement popping up in a generalized eigenproblem.

Sometimes, we want to compute with matrices but on a subspace of \( \mathbb{R}^n \) for which we do not have a basis at hand. The following principle is a slight generalization of [76, Lemma 5].

**PRINCIPLE 4.5** (projected eigenproblem). Let \( A, B \in \mathbb{R}^{n \times n} \), let \( \Pi : \mathbb{R}^n \rightarrow \mathbb{R}^n \) be some projection onto a subspace \( \text{range}(\Pi) \subset \mathbb{R}^n \), and let \( Q \in \mathbb{R}^{n \times n} \) be SPD on \( \text{range}(I - \Pi) \), e.g., \( Q = tI \) with \( t \in \mathbb{R} > 0 \).

Part (i): If (a) holds, then

\[
\Pi^T A \Pi y = \lambda \Pi^T B \Pi y \quad \text{and} \quad y \in \text{range}(\Pi)
\]

if and only if

\[
(b) \quad (\Pi^T A \Pi + (I - \Pi^T)Q(I - \Pi))y = \lambda \Pi^T B \Pi y.
\]

Part (ii): Assume that \( A \) is SPD on \( \text{range}(\Pi) \) and that

\[
\langle (\Pi^T A \Pi + (I - \Pi^T)Q(I - \Pi))y, y \rangle = 0.
\]

Then \( \langle A \Pi y, y \rangle = 0 \) and \( \langle Q(I - \Pi)y, (I - \Pi)y \rangle = 0 \). Due to the assumptions on \( A \) and \( Q \), we obtain \( \Pi y = 0 \) and \( (I - \Pi)y = 0 \), and finally \( y = 0 \).

**REMARK 4.6**. It is yet questionable, whether it is easier to construct a basis for a subspace of \( \mathbb{R}^n \) or a projection onto it. If the matrices \( S_i \) stem from sparse stiffness matrices, then we would like the basis transformation matrix to be sparse too in the sense that all rows and columns have \( O(1) \) non-zero entries except for \( O(1) \) rows/columns which may be dense.

**REMARK 4.7** ("saddle point" eigenproblem). With similar arguments as in the proof of the “Schur” Principle 4.4, one can show that the generalized eigenproblem

\[
\langle Ay, z \rangle = \lambda \langle By, z \rangle \quad \forall y, z \in V := \{v \in \mathbb{R}^n : Cv = 0\},
\]

with surjective \( C \in \mathbb{R}^{m \times n}, m < n \), is equivalent to

\[
\begin{bmatrix}
A & C^T \\
C & 0
\end{bmatrix}
\begin{bmatrix}
v \\
\mu
\end{bmatrix}
= \lambda
\begin{bmatrix}
B & 0 \\
0 & 0
\end{bmatrix}
\begin{bmatrix}
v \\
\mu
\end{bmatrix},
\]

up to some eigenvalues of infinity. The latter eigenproblem is posed on the simpler space \( \mathbb{R}^{n+m} \).

In the following two principles, the eigenvalues might change.

**PRINCIPLE 4.8** (eigenproblem on larger space). Let \( V, A, B \) be as in Lemma 4.2, and let \( \tilde{V} \supset V \) be a larger space with the natural embedding operator \( E : V \rightarrow \tilde{V} \). Suppose that there are SPD operators \( \tilde{A}, \tilde{B} : \tilde{V} \rightarrow \tilde{V}^* \) such that \( A = E^T \tilde{A} E \) and \( B = E^T \tilde{B} E \), and let \( (\tilde{\lambda}_k, \tilde{y}_k) \) be the eigenpairs of \( (\tilde{A}, \tilde{B}) \) according to Lemma 4.2. If \( \tilde{\lambda}_{k+1} \in (0, \infty) \), then for all \( z \in V \) with \( \langle E^T \tilde{B} \tilde{y}_k, z \rangle = 0 \), \( \ell = 1, \ldots, k \),

\[
\langle Bz, z \rangle = \tilde{\langle} \tilde{B} E z, E z \rangle \leq \frac{1}{\tilde{\lambda}_{k+1}} \tilde{\langle} \tilde{A} E z, E z \rangle = \frac{1}{\lambda_{k+1}} \langle Az, z \rangle.
\]
Principle 4.9 (nearby eigenproblem). Let $V$, $A$, $B$ be as in Lemma 4.2, and let $\tilde{A}$, $\tilde{B} : V \to V^*$ be two SPSD operators such that

$$\tilde{A} \leq c_1 A \text{ and } B \leq c_2 \tilde{B},$$

and let $(\tilde{\lambda}_k, \tilde{y}_k)$ be the eigenpairs of $(\tilde{A}, \tilde{B})$ according to Lemma 4.2. If $\tilde{\lambda}_{k+1} \in (0, \infty)$, then for all $z \in V$ with $\langle \tilde{B}\tilde{y}_k, z \rangle = 0$, $\ell = 1, \ldots, k$,

$$\langle Bz, z \rangle \leq c_2 \langle \tilde{B}z, z \rangle \leq \frac{c_2}{\tilde{\lambda}_{k+1}} \langle \tilde{A}z, z \rangle \leq \frac{c_1 c_2}{\tilde{\lambda}_{k+1}} \langle Az, z \rangle.$$

When $A$, $B$ have block structure, a special application of Principle 4.9 allows us to decouple the eigenproblem (at the price of an approximation).

Principle 4.10 (decoupling). Let $A, B : V^n \to (V^n)^*$ be SPSD block operators for a finite-dimensional Hilbert space $V$,

$$A = \begin{bmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{n1} & \cdots & A_{nn} \end{bmatrix}, \quad B = \begin{bmatrix} B_{11} & \cdots & B_{1n} \\ \vdots & \ddots & \vdots \\ B_{n1} & \cdots & B_{nn} \end{bmatrix},$$

and let $m \leq n$ be the maximal number of non-zero block-entries per row of $B$. For each $i = 1, \ldots, n$, let $S_i$ be the Schur complement of $A$ that eliminates all but the $i$-th block. Then

$$\begin{bmatrix} S_1 & \cdots & S_n \end{bmatrix} \leq n A, \quad B \leq m \begin{bmatrix} B_{11} & \cdots & B_{1n} \\ \vdots & \ddots & \vdots \\ B_{n1} & \cdots & B_{nn} \end{bmatrix}.$$

So if $(\lambda^{(i)}_k, y^{(i)}_k)$ are the eigenpairs of $(S_i, B_{ii})$ and if $\lambda^{(i)}_{k+1} \in (0, \infty)$, then for all $z \in V^n$ with $\langle B_{ii} y^{(i)}_\ell, z_i \rangle = 0$, for $\ell = 1, \ldots, k_i$,

$$\langle Bz, z \rangle \leq n m \max_{i=1, \ldots, n} \frac{1}{\lambda^{(i)}_{k+1}} \langle Az, z \rangle.$$

Of course, a different choice of the space splitting (leading to the block structure) can lead to different spectra in the decoupled eigenproblem.

Proof. The first spectral inequality follows from the minimizing property of the Schur complement (Lemma D.5), while the second one is simply a consequence of the Cauchy-Bunyakovsky-Schwarz inequality. The rest follows from Principle 4.9 (nearby eigenproblem).

We also provide a simple result to recombine decoupled eigenproblems $(A_i, B_i)$, $i = 1, \ldots, n$.

Principle 4.11 (recombination). Let $V$ be a finite-dimensional Hilbert space, and let $A_i, B_i : V \to V^*$, $i = 1, \ldots, n$, be SPSD operators. We consider the single eigenproblem

$$\langle (A_1 : A_2 : \ldots : A_N) y, z \rangle = \lambda \langle (B_1 + B_2 + \ldots + B_N) y, z \rangle \quad \text{for } y, z \in V,$$

with eigenpairs $(\bar{\lambda}_k, \bar{y}_k)$. Then, for $m < n$ with $\bar{\lambda}_{m+1} > 0$,

$$|z|^2_{B_i} \leq \frac{1}{\bar{\lambda}_{m+1}} |z|^2_{A_i} \quad \forall z \in \mathbb{R}^n : \langle B\bar{y}_k, z \rangle = 0 \quad \forall i = 1, \ldots, n.$$

The same result holds for any $A$ with $A \leq A_i$ for all $i = 1, \ldots, n$. 
Proof. For $i$ and $z$ as above, $$|z|_{2,i}^2 \leq |z|_{2,j}^2 \leq \frac{1}{\lambda_{m+1}} |z|_{2,i}^2 \leq \frac{1}{\lambda_{m+1}} |z|_{2,j}^2,$$

Finally, we need a result for general eigenproblems of a special structure.

LEMMA 4.12. Let $V$ be a finite-dimensional Hilbert space, $A : V \to V^*$ a linear SPSD operator, and $P : V \to V$ a projection ($P^2 = P$). Then for

$$B := \beta P^T A P,$$

with $\beta \in (0, \infty)$,

the following statements hold:

(i) The eigenspace of infinite generalized eigenvalues of $(A, B)$ is given by

$$\ker(B) = \ker(P) \oplus (\ker(A) \cap \range(P)),$$

and the ambiguous eigenspace by

$$\ker(A) \cap \ker(B) = (\ker(A) \cap \ker(P)) \oplus (\ker(A) \cap \range(P)).$$

(ii) If $\ker(A) \subset \ker(P)$, then $(A, B)$ has no genuine zero eigenvalues.

(iii) If $\ker(A) \cap \range(P) = \{0\}$ and if $(A, B)$ has no genuine zero eigenvalues, then

$$\ker(A) \subset \ker(P).$$

Proof. (i) Since $P$ is a projection, $V = \ker(P) \oplus \range(P)$. Assume that

$$v = v_1 + v_2 \in \ker(B), \quad \text{with } v_1 \in \ker(P), \quad v_2 \in \range(P).$$

From the definition of $B$ we see that $Bv_1 = 0$, and thus, if $v \in \ker(B)$, then

$$0 = \langle Bv, v \rangle = \langle Bv_1, v \rangle = \beta \langle Apv, pv \rangle = \beta \langle Av, v \rangle,$$

and so $Av = 0$. Conversely, if $v_1 \in \ker(P)$ and $v_2 \in \ker(A) \cap \range(P)$, then $v_1 + v_2 \in \ker(B)$. The formula for $\ker(A) \cap \ker(B)$ is then straightforward.

(ii) If $\ker(A) \subset \ker(P)$, then $\ker(A) \cap \range(P) = \{0\}$, and so $\ker(B) = \ker(P)$. Also, $\ker(A) \cap \ker(B) = \ker(A) \cap \ker(P)$ and $\ker(A) \setminus \ker(B) = \emptyset$.

(iii) Let $\lambda_1 \leq \ldots$ be the genuine eigenvalues of $(A, B)$ according to Lemma 4.2. If there are no genuine zero eigenvalues, then $\lambda_1 > 0$. Suppose $v \in \ker(A)$, then

$$\langle Bv, v \rangle \leq \frac{1}{\lambda_1} \langle Av, v \rangle = 0,$$

and so $v \in \ker(B) = \ker(P) \oplus (\ker(A) \cap \range(P))$, using Part (i). Due to our assumptions, $\ker(A) \cap \range(P) = \{0\}$, and so $v \in \ker(P)$. □

Let us apply the “Schur” Principle 4.4 to the generalized eigenproblem $(A, \beta P^T A P)$ and eliminate $\ker(P)$. If $\ker(A) \cap \range(P) = \{0\}$, then the reduced eigenproblem neither has ambiguous nor infinite eigenvalues. Under the stronger condition $\ker(A) \subset \ker(P)$ (see Lemma 4.12(ii) and (iii)), the reduced eigenproblem has only eigenvalues in $(0, \infty)$.

4.2. Generalized eigenproblems associated with the estimate (3.9). Let us fix a set of linearly independent primal dofs in the sense of Definition 2.16 (possibly an empty set), and let $G \in \mathcal{G}^*$. Recall the space $\tilde{W}_{N_G}$ from (3.3), and let $\tilde{I}_{N_G} : \tilde{W}_{N_G} \to W_{N_G}$ denote the natural embedding. Moreover define

$$S_{N_G} := \text{diag}(S_i)_{i \in N_G} : W_{N_G} \to W_{N_G},$$

and consider the
A UNIFIED FRAMEWORK FOR ADAPTIVE BDDC

(4.1) \[ \tilde{S}_G y = \lambda \tilde{B}_G y \quad (y \in \tilde{W}_G), \]
with \[ \tilde{S}_G := \widetilde{I}_G^T S_N \tilde{I}_G, \quad \tilde{B}_G := \tilde{P}_D G \tilde{S}_G \tilde{P}_D G = \tilde{I}_G^T P_D G S_N G P_D G \tilde{I}_G, \]
where \( \tilde{P}_{D,G} \) is the projection operator from Lemma 3.8(v). The next result immediately follows from Lemma 4.2.

**COROLLARY 4.13.** Let \( \lambda_{G,k} \) be the generalized eigenpairs of \( (\tilde{S}_G, \tilde{B}_G) \) according to Lemma 4.2 with \( 0 \leq \lambda_{G,1} \leq \lambda_{G,2} \leq \ldots \leq \infty. \)

(i) If there are no genuine zero eigenvalues \( (\lambda_{G,1} > 0) \), then the estimate (3.9) holds with
\[ \varpi_G = \frac{1}{\lambda_{G,1}}. \]

(ii) Let us fix a number \( m_G \) such that \( 0 < \lambda_{G,m_G+1} < \infty \), and let \( \Phi_{G,\text{add}} : \mathbb{R}^{m_G} \to \tilde{W}_G \) be the matrix whose columns are the first \( m_G \) eigenvectors,
\[ \Phi_{G,\text{add}} = [\ldots |\tilde{I}_G y_{G,k} | \ldots ]_{k=1}^{m_G}. \]
Then
\[ \sum_{i \in N_G} |(P_D G w)_i|^2 \leq \frac{1}{\lambda_{G,m_G+1}} \sum_{i \in N_G} |w_i|^2, \]
\[ \forall w \in \tilde{W}_G, \quad \Phi_{G,\text{add}}^T P_D G S_N G P_D G w = 0, \]
which is an improved estimate compared to (3.9).

**LEMMA 4.14.** If Assumption 3.4 \( (D_i \text{ block diagonal}) \) holds and Condition 3.1 (local kernel condition) is fulfilled, then (4.1) has no genuine zero eigenvalues.

**Proof.** Condition 3.1 is equivalent to \( \ker(\tilde{S}_N G) \cap \tilde{W}_N G \subset \ker(\tilde{P}_{D,G}) \), which, by using Lemma 3.8(v), is further equivalent to
\[ \ker(\tilde{S}_N G) \subset \ker(\tilde{P}_{D,G}). \]
Due to Assumption 3.4 and Lemma 3.8(v), \( \tilde{P}_{D,G} \) is a projection, and so the statement follows from Lemma 4.12(ii).

**REMARK 4.15.** The converse of Lemma 4.14 does not hold in general. In Section 7, we will formulate additional assumptions under which one can conclude Condition 3.1 (local kernel condition) from the positivity of the genuine eigenvalues.

One can now think of several strategies.

**Strategy 1.** We solve the generalized eigenproblem (4.1) right away.

**Strategy 2.** If each \( S_i \) is the Schur complement of a sparse stiffness matrix \( A_i \), then we can unroll the elimination and consider, by applying the “Schur” Principle 4.4, the associated sparse generalized eigenproblem, which has the same spectrum up to ambiguous and infinite eigenvalues. Applying additionally Principle 4.5 (projected eigenproblem) leads to the method in [76, 79] except that the eigenproblem therein is posed on the closed faces and that the roles of \( A \) and \( B \) are interchanged.
**Strategy 3.** In Strategies 1 and 2, we expect a series of infinite eigenvalues. To get rid of (some of) them, observe that

\[ \ker(\tilde{P}_{D,G}) = \tilde{X}_G \oplus \hat{Y}_G, \]
\[ \tilde{X}_G = \{ w \in \tilde{W}_{N_G} : \forall i \in N_G : R_{ig} w_i = 0 \}, \]
\[ \hat{Y}_G = \{ w \in \tilde{W}_{N_G} : \forall i \in N_G : R_{ig} w_i = 0 \text{ and } \forall j \in N_G : R_{ig} w_i = R_{gj} w_j \}, \]

where \( R_{ig} \) is the restriction matrix extracting all dofs not associated with glob \( G \) with the property

\[ (4.2) \quad R_{ig}^T R_{ig} + R_{ig}^T R_{ig} = I. \]

Functions from the space \( \tilde{X}_G \subset \tilde{W}_{N_G} \) vanish on \( G \), whereas functions from \( \hat{Y}_G \) are continuous on \( G \) and vanish on all other dofs. Using a change of basis, we can parametrize \( \tilde{W}_{N_G} \) and the two subspaces above explicitly. Forming the Schur eigenproblem according to Principle 4.4, eliminating \( \ker(\tilde{P}_{D,G}) \), we get rid of some ambiguous infinite eigenvalues, which may be important in practice.

**Strategy 4.** We apply Principle 4.8 and embed the eigenproblem into the larger space

\[ \tilde{W}_{N_G}^G := \{ w \in W_{N_G} : \forall i,j \in N_G : Q_{ij}^G (R_{ig} w_i - R_{gj} w_j) = 0 \} \]

from Remark 3.12. We warn the reader that by doing this, we **discard** any (good) influence of the primal constraints on the neighboring globs of \( G \). Defining the projection operator \( \tilde{P}^G_{D,G} \) analogously as \( \tilde{P}_{D,G} \) in Lemma 3.8(v), replacing \( \tilde{W}_{N_G} \) by \( \tilde{W}_{N_G}^G \), we find that the eigenproblem has the form \((\tilde{S}^G_{N_G}, (\tilde{P}^G_{D,G})^T \tilde{S}^G_{N_G} \tilde{P}^G_{D,G}) \). As an advantage,

\[ \ker(\tilde{P}^G_{D,G}) = X_G \oplus \hat{Y}, \]

where the first space

\[ X_G = \{ w \in W_{N_G} : \forall i \in N_G : R_{ig} w_i = 0 \} = \bigotimes_{i \in N_G} \{ w_i \in W_i : R_{ig} w_i = 0 \} \supseteq \tilde{X}_G \]

is much simpler than \( \tilde{X}_G \). Consequently, it is much simpler to implement the Schur complement operator of \( S_{N_G} \) on \( \tilde{W}_{N_G}^G \) eliminating \( \tilde{X}_G \oplus \hat{Y} \). Let us also note that if no primal constraints are enforced on the neighboring globs (\( G' \) with \( |N_G \cap N_{G'}| \geq 2 \)), then \( \tilde{W}_{N_G} = \tilde{W}_{N_G}^G \), i.e., the two eigenproblems are identical.

For all strategies, the underlying spaces are given implicitly, as subspaces of \( \mathbb{R}^n \). One can either explicitly parametrize them by \( \mathbb{R}^m \), \( m < n \) (i.e., constructing a basis), or construct a projection from \( \mathbb{R}^n \) to the subspace and apply Principle 4.5 (projected eigenproblem). As an alternative, one can use the constraints defining the subspace in the eigenproblem (Remark 4.7). Note also that for all the Strategies 1–4, the initially chosen primal constraints on the glob \( G \) are preserved. Modifying them means changing the eigenproblem; see also Remark 4.16 below.

No matter which of the four strategies we use, we will always get the statement of Corollary 4.13 (with some of the operators replaced):
1. If the minimal eigenvalue $\lambda_{G,1}$ of the respective generalized eigenproblem is positive, then the local glob estimate (3.9) holds with $\omega_G = \frac{1}{\lambda_{G,1}}$.

2. We can improve the estimate by enforcing additional constraints of the form

\[ \Phi_{G,\text{add}}^T P_{D,G}^T S_{N,G} P_{D,G} w_{N,G} = 0. \]

These constraints are of the more general form in Section 2.6 and fulfill the conditions (2.19)–(2.20) of locality and consistency.

**Remark 4.16 (orthogonality of constraints).** For each of the strategies, we consider a generalized eigenproblem of the form: find eigenpairs $(y, \lambda) \in \tilde{V} \times \mathbb{R}$:

\[ \langle Ay, z \rangle = \lambda \langle B y, z \rangle \quad \forall z \in \tilde{V} := \{ v \in V : Cv = 0 \}, \]

where $Cv = 0$ correspond to initially chosen constraints. An adaptively chosen constraint reads

\[ \langle q_k, u \rangle := \langle By_k, u \rangle = 0, \]

where $y_k$ is an eigenvector. Assume that $B$ is SPD. Then the functionals $q_k$ are pairwise orthogonal in the $B^{-1}$-inner product. Since $Cy_k = 0$, it follows that

\[ CB^{-1}q_k = 0, \]

so the new constraints $q_k$ are also pairwise orthogonal to the initial constraints in the $B^{-1}$-inner product. This pattern also applies to the simpler eigenproblems in the following section.

**5. Adaptive choice of the primal dofs.** In this section, we

(i) study in more detail the structure of the glob eigenproblem (4.1) for subdomain faces (Section 5.2) and general globs (Section 5.3),

(ii) show how to turn the constraints (4.3) originating from the local generalized eigenproblems into primal dofs (Section 5.4),

(iii) provide a way to rewrite the glob eigenproblem using a transformation of variables and to decouple it into $n - 1$ independent eigenproblems where $n$ is the number of subdomains shared by the glob (Section 5.5),

(iv) show that recombining the $n - 1$ problems into a single one leads to the eigenproblem proposed by Kim, Chung, and Wang (Section 5.6),

(v) comment on how the eigenproblems could be organized in an algorithm (Section 5.7).

To this end, we need further notation (given below) and the parallel sum of matrices (Section 5.1).

**Definition 5.1.** For $G \in \mathcal{G}_i$ let

\[ S_{iG} := R_{iG} S_i R_{iG}^T \]

(5.1)

denote the restriction of $S_i$ to the dofs on $G$, and let

\[ S_{iG^c} := R_{iG} S_i R_{iG^c}, \quad S_{iG^c} := R_{iG^c} S_i R_{iG}, \quad S_{iG^c} := R_{iG^c} S_i R_{iG^c} \]

denote the other subblocks of $S_i$, where $R_{iG^c}$ is the restriction matrix from (4.2). Finally, we define the (generalized) Schur complement

\[ S_{iG}^* := S_{iG} - S_{iG^c} S_{iG^c}^T S_{iG^c} G. \]

(5.2)
REMARK 5.2. In practice, the matrix $S_{kG}^*$ is usually linked to a problem on the subdomain $\kappa$ with fixed dofs on $G$ and homogeneous “Neumann” conditions on the remaining boundary dofs. Figure 5.1 shows that it may happen that $S_{kG}^*$, $S_{kG'}^*$, or $S_{kG}^*$ are singular.

REMARK 5.3. As we use these matrices in the subsequent eigenproblems, we spend some words on their handling in practice. Suppose that $S_1$ is the Schur complement of a sparse matrix $A_1$ eliminating interior dofs. Since $S_{1G}$ is a principal minor of $S_1$, its application can be realized by a subdomain solve. Some direct solvers, such as MUMPS [2] or PARDISO [66], offer the possibility of computing the dense matrix $S_{1G}$ directly. Since $G'$ usually contains many more dofs than $G$, computing $S_{1G''}$, $S_{1G''}$ in the same way would be inefficient. Instead, following Stefano Zampini [121], one can compute $S_1^*$ once and extract $S_{1G}^*$ as a principal minor of $S_1^*$; see also [84].

5.1. The parallel sum of matrices*. The following definition was originally introduced by Anderson and Duffin [3] for Hermitian positive semi-definite matrices.

**Definition 5.4 (parallel sum of matrices [3]).** For two SPSD matrices $A, B \in \mathbb{R}^{n \times n}$, the parallel sum of $A, B$ is given by

$$A : B = (A + B)^\dagger B,$$

where $(A + B)^\dagger$ is a generalized inverse, i.e., $(A + B)(A + B)^\dagger f = f$ for all $f \in \text{range}(A + B)$; cf. Definition D.1. The definition is independent of the particular choice of the generalized inverse (cf. [3, p. 579] and Proposition D.2), $A : B = B : A$ [3, Lemma 1], and $A : B$ is again SPSD [3, Lemma 2, Lemma 4]. Moreover, due to [3, Lemma 6], $(A : B) : C = A : (B : C)$.

REMARK 5.5. If $A$ and $B$ are both SPD, then $A : B = (A^{-1} + B^{-1})^{-1}$. Therefore, up to a factor of 2, the above matrix generalizes the harmonic mean value $\frac{2}{\frac{1}{a} + \frac{1}{b}}$ of two positive scalars $a, b$; cf. [5]. Moreover, it can be shown that for $A, B$ SPD,

$$x^\top (A : B)x = \inf_{x = x_1 + x_2} (x_1^\top Ax_1 + x_2^\top Bx_2) \quad \forall x \in \mathbb{R}^n,$$

i.e., $\|x\|_{A:B}$ is the natural norm on the sum of the Hilbert spaces $(\mathbb{R}^n, \|\cdot\|_A), (\mathbb{R}^n, \|\cdot\|_B)$; see also [10], [4, Theorem 9], and Corollary 5.11 below, as well as [112, Eqn. (4)] for a related result.

Let $A, B$ be as in Definition 5.4. We easily see that

$$A : A = \frac{1}{2} A, \quad (cA) : (cB) = c(A : B) \quad \forall c \in \mathbb{R}_0^+$$

(see also [3, Theorem 10]. Since $A, B$ are SPD, we have

$$\ker(A + B) = \ker(A) \cap \ker(B), \quad \text{range}(A + B) = \text{range}(A) + \text{range}(B),$$

and we can conclude that

$$\ker(A : B) = \ker(A) + \ker(B), \quad \text{range}(A : B) = \text{range}(A) \cap \text{range}(B);$$

Fig. 5.1. Left: Laplace/linear elasticity: $S_{kF}^*$ empty, $S_{kF}^*$ singular. Right: Linear elasticity, straight edge $E$: $S_{kE}^* = S_{kF}^*$ singular, $S_{kE}^* = S_{kE}^*$ non-singular, $S_{kF}^*, S_{kE}^*$ singular.
cf. [3, Lemma 3]. From Definition 5.4 and Proposition D.2, one easily shows

\[(5.5)\quad A : B = A - A(A + B)^\dagger A = B - B(A + B)^\dagger B;\]

see also [4, Eq. (10)]. Next, let us consider the generalized eigenproblem

\[(5.6)\quad Ap = \lambda(A + B)p\]

in the sense of Section 4.1. With the above relations, it is straightforward to verify that, if \((p, \lambda)\) is an eigenpair of (5.6) (and \(p \notin \text{ker}(A) \cap \text{ker}(B)\)), then \(\lambda \in [0, 1]\) and

\[(5.7)\quad Bp = (1 - \lambda)(A + B)p,\]

\[(A : B)p = \lambda(1 - \lambda)(A + B)p.\]

From (5.4) and (5.7), we easily conclude that

\[(5.8)\quad A : B \leq A, \quad A : B \leq B,\]

which is a special case of [3, Lemma 18] (as usual, \(A \leq B\) stands for \(y^\top Ay \leq y^\top By\) for all \(y \in \mathbb{R}^n\)). Anderson and Duffin also show an important transitivity property:

**Lemma 5.6 ([3, Corollary 21]).** Let \(D, E, F \in \mathbb{R}^{n \times n}\) be SPSD matrices. Then \(D \leq E \implies D : F \leq E : F\).

As the next proposition shows, the parallel sum \(A : B\) is—up to a factor of two—a "sharp lower bound matrix" of \(A\) and \(B\).

**Proposition 5.7.** Let \(A, B\) be as in Definition 5.4 and let the matrix \(C \in \mathbb{R}^{n \times n}\) be SPSD with \(C \leq A\) and \(C \leq B\). Then

\[C \leq 2(A : B).\]

**Proof.** Due to Lemma 5.6, \(C \leq A\) implies \(\frac{1}{2}C = C : C \leq A : C\). With the same Lemma, \(C \leq B\) implies \(A : C \leq A : B\).

The following result states that the parallel sum of two spectrally equivalent matrices is spectrally equivalent to the parallel sum of the original matrices.

**Proposition 5.8.** Let \(A, \tilde{A}, B, \tilde{B} \in \mathbb{R}^{n \times n}\) be SPSD, and assume that

\[\alpha A \leq \tilde{A} \leq \overline{\alpha} A, \quad \beta B \leq \tilde{B} \leq \overline{\beta} B,\]

with strictly positive constants \(\alpha, \overline{\alpha}, \beta, \overline{\beta}\). Then

\[\min(\alpha, \beta)(A : B) \leq \tilde{A} : \tilde{B} \leq \max(\overline{\alpha}, \overline{\beta})(A : B).\]

**Proof.** Firstly, we set \(\gamma := \min(\alpha, \beta), \overline{\gamma} := \max(\overline{\alpha}, \overline{\beta})\) and observe that

\[\gamma A \leq \tilde{A} \leq \overline{\gamma} A, \quad \gamma B \leq \tilde{B} \leq \overline{\gamma} B.\]

Secondly, from Lemma 5.6 and (5.3) we obtain

\[\tilde{A} : \tilde{B} \leq (\overline{\gamma} A) : (\overline{\gamma} B) = \overline{\gamma}(A : B)\]

as well as the analogous lower bound \(\tilde{A} : \tilde{B} \geq \gamma(A : B)\).

**Proposition 5.9.** For non-negative constants \(c_1, c_2\) and a SPSD matrix \(A,\)

\[(c_1 A) : (c_2 A) = c_1 (c_1 + c_2)^\dagger c_2 A.\]
The last lemma of this section appears to be new (for earlier versions see [52, 95]) and generalizes the elementary identity and inequality
\[
\frac{a b^2}{(a + b)^2} + \frac{a^2 b}{(a + b)^2} = \frac{a b}{a + b}, \quad \frac{a b^2}{(a + b)^2} \leq \min(a, b)
\]
for non-negative scalars \(a, b\) with \(a + b > 0\), cf. [113, (6.19), p. 141].

**Lemma 5.10.** Let \(A, B \in \mathbb{R}^{n \times n}\) be SPD. Then
\[
B(A + B)A(A + B)^\top B + A(A + B)^\top B(A + B)^\top A = A : B.
\]
In particular,
\[
B(A + B)^\top A(A + B)^\top B, \quad A(A + B)^\top B(A + B)^\top A \leq A : B \leq \begin{cases} A, \\ B. \end{cases}
\]

**Proof.** Since \(A : B = B(A + B)^\top A = A(A + B)^\top B\) (see Definition 5.4),
\[
\begin{align*}
&\quad \quad \quad \quad = (A : B)(A + B)^\top B + (A : B)(A + B)^\top A \\
&\quad \quad \quad \quad = (A : B)(A + B)^\top (A + B) = A : B.
\end{align*}
\]
The last identity holds because for any \(v \in \mathbb{R}^n\), \((A + B)^\top (A + B)v = v + v_K\) for some \(v_K \in \ker(A + B) = \ker(A) \cap \ker(B) \subset \ker(A) + \ker(B) = \ker(A : B)\); cf. (D.1), Appendix D, and (5.4). So \(H_1 + H_2 = A : B\). Since \(H_1\) and \(H_2\) are both SPD, we have that \(H_1, H_2 \leq A : B\). Due to (5.8), \(A : B \leq A, B\), which concludes the proof.

The following corollary appears to be a special case of [4, Theorem 9].

**Corollary 5.11.** For SPD matrices \(A, B \in \mathbb{R}^n\),
\[
|x|_{A:B}^2 = \inf_{x_1 + x_2 \in \mathbb{R}^p} |x_1|_A^2 + |x_2|_B^2 \quad \forall x \in \mathbb{R}^n.
\]

**Proof.** Minimization yields the first order condition \(x^+_1 = (A + B)^\top Bx + x_{1K}\) for some \(x_{1K} \in \ker(A + B) = \ker(A) \cap \ker(B)\) and \(x^+_2 = (A + B)^\top Ax + x_{2K}\) for a suitable vector \(x_{2K} \in \ker(A) \cap \ker(B)\). The Hessian is given by \(A + B\), so all these solutions are minimizers. Due to Lemma 5.10, \(|x|_{A:B}^2 = x^\top (A : B)x = |x|_{A,B}^2\). ⊡

**Remark 5.12.** Unfortunately, Lemma 5.10 cannot be generalized to three matrices (see also [15]), in the sense that already for SPD matrices \(A, B, C\),
\[
B(A + B + C)^{-1} A(A + B + C)^{-1} B \not\leq A : B \quad \text{in general!}
\]

Our counterexample in Appendix E shows that \(B(A + B + C)^{-1} A(A + B + C)^{-1} B \not\leq A\). Since \(A : B \leq A\), the above inequality cannot hold.

### 5.2. Subdomain faces

Suppose that \(F\) is a face shared by the subdomains \(\mathcal{N}_F = \{i, j\}\).

Firstly, we have a look at the right-hand side of eigenvalue problem (4.1).

**Lemma 5.13.** Under Assumption 3.4, for a face \(F\) with \(\mathcal{N}_F = \{i, j\}\), we have
\[
z^\top P_{D,F}^T S_{\mathcal{N}_F} P_{D,F} y = (R_i F z_i - R_j F z_j)^\top M_F (R_i F y_i - R_j F y_j) \quad \forall y, z \in W_{\mathcal{N}_F},
\]
with
\[
M_F := D_j^T F S_{i,F} D_i F + D_i^T F S_{j,F} D_j F.
\]
Proof. We obtain from (3.8) that
\[
(P_{D,F}w)_i = R_{iF}^T D_{jF}(R_{iF}w_i - R_{jF}w_j),
(P_{D,F}w)_j = R_{jF}^T D_{iF}(R_{jF}w_j - R_{iF}w_i).
\]
Hence, using that \( R_{kF}^t S_k R_{kF}^t = S_{kF} \) for \( k \in \{i, j\} \),
\[
z^\top P_{D,F}^t S_{N_F} P_{D,F} z = (R_{iF}z_i - R_{jF}z_j)^\top [D_{jF}^t - D_{iF}^t] \begin{bmatrix} S_{iF} & 0 \\ 0 & S_{jF} \end{bmatrix} \begin{bmatrix} D_{jF} \\ -D_{iF} \end{bmatrix} (R_{iF}y_i - R_{jF}y_j). \]
\[
= (D_{jF}^t S_{iF} D_{jF} + D_{iF}^t S_{jF} D_{iF}) \]

The lemma shows that the constraints \( z^\top P_{D,F}^t S_{N_F} P_{D,F} w = 0 \) are classical primal constraints (Definition 2.16), and so for each column \( z \) of \( \Phi_{G,add} \) in (4.3), we can use
\[
M_F(R_{iF}z_i - R_{jF}z_j)
\]
as an additional column in \( Q_G \) (after a modified Gram-Schmidt orthonormalization).

Secondly, we investigate the structure of the eigenproblem (4.1). The next lemma reduces the eigenproblem on \( \tilde{W}_{N_F}^F \) (Strategy 4) to an eigenproblem on a subspace of \( U_G \).

**Lemma 5.14.** Let \( F \) be a face shared by subdomains \( \mathcal{N}_F = \{i, j\} \). Then the corresponding generalized eigenproblem of Strategy 4, i.e., finding eigenpairs \((y, \lambda) \in \tilde{W}_{N_F}^F \times \mathbb{R}, \)
\[
\langle S_{N_F} y, z \rangle = \lambda \langle P_{D,F}^t S_{N_F} P_{D,F} y, z \rangle \quad \forall z \in \tilde{W}_{N_F}^F
\]
is (up to infinite eigenvalues) equivalent to finding eigenpairs \((\tilde{y}_F, \lambda) \in U_{F_{\Delta}} \times \mathbb{R}, \)
\[
\langle (S_{iF}^* : S_{jF}^*) \tilde{y}_F, \tilde{z}_F \rangle = \lambda \langle M_F \tilde{y}_F, \tilde{z}_F \rangle \quad \forall \tilde{z}_F \in U_{F_{\Delta}},
\]
where \( \tilde{y}_F := R_{iF}y_i - R_{jF}y_j, \tilde{z}_F := R_{iF}z_i - R_{jF}z_j, \) \( U_{F_{\Delta}} = \{ q \in U_F : Q_F^t q = 0 \} \), and \( M_F \) is the matrix from (5.9).

**Proof.** Let us first rewrite (5.10) using Lemma 5.13:
\[
\begin{bmatrix} z_i \\ z_j \end{bmatrix}^\top \begin{bmatrix} S_i & 0 \\ 0 & S_j \end{bmatrix} \begin{bmatrix} y_i \\ y_j \end{bmatrix} = \lambda (R_{iF}z_i - R_{jF}z_j)^\top M_F(R_{iF}y_i - R_{jF}y_j).
\]
Due to Definition 5.1, we have
\[
y_k = R_{kF}^t y_k + R_{kF}^t y_k \quad \text{for } k \in \{i, j\},
\]
for some vectors \( y_k, y_k \). Since \( y \in \tilde{W}_{N_F}^F \) (not \( \tilde{W}_{N_F}^F \)), we do not get any constraints on \( y_i, y_j \). Moreover, since \( P_{D,F} y \) is independent of \( y_i, y_j \), we can use the “Schur” Principle 4.4. With (5.2), we obtain that the eigenproblem (5.10) is (up to infinite eigenvalues) equivalent to
\[
\begin{bmatrix} z_i \\ z_j \end{bmatrix}^\top \begin{bmatrix} S_{iF}^* & 0 \\ 0 & S_{jF}^* \end{bmatrix} \begin{bmatrix} y_i \\ y_j \end{bmatrix} = \lambda (z_i - z_j)^\top M_F(y_i - y_j),
\]

where the eigenvectors and test vectors fulfill $Q_F^T(y_{iF} - y_{jF}) = 0$, $Q_F^T(z_{iF} - z_{jF}) = 0$, respectively. To get the last side condition explicitly, we use a simple transformation of variables:

$$
y_{iF} = \tilde{y}_{iF} + \frac{1}{2}\tilde{y}_{F}, \quad z_{iF} = \tilde{z}_{iF} + \frac{1}{2}\tilde{z}_{F},$
$$
y_{jF} = \tilde{y}_{jF} - \frac{1}{2}\tilde{y}_{F}, \quad z_{jF} = \tilde{z}_{jF} - \frac{1}{2}\tilde{z}_{F}.
$$

Since

$$y_{iF} - y_{jF} = \tilde{y}_{F}, \quad z_{iF} - z_{jF} = \tilde{z}_{F},$$

the condition $y, z \in \hat{W}_F$, is equivalent to

$$\tilde{y}_{F}, \tilde{z}_{F} \in U_F, \quad \tilde{y}_{F}, \tilde{z}_{F} \in U_{F\Delta}.$$

A straightforward calculation shows that

$$
\begin{bmatrix}
  z_{iF} \\
  z_{jF}
\end{bmatrix}^T
\begin{bmatrix}
  S^*_{iF} & 0 \\
  0 & S^*_{jF}
\end{bmatrix}
\begin{bmatrix}
  y_{iF} \\
  y_{jF}
\end{bmatrix} =
\begin{bmatrix}
  \tilde{z}_{F} \\
  \tilde{z}_{F}
\end{bmatrix}
+ \begin{bmatrix}
  \frac{1}{2}(S^*_{iF} + S^*_{jF}) & \frac{1}{2}(S^*_{iF} - S^*_{jF}) \\
  \frac{1}{2}(S^*_{iF} - S^*_{jF}) & \frac{1}{2}(S^*_{iF} + S^*_{jF})
\end{bmatrix}
\begin{bmatrix}
  \tilde{y}_{G} \\
  \tilde{y}_{G}
\end{bmatrix}.
$$

Hence, we can use the “Schur” Principle 4.4 once again and eliminate $\tilde{y}_{F}, \tilde{z}_{F}$ from the eigenproblem (5.12). The corresponding Schur complement of the matrix in (5.14) is given by

$$
\begin{align*}
\frac{1}{4} & [ (S^*_{iF} + S^*_{jF}) - (S^*_{iF} - S^*_{jF})(S^*_{iF} + S^*_{jF})^\dagger (S^*_{iF} - S^*_{jF}) ] \\
= & \frac{1}{4} [ (S^*_{iF} + S^*_{jF})(S^*_{iF} + S^*_{jF})^\dagger - (S^*_{iF} - S^*_{jF})(S^*_{iF} + S^*_{jF})^\dagger (S^*_{iF} - S^*_{jF}) ] \\
= & \frac{1}{4} [ 2S^*_{iF}(S^*_{iF} + S^*_{jF})^\dagger S^*_{jF} + 2S^*_{jF}(S^*_{iF} + S^*_{jF})^\dagger S^*_{iF} ] \\
= & S^*_{iF} : S^*_{jF};
\end{align*}
$$

cf. Definition 5.4 (Section 5.1).

**Remark 5.15.**

(i) The generalized eigenproblem $(S^*_{iF} : S^*_{jF})v = \lambda M_Fv$ has been used in [52, 54] and in [15, 95, 123, 124] for the deluxe scaling, which we further investigate in Section 6.1 below.

(ii) If we consider the original glob eigenproblem (4.1) (on $\hat{W}_{N_G}$), then we can still apply the “Schur” Principle 4.4 to the splitting (5.11). But the primal constraints enforced on the globs neighboring $F$ (i.e., globs $G \neq F$ with $|N_G \cap N_F| \geq 2$) result in an equivalent eigenproblem of the form

$$
\begin{bmatrix}
  z_{iF} \\
  z_{jF}
\end{bmatrix}^T
\begin{bmatrix}
  T_{iF} & T_{jF} \\
  T_{jF} & T_{jF}
\end{bmatrix}
\begin{bmatrix}
  y_{iF} \\
  y_{jF}
\end{bmatrix} = \lambda(z_{iF} - z_{jF})^T M_F(y_{iF} - y_{jF}),
$$

in general with $T_{iF} \neq 0, T_{jF} \neq 0$. In that case, the transformation (5.13) will lead to a matrix different than $(S^*_{iF} : S^*_{jF})$.

### 5.3. Globs shared by more than two subdomains

Recall that

$$(P_{D,G} y)_i = R^T_{iG} \sum_{j \in N_G(i)} D_{jG}(R_{iG} y_i - R_{jG} y_j).$$

Therefore, any new constraint of the form

$$(P_{D,G} z)^T S_{N_G} P_{D,G} y = 0$$
that we wish to impose, rewrites as

\[(5.15) \sum_{i \in N_G} \left( \sum_{j \in N_G \setminus \{i\}} (R_{iG} z_i - R_{jG} z_j) \right)^T D_{iG}^T S_{iG} \left( \sum_{j \in N_G \setminus \{i\}} D_{jG} (R_{iG} y_i - R_{jG} y_j) \right) = 0.\]

(The matrix on the left-hand side is related to but substantially different from the matrix \(A_E\) in [46, 47].) It is not hard to show that (5.15) has the form

\[(5.16) \sum_{j \in N_G} L_{jG} R_{jG} y_j = 0\]

(Condition (2.19) from Section 2.6). From Lemma 3.8(iii), we know that \(P_{D,G}\) vanishes for functions that are continuous across \(G\), from which we obtain \(\sum_{j \in N_G} L_{jG} = 0\) (Condition (2.21) from Section 2.6).

Appendix C shows that such generalized primal constraints (5.15) can be cast into an algorithm very similar to the original BDDC method [20], leading to independent subdomain problems and a sparse SPD coarse problem. Alternatively, in a FETI-DP framework, one can enforce the generalized primal constraints by deflation [41, 53, 58]; see also [46, 47]. In the next section, we suggest for BDDC to convert the constraints (5.16) into (stronger) classical primal constraints and show that this is more favorable.

5.4. Enforcing generalized primal constraints by (stronger) classical primal constraints*. In this section, we assume that we are given generalized primal constraints of the form (2.19) (or (5.16)). We show first how these can be enforced by classical primal constraints (cf. Definition 2.16). Although this can increase the total number of constraints, we are able to show in a second step that the coarse problem underlying the classical constraints is smaller or equal in its dimension to the coarse problem underlying the generalized constraints (while the condition number bound that we obtain for the generalized constraints also holds for the classical constraints).

Let \(G\) be an arbitrary but fixed glob \(G\) and consider one of the rows of the equation \(\sum_{j \in N_G} L_{jG} R_{jG} w_j = 0\), which we rewrite as

\[\sum_{j \in N_G} \ell_{jG}^T R_{jG} w_j = 0,\]

where \(\ell_{jG}\) is the column vector with the same entries as the selected row of \(L_{jG}\). Since the constraint above is non-trivial and because of (2.21), at least two of the vectors \(\ell_{jG}\) are non-zero. We select \(j^* \in N_G\) such that \(\ell_{j^*G}\) is non-zero. Without loss of generality, we assume that

\[N_G = \{1, \ldots, n\}, \quad j^* = 1,\]

and introduce the simplified notation

\[w_{jG} := R_{jG} w, \quad j \in N_G.\]

Next, we define a transformation of variables:

\[(5.17) \begin{cases} \tilde{w}_{1G} := \frac{1}{n} \sum_{j=1}^{n} w_{jG}, \\ \tilde{w}_{jG} := w_{jG} - \tilde{w}_{1G}, \quad \forall j = 2, \ldots, n. \end{cases}\]

The inverse transformation is given by

\[(5.18) \begin{cases} w_{1G} = \tilde{w}_{1G} - \frac{1}{n} \sum_{k=2}^{n} \tilde{w}_{kG}, \\ w_{jG} = \tilde{w}_{1G} - \frac{1}{n} \sum_{k=2}^{n} \tilde{w}_{kG} + \tilde{w}_{jG}, \quad \forall j = 2, \ldots, n. \end{cases}\]
Using (2.21) one can show that $\sum_{j\in N_G} \ell^T_G w_j G = \sum_{j=2}^n \ell^T_G w_j G$, and so,

\begin{align}
\ell^T_{kG}(w_i G - w_j G) &= 0 & \forall i, j \in N_G \forall k \in N_G \setminus \{j^*\} \\
\implies \ell^T_{jG} \tilde{w}_j G &= 0 & \forall j = 2, \ldots, n \\
\sum_{j\in N_j} \ell^T_{jG} w_j G &= 0.
\end{align}

(5.19)  (5.20)

The first line is in a suitable form for classical primal constraints only that we should orthonormalize the vectors $\ell_{kG} u_k G$ and possibly drop some of them. Because of (2.21), the space of vectors $\ell_{w_i G} u_i G$ fulfilling (5.19) is independent of the choice of the distinguished index $j^*$. If $|N_G| = 2$, then (5.19) and (5.20) are equivalent.

From the development above, it becomes clear that in any case we end up with a matrix $Q^T_G$ of full row rank such that for some matrix $T_G$ of full column rank,

\begin{align}
L^C_G := \begin{bmatrix}
\vdots \\
L_{jG} \\
\vdots \\
\end{bmatrix}
= T_G Q^T_G, & \quad \text{rank}(L^C_G) = \text{rank}(Q^T_G) \\
\end{align}

(5.21)

($L^C_G$ is a block column vector). A primal dof matrix $Q^T_G$ fulfilling the above can be obtained in various ways. Theoretically, we just have to remove linearly dependent rows from $L^C_G$. In practice, one can use the (thin) QR factorization (either implemented via Householder, Givens, or (modified) Gram-Schmidt; cf. [37, Section 5.2]):

$$L^C_G = \begin{bmatrix} Q_1 | Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix} = Q_1 R_1,$$

set $T_G := Q_1$, $Q^T_G := R_1$

such that $Q^T_G$ is even upper triangular. Note that the QR factorization is also used in the algorithm proposed in [79, Section 5]. In any case, the number of classical primal dofs on the glob $G$ is given by

$$n_{\Pi G} = \dim(U_{\Pi G}) = \text{rank}(L^C_G).$$

Following this construction for all globs results in classical primal dofs $\{Q^T_G\}_{G \in \mathcal{G}}$ and the corresponding space from Definition 2.22, which we denote by $\tilde{W}$ in order to distinguish it from $\hat{W}$ defined by (2.16). From Proposition 2.24, we obtain:

**Corollary 5.16.** Let $\tilde{W}$ be as in Definition 2.22 based on the classical primal dofs $Q^T_G$ from (5.21), and let

$$\tilde{W}_\Delta = \bigotimes_{i=1}^N \tilde{W}_i \Delta, \quad \tilde{W}_i \Delta := \{ w_i \in W_i : \forall G \in \mathcal{G}_i : Q^T_G R_{iG} w_i = 0 \}$$

(cf. (2.13)). Then for any space $\tilde{W}_\Pi$ fulfilling $\tilde{W} = \tilde{W}_\Pi \oplus \tilde{W}_\Delta$,

$$\dim(\tilde{W}_\Pi) = n_\Pi = \dim(U_{\Pi}) = \sum_{G \in \mathcal{G}} n_{\Pi G} = \sum_{G \in \mathcal{G}} \text{rank}(L^C_G).$$

Before we can state the main theorem of this section, we need to discuss the dimension of the more general space $\hat{W}$ of the form (2.16), (2.19). Let $r_{\Pi G}$ denote the number of (linearly
independent) constraints on \( G \), i.e., the number of linearly independent rows of the equation

\[
\sum_{j \in G} L_{jG} R_{jG} w_j = 0.
\]

Since each \( R_{jG} \) is surjective,

\[
\Pi^G_j \quad \text{rank} \quad L_{jG} R_{jG} q_{\Pi G},
\]

where \( L_{jG} R_{jG} \) is a block row vector, opposed to \( L_{jG} \). Moreover, it is easily seen that

\[
\dim(P) \Delta \geq \sum_{i = 1}^N \dim(W_i) - \sum_{G \in G} \Pi^G_j.
\]

We define the generalized dual spaces

\[
W_i \Delta := \{ w_i \in W_i : \forall G \in G_i : L_{iG} R_{iG} w_i = 0 \}, \quad W_\Delta := \oplus_{i = 1}^N W_i \Delta
\]
as well as the numbers

\[
\Pi^G_j := \text{rank}(L_{jG}), \quad \Pi^G_i := \sum_{G \in G_i} \Pi^G_j.
\]

**Proposition 5.17.** Let \( \tilde{W} \) be the space based on generalized primal constraints given by (2.16), assume that (2.19), (2.21) hold, and let \( W_\Delta \) as in (5.24). Then

(i) \( W_\Delta \subset \tilde{W} \);  
(ii) the space \( W_\Delta \) in (5.24) is the maximal subspace of \( \tilde{W} \) which has the form \( \otimes_{i = 1}^N V_i \);  
(iii) \( \dim(W_\Delta) = \dim(W_i) - \Pi^G_i \) with \( \Pi^G_i \) from (5.25).

(iv) for any complementary space \( \tilde{W}_\Pi \) fulfilling \( \tilde{W} = \tilde{W}_\Pi \oplus W_\Delta \),

\[
\dim(\tilde{W}_\Pi) = \sum_{i = 1}^N \Pi^G_i - \sum_{G \in G} \Pi^G_j \quad \text{with} \quad \Pi^G_j \text{from (5.22)},
\]

**Proof.** Parts (i)–(iii) can easily be verified. Since the sum in Part (iv) is direct, we obtain from (5.23) and Part (iii) that

\[
\dim(\tilde{W}_\Pi) = \dim(\tilde{W}) - \sum_{i = 1}^N \dim(W_i \Delta)
\]

\[
= \sum_{i = 1}^N \dim(W_i) - \sum_{G \in G} \Pi^G_j - \sum_{i = 1}^N \left( \dim(W_i) - \sum_{G \in G_i} \Pi^G_j \right)
\]

\[
= \sum_{G \in G} \Pi^G_j + \sum_{G \in G} \sum_{j \in G} \Pi^G_j = \sum_{i = 1}^N \sum_{G \in G_i} \Pi^G_i.
\]

We next state the main result of this section.

**Theorem 5.18.** Let \( \tilde{W} \) be the space based on generalized glob constraints given by (2.16), and let \( \tilde{W}_\Delta \) denote the corresponding dual space from (5.24). Then, for \( \tilde{W}, \tilde{W}_\Delta \) as in Corollary 5.16,

\[
\tilde{W} \subset \tilde{W}, \quad \dim(\tilde{W}_\Delta) \leq \dim(\tilde{W}_\Delta),
\]
and for any complementary spaces \( \tilde{W}_\Pi, \tilde{W}_\Pi \) with \( \tilde{W} = \tilde{W}_\Pi \oplus \tilde{W}_\Delta \) and \( \tilde{W} = \tilde{W}_\Pi \oplus \tilde{W}_\Delta \),

\[
\dim(\tilde{W}_\Pi) \leq \dim(\tilde{W}_\Pi).
\]

Let us first rephrase the statement of Theorem 5.18 based on the following observation. According to [20, 75] (or Appendix C), the action of \( \tilde{I} \tilde{S}^{-1} \tilde{I}^\top \) can be performed by independent subdomain problems and a sparse SPD coarse problem of dimension \( \dim(\tilde{W}_\Pi) \). Correspondingly, the operator \( \tilde{I} \tilde{S}^{-1} \tilde{I}^\top \) involving the more general space \( \tilde{W} \) leads to a coarse problem of size at least \( \dim(\tilde{W}_\Pi) \). Actually, we show in Appendix C that the coarse problem is of size exactly equal to \( \dim(\tilde{W}_\Pi) \). So,

(i) although in \( \tilde{W} \) more constraints are enforced than in \( \tilde{W} \), working with the space \( \tilde{W} \)
leads to a coarse problem of lower dimension (thus more efficiently solvable) than
for \( \tilde{W} \).

(ii) At the same time, we obtain from Remark 2.11 that at high probability, the smaller
space \( \tilde{W} \) leads to a smaller condition number as well.

\[\begin{align*}
\text{Summarizing, the advantages of using the (stronger) classical primal dofs from (5.21) clearly prevails.}
\end{align*}\]

The simple example in Figure 5.2 shows that this is (although counter-intuitive) indeed possible.

Remark 5.19. If the constraints are imposed by deflation in a FETI-DP framework [41, 46, 47, 53, 58], things turn around: Since there, the number of dofs in the second coarse problem equals the number of constraints, it is better to use the original constraints (2.19) (or (5.16)) in the deflation process.

Proof of Theorem 5.18. The first two statements follow from Definition 2.16, (2.16), and (5.19)–(5.20). The remainder of the proof is devoted to the inequality relating the primal space dimensions. Beforehand, recall the matrices \( L_G^j \) from (5.21). From Corollary 5.16 and Proposition 5.17, we obtain

\[
\dim(\tilde{W}_\Pi) = \sum_{G \in \mathcal{G}} \text{rank}(L_G^j),
\]

\[
\dim(\tilde{W}_\Pi) = \sum_{G \in \mathcal{G}} (q_{1G} - r_{1G}), \quad q_{1G} = \sum_{j \in N_G} q_{1jG}.
\]

We will show that each of the summands in the first line is less than or equal to the corresponding one in the second line. Therefore, we can consider a single glob \( G \in \mathcal{G} \) at a time. For a clearer presentation, we assume that \( N_G = \{1, \ldots, n\} \) and omit the subscripts \( G \) and \( \Pi \).
For each $j \in \mathcal{N}_G$ we consider the matrix $L_{jG}$, which may have incomplete row rank. However, we can find matrices $\bar{L}_{jG}$ of full row rank such that

\begin{equation}
L_{jG} = K_{jG} \bar{L}_{jG}, \quad q_{1jG} := \text{rank}(L_{jG}) = \text{rank}(\bar{L}_{jG}) \leq r_{1jG},
\end{equation}

for some matrix $K_{jG} \in \mathbb{R}^{q_{1jG} \times q_{1jG}}$, e.g., via the thin QR factorization [37, Section 5.2]. It is easy to see that

\begin{equation}
\text{rank}(L^C) = \text{rank}(\bar{L}^C), \quad \text{where } \bar{L}^C := \begin{bmatrix} \bar{L}_1 \\ \vdots \\ \bar{L}_n \end{bmatrix}.
\end{equation}

Therefore, we only have to show that

\begin{equation}
\text{rank}(\bar{L}^C) \leq q - r, \quad \text{where } q = \sum_{j=1}^n q_j.
\end{equation}

Recall that $L^R := [L_1|\cdots|L_n]$ and $r = \text{rank}(L^R)$; cf. (5.22). If $m$ is the number of dofs on the glob $G$, then

\begin{equation}
\text{dim}(\text{ker}(L^R)) = nm - r.
\end{equation}

A different characterization is related to the matrices $\{K_j\}_{j=1}^n$ from (5.26). From $L_j = K_j \bar{L}_j$ we derive

\begin{equation}
L^R = \begin{bmatrix} K_1 | \cdots | K_n \end{bmatrix}_{=:K} \begin{bmatrix} \bar{L}_1 \\ \vdots \\ \bar{L}_n \end{bmatrix}_{=:\bar{L}^D}.
\end{equation}

Since each $\bar{L}_j$ is surjective, so is $\bar{L}^D$, and we can conclude that

\begin{equation}
\text{dim}(\text{ker}(L^R)) = \text{dim}(\text{ker}(L^D)) + \text{dim}(\text{ker}(K)).
\end{equation}

From $\text{rank}(\bar{L}_j) = q_j$ it follows that $\text{dim}(\text{ker}(\bar{L}^D)) = nm - \sum_{j=1}^n q_j = nm - q$. Combining with (5.27), we obtain

\begin{equation}
\text{dim}(\text{ker}(K)) = q - r.
\end{equation}

Finally, recall that $\sum_{j=1}^n L_j = 0$, which can be rewritten as

\begin{equation}
K \bar{L}^C = \begin{bmatrix} K_1 | \cdots | K_n \end{bmatrix} \begin{bmatrix} \bar{L}_1 \\ \vdots \\ \bar{L}_n \end{bmatrix} = 0.
\end{equation}

In other words, the columns of $\bar{L}^C$ are in $\text{ker}(K)$, and so there can only be as many linearly independent columns as the dimension of $\text{ker}(K)$. To summarize,

\begin{equation}
\text{rank}(L^C) = \text{rank}(\bar{L}^C) \leq \text{dim}(\text{ker}(K)) = q - r.
\end{equation}
5.5. Alternative eigenproblem for subdomain edges. In this section, we show that using the transformation of variables (5.17)–(5.18) and Principle 4.9 (nearby eigenproblem), one can decouple the glob eigenproblem of Strategy 4 into \(|N_G| - 1\) independent eigenproblems, similar to Principle 4.10. The price to pay is a potentially larger set of constraints because (i) we use Strategy 4 and neglect the neighboring globs (cf. Definition 2.13) and (ii) replace the coupled eigenproblem by a decoupled one.

Let \(G\) be an arbitrary but fixed glob and assume without loss of generality that \(N_G = \{1, \ldots, n\}\). Recall the shortcut \(w_{iG} = R_j w_i\) as well as the transformation (5.17):

\[
\begin{align*}
\hat{w}_{1G} &= \frac{1}{n} \sum_{j=1}^{n} w_{jG}, \\
\tilde{w}_{jG} &= w_{jG} - w_{1G} \quad \forall j = 2, \ldots, n.
\end{align*}
\]

Notice for \(|N_G| = 2\) that this transformation (up to a positive or negative sign) is not biased towards either the first or second subdomain in \(G\). In contrast, for \(|N_G| > 2\), there is a clear bias towards the first subdomain.

**Lemma 5.20.** Under the assumptions above,

\[
\sum_{i \in N_G} |(P_{D,G} w)_i|^2 \leq (|N_G| - 1) \sum_{i=2}^{n} \tilde{w}_{iG} M_{iG} \tilde{w}_{iG},
\]

where

\[
M_{iG} := D_{iG}^T \left( \sum_{j \in N_G \setminus \{i\}} S_{jG} \right) D_{iG} + \left( \sum_{j \in N_G \setminus \{i\}} D_{jG}^T \right) S_{iG} \left( \sum_{j \in N_G \setminus \{i\}} D_{jG} \right),
\]

for \(i = 2, \ldots, n\).

For a face, i.e., \(G = F \in \mathcal{F}\), we have equality and \(M_{2F} = M_F\).

**Proof.** Firstly, observe that

\[
(P_{D,G} w)_i = R_i^T \sum_{j \in N_G \setminus \{i\}} D_{jG} (w_{iG} - w_{jG})
\]

\[
= R_i^T \begin{cases}
\sum_{j \in N_G \setminus \{i\}} -D_{jG} \tilde{w}_{jG} & i = 1, \\
(D_{1G} \tilde{w}_{iG} + \sum_{j \in N_G \setminus \{i\}} D_{jG} (\tilde{w}_{iG} - \tilde{w}_{jG})) & i \neq 1.
\end{cases}
\]

Using the above, we rewrite the expression

\[
\sum_{i \in N_G} (P_{D,G} w)_i^T S_i (P_{D,G} w)_i
\]

in the new variables \((\hat{w}_{1G}, \hat{w}_{2G}, \ldots, \hat{w}_{nG}), (\tilde{z}_{1G}, \tilde{z}_{2G}, \ldots, \tilde{z}_{nG})\). The whole expression is independent of \((\hat{w}_{1G}, \tilde{z}_{1G})\); in particular, the diagonal entry corresponding to \(\hat{w}_{1G}, \tilde{z}_{1G}\) is simply zero. The diagonal entry corresponding to \(\tilde{w}_{kG}, \tilde{z}_{kG}\) computes as

\[
D_{kG}^T S_{kG} D_{kG} + \sum_{i=2}^{n} \left( D_{1G} \delta_{ik} + \sum_{j=2}^{n} D_{jG} (\delta_{ik} - \delta_{jk}) \right)^T S_{ik} \left( D_{1G} \delta_{ik} + \sum_{j=2}^{n} D_{jG} (\delta_{ik} - \delta_{jk}) \right)
\]

\[
= \begin{bmatrix}
\sum_{j \neq i}^{n} D_{jG} & k = i \\
D_{kG} & k \neq i
\end{bmatrix}
\]

\[
= D_{kG} \left( \sum_{j \neq k}^{n} S_{jG} \right) D_{kG} + \left( \sum_{j \neq k}^{n} D_{jG} \right)^T S_{kG} \left( \sum_{j \neq k}^{n} D_{jG} \right).
\]
The second inequality in Principle 4.10 yields the desired inequality.

Applying the whole idea of Principle 4.10 (decoupling), we have to compute the Schur complement of \( \text{diag}(S^*_G)_{i \in N_G} \) but in the transformed variables \((\tilde{w}_1, \tilde{w}_2, \ldots, \tilde{w}_n)\) eliminating \( \tilde{w}_{kG} \) for \( k = 2, \ldots, n \), which we call \( \tilde{S}^*_{kG} \) in the sequel. From Lemma D.6, we know that \( \tilde{S}^*_{kG} \) does not depend on the complementary space. Therefore, we may use the simpler transformation \((w_1, \ldots, w_n) \mapsto (w_1, \ldots, w_{(k-1)}, \tilde{w}_k, w_k, \ldots, w_n)\), where \( w_{kG} = w_k + \tilde{w}_{kG} \). When we write the operator \( \text{diag}(S^*_G)_{i \in N_G} \) in the new variables, then \((w_1, \hat{w}_k)\) are decoupled from the remaining variables. So, if we form the Schur complement eliminating \( w_j, j = 1, \ldots, n, j \neq k \), it suffices to take the Schur complement of \[
\begin{bmatrix}
S^*_{1G} + S^*_{kG} & S^*_{kG} \\
S^*_{kG} & S^*_{kG}
\end{bmatrix},
\]
which is
\[
\tilde{S}^*_{kG} = S^*_{kG} - S^*_{kG}(S^*_{1G} + S^*_{kG})^T S^*_{kG} = S^*_{kG} : S^*_G,
\]
where in the last step, we have used (5.5). Principle 4.10 implies
\[
\sum_{k=2}^{n} w^T_{kG}(S^*_{1G} : S^*_G) \tilde{w}_{kG} \leq (n - 1) \sum_{j=1}^{n} w^T_{jG} S^*_j w_{jG},
\]
and we may alternatively study \( n - 1 \) decoupled eigenproblems of the form
\[(5.29)\]
\[
\tilde{z}^T_{iG}(S^*_{1G} : S^*_G) \tilde{y}_{iG} = \lambda \tilde{z}^T_{iG} M_G \tilde{y}_{iG} \quad \text{for } \tilde{y}_{iG}, \tilde{z}_{iG} \in U G \Delta,
\]
for \( i = 2, \ldots, n \), and with the matrix \( M_G \) from Lemma 5.20. Apparently, there is a bias towards the first subdomain.

**Remark 5.21.** If we compute the decoupled eigenproblems independently to form primal constraints, we have to orthonormalize eigenvectors originating from different eigenproblems. This can, however, lead to many unnecessary constraints. A more attractive strategy could be the following:

- Compute the eigenproblem for \( i = 2 \) and get adaptive constraints \( Q_{G2} \).
- For \( i = 3, \ldots, n \):
  - Project the eigenproblem \( i \) onto the space orthogonal to \( Q_{G2}, \ldots, Q_{G(i-1)} \).
  - Compute the constraints \( Q_{Gi} \).
- Use \( Q_{G2}, \ldots, Q_{Gn} \) as set of adaptive constraints.

(This corresponds to updating \( U G \Delta \) each time in the spirit of a Gauss-Seidel iteration.)

**5.6. A recombined edge eigenproblem.** A different recipe is to use Principle 4.11 and recombine the decoupled eigenproblems (5.29) into a single one:
\[
\tilde{z}^T_G(S^*_{1G} : S^*_G : \ldots : S^*_n) \tilde{y}_G = \lambda \tilde{z}^T_G (M_{2G} + \ldots + M_n) \tilde{y}_G \quad \text{for } \tilde{y}_G, \tilde{z}_G \in U G \Delta,
\]
where \( M_{iG} \) is the matrix from Lemma 5.20. Due to the Cauchy-Bunyakovsky-Schwarz inequality,
\[
M_{iG} \leq \sum_{j \in N_G \setminus \{i\}} D^T_{jG} S_{jG} D_{jG} + (|N_G| - 1) \sum_{j \in N_G \setminus \{i\}} \underbrace{D^T_{jG} S_{jG} D_{jG}}_{=: A_{iG}}.
\]
Therefore,
\[
\sum_{i=2}^{n} M_{iG} \leq \sum_{i=2}^{n} \sum_{j \in N_G \setminus \{i\}} D^T_{jG} S_{jG} D_{jG} + (n - 1) \sum_{i=2}^{n} A_{iG} \leq |N_G| \sum_{i=1}^{n} A_{iG}.
\]
Applying Principle 4.9 (nearby eigenproblem with constant $c_2 = |\mathcal{N}_G| = n$) yields the eigenproblem

$$
\bar{z}_G^T (S_{1G} \cdot S_{2G} \cdot \ldots \cdot S_{nG}) y_G = \lambda \bar{z}_G^T (A_{1G} + \ldots + A_{nG}) y_G \quad \text{for } y_G, \bar{z}_G \in \mathbb{U}_{\Delta G},
$$

which is the one proposed by Kim, Chung, and Wang [46, 47].

5.7. Comments on the adaptive algorithm. In general, adaptively chosen constraints can be enforced in several ways. Firstly, one can just add them to the previously chosen ones (if there are any) and recompute some components of BDDC. Secondly, for FETI-DP, the newly chosen constraints can be enforced by deflation; see [41, 46, 47, 53, 58]. Suppose, we want to add adaptively chosen constraints to the existing primal constraints, then we fall into one of the two cases below.

(i) If the chosen glob eigenproblems discard the influence of their neighboring globs (or if the neighboring globs are all totally primal), then they can be computed independently from each other.

(ii) Otherwise, one has to make an additional choice where either after computing the adaptive constraints on a single glob, one would update at once the global set of primal constraints (in the spirit of a Gauss-Seidel iteration), or not (like a Jacobi iteration). In the first case, of course the ordering of the globs matters.

In several publications [45, 53, 54, 76, 79], it is proposed to use a fixed tolerance as bound for the eigenvalues and use all the corresponding eigenvectors simultaneously for constraints. A different option is to impose one constraint at a time and update the neighboring eigenproblems at once; see also Remark 5.21.

6. The deluxe scaling. The **deluxe scaling** was originally introduced in [23] for 3D $H(\text{curl})$ problems and further used in, e.g., [8, 9, 14, 18, 23, 24, 54, 69, 88]. Recall the definition of $S_{iG}$ from (5.1), and set

$$
\bar{S}_G := \sum_{j \in \mathcal{N}_G} S_{jG}.
$$

The **deluxe scaling** is the following choice of the scaling matrices $D_{iG}$ from Assumption 3.4:

$$
D_{iG} = \bar{S}_G^{-1} S_{iG}.
$$

It is easily seen that $\bar{S}_G$ is a principal minor of the original problem matrix $\bar{S}$ and as such non-singular. The application of the inverse $\bar{S}_G^{-1}$ can be realized in several ways. Firstly, applying $\bar{S}_G^{-1}$ is equivalent to solving an SPD matrix problem on the subdomains $\mathcal{N}_G$ sharing the glob $G$ [8]. Secondly, some sparse direct solvers such as MUMPS [2] or PARDISO [66] offer a Schur complement option to compute the dense matrices $S_{jG}$ in a complexity comparable to a direct subdomain solve (see also Remark 5.3). The latter option might be quite interesting for computations on a large number of cores [123, 124].

By construction, choice (6.1) fulfills the glob-wise partition of unity property (Condition 3.5). Note that it is not guaranteed that each single matrix $S_{jG}$ is non-singular. For example, for the standard FEM-discretization of Poisson’s problem or linear elasticity, the matrix $S_{kF}$ corresponding to Figure 5.1(left), is singular.

6.1. Deluxe scaling on faces. Recall that for a face $F$ with $\mathcal{N}_F = \{i, j\}$,

$$
z^T P_{D,F} S_{\mathcal{N}_F} P_{D,F} y = (R_{iF} z_i - R_{jF} z_j)^T M_F (R_{iF} y_i - R_{jF} y_j)
$$

with

$$
M_F = D_{iF}^T S_{jF} D_{iF} + D_{jF}^T S_{iF} D_{iF} = S_{iF} \bar{S}_F^{-1} S_{jF} \bar{S}_F^{-1} S_{iF} + S_{jF} \bar{S}_F^{-1} S_{iF} \bar{S}_F^{-1} S_{jF}
$$

for the deluxe scaling.
Whereas it has been shown in many references [8, 18, 24] that $D_i F S_{iF} D_j F \leq S_{iF}$ and $D_i F S_{iF} D_j F \leq S_{jF}$. The inequality in Lemma 5.10 implies $D_j F^T S_{iF} D_j F \leq S_{iF} : S_{jF}$; see also [53], and so

$$M_F \leq 2(S_{iF} : S_{jF}).$$

The core of Lemma 5.10, however, implies the surprising result:

**Corollary 6.1.** If $F$ is a face with $N_F = \{i, j\}$ and if $D_i F, D_j F$ are chosen according to the deluxe scaling (6.1), then the following identity holds for $M_F$ (defined in (5.9)):

$$M_F = S_{iF} : S_{jF}. $$

Using Corollary 6.1, the eigenproblem in Lemma 5.14 (under the stated assumptions!) rewrites as

$$Z_F^T (S_{iF}^* : S_{jF}^*) \tilde{y}_F = \lambda Z_F^T (S_{iF} : S_{jF}) \tilde{y}_F \quad \text{for } \tilde{y}_F, \tilde{z}_F \in U_{F, \Delta}.$$

We warn the reader that possible constraints enforced on globs neighboring $G$ are ignored in the above eigenproblem, whereas they are present in the original eigenproblem (4.1).

**Remark 6.2.** Assume that $S_{iF}$ is spectrally equivalent to $\alpha_k S_{iF}, k \in \{i, j\}$, and $S_{jF}^*$ to $\alpha_k S_{jF}^*, k \in \{i, j\}$, with constant coefficients $\alpha_k > 0$ and with benign equivalence constants. Due to Proposition 5.9,

$$(\alpha_i S_{iF}) : (\alpha_j S_{jF}) = \frac{\alpha_i \alpha_j}{\alpha_i + \alpha_j} S_{iF}, \quad (\alpha_i S_{iF}^*) : (\alpha_j S_{jF}^*) = \frac{\alpha_i \alpha_j}{\alpha_i + \alpha_j} S_{iF}^*.$$

Together with Proposition 5.8 we can instead study the eigenproblem

$$Z_F^T S_{iF}^* \tilde{y}_F = \lambda Z_F^T S_{iF} \tilde{y}_F \quad \text{for } \tilde{y}_F, \tilde{z}_F \in U_{F, \Delta}.$$

For the case of scalar diffusion, $S_{iF}^*$ corresponds to the $H^{1/2}(F)$-norm and $S_{iF}$ to the $H^{1/2}(F)$-norm; see [113]. The coefficient-dependent scaling $D_{kF} = \frac{\alpha_k}{\alpha_i + \alpha_j} I$ (sometimes called $\rho$-scaling, cf. [100, 113]) leads to the same eigenproblem.

**Remark 6.3.** As noted by Stefano Zampini [123, 124], if we compute the eigenproblem on the space $U_F$ instead of $U_{F, \Delta}$ and if $S_{iF}, S_{jF}, S_{iF}^*, S_{jF}^*$ are all definite, then one can apply the formula from Remark 5.5 and rewrite the eigenproblem as

$$(S_{iF}^{-1} + S_{jF}^{-1}) v = \lambda (S_{iF}^* S_{iF}^{-1} + S_{jF}^* S_{jF}^{-1}) v.$$

### 6.2. Optimality of the deluxe scaling for subdomain faces*

The following lemma can be seen as a matrix version of Corollary 5.11.

**Lemma 6.4.** Let $A, B \in \mathbb{R}^{n \times n}$ be SPSD matrices with $A + B$ definite and define

$$M_{A,B}(X) := X^T A X + (I - X)^T B(I - X).$$

Then for any (fixed) symmetric positive definite matrix $C \in \mathbb{R}^{n \times n}$, the functional

$$J_{A,B,C}(X) = \text{tr}(C M_{A,B}(X) C)$$

attains its global minimum at

$$X_* = (A + B)^{-1} B,$$

where $\text{tr}(M) := \sum_{i=1}^n M_{ii}$ denotes the trace of the matrix $M \in \mathbb{R}^{n \times n}$. 


The term on the left-hand side is the sum over factors that we could potentially obtain in the
rest follows from Lemma 6.4.

Lemma 5.14. Assume further that
quadratic condition number bound, so minimizing the
optimization problem. But under the outlier assumption,
liers where

Proof. Let us first assume that \( C = I \). From the properties of the trace, we see that for
any \( X, Y \in \mathbb{R}^{n \times n} \),

\[
J_{A,B,I}(X + Y) = J_{A,B,I}(X) + 2\text{tr}(Y^TAX + Y^TB(X - I)) + \text{tr}(Y^T(A + B)Y).
\]

Since \( \langle M_1, M_2 \rangle_F := \text{tr}(M_1^T M_2) \) is an inner product on \( \mathbb{R}^{n \times n} \), we find that the gradient of
\( J_{A,B,I} \) at \( X \) is given by \( AX + B(X - I) \). The gradient vanishes if and only if

\[
(A + B)X = B.
\]

Since the expression \( \text{tr}(Y^T(A + B)Y) \) is positive unless \( Y = 0 \), we have the global minimum.
For a general SPD matrix \( C \), one easily sees that

\[
CM_{A,B}(X)C = M_{\tilde{A}, \tilde{B}}(\tilde{X}),
\]

where \( \tilde{A} = CAC, \tilde{B} = CBC \), and \( \tilde{X} = C^{-1}XC \). From the earlier case, the minimum of
\( J_{A,B,C}(X) = J_{\tilde{A}, \tilde{B}, I}(\tilde{X}) \) is attained at \( \tilde{X}_* = (\tilde{A} + \tilde{B})^{-1}\tilde{B} \). Transforming back reveals the
formula for \( X_* \).

Corollary 6.5. Let \( F \) be a subdomain face, and let \( 0 \leq \lambda_1(X) \leq \cdots \leq \lambda_n(X) \leq \infty \)
denote the generalized eigenvalues of

\[
(S_i^F : S_j^F)y = \lambda M_{S_i^F,S_j^F}(X)y \quad \text{for } y \in U_F,
\]

so for \( X = D_{jF} \) and \( I - X = D_{iF} \), the matrix on the right-hand side equals \( M_F \) from
Lemma 5.14. Assume further that \( S_i^F : S_j^F \) is non-singular such that \( \lambda_1(X) > 0 \). Then the
choice \( X = D_{jF} = (S_i^F + S_j^F)^{-1}S_j^F \) according to the deluxe scaling minimizes

\[
\mathcal{J}(X) := \sum_{i=1}^n \lambda_i(X)^{-1}.
\]

Proof. We set \( C = (S_i^F : S_j^F)^{-1/2} \), where \( (S_i^F : S_j^F)^{1/2} \) is the SPD matrix square
root. Then \( 0 \leq \lambda_n(X)^{-1} \leq \cdots \leq \lambda_1(X)^{-1} < \infty \) are the regular eigenvalues of the matrix
\( CM_{S_i^F,S_j^F}(X)C \) (recall that we have set \( \infty^{-1} := 0 \)) and,

\[
\mathcal{J}(X) = \text{tr}(CM_{S_i^F,S_j^F}(X)C).
\]

The rest follows from Lemma 6.4.

In a practical algorithm, one would actually like to minimize the number \( m \) of out-
liers where \( \lambda_1(X) \leq \cdots \leq \lambda_m(X) \leq \lambda_{m+1}(X) \), but this would lead to a non-quadratic
optimization problem. But under the outlier assumption,

\[
\sum_{i=1}^m \lambda_i(X)^{-1} \approx \sum_{i=1}^n \lambda_i(X)^{-1} = \mathcal{J}(X).
\]

The term on the left-hand side is the sum over factors that we could potentially obtain in the
condition number bound, so minimizing the quadratic functional \( \mathcal{J}(X) \) appears to be a good
alternative.

Remark 6.6. The case of singular \( (S_i^F : S_j^F) \) is harder and left for future research.
Applying Principle 4.11 (recombination), we obtain the single
\[ S_{IF} \leq S_{IF\eta}, \quad S^*_IF \leq S^*_IF; \]
for details see [54]. The economic deluxe scaling (on face the \( F \) shared by the subdomains \( i \) and \( j \)) is given by
\[ D_{IF} := (S_{IF\eta} + S_{JF\eta})^{-1} S_{IF\eta}. \]

For sufficiently small \( \eta \), the computation of this matrix or its application to a vector is much cheaper than for the original deluxe scaling. In [24], only one layer of elements is used (\( \eta = h \)). From (6.2) and Lemma 5.10, we obtain
\[ M_F = D_{IF}^T S_{IF} D_{IF} + D_{JF}^T S_{IF} D_{JF} \leq D_{IF}^T S_{IF\eta} D_{IF} + D_{JF}^T S_{IF\eta} D_{JF} = S_{IF\eta} : S_{JF\eta}. \]

From (6.2) and Proposition 5.8, we obtain
\[ (S^*_{IF\eta} : S^*_{JF\eta}) \leq (S^*_{IF} : S^*_{JF}). \]

In [54], it is proposed to consider the face eigenproblem
\[ (S^*_{IF\eta} : S^*_{JF\eta}) v = \lambda (S_{IF\eta} : S_{JF\eta}) v. \]
In view of (6.2)–(6.4), this is an implicit application of Principle 4.9 (nearby eigenproblem).

### 6.3. Economic deluxe scaling on faces

**Economic versions of the deluxe scaling have been proposed in [24, 54].** Recall that in the typical application, the matrix \( S_i \) and the derived matrices \( S_{IF}, S^*_{IF} \) stem from the elimination of interior subdomain dofs. Replacing the original stiffness matrix \( K_i \) by the one just assembled over the elements at a distance \( \leq \eta \) from the face \( F \), one arrives at matrices \( S_{IF\eta}, S^*_{IF\eta} \) with the properties

\[ S_{IF} \leq S_{IF\eta}, \quad S^*_{IF} \leq S^*_{IF\eta}; \]

Next, we investigate the decoupled eigenproblem from Section 5.5. Suppose again that \( N_G = \{1, \ldots, n\} \), and set \( S^e_{IG} : = \sum_{j \in N_G \setminus \{i\}} S_j G = \bar{S}_G - S_i G \). Then, due to Lemma 5.10,
\[ M_{IG} = S_i G \bar{S}_G^{-1} S^e_{IG} \bar{S}_G^{-1} S_i G + S^e_{IG} \bar{S}_G^{-1} S_i G \bar{S}_G^{-1} S^e_{IG} = S_i G : S^e_{IG}. \]

Hence, the \( n - 1 \) decoupled eigenproblems from (5.29) rewrite as
\[ \bar{z}_{iG}(S^e_{IG} : S^e_{IG}) \bar{y}_{iG} = \lambda \bar{z}_{iG}(S^e_{IG} : S^e_{IG}) \bar{y}_{iG}, \quad \forall i = 2, \ldots, n. \]

Applying Principle 4.11 (recombination), we obtain the single eigenproblem
\[ \bar{z}_{iG}(S^e_{IG} : S^e_{2G} : \ldots : S^e_{nG}) \bar{y}_{iG} = \lambda \bar{z}_{iG}(T_{2G} + \ldots + T_{nG}) \bar{y}_{iG}, \quad \forall i = 2, \ldots, n, \]
where \( T_{iG} := S_{iG} : S^e_{IG} \). Applying Principle 4.9 (nearby eigenproblem), replacing the matrix on the right-hand side by \( T_{iG} + \ldots + T_{nG} \), results in the eigenproblem proposed by Calvo and Widlund [15, 119].

**REMARK 6.7.** Recall the eigenproblem (5.30),
\[ \bar{z}_{G}(S^*_{IG} : S^*_{2G} : \ldots : S^*_{nG}) \bar{y}_{iG} = \lambda \bar{z}_{G}(A_{1G} + \ldots + A_{nG}) \bar{y}_{iG}, \quad \forall i = 2, \ldots, n, \]
where \( A_{iG} := S_{iG} : S^*_{IG} \).
We treat two cases. Firstly, assume that $G = \sum_{j=1, j \neq i}^n D_{jG}^T S_{jG} D_{jG}$. For the deluxe scaling,

$$
(6.5) \sum_{i=1}^n A_{iG} = \sum_{i=1}^n \sum_{j=1, j \neq i}^n S_{jG} \tilde{S}_G^{-1} S_{jG} \tilde{S}_G^{-1} S_{jG} = \sum_{j=1}^n S_{jG} \tilde{S}_G^{-1} S_{jG} \tilde{S}_G^{-1} S_{jG} = \sum_{j=1}^n T_{jG},
$$

where in the last step, we have used Lemma 5.10. That means, for the deluxe scaling, one can get from the Kim-Chung-Wang eigenproblem to the Calvo-Widlund eigenproblem by Principle 4.9 (nearby eigenproblem) using the spectral inequality (6.5).

7. Achieving definiteness of $\tilde{S}$. In this section, we show that under the following mild assumptions, we can guarantee the definiteness of $\tilde{S}$ algorithmically.

**Assumption 7.1.** Each subdomain has at least one face.

**Assumption 7.2.** If $F$ is a face of the subdomain $k$ then

$$(S_k w_k = 0, \quad R_k F w_k = 0) \Rightarrow w_k = 0.$$

**Assumption 7.3.** For each $k = 1, \ldots, N$ either

1. ker$(S_k) = \{0\}$, or
2. the subdomain $k$ has two faces, or
3. the subdomain $k$ has only one face $F$, $N_F = \{k, \ell\}$, and the matrix

$$M_F = D_{kF}^T S_{\ell F} D_{kF} + D_{\ell F}^T S_{k F} D_{\ell F}$$

is definite on $U_{F\Delta} := \{u \in U_F : Q_{k F}^u u = 0\}$.

**Lemma 7.4.** If Assumptions 7.1–7.3 hold, then for each $G \in \mathcal{G}^*$,

$$\ker(S_{N_G}) \cap \text{range}(P_{D,G}) \cap \tilde{W}_{N_G} = \{0\}.$$

**Proof.** Throughout the proof, let $w \in \ker(S_{N_G}) \cap \text{range}(P_{D,G}) \cap \tilde{W}_{N_G}$ be arbitrary but fixed. From $w \in \text{range}(P_{D,G}) \cap \tilde{W}_{N_G}$ and Lemma 3.8, we obtain that

$$w = P_{D,G} y \quad \text{for some} \ y \in \tilde{W}_{N_G}.$$

We treat two cases. Firstly, assume that $G$ is a face shared by the subdomains $i$ and $j$ such that

$$w_i = (P_{D,G} y)_i = R_{iG}^T D_{jG} (R_{iG} y_i - R_{jG} y_j),$$

$$w_j = (P_{D,G} y)_j = -R_{jG}^T D_{iG} (R_{iG} y_i - R_{jG} y_j).$$

Assume now that $S_i w_i = 0$ and $S_j w_j = 0$. For $k \in \{i, j\}$ we apply Assumption 7.3:

1. If ker$(S_k) = \{0\}$, then $w_k = 0$.
2. If the subdomain $k$ has two faces, namely $G$ and $F'$, then we see from (7.2) that $R_{kF} w_k = 0$, and Assumption 7.2 implies that $w_k = 0$.
3. Finally, if the subdomain $k$ has only one face (namely $G$) and if $M_G$ is definite on $U_{G\Delta}$, then we have (using (7.2) and the fact that $S_{iG} = R_{iG} S_i R_{iG}^T$, etc.)

$$0 = |w_i|_{S_i}^2 + |w_j|_{S_j}^2$$

$$= (R_{iG} y_i - R_{jG} y_j)^T (D_{iG}^T S_{iG} D_{iG} + D_{jG}^T S_{jG} D_{jG}) (R_{iG} y_i - R_{jG} y_j).$$

Since $R_{iG} y_i - R_{jG} y_j \in U_{G\Delta}$ and since $M_G$ is definite on that space, we can conclude that $R_{iG} y_i - R_{jG} y_j = 0$. This is sufficient to conclude (from (7.2)) that $w_i = w_j = 0$. 

Secondly, assume that $G \in G^* \setminus F$. Due to our assumptions, $S_k w_k = 0$, for all $k \in N_G$. For any such $k$, since the subdomain $k$ has a face $F$ (cf. Assumption 7.1), we see from (7.1) and the formula for $P_{D,G}$ that $R_k F w_k = 0$. Assumption 7.2 implies that $w_k = 0$.

**Theorem 7.5.** Let Assumptions 7.1–7.3 hold. Assume further that for each glob $G \in G^*$ the glob eigenproblem

$$(\tilde{S}_{N_G}, \tilde{P}_{D,G}^T \tilde{S}_{N_G} \tilde{P}_{D,G}) \quad \text{on} \quad \tilde{W}_{N_G}$$

has no zero eigenvalues. Then $S$ is definite on $\tilde{W}$.

**Proof.** Let $G \in G^*$ be arbitrary but fixed, and set $A = \tilde{S}_{N_G}$, $P = \tilde{P}_{D,G}$, and $B = \tilde{B}_G$. Thanks to Lemma 7.4,

$$\ker(A) \cap \text{range}(P) = \{0\},$$

and due to our assumptions, $(A, B)$ has no genuine zero eigenvalues. Lemma 4.12(iii) implies that $\ker(A) \subset \ker(P)$, which means

$$\forall w \in \tilde{W}_{N_G} : (\forall j \in N_G : S_j w_j = 0 \implies \tilde{P}_{D,G} w = 0).$$

Due to Lemma 3.8(iii) the last identity implies

$$R_{iG} w_i - R_{jG} w_j = 0 \quad \forall i, j \in N_G.$$

Since $G \in G^*$ was arbitrary, Condition 3.1 is fulfilled, and Lemma 3.2 concludes the proof.

**Remark 7.6.**
(i) Assumption 7.1 usually holds in practice, otherwise we would have subdomains joined to the rest only by an edge or a vertex, which is somewhat unphysical.

(ii) Assumption 7.2 is fulfilled for the typical finite element discretizations and for the typical differential operators, provided that
- the face $F$ is large enough and
- each subdomain is connected.

Note that connectivity is a geometric concept that can, nevertheless, be made accessible via the matrix graph of the underlying sparse matrix; cf. [125].

(iii) Should neither Item 1 nor Item 2 of Assumption 7.3 hold, then Item 3 can be fulfilled by computing the eigenproblem

$$M_F = \lambda I$$

first and then converting any zero modes into primal constraints.

**Acknowledgements.** We would like to thank Walter Zulehner for helpful hints that have led us to the parallel sum of matrices. We are also grateful to Stefano Zampini, who provided us many hints and suggestions that improved the manuscript in a late stage; in particular he motivated us to introduce Section 5.6 and the recombinied edge eigenproblem in Section 6.4. Furthermore, we wish to thank Olof Widlund for fruitful discussions and helpful hints on the whole topic, Robert Scheichl and Marcus Sarkis for early discussions on FETI-DP and multiscale problems, and Johannes Kraus for detailed explanations and discussions on [65]. Finally, the first author is grateful for the good atmosphere at the Institute of Numerical
we find that there exists a projection \( r \). In addition, let \( T \). Using the Rayleigh quotient we obtain the lower bound for \( \lambda \). If (A.2) holds, then

\[
\lambda_{\max}(\Pi \tilde{A}^{-1} \Pi^T A) = \sup_{v \in H \setminus \{0\}} \frac{\langle A \Pi \tilde{v}, \Pi \tilde{v} \rangle}{\langle \Pi \tilde{v}, \Pi \tilde{v} \rangle} =: \gamma_2.
\]

In addition, let \( T : H \to \tilde{H} \) be a linear operator such that

\[
\Pi T = v \quad \text{and} \quad \gamma_1 \langle \tilde{A} T v, T v \rangle \leq \langle Av, v \rangle \quad \forall v \in H,
\]

for some constant \( \gamma_1 > 0 \). Then \( \lambda_{\min}(\Pi \tilde{A}^{-1} \Pi^T A) \geq \gamma_1 \). Summarizing,

\[
\kappa(\Pi \tilde{A}^{-1} \Pi^T A) \leq \gamma_2 / \gamma_1.
\]

**Proof.** With \( \|v\|_B := \langle B v, v \rangle^{1/2} \) for a positive definite \( B \) and basic functional analysis, we obtain \( \|v\|_{B^{-1}} = \sup_{v \in V \setminus \{0\}} \langle \psi, v \rangle \|v\|_B \) and \( \langle \psi, v \rangle \leq \|v\|_{B^{-1}} \|v\|_B \). Since the operator \( \Pi \tilde{A}^{-1} \Pi^T A \) is self-adjoint with respect to the inner product \( \langle Av, v \rangle \), its spectrum is real. We show (A.1) using the Rayleigh quotient and simply omit "\( \setminus \{0\} \)" in all suprema:

\[
\lambda_{\max}(\Pi \tilde{A}^{-1} \Pi^T A) = \sup_{v \in H} \frac{\langle \Pi \tilde{A}^{-1} \Pi^T A v, Av \rangle}{\langle Av, v \rangle} = \sup_{\psi \in H^*} \frac{\|\Pi^T \psi\|_{A^{-1}}^2}{\|\psi\|_{A^{-1}}^2} = \sup_{\psi \in H} \frac{\|\Pi^T \psi\|_{A^{-1}}^2}{\|\psi\|_{A^{-1}}^2}.
\]

If (A.2) holds, then

\[
\|\psi\|_{A^{-1}}^2 = \sup_{v \in H} \frac{\langle v, \Pi T v \rangle^2}{\|v\|_A^2} = \sup_{v \in H} \frac{\langle \Pi^T \psi, T v \rangle^2}{\|v\|_A^2} \leq \|\Pi^T \psi\|_{A^{-1}}^2 \sup_{v \in H} \frac{\|T v\|_A^2}{\|v\|_A^2} \leq \frac{1}{\gamma_1} \|\Pi^T \psi\|_{A^{-1}}^2.
\]

Using the Rayleigh quotient we obtain the lower bound for \( \lambda_{\min}(\Pi \tilde{A}^{-1} \Pi^T A) \). \( \Box \)
To get the BDDC condition number bound (see also [65]), we set $H := U \tilde{H} := \tilde{W}$, $A := \tilde{S}$, $\tilde{A} = \tilde{S}$, and $\Pi := E_D \tilde{I}$. Then bound (2.4) is equivalent to $\lambda_{\text{max}}(M^{-1}_{\text{BDDC}}\tilde{S}) \leq \omega$. To get (A.2), we first define $T: U \rightarrow \tilde{W}$ by $Tv := Rv$ for $v \in U$, which is well-defined since range$(R) \subset \tilde{W}$. From $E_D R = I$ we conclude that $\Pi T = E_D \tilde{I} T = I$. Finally, since $RE_D$ is a projection,

$$\langle \tilde{A} T v, T v \rangle = \langle S R v, R v \rangle = \langle \hat{S} v, v \rangle = \langle A v, v \rangle \quad \forall v \in \tilde{H} = U,$$

so the inequality in (A.2) holds with $\gamma_1 = 1$ and $\lambda_{\text{min}}(M^{-1}_{\text{BDDC}}\tilde{S}) \geq 1$.

**Appendix B. The related FETI-DP method.** Let $\Lambda$ be an Euclidean space (usually called space of Lagrange multipliers) and $B: \tilde{W} \rightarrow \Lambda$ be a matrix (usually called the jump operator) such that

$$\tilde{W} = \ker(B).$$

**Remark B.1.** Identity (2.3) already implies the existence of a matrix $B$ with $\tilde{W} = \ker(B)$. For standard choices of $B$, see, e.g., [34, 113]. Furthermore, $\tilde{W} \subset \tilde{W} \subset W$ (Condition 2.6) implies the existence of a matrix $\tilde{L}$ of full rank such that $\tilde{W} = \ker(\tilde{L})B$; see also [76, Section 2.3] and Remark 2.30.

With $\tilde{B} := B \tilde{I}: \tilde{W} \rightarrow \Lambda$, problem (2.1) can be rewritten as

$$\text{(B.1) } \text{find } (\tilde{u}, \lambda) \in \tilde{W} \times \Lambda: \begin{bmatrix} \tilde{S} & \tilde{B}^T \\ \tilde{B} & 0 \end{bmatrix} \begin{bmatrix} \tilde{u} \\ \lambda \end{bmatrix} = \begin{bmatrix} \tilde{g} \\ 0 \end{bmatrix},$$

where $\tilde{g} := \tilde{I}^T g$. Since the restriction of $\tilde{S}$ to $\ker(\tilde{B})$ is isomorphic to $\tilde{S}$, which was assumed to be definite, problem (B.1) is uniquely solvable up to adding an element from $\ker(\tilde{B}^T)$ to $\lambda$. If $S$ is definite on $\tilde{W}$ (Condition 2.7), then we can eliminate the variable $\tilde{u}$ and obtain the dual equation

$$\text{(B.2) } F \lambda = d,$$

where $F := B \tilde{I} \tilde{S}^{-1} \tilde{I}^T \tilde{B}^T$ and $d := B \tilde{I} \tilde{S}^{-1} \tilde{I}^T g$. We assume that there exists a matrix $B_D: \tilde{W} \rightarrow \Lambda$ such that

$$B_D^T B = P_D = I - RE_D.$$

**Remark B.2.** Under Assumption 3.4 and for fully redundant Lagrange multipliers, $B_D$ indeed exists. For the fully redundant setting, $\Lambda = \bigotimes_{G \in \mathcal{G}} \bigotimes_{i,j \in N_G, i > j} U_G$. We denote the components of $\lambda \in \Lambda$ by $\lambda_{G,ij}$, for $G \in \mathcal{G}$, $i > j \in N_G$, and define

$$(B u)_{G,ij} := R_{iG} u_i - R_{jG} u_j$$

(cf. (2.11)). The definition of $B_D$ then reads

$$(B_D w)_{G,ij} := D_{ijG} R_{iG} w_i - D_{ijG} R_{jG} w_j.$$ 

This generalizes the well-known formula for diagonal matrices $D_i$; see [113, Section 6.3.3] or [93, Section 2.2.4.2]. The transpose is given by

$$(B_D^T \mu)_i = \sum_{G \in \mathcal{G}} \sum_{j \in N_G \setminus \{i\}} \text{sign}(i - j) R_{iG}^T D_{jG} \mu_{G,ij},$$
from which one can infer that $B_D^T B = P_D$. The FETI-DP preconditioner (for problem (B.2)) is defined as

$$M^{-1}_{\text{FETI-DP}} := B_D S B_D^T : \Lambda \to \Lambda.$$  

(B.3)

In [12, 72, 75, 77] it was shown that the bound (2.5) (or equally (2.4)) implies

$$\kappa_{\text{FETI-DP}} := \kappa(M^{-1}_{\text{FETI-DP}} F_{\Lambda_i \ker(\tilde{\Lambda}_i)}) \leq \omega$$

and that the spectra of BDDC and FETI-DP (with corresponding components) are identical except for possible eigenvalues equal to 1.

Appendix C. Realization of $I \tilde{S}^{-1} I^T$. The method in Section C.1–C.2 treats the case of classical primal dofs (Section 2.5) and was introduced in [20]. For similar approaches see, e.g., [33], [113, Section 6.4], [72], [62, Section 4.2], and [93, Section 5.3]. In Section C.3, we extend the method to the generalized primal constraints from Section 2.6.

C.1. The energy minimizing basis of $\tilde{W}_\Pi$ for classical primal dofs. Let the matrices $C_i : W_i \to U_{i\Pi}$ fulfill $\ker(C_i^\top) = \{0\}$ (Condition 2.19), let $\tilde{W}$ be defined via (2.15), i.e.,

$$\tilde{W} = \{w \in W : \exists u_{\Pi} \in U_{\Pi} \forall i = 1, \ldots, N : C_i w_i = R_{i\Pi} u_{\Pi}\},$$

and $W_{i\Delta} = \{w_i \in W_i : C_i w_i = 0\}$, $W_{\Delta} := \bigotimes_{i=1}^N W_{i\Delta}$. Let $\Psi_i : W_{i\Pi} \to W_i$ fulfill

$$C_i \Psi_i = I.$$  

(C.1)

Such matrices $\Psi_i$ exist because $C_i$ is surjective, e.g., we could use $\Psi_i = C_i^\top (C_i C_i^\top)^{-1}$. A distinguished choice is defined by the linear saddle point system

$$\begin{bmatrix} S_i & C_i^\top \\ C_i & 0 \end{bmatrix} \begin{bmatrix} \Psi_i \\ \Lambda_i \end{bmatrix} = \begin{bmatrix} 0 \\ I \end{bmatrix},$$

with Lagrange parameters $\Lambda_i : W_{i\Pi} \to W_{i\Pi}$. Assume that $S_i$ is definite on $\ker(C_i) = W_{i\Delta}$ (cf. Condition (3.1)). Due to $\ker(C_i^\top) = \{0\}$ (Condition 2.19), problem (C.2) is guaranteed to have a unique solution.

The columns of $\Psi_i$ can be regarded as shape functions on the subdomain $i$. Condition (C.1) states that the primal dof $k$ of the shape function $j$ evaluates to $\delta_{kj}$.

Proposition C.1.

(i) $\Psi_i$ has full column rank,

(ii) $\text{range}(\Psi_i) \cap W_{i\Delta} = \{0\},$

(iii) if (C.2) holds then even

$$\langle S_i z_i, w_i \rangle = 0 \quad \forall z_i \in \text{range}(\Psi_i), w_i \in W_{i\Delta}.$$

Proof. Part (i) follows directly from (C.1).

Part (ii). If $w_i = \Psi_i v \in W_{i\Delta}$, then $0 = C_i w_i = C_i \Psi_i v = v$, so $v = 0$ and $w_i = 0$.

Part (iii). From the first line of (C.2) we derive that for any $w_i \in W_{i\Delta}$,

$$w_i^\top S_i \Psi_i = -w_i^\top C_i^\top \Lambda_i = -\Lambda_i^\top \left[ \sum_{j=1}^N C_j w_j \right]_{=0} = 0.$$  

[]

For each $i = 1, \ldots, N$, choose $\Psi_i : W_{i\Pi} \to W_i$ such that (C.1) holds. We set

$$\Psi := \text{diag}(\Psi_i)_{i=1}^N : W_{i\Pi} \to W$$
and define, in a finite element spirit, the assembled basis

\[(C.3) \quad \tilde{\Psi} : U_{\Pi} \to W, \quad \tilde{\Psi} := \Psi R_{\Pi} .\]

where \( R_{\Pi} : U_{\Pi} = \mathbb{R}^{n_\Pi} \to W_{\Pi} \) is the matrix from (2.12) and has full column rank.

**Lemma C.2.** Let \( \tilde{\Psi} \) be given as in (C.3). Then

(i) \( \tilde{\Psi} \) has full column rank, in particular, \( \dim(\text{range}(\tilde{\Psi})) = n_{\Pi} \).

(ii) \( \text{range}(\tilde{\Psi}) \subset \tilde{W} \).

(iii) \( \tilde{W} = \text{range}(\tilde{\Psi}) \oplus W_\Delta \).

(iv) If for each \( i = 1, \ldots, N \) (C.2) holds, then even

\[ \langle Sw, z \rangle = 0 \quad \forall w \in W_\Delta, \ z \in \text{range}(\tilde{\Psi}). \]

**Proof.** Part (i). Due to Proposition C.1(i), \( \Psi \) is injective. Since \( R_{\Pi} \) is injective, the composition \( \tilde{\Psi} \) is injective too.

Part (ii). Due to (C.1), for any \( G \in \mathcal{G} \) and \( i \in \mathcal{N}_G \):

\[
Q_G^T R_{G_G} (\tilde{\Psi})_i = \sum_{G' \in \mathcal{G}} R_{G_G} R_{G_G'} Q_{G'}^T R_{G_G} \Psi_i R_{\Pi} = R_{G_G} C_{G_G} R_{\Pi} = \hat{R}_{G_G},
\]

and so \( Q_G^T R_{G_G} (\tilde{\Psi})_i = Q_G^T R_{G_G} (\tilde{\Psi})_j \) for all \( i, j \in \mathcal{N}_G \).

Part (iii). From Proposition C.1(ii) we obtain \( \text{range}(\Psi) \cap W_\Delta = \{0\} \), so the sum is direct. Thanks to Part (i) and Proposition 2.24,

\[
\dim(\text{range}(\tilde{\Psi})) + \dim(W_\Delta) = n_{\Pi} + \dim(W_\Delta) = \dim(\tilde{W}),
\]

so together with Part (ii), the direct sum must equal \( \tilde{W} \).

Part (iv) follows directly from Proposition C.1(iii). \( \square \)

**C.2. Realization of \( \tilde{T} \tilde{S}^{-1} \tilde{T}^\top \).** For this section, we only make two assumptions. Firstly,

\[
\hat{W} = \text{range}(\tilde{\Psi}) \oplus W_\Delta,
\]

where \( \tilde{\Psi} : U_{\Pi} \to \hat{W} \) is injective and \( W_\Delta = \bigotimes_{i=1}^{N} W_i \) with \( W_i = \{ w_i \in U_i : C_i w_i = 0 \} \).

Secondly, we assume that \( \text{range}(\tilde{\Psi}) \) and \( W_\Delta \) are \( \tilde{S} \)-orthogonal (see Remark C.6 for the non-orthogonal case). Since the sum is direct, we can identify \( \hat{W} \) with the product space \( \hat{W} := U_{\Pi} \times W_\Delta \) and obtain

\[
\tilde{T} : \hat{W} \to W : \begin{bmatrix} u_{\Pi} \\ w_\Delta \end{bmatrix} \mapsto \tilde{\Psi} w_\pi + w_\Delta, \quad \tilde{T}^\top : W^* \to \hat{W}^* : f \mapsto \begin{bmatrix} \tilde{\Psi}^T f \end{bmatrix}^T.
\]

The operator \( \tilde{S} \) can then be identified with \( \tilde{S} : \hat{W} \to \hat{W}^* \) given by

\[
\begin{bmatrix} u_{\Pi} \\ v_\Delta \end{bmatrix}^T \tilde{S} \begin{bmatrix} u_{\Pi} \\ w_\Delta \end{bmatrix} = u_{\Pi}^T (\tilde{\Psi}^T S \tilde{\Psi}) u_{\Pi} + v_\Delta S w_\Delta,
\]

\[4\]To be strict, we actually add the embedded function \( w_\Delta \in W_\Delta \subset W \), and correspondingly, in the second component of \( \tilde{T}^\top f \), we would have to write the embedding of \( f \in W^* \subset W_\Delta^* \).
which is a block-diagonal operator. Its inverse is given by

\[ \tilde{S}^{-1} \begin{bmatrix} r_{\Pi} \\ r \end{bmatrix} = \begin{bmatrix} (\tilde{\Psi}^T S \tilde{\Psi})^{-1} r_{\Pi} \\ z_\Delta \end{bmatrix}, \]

where \( z_\Delta \in W_\Delta \) is such that \( \langle S z_\Delta, v_\Delta \rangle = \langle r, v_\Delta \rangle \) for all \( v_\Delta \in W_\Delta \). The latter can be obtained by solving the saddle point problem

\[ \begin{bmatrix} S & C^T \\ C & 0 \end{bmatrix} \begin{bmatrix} z_\Delta \\ \mu \end{bmatrix} = \begin{bmatrix} r \\ 0 \end{bmatrix}, \]

whose system matrix is block-diagonal with blocks identical to (C.2).

To summarize, the application \( v = \tilde{I} \tilde{S}^{-1} \tilde{I}^T r, r \in W \), is now realized by

\[ v = \tilde{\Psi} w_{\Pi} + z_\Delta, \]

where \( w_{\Pi} \in \mathbb{R}^{n_E} \) solves the (global) coarse problem

\[ (\tilde{\Psi}^T S \tilde{\Psi}) w_{\Pi} = \tilde{\Psi}^T r, \]

and the components \( z_i \) of \( z_\Delta \) solve the local (and independent) saddle point problems

\[ \begin{bmatrix} S_i & C_i^T \\ C_i & 0 \end{bmatrix} \begin{bmatrix} z_i \\ \mu_i \end{bmatrix} = \begin{bmatrix} r_i \\ 0 \end{bmatrix}. \]

**Remark C.3.** Certainly, the saddle point problems (C.2), (C.5) can either (i) be solved as they are, (ii) be reformulated by penalty techniques, or (iii) by using a transformation of basis [55, 72] one can enforce the constraints explicitly, eliminate some dofs, and reduce the saddle point problem to an SPD problem.

**Remark C.4.** For the energy minimizing construction (C.2), the coarse matrix in (C.4) can be assembled from the subdomain contributions \( \Psi_i^T S_i \Psi_i = -\Psi_i^T C_i^T \Lambda_i = -\Lambda_i \); cf. [93, Section 5.3.4.2].

**Remark C.5.** If \( S_i \) is a Schur complement of a matrix \( K_i \) eliminating interior dofs, then the saddle point problems (C.2) and (C.5) can easily be rewritten in terms of \( K_i \) and are thus amenable to sparse direct solvers. In that context, however, it is recommended to suitably scale the second line and to check for the right parameters such that the solver can cope with the zero block on the lower right (e.g., weighted matching [101] in case of PARDISO).

**Remark C.6.** Based on the block Cholesky factorization, a similar algorithm can also be given for the case that \( \text{range}(\tilde{\Psi}) \) is not \( S \)-orthogonal to \( W_\Delta \). Then, however, the coarse and the local problems are not anymore independent, and two local problems have to be solved; see [72] and [93, Section 5.3].

**C.3. A basis of \( \tilde{W}_{\Pi} \) for generalized primal constraints.** Let \( \tilde{W} \) be a space generated from generalized primal constraints, i.e., (2.16), (2.19). We give an algorithm by computing a basis of \( \tilde{W}_{\Pi} \) that has local support such that \( \tilde{W} = \tilde{W}_{\Pi} \oplus W_\Delta \), with \( W_\Delta \) defined in (5.24). We only require that \( S \) is definite on \( \tilde{W} \) (Condition 2.7).

**Step 1.** For each subdomain \( i \) and glob \( G \in \mathcal{G}_i \) we construct a matrix \( \tilde{L}_{iG} \in \mathbb{R}^{n_{iG} \times n_{\Pi G}} \) of full row rank such that

\[ \tilde{L}_{iG} = K_{iG} \tilde{L}_{iG}, \quad q_{\Pi G} = \text{rank}(\tilde{L}_{iG}) = \text{rank}(L_{iG}). \]
This can, e.g., be achieved by the QR factorization \[37, \text{Section 5.2}\] (see also the proof of Theorem 5.18). We collect them into a subdomain constraint matrix

\[
\bar{L}_i = \begin{bmatrix}
\vdots \\
\bar{L}_{iG}R_{iG} \\
\vdots
\end{bmatrix}_{G \in \mathcal{G}_i},
\]

which again has full row rank \(q_{\Pi_i} := \sum_{G \in \mathcal{G}_i} q_{\Pi_iG}\). The space from (5.24) rewrites as

\[
W_i = \{w_i \in W_i : \bar{L}_i w_i = 0\}.
\]

**Step 2.** For each subdomain \(i\), we construct a matrix \(\Psi_i : \mathbb{R}^{q_{\Pi_i}} \rightarrow W_i\) such that

\[
\bar{L}_i \Psi_i = I,
\]

e.g., we could use \(\Psi_i = \bar{L}_i^T (\bar{L}_i \bar{L}_i^T)^{-1}\). A distinguished choice are the energy-minimizing functions given by the solution of the saddle point system

\[
\begin{bmatrix}
S_i & \bar{L}_i^T \\
\bar{L}_i & 0
\end{bmatrix}
\begin{bmatrix}
\Psi_i \\
\Lambda_i
\end{bmatrix}
= 
\begin{bmatrix}
0 \\
I
\end{bmatrix}
\]

with Lagrange multipliers \(\Lambda_i \in \mathbb{R}^{q_{\Pi_i} \times q_{\Pi_i}}\).

**PROPOSITION C.7.** For a matrix \(\Psi_i\) fulfilling (C.9), the following statements hold:

(i) The columns of \(\Psi_i\) are linearly independent.

(ii) The system matrix in (C.10) is invertible.

(iii) If \(\Psi_i\) is constructed via (C.10), then \(\langle S_i \Psi_i, z_i \rangle = 0, \forall z_i \in W_i\).

**Proof.** Part (i) follows immediately from (C.9).

Part (ii). \(S_i\) is definite on \(\ker(\bar{L}_i) = W_i\Delta\) (cf. (3.1)), and \(\ker(\bar{L}_i^T) = \{0\}\).

Part (iii). From the first line in (C.10) and from (C.8) we derive for \(z_i \in W_i\Delta\),

\[
\langle S_i \Psi_i, z_i \rangle = -\langle \bar{L}_i^T \Lambda_i, z_i \rangle = -\langle \Lambda_i, \bar{L}_i z_i \rangle \overset{(C.8)}{=} 0.
\]

**Step 3.** Corresponding to (C.7), the shape functions are arranged into groups corresponding to the globs:

\[
\Psi_i = \left[ \cdots | \Psi_i^{(G)} | \cdots \right]_{G \in \mathcal{G}_i}.
\]

One easily shows the property

\[
\bar{L}_{iG} R_{iG} \Psi_i^{(G)} = \delta_{GG} I.
\]

**Step 4.** Next, we loop over all globs \(G \in \mathcal{G}\) and return to the original constraint matrices \(\{L_{jG}\}_{j \in \mathcal{N}_G}\). We form the matrix

\[
K_G = \left[ \cdots | L_{jG} R_{jG} \Psi_j^{(G)} | \cdots \right]_{j \in \mathcal{N}_G} = \left[ \cdots | K_{jG} | \cdots \right]_{j \in \mathcal{N}_G} \in \mathbb{R}^{q_{\Pi_G} \times q_{\Pi_G}}
\]

and compute a coefficient matrix

\[
Y_G := \begin{bmatrix}
y_j G \\
\vdots \\
y_j G
\end{bmatrix}_{j \in \mathcal{N}_G} \in \mathbb{R}^{q_{\Pi_G} \times q_{\Pi_G}},
\]
whose columns form a basis of $\ker(K_G)$, i.e.,

\[(C.13) \quad K_G Y_G = 0, \quad n_{11G} = \text{rank}(Y_G) = \dim(\ker(K_G)).\]

This can, e.g., be done by a singular value decomposition (SVD); see [37, Section 2.5]. As we have shown in (5.28) in the proof of Theorem 5.18,

\[(C.14) \quad \dim(\ker(K_G)) = q_{11G} - r_{11G}.\]

**Step 5.** The number $n_{11G}$ will be the number of coarse basis functions used on the glob $G$. Therefore, the global space of coarse dofs is given by $U_\Pi := \mathbb{R}^{n_\Pi}$ with $n_\Pi = \sum_{G \in \mathcal{G}} n_{11G}$. The coarse basis itself is given by

\[
\tilde{\Psi} : U_\Pi \rightarrow W, \quad \tilde{\Psi} := \left[ \cdots | \tilde{\Psi}_{1}^{(G)} | \cdots \right]_{G \in \mathcal{G}},
\]

where

\[
\tilde{\Psi}_{i}^{(G)} : \mathbb{R}^{n_{11G}} \rightarrow W : \tilde{\Psi}_{i}^{(G)} := \begin{cases} 
\Psi_{i}^{(G)} Y_G & i \in \mathcal{N}_G, \\
0 & \text{otherwise.}
\end{cases}
\]

**Theorem C.8.** For the construction above the following statements hold:

(i) range($\tilde{\Psi}$) $\subset \tilde{W}$

(ii) the columns of $\tilde{\Psi}$ are linearly independent and $\dim(\text{range}(\tilde{\Psi})) = n_\Pi$

(iii) $\tilde{W} = \text{range}(\tilde{\Psi}) \oplus W_\Delta$.

(iv) if all matrices $\Psi_i$ are constructed via (C.10), then

\[\langle Sw, z \rangle = 0 \quad \forall w \in \text{range}(\tilde{\Psi}), \quad z \in W_\Delta.\]

**Proof.** Part (i). We simply show that range($\tilde{\Psi}$) $\subset \tilde{W}$ for an arbitrary but fixed glob $G \in \mathcal{G}$. From the definition of $\tilde{\Psi}_{i}^{(G)}$ and property (C.12) we derive that for any glob $G' \in \mathcal{G}$ and any $j \in \mathcal{N}_{G'}$,

\[
\tilde{L}_{jG'} R_{jG'} \tilde{\Psi}_{j}^{(G)} = \begin{cases} 
\tilde{L}_{jG'} R_{jG'} \Psi_{j}^{(G)} Y_{G'} = \delta_{GG'} Y_{jG} & \text{if } j \in \mathcal{N}_G, \\
0 & \text{otherwise.}
\end{cases}
\]

From (C.6) and the above we conclude that

\[
\sum_{j \in \mathcal{N}_G} \tilde{L}_{jG'} R_{jG'} \tilde{\Psi}_{j}^{(G)} = \sum_{j \in \mathcal{N}_G} \begin{cases} 
\delta_{GG'} K_{jG} Y_{jG} & \text{if } j \in \mathcal{N}_G, \\
0 & \text{otherwise}
\end{cases}
\]

\[
= \begin{cases} 
\sum_{j \in \mathcal{N}_G} K_{jG} Y_{jG} & \text{if } G' = G, \\
0 & \text{otherwise,}
\end{cases}
\]

but due to (C.13), this expression always evaluates to zero.

Part (ii). Firstly, we define

\[
\Psi_{i}^{(G)} := \text{diag}(\Psi_{i}^{(G)})_{j=1}^{N} : \mathbb{R}^{n_{11G}} \rightarrow W, \quad \Psi_{i}^{(G)} := 0 \in \mathbb{R}^{\dim(W_i) \times 0} \text{ if } i \notin \mathcal{N}_G,
\]

\[
\Psi := [ \cdots | \Psi_{i}^{(G)} | \cdots ]_{G \in \mathcal{G}}
\]

such that we can write

\[
\tilde{\Psi} = \Psi \text{ diag}(Y_G)_{G \in \mathcal{G}}.
\]
From (C.11), we observe that the columns of $\Psi$ are just columns of some matrix $\Psi_1$ extended by zero to the remaining subdomains. Hence, Proposition C.7 implies that $\Psi$ is injective. Since each $Y_G$ is injective, $\Psi$ is injective as well.

Part (iii). Let $\Psi$ be as above. From Proposition C.7 we obtain that $\text{range}(\Psi) \cap W_\Delta = \{0\}$, which implies that $\text{range}(\tilde{\Psi}) \cap W_\Delta = \{0\}$, so the sum $\text{range}(\tilde{\Psi}) + W_\Delta$ is direct and

$$\dim(\text{range}(\tilde{\Psi}) + W_\Delta) = n_\Omega + \dim(W_\Delta).$$

From (C.14) and Proposition 5.17 we obtain that $\dim(\tilde{\Psi}) = \dim(W_\Delta) + n_\Omega$, therefore the sum must equal $\tilde{\Psi}$.

Part (iv). Proposition C.7(iii) implies that for $z_i \in W_\Delta$, $\langle S_i \tilde{\Psi}_i, z_i \rangle = \langle S_i \sum_{G \in G_\Omega} \Psi_1^{(G)} R_{1G}, z_i \rangle = 0$. 

Based on the direct sum $\tilde{\Psi} = \text{range}(\tilde{\Psi}) \oplus W_\Delta$, the operator $\tilde{\Psi}^{-1}$ can be realized as in Section C.2.

**Appendix D. Generalized inverse and Schur complement.**

Throughout this section, $V$ is a finite-dimensional vector space and $A : V \to V^*$ a linear operator.

**DEFINITION D.1 (generalized inverse).** $A^\dagger : V^* \to V$ is a generalized inverse of $A$ if

$$A A^\dagger f = f \quad \forall f \in \text{range}(A).$$

From this definition, one easily derives

$$A^\dagger A x = x + vK \text{ for some } vK \in \ker(A) \quad \forall x \in V,$$

as well as the following statement.

**PROPOSITION D.2.** For linear operators $A$, $C$, $D : V \to V^*$ with with $\ker(A) \subset \ker(C)$ and $\text{range}(D) \subset \text{range}(A)$, the expression $CA^\dagger D$ is invariant under the particular choice of the generalized inverse $A^\dagger$. Moreover, if $D = A$, then $CA^\dagger A = C$, and if $C = A$, then $AA^\dagger D = D$.

For the following, let $V = V_1 \times V_2$ and

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}. 

**LEMMA D.3.** If $A$ is SPD then $\ker(A_{22}) \subset \ker(A_{12})$ and $\text{range}(A_{21}) \subset \text{range}(A_{22})$. In particular, $A_{22} A_{22}^\dagger A_{21} = A_{21}$.

**Proof.** Suppose that there exists an element $v_2 \in \ker(A_{22}) \setminus \{0\}$ with $A_{12} v_2 \neq 0$. Then there exists $v_1 \in V_1$ with $\langle A_{12} v_2, v_1 \rangle < 0$. From the assumption on $A$ we get for any $\beta \in \mathbb{R}^+$,

$$0 < \left\langle A \begin{bmatrix} v_1 \\ \beta v_2 \end{bmatrix}, \begin{bmatrix} v_1 \\ \beta v_2 \end{bmatrix} \right\rangle = \left\langle A_{11} v_1, v_1 \right\rangle + 2\beta \left\langle A_{12} v_2, v_1 \right\rangle,$$

which is a contradiction. From functional analysis we know that $\text{range}(A^\dagger) = \ker(A)^\perp$ where $W^\perp := \{ \psi \in V^* : \langle \psi, w \rangle = 0 \forall w \in W \}$ (for $W \subset V$) is the annihilator (see, e.g., [83, p. 23]). This shows $\text{range}(A_{22})^\perp \subset \text{range}(A_{21})^\perp$ which implies the second assertion. 

**DEFINITION D.4.** Let $A$, $V_1$, $V_2$ be as in (D.2). Then the generalized Schur complement (eliminating the components in $V_2$) is given by

$$S_1 := A_{11} - A_{12} A_{22}^\dagger A_{21},$$

\[\text{If, additionally, } A^\dagger A A^\dagger = A^\dagger \text{, then } A^\dagger \text{ is called reflexive generalized inverse, but we do not need this property.}\]
where \( A_{22}^\dagger \) is a generalized inverse of \( A_{22} \). If the conditions \( \ker(A_{22}) \subset \ker(A_{12}) \) and \( \operatorname{range}(A_{21}) \subset \operatorname{range}(A_{22}) \) hold, then this definition is independent of the particular choice of \( A_{12}^\dagger \).

The generalized Schur complement is closely related to the shorted operator in [4, Theorem 1].

**Lemma D.5.** Let \( A, V_1, V_2 \) be as in (D.2) and assume that \( A \) is SPD. Then the generalized Schur complement \( S_1 \) has the following properties:

\[
\begin{align*}
\langle S_1 v_1, v_1 \rangle & \leq \langle Av, v \rangle \quad \forall v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \in V, \\
\langle S_1 v_1, v_1 \rangle & = \langle Av, v \rangle \quad \forall v = \begin{bmatrix} -A_{22}^\dagger A_{21} v_1 + v_2^K \\ v_1 \end{bmatrix}, \quad v_1 \in V_1, \quad v_2^K \in \ker(A_{22}).
\end{align*}
\]

**Proof.** Minimization of the quadratic functional \( \langle Av, v \rangle \) with respect to \( v_2 \) for fixed \( v_1 \) leads to the first-order condition

\[
A_{22} v_2 + A_{21} v_1 = 0.
\]

By Lemma D.3, \( A_{21} v_1 \in \operatorname{range}(A_{22}) \), and so all solutions of (D.5) have the form

\[
v_2 = -A_{22}^\dagger A_{21} v_1 + v_2^K, \quad \text{with} \quad v_2^K \in \ker(A_{22}).
\]

The Hessian is given by \( A_{22} \) and is by assumption positive semi-definite, so all these solutions are minimizers. We verify (D.4):

\[
\begin{align*}
\langle Av, v \rangle & = \langle A_{11} v_1, v_1 \rangle + \langle -A_{12} A_{22}^\dagger A_{21} v_1, v_1 \rangle + \langle A_{12} v_2^K, v_1 \rangle + \langle A_{21} v_1, -A_{12} A_{22}^\dagger A_{21} v_1 \rangle \\
& \quad + \langle A_{21} v_1, v_2^K \rangle + \langle A_{22}(A_{22}^\dagger A_{21} v_1 + v_2^K), A_{22}^\dagger A_{21} v_1 + v_2^K \rangle = \langle S_1 v_1, v_1 \rangle,
\end{align*}
\]

where we have used Lemma D.3 and Definition D.1. Now (D.3) follows.

**Lemma D.6.** Let \( A: V \to V^* \) be SPD and let \( V = V_1 \oplus V_2 = V_1 \oplus V_2' \) be two (direct) space splittings. Let \( S_1, S_1' \) be the generalized Schur complements corresponding to the first and second splitting, respectively. Then \( S_1 = S_1' \).

**Proof.** For \( v \in V \) let \((v_1, v_2), (v_1', v_2')\) be the components corresponding to the first and second splitting, respectively. From the properties of the direct sum, we see that there exists mappings \( T_1, T_2 \) with \( T_2 \) non-singular such that \( v_2' = T_1 v_1 + T_2 v_2 \) and \( v_2 = T_2^{-1}(v_2' - T_1 v_1) \).

Let \( A_{11}, A_{12}, A_{21}, A_{22} \) be the components of \( A \) corresponding to the first space splitting, such that the components corresponding to the second splitting are given by

\[
\begin{bmatrix}
A_{11} - A_{12} T_2^{-1} T_1 & -T_1^\top T_2^{-1} A_{21} \\
T_2^{-\top} A_{22} + A_{21} T_2^{-1} T_1 & A_{22} T_2^{-1} - T_2^\top T_2^{-1} A_{22} T_2^{-1}
\end{bmatrix}.
\]

Computing the generalized Schur complement eliminating the second component \( (v_2') \) and using Lemma D.3, one can easily verify that \( S_1' = S_1 \). ⊣

**Appendix E. A counterexample:** \( B(A + B + C)^{-1} A (A + B + C)^{-1} B \not\preceq A \).

We set

\[
A = I, \quad B = \begin{bmatrix} 2.5 \cdot 10^{-5} & 0.0275 \\
0.0275 & 838.6 \end{bmatrix}, \quad C = \begin{bmatrix} 7.2 & -29 \\
-29 & 225 \end{bmatrix}.
\]
Clearly, $A$, $B$, and $C$ are SPD as the diagonal entries are strictly positive and 
\[ \det(B) = 0.0134025, \quad \det(C) = 779. \]

However,
\[ \sigma(A - B(A + B + C)^{-1}A(A + B + C)^{-1}B) = \{-9.26834, 1\} \]
\[ \sigma(10A - B(A + B + C)^{-1}A(A + B + C)^{-1}B) = \{-0.248337, 10\}. \]

So, for this particular example,
\[ B(A + B + C)^{-1}A(A + B + C)^{-1}B \preceq A \]
\[ B(A + B + C)^{-1}A(A + B + C)^{-1}B \preceq 10A. \]

However, from Lemma 5.10 we obtain
\[ B(A + B + C)^{-1}A(A + B + C)^{-1}B \leq B(A + B + C)^{-1}(A + C)(A + B + C)^{-1}B \leq B : (A + C) \leq B. \]

So it is really the inequality with $A$ on the right-hand side that fails to hold in general.

REFERENCES


http://www.mcs.anl.gov/petsc/petsc-current/docs/manualpages/PC/PCBDDC.html

[123] ———, PCBBDC: dual-primal preconditioners in PETSc, Talk at the Conference: Celebrating 20 Years of  
Computational Science with PETSc, June 2015.  


[125] ———, Adaptive BDCC deluxe methods for H(curl), in Domain Decomposition Methods in Science and  