



Highly scalable hybrid domain decomposition method for the solution of huge scalar variational inequalities

Zdeněk Dostál¹ · David Horák² · Jakub Kružík² · Tomáš Brzobohatý³ · Oldřich Vlach¹

Received: 18 May 2021 / Accepted: 11 February 2022 / Published online: 18 April 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

The unpreconditioned hybrid domain decomposition method was recently shown to be a competitive solver for linear elliptic PDE problems discretized by structured grids. Here, we plug H-TFETI-DP (hybrid total finite element tearing and interconnecting dual primal) method into the solution of huge boundary elliptic variational inequalities. We decompose the domain into subdomains that are discretized and then interconnected partly by Lagrange multipliers and partly by edge averages. After eliminating the primal variables, we get a quadratic programming problem with a well-conditioned Hessian and bound and equality constraints that is effectively solvable by specialized algorithms. We prove that the procedure enjoys optimal, i.e., asymptotically linear complexity. The analysis uses recently established bounds on the spectrum of the Schur complements of the clusters interconnected by edge/face averages. The results extend the scope of scalability of massively parallel algorithms for the solution of variational inequalities and show the outstanding efficiency of the H-TFETI-DP coarse grid split between the primal and dual variables.

Keywords Domain decomposition · Variational inequality · Scalability · Massively parallel algorithms · Hybrid TFETI-DP

Mathematics Subject Classification (2010) MSC 65K15 · MSC 65Y05 · 90C06

1 Introduction

Variants of the FETI (finite element tearing and interconnecting) methods introduced by Farhat and Roux [23, 24] belong to the most powerful methods for a massively

✉ Zdeněk Dostál
zdenek.dostal@vsb.cz

parallel solution of large discretized elliptic partial differential equations. The basic idea is to decompose the domain into subdomains interconnected by Lagrange multipliers and then eliminate the primal variables to get a small coarse problem and many local problems solvable in parallel. After Farhat, Mandel, and Roux [25] proved that the condition number of the dual stiffness matrix is uniformly bounded on the subspace defined by the kernels of subdomain stiffness matrices, FETI became the first theoretically supported massively parallel scalable solver for elliptic PDE. The subspace is also called the natural coarse grid. Later, Farhat, Lesoinne, and Pierson [26] introduced a modification of FETI called FETI-DP (dual-primal) that enforces some constraints on the primal level and avoids manipulations with singular matrices. More on FETI methods for linear systems can be found, e.g., in Tosseli and Widlund [40] or Pechstein [38].

If we apply FETI to elliptic variational inequalities, such as those describing the equilibrium of a system of elastic bodies in contact, then the duality transforms the inequality constraints into bound constraints. Moreover, the dual problem's solution is guaranteed to be in the subspace defined by the natural coarse grid. These observations led to the development of massively parallel scalable algorithms for solving elliptic boundary variational inequalities [11] (currently tens of thousands of cores for billions of nodal variables [18]).

The scope of scalability of the original FETI methods is limited by the coarse problem's dimension, which is proportional to the number of subdomains. A direct solver typically solves the coarse problem at the cost proportional to the square of its dimension—it starts to dominate when the number of subdomains is large, currently some tens of thousands of subdomains. Klawonn and Rheinbach [30, 31], Klawonn et al. [33] used the idea of FETI-DP to interconnect groups of subdomains into *clusters* by enforcing some constraints on the primal level so that the defect of each cluster is the same as that of each of its subdomains (see also Brzobohatý et al. [4] or Jungho Lee [35, 36]). The latter authors used a variant of FETI called TFETI (total FETI) that enforces the Dirichlet conditions by Lagrange multipliers [16] so that all subdomains are floating and their stiffness matrices have a priori known kernels. The latter properties considerably simplify stable elimination of primal variables [3]. The methods which combine FETI-DP with TFETI are called H-TFETI-DP (hybrid TFETI-DP).

Both original FETI-DP and H-FETI-DP were combined with preconditioners, and their scalability was proved in the context of preconditioned methods. Though preconditioning is a standard tool for the solution of linear problems, it becomes a complication when we try to apply it to variational inequalities. The reason is that the preconditioning turns the bound constraints into more general inequality constraints, destroying the favorable structure of the resulting quadratic programming (QP) problem and excluding the possibility to use specialized QP solvers [10]. However, experimental results by Klawonn and Rheinbach [31] and Jungho Lee [35] indicated that the scalability could be observed even without a special preconditioner, using only the projector to the natural coarse grid. We recently confirmed these observations by establishing H/h -bounds on the spectrum of the Schur complements of the clusters interconnected by edge averages [20] or face averages [21].

Here, we enhance the above results to the development of a scalable massively parallel H-TFETI-DP based algorithm. We introduce two variants of a model problem,

coercive and semi-coercive scalar boundary variational inequalities proposed by Ivan Hlaváček. Then, we describe their discretization and decomposition into subdomains and clusters, use the duality to reduce the solution to bound and equality constrained QP problems, and review relevant algorithms for solving the resulting QP problems. Finally, we prove the algorithms’ numerical scalability and demonstrate their numerical and parallel scalability by experiments. The results extend the scope of scalability of powerful massively parallel algorithms for the solution of variational inequalities [18, Chap. 10] and confirm an outstanding efficiency of the H-TFETI-DP coarse grid that is split between the primal and dual variables.

Throughout the paper, we shall use the following notation. For any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and subsets $\mathcal{I} \subseteq \{1, \dots, m\}$ and $\mathcal{J} \subseteq \{1, \dots, n\}$, we shall denote by $\mathbf{A}_{\mathcal{I}\mathcal{J}}$ a submatrix of \mathbf{A} with the rows $i \in \mathcal{I}$ and columns $j \in \mathcal{J}$. The full set of indices can be replaced by $*$, so that $\mathbf{A} = \mathbf{A}_{**}$ and \mathbf{A}_{i*} denotes the i th row of \mathbf{A} .

If $m = n$ and \mathbf{A} is symmetric, then $\lambda_i(\mathbf{A})$, $\lambda_{\min}(\mathbf{A})$, and $\lambda_{\max}(\mathbf{A})$ denote the eigenvalues of \mathbf{A} ,

$$\lambda_{\max}(\mathbf{A}) = \lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \dots \lambda_n(\mathbf{A}) = \lambda_{\min}(\mathbf{A}).$$

If $\lambda_{\min}(\mathbf{A}) = 0$, then we denote the *regular condition number of \mathbf{A}* by

$$\bar{\kappa}(\mathbf{A}) = \lambda_{\max}(\mathbf{A}) / \bar{\lambda}_{\min}(\mathbf{A}),$$

where $\bar{\lambda}_{\min}(\mathbf{A})$ denotes the least nonzero eigenvalue of \mathbf{A} .

The same matrices and vectors can be indexed in various places alternatively by lower and upper indices in order to simplify the notations, in particular to avoid superfluous brackets. Thus,

$$\mathbf{v}^i = \mathbf{v}_i, \quad \mathbf{K}^i = \mathbf{K}_i, \quad \text{etc.}$$

The Euclidean norm is denoted by $\|\cdot\|$.

2 Model problem

We shall reduce our analysis to two simple model problems, but our reasoning is also valid for more general cases. Let $\Omega = \Omega^1 \cup \Omega^2$, $\Omega^1 = (0, 1) \times (0, 1)$, and $\Omega^2 = (1, 2) \times (0, 1)$ denote open domains with boundaries Γ^1, Γ^2 and their parts $\Gamma_U^i, \Gamma_F^i, \Gamma_C^i = \Gamma_C$ formed by the sides of Ω^i , $i = 1, 2$, as in Fig. 1. Let $H^1(\Omega^i)$, $i = 1, 2$, denote the Sobolev spaces of the first order in the space $L^2(\Omega^i)$ of the functions on Ω^i whose squares are integrable in the sense of Lebesgue. Let

$$V^i = \left\{ v^i \in H^1(\Omega^i) : v^i = 0 \quad \text{on} \quad \Gamma_U^i \right\}$$

denote the closed subspaces of $H^1(\Omega^i)$, $i = 1, 2$. Let $\mathcal{H} = H^1(\Omega^1) \times H^1(\Omega^2)$, and let

$$V = V^1 \times V^2 \quad \text{and} \quad \mathcal{K} = \left\{ (v^1, v^2) \in V : v^2 - v^1 \geq 0 \quad \text{on} \quad \Gamma_C \right\}$$

denote the closed subspace and the closed convex subset of \mathcal{H} , respectively. The relations on the boundaries are in terms of traces. On \mathcal{H} , we shall define a symmetric bilinear form

$$a(u, v) = \sum_{i=1}^2 \int_{\Omega^i} \left(\frac{\partial u^i}{\partial x} \frac{\partial v^i}{\partial x} + \frac{\partial u^i}{\partial y} \frac{\partial v^i}{\partial y} \right) d\Omega$$

and a linear form

$$\ell(v) = \sum_{i=1}^2 \int_{\Omega^i} f^i v^i d\Omega,$$

where $f^i \in L^2(\Omega^i)$, $i = 1, 2$, are the restrictions of

$$f(x, y) = \begin{cases} -1 & \text{for } (x, y) \in (0, 1) \times [0.75, 1) \\ 0 & \text{for } (x, y) \in (0, 1) \times [0, 0.75) \cup (1, 2) \times [0.25, 1) \\ -3 & \text{for } (x, y) \in (1, 2) \times [0, 0.25) \end{cases}$$

Thus, we can define a problem to find

$$\min q(u) = \frac{1}{2}a(u, u) - \ell(u) \text{ subject to } u \in \mathcal{K}. \tag{2.1}$$

We shall consider two variants of the Dirichlet data. In the first case, both membranes are fixed on the outer edges as in Fig. 1 left, so that

$$\Gamma_U^1 = \{(0, y) \in \mathbb{R}^2 : y \in [0, 1]\}, \quad \Gamma_U^2 = \{(2, y) \in \mathbb{R}^2 : y \in [0, 1]\}.$$

Since the Dirichlet conditions are prescribed on the parts of the boundaries of both membranes with positive measure, the quadratic form a is positive definite, which guarantees the existence and the uniqueness of the solution [27]. In the second case, only the left membrane is fixed on the outer edge, and the right membrane has no prescribed displacement as in Fig. 1 right, so that

$$\Gamma_U^1 = \{(0, y) \in \mathbb{R}^2 : y \in [0, 1]\}, \quad \Gamma_U^2 = \emptyset.$$

Even though a is in this case only semidefinite, the cost function q is still coercive due to the choice of f , so the solution exists; it can be proved unique [27].

The model problem’s solution may be interpreted as the displacement of two membranes under the traction f . The left edge of the right membrane cannot penetrate below the right edge of the left membrane.

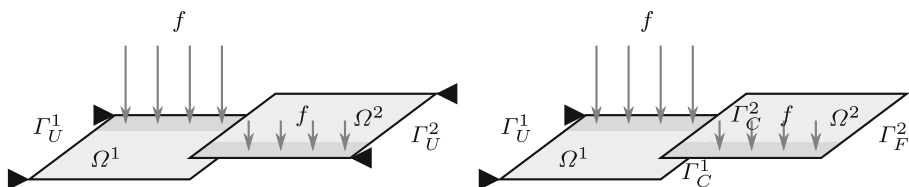


Fig. 1 Coercive (left) and semicoercive (right) model problems

3 Domain decomposition

So far, we have used only the natural decomposition of the spatial domain Ω into Ω^1 and Ω^2 . However, to enable efficient application of domain decomposition methods, we can optionally decompose each Ω^i into $p = 1/H_s \times 1/H_s, i = 1, 2$, square subdomains $\Omega^{i1}, \dots, \Omega^{ip}$ as in Fig. 2. We shall call H_s a *decomposition parameter*.

The continuity of a global solution in Ω^1 and Ω^2 can be enforced by the conditions

$$u^{ij}(x) = u^{ik}(x), \tag{3.1}$$

$$\nabla u^{ij} \cdot n^{ij} = -\nabla u^{ik} \cdot n^{ik}, \tag{3.2}$$

which should be satisfied by relevant traces of u^{ij} and u^{ik} on

$$\Gamma^{ij,ik} = \Gamma^{ij} \cap \Gamma^{ik}.$$

Let us simplify the notation by assigning to each subdomain and its boundary number $k = 1, \dots, s = 2p$,

$$\Omega_k = \Omega^{ij}, \Gamma_k = \Gamma^{ij}, k = k(i, j) = (i - 1)p + j.$$

The boundary between Ω_i and Ω_j is denoted by Γ_{ij} . Let

$$V_D^k = \left\{ v^k \in H^1(\Omega_k) : v^k = 0 \text{ on } \Gamma_U \cap \Gamma_k \right\}, k = 1, \dots, s,$$

denote the closed subspaces of $H^1(\Omega_i)$, and let

$$\begin{aligned} V_D &= V_D^1 \times \dots \times V_D^s, \\ \mathcal{K}_D^C &= \left\{ v \in V_D : v^j - v^i \geq 0 \text{ on } \Gamma_C \cap \Gamma_{ij}, i \leq p < j \right\}, \\ \mathcal{K}_D &= \left\{ v \in \mathcal{K}_D^C : v^i = v^j \text{ on } \Gamma_{ij} \right\}. \end{aligned} \tag{3.3}$$

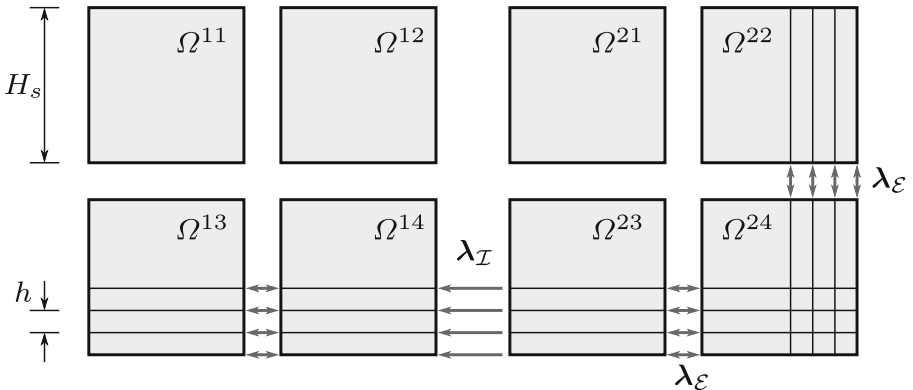


Fig. 2 Domain decomposition and discretization

The relations on the boundaries are again in terms of traces. On V_D , we define the broken scalar product

$$(u, v)_D = \sum_{i=1}^s \int_{\Omega_i} u^i v^i \, d\Omega,$$

the symmetric bilinear form

$$a_D(u, v) = \sum_{i=1}^s \int_{\Omega_i} \left(\frac{\partial u^i}{\partial x_1} \frac{\partial v^i}{\partial x_1} + \frac{\partial u^i}{\partial x_2} \frac{\partial v^i}{\partial x_2} \right) d\Omega,$$

and the linear form

$$\ell_D(v) = (f, v)_D = \sum_{i=1}^s \int_{\Omega_i} f^i v^i \, d\Omega,$$

where $f^i \in L^2(\Omega_i)$ denotes the restriction of f to Ω_i .

Using the above notation, it is a standard exercise (see, e.g., the book [18, Sect. 10.2]) to prove that (2.1) is equivalent to the problem to find $u \in \mathcal{K}_D$ such that

$$q_D(u) \leq q_D(v), \quad q_D(v) = \frac{1}{2} a_D(v, v) - \ell_D(v), \quad v \in \mathcal{K}_D. \quad (3.4)$$

4 Discretization

To reduce (3.4) to a finite-dimensional problem, let us introduce on each subdomain Ω_i a regular grid with the step h as in Fig. 2 so that the grids match across the interfaces Γ_{ij} of the adjacent subdomains, index contiguously the nodes and entries of corresponding vectors in the subdomains, and define piece-wise linear functions

$$\phi_\ell^i, \quad \ell = 1, \dots, n_s,$$

where $n_s = (H_s/h + 1)^2$ denotes a number of nodes in $\overline{\Omega}_i$, $i = 1, \dots, s$. We shall look for an approximate solution u_h in a trial space V_h which is spanned by the basis functions ϕ_ℓ^i (Fig. 3),

$$\begin{aligned} V_h &= V_h^1 \times \dots \times V_h^s, \\ V_h^i &= \text{Span}\{\phi_\ell^i : \ell = 1, \dots, n_s\}. \end{aligned}$$

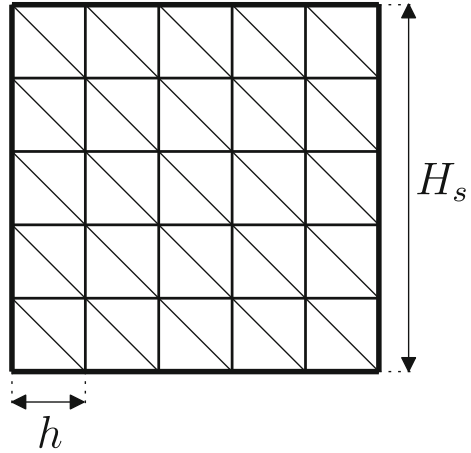
Decomposing u_h into the components which comply with the decomposition, i.e.,

$$\begin{aligned} u_h &= (u_h^1, \dots, u_h^s), \\ u_h^i &= \sum_{\ell=1}^{n_s} u_\ell^i \phi_\ell^i(\mathbf{x}), \end{aligned}$$

we get

$$a(u_h, v_h) = \sum_{i=1}^s a_i(u_h^i, v_h^i), \quad (4.1)$$

Fig. 3 Subdomain Ω_i and its triangularization



$$a^i(u_h^i, v_h^i) = \sum_{\ell=1}^{n_s} \sum_{m=1}^{n_s} a_i(\phi_\ell^i, \phi_m^i) v_\ell^i v_m^i = \mathbf{u}_i^T \mathbf{K}_i \mathbf{v}_i, \tag{4.2}$$

$$[\mathbf{K}_i]_{\ell m} = a_i(\phi_\ell^i, \phi_m^i), \quad [\mathbf{u}_i]_\ell = u_\ell^i, \quad [\mathbf{v}_i]_\ell = v_\ell^i. \tag{4.3}$$

Similarly

$$\begin{aligned} \ell(u_h) &= \sum_{i=1}^s (f^i, u_h^i), \\ (f^i, u_h^i) &= \sum_{\ell=1}^{n_s} (f^i, u_\ell^i \phi_\ell^i) = \mathbf{f}_i^T \mathbf{u}_i, \\ [\mathbf{f}_i]_\ell &= (f^i, \phi_\ell^i). \end{aligned}$$

Using the above notation, we get the discretized version of problem (3.4) with auxiliary domain decomposition

$$\min \frac{1}{2} \mathbf{u}^T \mathbf{K} \mathbf{u} - \mathbf{f}^T \mathbf{u} \quad \text{s.t.} \quad \mathbf{B}_I \mathbf{u} \leq \mathbf{0} \quad \text{and} \quad \mathbf{B}_E \mathbf{u} = \mathbf{0}. \tag{4.4}$$

In (4.4), $\mathbf{K} \in \mathbb{R}^{n \times n}$, $n = sn_s$, denotes a block diagonal symmetric positive semidefinite (SPS) stiffness matrix, the full rank matrices \mathbf{B}_I and \mathbf{B}_E describe the discretized non-penetration and gluing conditions, respectively, and \mathbf{f} represents the discrete analog of the linear form $\ell(u)$. We can write the stiffness matrix and the vectors in the block form

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}^1 & \mathbf{O} & \dots & \mathbf{O} \\ \mathbf{O} & \mathbf{K}^2 & \dots & \mathbf{O} \\ \dots & \dots & \dots & \dots \\ \mathbf{O} & \mathbf{O} & \dots & \mathbf{K}^s \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \dots \\ \mathbf{u}_s \end{bmatrix}, \quad \mathbf{f} = \begin{bmatrix} \mathbf{f}_1 \\ \dots \\ \mathbf{f}_s \end{bmatrix}.$$

The rows of \mathbf{B}_E and \mathbf{B}_I are filled with zeros except 1 and -1 in the positions corresponding to the nodes on the subdomain boundaries. We get three types of equality constraints.

If \mathbf{b}_i denotes a row of \mathbf{B}_I or \mathbf{B}_E , then \mathbf{b}_i does not have more than four nonzero entries. The continuity of the solution in the “wire basket” (displacements u_i, u_j, u_k , and u_l of four corners of adjacent subdomains), on the interface (displacements u_i and u_j of adjacent nodes in the interior of adjacent edges), or on the Dirichlet’s boundary satisfy

$$u_i = u_j, u_k = u_l, u_i + u_j = u_k + u_l; \quad u_i = u_j; \quad u_i = 0;$$

respectively. The identification can be expressed by the vectors

$$\mathbf{b}_{ij} = (\mathbf{s}_i - \mathbf{s}_j), \mathbf{b}_{kl} = (\mathbf{s}_k - \mathbf{s}_l), \mathbf{b}_{ijkl} = (\mathbf{s}_i + \mathbf{s}_j - \mathbf{s}_k - \mathbf{s}_l); \quad \mathbf{b}_i = \mathbf{s}_i;$$

where \mathbf{s}_i denotes the i th column of the identity matrix \mathbf{I}_n . The continuity of the solution across the interior of subdomains interface is implemented by

$$\mathbf{b}_{ij}^T \mathbf{u} = 0,$$

The non-penetration is enforced similarly. If i and j are the indices of matching nodes on Γ_C^1 and Γ_C^2 , respectively, then any feasible nodal displacements satisfy

$$\mathbf{b}_{ij}^T \mathbf{u} \leq 0.$$

If u_i, u_j and u_k, u_l introduced above are the displacement of the adjacent corners on Γ_C^1 and Γ_C^2 , respectively, then we place \mathbf{b}_{ij}^T and \mathbf{b}_{kl}^T into the rows of \mathbf{B}_E and \mathbf{b}_{ijkl}^T into the rows of \mathbf{B}_I .

5 TFETI problem

Our next step is to simplify the problem by using the duality theory; in particular, we replace the general inequality constraints

$$\mathbf{B}_I \mathbf{u} \leq \mathbf{o}$$

by the nonnegativity constraints. To this end, let us define the Lagrangian associated with problem (4.4) by

$$L(\mathbf{u}, \lambda_I, \lambda_E) = \frac{1}{2} \mathbf{u}^T \mathbf{K} \mathbf{u} - \mathbf{f}^T \mathbf{u} + \lambda_I^T \mathbf{B}_I \mathbf{u} + \lambda_E^T \mathbf{B}_E \mathbf{u}, \quad (5.1)$$

where λ_I and λ_E are the Lagrange multipliers associated with the inequalities and equalities, respectively. Introducing the notation

$$\lambda = \begin{bmatrix} \lambda_I \\ \lambda_E \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} \mathbf{B}_I \\ \mathbf{B}_E \end{bmatrix},$$

we can observe that $\mathbf{B} \in \mathbb{R}^{m \times n}$ is a full rank matrix and write the Lagrangian briefly as

$$L(\mathbf{u}, \lambda) = \frac{1}{2} \mathbf{u}^T \mathbf{K} \mathbf{u} - \mathbf{f}^T \mathbf{u} + \lambda^T \mathbf{B} \mathbf{u}.$$

The solution satisfies the KKT conditions, including

$$\mathbf{K} \mathbf{u} - \mathbf{f} + \mathbf{B}^T \lambda = \mathbf{o}. \quad (5.2)$$

Equation (5.2) has a solution if and only if

$$\mathbf{f} - \mathbf{B}^T \boldsymbol{\lambda} \in \text{Im} \mathbf{K}, \tag{5.3}$$

which can be expressed more conveniently by means of a matrix \mathbf{R} the columns of which span the null space of \mathbf{K} as

$$\mathbf{R}^T (\mathbf{f} - \mathbf{B}^T \boldsymbol{\lambda}) = \mathbf{o}. \tag{5.4}$$

The matrix \mathbf{R} can be formed directly so that each floating subdomain is assigned to a column of \mathbf{R} with ones in the positions of the nodal variables that belong to the subdomain and zeros elsewhere. It may be checked that $\mathbf{R}^T \mathbf{B}^T$ is a full rank matrix.

Now assume that $\boldsymbol{\lambda}$ satisfies (5.3), so that we can evaluate $\boldsymbol{\lambda}$ from (5.2) by means of any (left) generalized matrix \mathbf{K}^+ which satisfies

$$\mathbf{K} \mathbf{K}^+ \mathbf{K} = \mathbf{K}. \tag{5.5}$$

It may be verified directly that if \mathbf{u} solves (5.2), then there is a vector α such that

$$\mathbf{u} = \mathbf{K}^+ (\mathbf{f} - \mathbf{B}^T \boldsymbol{\lambda}) + \mathbf{R} \alpha. \tag{5.6}$$

For the effective evaluation of the action of the generalized inverse, notice that \mathbf{K}_i can be written in the form

$$\mathbf{K}_i = \begin{bmatrix} a_i & \mathbf{b}_i^T \\ \mathbf{b}_i & \mathbf{A}_i \end{bmatrix}$$

with symmetric positive definite \mathbf{A}_i , so that we can use

$$\mathbf{K}^+ = \text{diag}(\mathbf{K}_1^+, \dots, \mathbf{K}_s^+), \quad \mathbf{K}_i^+ = \begin{bmatrix} 0 & \mathbf{o}^T \\ \mathbf{o} & \mathbf{A}_i^{-1} \end{bmatrix}.$$

After eliminating the primal variables \mathbf{u} , we can find $\boldsymbol{\lambda}$ by solving the minimization problem

$$\min \theta(\boldsymbol{\lambda}) \quad \text{s.t.} \quad \boldsymbol{\lambda}_{\mathcal{I}} \geq \mathbf{o} \quad \text{and} \quad \mathbf{R}^T (\mathbf{f} - \mathbf{B}^T \boldsymbol{\lambda}) = \mathbf{o}, \tag{5.7}$$

where

$$\theta(\boldsymbol{\lambda}) = \frac{1}{2} \boldsymbol{\lambda}^T \mathbf{F} \boldsymbol{\lambda} - \boldsymbol{\lambda}^T \mathbf{B} \mathbf{K}^+ \mathbf{f}, \quad \mathbf{F} = \mathbf{B} \mathbf{K}^+ \mathbf{B}^T. \tag{5.8}$$

Once the solution $\widehat{\boldsymbol{\lambda}}$ of (5.7) is known, the vector $\widehat{\mathbf{u}}$ which solves (4.4) can be evaluated by (5.6) and

$$\boldsymbol{\alpha} = -(\mathbf{R}^T \widehat{\mathbf{B}}^T \widehat{\mathbf{B}} \mathbf{R})^{-1} \mathbf{R}^T \widehat{\mathbf{B}}^T \widehat{\mathbf{B}} \mathbf{K}^+ (\mathbf{f} - \widehat{\mathbf{B}}^T \widehat{\boldsymbol{\lambda}}), \tag{5.9}$$

where $\widehat{\mathbf{B}} = [\widehat{\mathbf{B}}_I^T, \widehat{\mathbf{B}}_E^T]^T$, and the matrix $\widehat{\mathbf{B}}_I$ is formed by the rows \mathbf{b}_i of \mathbf{B}_I that correspond to the positive components of the solution $\widehat{\boldsymbol{\lambda}}_I$ characterized by $\widehat{\lambda}_i > 0$. A more effective procedure avoiding manipulation with $\widehat{\mathbf{B}}$ can be found in Horák, Dostál, and Sojka [19].

Using the orthogonal projectors on the kernel of $\mathbf{G} = \mathbf{R}^T \mathbf{B}^T$ and its complement, we can modify (5.7) into the form that can be solved very efficiently by a combination of the active set strategy and the augmented Lagrangian method. We shall give the details with the description of H-TFETI-DP in Section 7.

6 Connecting subdomains into clusters

The bottleneck of classical FETI methods is the dual coarse grid dimension, which is equal to the defect of the stiffness matrix \mathbf{K} , in our case, to the number of subdomains. To increase the rank of \mathbf{K} , we shall use the idea of Klawonn and Rheinbach [30] to interconnect some subdomains on the primal level into *clusters* so that the defect of the stiffness matrix of the cluster is equal to the defect of one of the subdomain stiffness matrices.

6.1 Prolog: interconnecting four subdomains in a common corner

For example, to join four adjacent subdomains in the only common node

$$\mathbf{x} \in \overline{\Omega}_i \cap \overline{\Omega}_j \cap \overline{\Omega}_k \cap \overline{\Omega}_\ell,$$

we can transform the nodal variables associated with

$$\tilde{\Omega}_q = \overline{\Omega}_i \times \overline{\Omega}_j \times \overline{\Omega}_k \times \overline{\Omega}_\ell$$

by the expansion matrix

$$\mathbf{L}_q \in \mathbb{R}^{n_c \times \tilde{n}_c}, \quad \mathbf{L}_q^T \mathbf{L}_q = \mathbf{I}, \quad \tilde{n}_c = n_c - 3, \quad n_c = 4n_s.$$

The matrix \mathbf{L}_q can be obtained by replacing appropriate four columns of the identity matrix by their normalized sum. *Feasible* variables \mathbf{u}^q of the cluster are related to *global* variables $\tilde{\mathbf{u}}^q$ by

$$\mathbf{u}^q = \mathbf{L}_q \tilde{\mathbf{u}}^q.$$

The stiffness matrix $\tilde{\mathbf{K}}_q$ of such cluster in global variables is defined by

$$\tilde{\mathbf{K}}_q = \mathbf{L}_q^T \text{diag}(\mathbf{K}_i, \mathbf{K}_j, \mathbf{K}_k, \mathbf{K}_\ell) \mathbf{L}_q.$$

The kernel of $\tilde{\mathbf{K}}_q$ is spanned by a vector $\tilde{\mathbf{e}}^q$ which can be obtained from the unit vector $\mathbf{e}^q \in \text{Im } \mathbf{L}_q$ using

$$\tilde{\mathbf{e}}^q = \mathbf{L}_q^T \mathbf{e}^q.$$

Assuming that the set of all subdomains is decomposed into c clusters comprising four adjacent subdomains with a common vertex, we can use the global expansion matrix with orthonormal columns

$$\mathbf{L} = \text{diag}(\mathbf{L}_1, \dots, \mathbf{L}_c)$$

to get partially assembled global stiffness matrix

$$\tilde{\mathbf{K}} = \mathbf{L}^T \mathbf{K} \mathbf{L} = \text{diag}(\tilde{\mathbf{K}}_1, \dots, \tilde{\mathbf{K}}_c)$$

and the matrices

$$\tilde{\mathbf{B}} = \mathbf{E} \mathbf{B} \mathbf{L}, \quad \tilde{\mathbf{R}} = \text{diag}(\tilde{\mathbf{e}}^1, \dots, \tilde{\mathbf{e}}^c),$$

where \mathbf{E} denotes a matrix obtained from the identity matrix by deleting the rows corresponding to zero rows of $\mathbf{B}_E \mathbf{L}$. However, the procedure made the regular condition number of $\tilde{\mathbf{F}}$ to deteriorate to

$$\bar{\kappa} \left(\tilde{\mathbf{F}} | \text{Ker}(\tilde{\mathbf{R}}^T \tilde{\mathbf{B}}^T) \right) \leq C H_s \left(1 + \ln \frac{H_s}{h} \right) / h$$

with C independent of H and h (see Vodstrčil et al. [41]). In the next two subsections we show how to interconnect the subdomains by edge averages.

6.2 Coupling two subdomains by edge averages

In what follows, we shall denote by \mathbf{e} and $\bar{\mathbf{e}}$ the vectors with all components equal to 1 and $1/\|\mathbf{e}\|$, respectively, where $\|\cdot\|$ denotes the Euclidean norm. To simplify the notations, we shall often avoid specification of the dimension of vectors and matrices when we can deduce it from the assumption that they appear in well-defined expressions or when we introduce a generic object as above. It is thus possible that one symbol can represent in one formula two objects of different dimensions.

To describe the coupling by averages, we shall use the transformation of bases proposed by Klawonn and Widlund [32], see also Klawonn and Rheinbach [29] and Li and Widlund [37]. The basic idea is a rather trivial observation that if $\mathbf{x} \in \mathbb{R}^p$ denotes any vector, then

$$\mathbf{x}^T \mathbf{e} = \sum_{i=1}^p x_i,$$

so if

$$[\mathbf{c}_1, \dots, \mathbf{c}_{p-1}, \bar{\mathbf{e}}], \quad \bar{\mathbf{e}} = \frac{1}{\sqrt{p}} \mathbf{e},$$

denote an orthonormal basis of \mathbb{R}^p , then the last coordinate of a vector $\mathbf{x} \in \mathbb{R}^p$ in this basis is given by $x_p = \bar{\mathbf{e}}^T \mathbf{x}$.

To find the basis and transformation, denote by $\mathbf{T} \in \mathbb{R}^{p \times p}$ an orthogonal matrix that can be obtained by the application of the Gram–Schmidt procedure to the columns of

$$\mathbf{T}_0 = \begin{bmatrix} \mathbf{I} & \mathbf{e} \\ -\mathbf{e}^T & 1 \end{bmatrix} \in \mathbb{R}^{p \times p},$$

starting from the last one. We get

$$\mathbf{T} = [\mathbf{C}, \bar{\mathbf{e}}], \quad \mathbf{C}^T \mathbf{C} = \mathbf{I}, \quad \mathbf{C}^T \bar{\mathbf{e}} = \mathbf{o}, \quad \|\bar{\mathbf{e}}\| = 1, \quad \mathbf{C} \in \mathbb{R}^{p \times (p-1)}, \quad \bar{\mathbf{e}} \in \mathbb{R}^p, \quad (6.1)$$

and if $\mathbf{x} = \mathbf{T}\mathbf{y}$, $\mathbf{y} \in \mathbb{R}^p$, then

$$y_p = \bar{\mathbf{e}}^T \mathbf{x} = \frac{1}{\sqrt{p}} \sum_{i=1}^p x_i, \quad \bar{\mathbf{e}} = \frac{1}{\sqrt{p}} \mathbf{e}.$$

If we apply the transformation to variables associated with the interiors of adjacent edges, we can join them by the extension matrix \mathbf{L} as in Section 6.1.

Let us first show how to join two subdomains Ω^1 and Ω^2 by the averages of variables associated with the interior of adjacent edges. The basis normalized constant vectors associated with the interior of edge Γ^{ij} of Ω^i adjacent to Ω^j will be denoted by $\bar{\mathbf{e}}^{ij}$ as in Fig. 4, $\bar{\mathbf{e}}^{ij} \in \mathbb{R}^{n_e}$.

On the interiors of edges Γ^{12} and Γ^{21} , we shall introduce the transformation matrices $\mathbf{T}^{12}, \mathbf{T}^{21} \in \mathbb{R}^{n_e \times n_e}$,

$$\mathbf{x}^{12} = \mathbf{T}^{12} \mathbf{y}^{12}, \quad \mathbf{x}^{21} = \mathbf{T}^{21} \mathbf{y}^{21}, \quad \mathbf{T}^{ij} = [\mathbf{C}^{ij}, \bar{\mathbf{e}}^{ij}] \in \mathbb{R}^{n_e \times n_e}, \quad (\mathbf{T}^{ij})^T \mathbf{T}^{ij} = \mathbf{I},$$

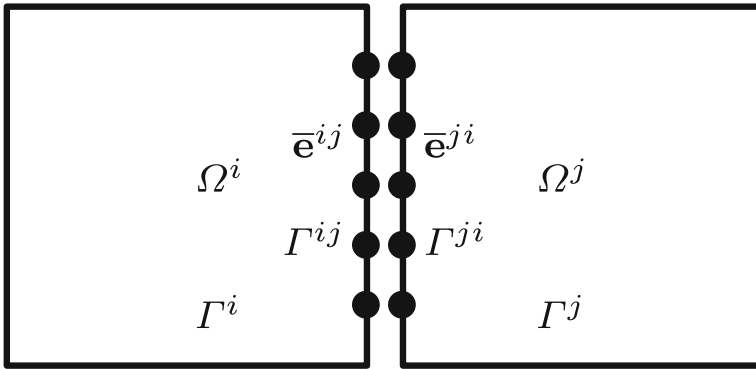


Fig. 4 Joining two subdomains by the edge averages

so that we can define orthogonal transformations $\mathbf{T}^1, \mathbf{T}^2 \in \mathbb{R}^{n_s \times n_s}$ acting on Ω^1 and Ω^2 by

$$\mathbf{T}^1 = \begin{bmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{O} & \mathbf{T}^{12} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{O} & \mathbf{o} \\ \mathbf{O} & \mathbf{C}^{12} & \bar{\mathbf{e}}^{12} \end{bmatrix}, \quad \mathbf{T}^2 = \begin{bmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{O} & \mathbf{T}^{21} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{O} & \mathbf{o} \\ \mathbf{O} & \mathbf{C}^{21} & \bar{\mathbf{e}}^{21} \end{bmatrix}.$$

The identity matrix is associated with the variables that are not affected by coupling. The global transformation $\mathbf{T} \in \mathbb{R}^{n \times n}$, $n = 2n_s$, then reads

$$\mathbf{T} = \begin{bmatrix} \mathbf{T}^1 & \mathbf{O} \\ \mathbf{O} & \mathbf{T}^2 \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{O} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{T}^{12} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \mathbf{I} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \mathbf{O} & \mathbf{T}^{21} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{O} & \mathbf{o} & \mathbf{O} & \mathbf{O} & \mathbf{o} \\ \mathbf{O} & \mathbf{C}^{12} & \bar{\mathbf{e}}^{12} & \mathbf{O} & \mathbf{O} & \mathbf{o} \\ \mathbf{O} & \mathbf{O} & \mathbf{o} & \mathbf{I} & \mathbf{O} & \mathbf{o} \\ \mathbf{O} & \mathbf{O} & \mathbf{o} & \mathbf{O} & \mathbf{C}^{21} & \bar{\mathbf{e}}^{21} \end{bmatrix}.$$

Since we are especially interested in the columns corresponding to averages, it is convenient to move the columns with $\bar{\mathbf{e}}^{ij}$ to the right to get

$$\mathbf{T}_\Pi = \begin{bmatrix} \mathbf{I} & \mathbf{O} & \mathbf{O} & \mathbf{O} & \mathbf{o} & \mathbf{o} \\ \mathbf{O} & \mathbf{C}^{12} & \mathbf{O} & \mathbf{O} & \bar{\mathbf{e}}^{12} & \mathbf{o} \\ \mathbf{O} & \mathbf{O} & \mathbf{I} & \mathbf{O} & \mathbf{o} & \mathbf{o} \\ \mathbf{O} & \mathbf{O} & \mathbf{O} & \mathbf{C}^{21} & \mathbf{o} & \bar{\mathbf{e}}^{21} \end{bmatrix} = \begin{bmatrix} \mathbf{C}^1 & \mathbf{O} & \tilde{\mathbf{e}}^{12} & \mathbf{o} \\ \mathbf{O} & \mathbf{C}^2 & \mathbf{o} & \tilde{\mathbf{e}}^{21} \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

The coupling can then be implemented by the normalized expansion matrix

$$\mathbf{L} = \begin{bmatrix} \mathbf{I} & \mathbf{O} & \mathbf{o} \\ \mathbf{O} & \mathbf{I} & \mathbf{o} \\ \mathbf{o}^T & \mathbf{o}^T & 1/\sqrt{2} \\ \mathbf{o}^T & \mathbf{o}^T & 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{o} \\ \mathbf{o}^T & 1/\sqrt{2} \\ \mathbf{o}^T & 1/\sqrt{2} \end{bmatrix} \in \mathbb{R}^{n \times (n-1)}.$$

Using \mathbf{T}_Π and \mathbf{L} , we can define

$$\mathbf{Z} = \mathbf{T}_\Pi \mathbf{L} = \begin{bmatrix} \mathbf{I} & \mathbf{O} & \mathbf{O} & \mathbf{O} & \mathbf{o} \\ \mathbf{O} & \mathbf{C}^{12} & \mathbf{O} & \mathbf{O} & 1/\sqrt{2} \bar{\mathbf{e}}^{12} \\ \mathbf{O} & \mathbf{O} & \mathbf{I} & \mathbf{O} & \mathbf{o} \\ \mathbf{O} & \mathbf{O} & \mathbf{O} & \mathbf{C}^{21} & 1/\sqrt{2} \bar{\mathbf{e}}^{21} \end{bmatrix} = \begin{bmatrix} \mathbf{C}^1 & \mathbf{O} & 1/\sqrt{2} \tilde{\mathbf{e}}^{12} \\ \mathbf{O} & \mathbf{C}^2 & 1/\sqrt{2} \tilde{\mathbf{e}}^{21} \end{bmatrix} = [\mathbf{C} \mathbf{E}],$$

$\mathbf{Z} \in \mathbb{R}^{n \times (n-1)}$, $\mathbf{C} \in \mathbb{R}^{n \times (n-2)}$, $\mathbf{E} \in \mathbb{R}^n$, the columns of which span the subspace of feasible vectors. Indeed, using (6.1) and the definition of \mathbf{T}_Π , we can check that the interior edge variables \mathbf{x}^{12} and \mathbf{x}^{21} of any vector $\mathbf{x} = \mathbf{Z}\mathbf{y}$,

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}^1 \\ \mathbf{x}^{12} \\ \mathbf{x}^2 \\ \mathbf{x}^{21} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{O} & \mathbf{O} & \mathbf{O} & \mathbf{o} \\ \mathbf{O} & \mathbf{C}^{12} & \mathbf{O} & \mathbf{O} & 1/\sqrt{2}\bar{\mathbf{e}}^{12} \\ \mathbf{O} & \mathbf{O} & \mathbf{I} & \mathbf{O} & \mathbf{o} \\ \mathbf{O} & \mathbf{O} & \mathbf{O} & \mathbf{C}^{21} & 1/\sqrt{2}\bar{\mathbf{e}}^{21} \end{bmatrix} \begin{bmatrix} \mathbf{y}^1 \\ \mathbf{y}^{12} \\ \mathbf{y}^2 \\ \mathbf{y}^{21} \\ \mathbf{y}_{n-1} \end{bmatrix},$$

satisfy (with $\bar{\mathbf{e}} \in \mathbb{R}^{n_e}$, $\bar{\mathbf{e}}^T \bar{\mathbf{e}} = 1$)

$$\begin{aligned} \bar{\mathbf{e}}^T \mathbf{x}^{12} &= \bar{\mathbf{e}}^T \left(\mathbf{C}^{12} \mathbf{y}^{12} + 1/\sqrt{2} \mathbf{y}_{n-1} \bar{\mathbf{e}}^{12} \right) = 1/\sqrt{2} \mathbf{y}_{n-1}, \\ \bar{\mathbf{e}}^T \mathbf{x}^{21} &= \bar{\mathbf{e}}^T \left(\mathbf{C}^{21} \mathbf{y}^{21} + 1/\sqrt{2} \mathbf{y}_{n-1} \bar{\mathbf{e}}^{21} \right) = 1/\sqrt{2} \mathbf{y}_{n-1}. \end{aligned}$$

Recall that $\bar{\mathbf{e}}^{ij} \in \mathbb{R}^{n_e}$ denotes a vector with n_e entries equal $1/\sqrt{n_e}$ associated with the interior of the part of Γ^{ij} and $\mathbf{C}^{ij} \in \mathbb{R}^{n_e \times (n_e-1)}$. Notice that the feasible vectors can be described in a much simpler way using

$$\text{Im}\mathbf{Z} = \{ \mathbf{x} \in \mathbb{R}^n : \mathbf{x}^T \mathbf{Z}_\perp = 0 \} = (\text{Im}\mathbf{Z}_\perp)^\perp, \quad \mathbf{Z}_\perp = \begin{bmatrix} 1/\sqrt{2}\tilde{\mathbf{e}}^{12} \\ -1/\sqrt{2}\tilde{\mathbf{e}}^{21} \end{bmatrix}.$$

The latter observation is a key ingredient of the analysis of bounds on the spectra of H-TFETI-DP clusters [20].

6.3 Connecting square clusters by edge averages

The procedure can be generalized to specify the feasible vectors of any cluster connected by the averages of any set of adjacent edges. Here, we consider the clusters formed by m^2 square subdomains joined by edge averages. Using a proper numbering of variables by subdomains, in each subdomain setting first the variables that are not affected by the interconnecting, then the variables associated with the averages ordered by edges, we get

$$\mathbf{Z} = [\mathbf{C} \ \mathbf{E}], \quad \mathbf{C} = \text{diag}(\mathbf{C}^1, \dots, \mathbf{C}^{s_c}), \quad \mathbf{E} = 1/\sqrt{2} \begin{bmatrix} \cdot & \cdot & \cdot \\ \cdot & \tilde{\mathbf{e}}^{ij} & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \tilde{\mathbf{e}}^{ji} & \cdot \\ \cdot & \cdot & \cdot \end{bmatrix}, \quad (i, j) \in \mathcal{C}, \quad (6.2)$$

where \mathcal{C} denotes a set of ordered couples of the subdomains' indices that define connecting of the interiors of adjacent edges by averages, and s_c here denotes the number of subdomains in the cluster.

For example, let us show how to interconnect four adjacent subdomains Ω^i , Ω^j , Ω^k , and Ω^l of the left membrane of Fig. 2, $i = 11$, $j = 12$, $k = 13$, $l = 14$. The corresponding coupling set is defined by

$$\mathcal{C} = \{(i, j), (k, l), (i, k), (j, l)\}.$$

The cluster is defined on

$$\tilde{\Omega}^q = \bar{\Omega}^i \times \bar{\Omega}^j \times \bar{\Omega}^k \times \bar{\Omega}^\ell.$$

The procedure is very similar to that described in Section 6.1; the only difference is that we shall replace the expansion matrix \mathbf{L}^q by the basis of feasible displacements of the cluster \mathbf{Z}^q . The feasible variables of the cluster are related to global variables $\tilde{\mathbf{u}}^q$ by

$$\mathbf{u}^q = \mathbf{Z}^q \tilde{\mathbf{u}}^q$$

and the stiffness matrix $\tilde{\mathbf{K}}^q$ of such cluster in global variables can be obtained by

$$\tilde{\mathbf{K}}^q = (\mathbf{Z}^q)^T \text{diag}(\mathbf{K}^i, \mathbf{K}^j, \mathbf{K}^k, \mathbf{K}^\ell) \mathbf{Z}^q.$$

The kernel of $\tilde{\mathbf{K}}^q$ is spanned by a vector $\tilde{\mathbf{e}}^q$ which can be obtained from the unit vector $\mathbf{e}^q \in \text{Im } \mathbf{Z}^q$ using

$$\tilde{\mathbf{e}}^q = (\mathbf{Z}^q)^T \mathbf{e}^q.$$

Assuming that the set of all subdomains is decomposed into c clusters interconnected by the edge averages, we can use the matrix

$$\mathbf{Z} = \text{diag}(\mathbf{Z}^1, \dots, \mathbf{Z}^c)$$

with orthonormal columns to connect the groups of $m \times m$ subdomains into clusters to get the stiffness matrix

$$\tilde{\mathbf{K}} = \mathbf{Z}^T \mathbf{K} \mathbf{Z} = \text{diag}(\tilde{\mathbf{K}}^1, \dots, \tilde{\mathbf{K}}^c). \tag{6.3}$$

7 H-TFETI-DP problem

To define H-TFETI-DP problem, we shall proceed as in Section 6.1. The matrix $\tilde{\mathbf{B}}$ can be assembled in the same way as the matrix \mathbf{B} which enforces the constraints on variables \mathbf{x} , so we can achieve that

$$\tilde{\mathbf{B}} \tilde{\mathbf{B}}^T = \mathbf{I}. \tag{7.1}$$

Notice that $\tilde{\mathbf{B}}$ enforces both constraints that connect the subdomains into clusters and those connecting the clusters. Moreover, $\text{Ker } \tilde{\mathbf{B}} = \text{Ker } \mathbf{B} \mathbf{Z}$, but $\mathbf{B} \mathbf{Z}$ need not have orthonormal rows.

The kernel $\tilde{\mathbf{R}}$ is defined by

$$\tilde{\mathbf{R}} = \text{diag}(\tilde{\mathbf{e}}^1, \dots, \tilde{\mathbf{e}}^c), \quad \tilde{\mathbf{e}}^i = \|\mathbf{Z}_i^T \mathbf{e}\|^{-1} \mathbf{Z}_i^T \mathbf{e}, \quad \mathbf{e} = (1, \dots, 1) \in \text{Im } \mathbf{Z}_i.$$

Let us denote

$$\begin{aligned} \tilde{\mathbf{F}} &= \tilde{\mathbf{B}} \tilde{\mathbf{K}} + \tilde{\mathbf{B}}^T, & \tilde{\mathbf{d}} &= \tilde{\mathbf{B}} \tilde{\mathbf{K}} + \mathbf{f}, \\ \tilde{\mathbf{G}} &= \tilde{\mathbf{R}}^T \tilde{\mathbf{B}}^T, & \tilde{\mathbf{e}} &= \tilde{\mathbf{R}}^T \mathbf{f}, \end{aligned}$$

and let \mathbf{N} denote a regular matrix that defines orthonormalization of the rows of $\tilde{\mathbf{G}}$ so that the matrix

$$\tilde{\mathbf{G}} = \mathbf{N} \tilde{\mathbf{G}}$$

has orthonormal rows. After denoting

$$\tilde{\mathbf{e}} = \mathbf{N} \tilde{\mathbf{e}},$$

problem (5.7) reads

$$\min \frac{1}{2} \lambda^T \tilde{\mathbf{F}} \lambda - \lambda^T \tilde{\mathbf{d}} \quad \text{s.t.} \quad \lambda_I \geq \mathbf{0} \quad \text{and} \quad \tilde{\mathbf{G}} \lambda = \tilde{\mathbf{e}}. \tag{7.2}$$

Next, we shall transform the problem of minimization on the subset of the affine space to that on the subset of a vector space by looking for the solution of (7.2) in the form

$$\lambda = \mu + \tilde{\lambda}, \quad \text{where} \quad \tilde{\mathbf{G}} \tilde{\lambda} = \tilde{\mathbf{e}}.$$

Notice that $\tilde{\lambda}$ which satisfies $\tilde{\lambda}_I \geq \mathbf{0}$ exists and can be obtained by

$$\tilde{\lambda} = \arg \min \frac{1}{2} \|\lambda\|^2 \quad \text{s.t.} \quad \lambda_I \geq \mathbf{0} \quad \text{and} \quad \tilde{\mathbf{G}} \lambda = \tilde{\mathbf{e}}.$$

To carry out the transformation, denote $\lambda = \mu + \tilde{\lambda}$, so that

$$\frac{1}{2} \lambda^T \tilde{\mathbf{F}} \lambda - \lambda^T \tilde{\mathbf{d}} = \frac{1}{2} \mu^T \tilde{\mathbf{F}} \mu - \mu^T (\tilde{\mathbf{d}} - \tilde{\mathbf{F}} \tilde{\lambda}) + \frac{1}{2} \tilde{\lambda}^T \tilde{\mathbf{F}} \tilde{\lambda} - \tilde{\lambda}^T \tilde{\mathbf{d}},$$

and problem (7.2) is, after returning to the old notation, equivalent to

$$\min \frac{1}{2} \lambda^T \tilde{\mathbf{F}} \lambda - \lambda^T \tilde{\mathbf{d}} \quad \text{s.t.} \quad \tilde{\mathbf{G}} \lambda = \mathbf{0} \quad \text{and} \quad \lambda_I \geq -\tilde{\lambda}_I \tag{7.3}$$

with $\tilde{\mathbf{d}} = \tilde{\mathbf{d}} - \tilde{\mathbf{F}} \tilde{\lambda}$ and we can achieve that $\tilde{\lambda}_I \geq \mathbf{0}$.

Our final step is based on the observation that problem (7.3) is equivalent to

$$\min \tilde{\theta}_\varrho(\lambda) \quad \text{s.t.} \quad \tilde{\mathbf{G}} \lambda = \mathbf{0} \quad \text{and} \quad \lambda_I \geq -\tilde{\lambda}_I, \tag{7.4}$$

where ϱ is a positive constant and

$$\tilde{\theta}_\varrho(\lambda) = \frac{1}{2} \lambda^T \tilde{\mathbf{H}}_\varrho \lambda - \lambda^T \tilde{\mathbf{P}} \tilde{\mathbf{d}}, \quad \tilde{\mathbf{H}}_\varrho = \tilde{\mathbf{P}} \tilde{\mathbf{F}} \tilde{\mathbf{P}} + \varrho \tilde{\mathbf{Q}}, \tag{7.5}$$

$$\tilde{\mathbf{Q}} = \tilde{\mathbf{G}}^T \tilde{\mathbf{G}}, \quad \tilde{\mathbf{P}} = \mathbf{I} - \tilde{\mathbf{Q}}. \tag{7.6}$$

The matrices $\tilde{\mathbf{P}}$ and $\tilde{\mathbf{Q}}$ are the orthogonal projectors on the kernel of $\tilde{\mathbf{G}}$ and the image space of $\tilde{\mathbf{G}}^T$, respectively. The regularization term is introduced in order to enable the reference to the results on strictly convex QP problems. In what follows, we assume that

$$\varrho \approx \|\tilde{\mathbf{F}}\|. \tag{7.7}$$

Notice that the number of rows of \mathbf{G} is m^2 times larger than that of $\tilde{\mathbf{G}}$.

8 Bounds on the spectrum of $\tilde{\mathbf{P}} \tilde{\mathbf{F}} \tilde{\mathbf{P}}$

Using that $\text{Im} \tilde{\mathbf{P}}$ and $\text{Im} \tilde{\mathbf{Q}}$ are invariant subspaces of $\tilde{\mathbf{H}}_\varrho$, it is easy to check that

$$\sigma(\tilde{\mathbf{H}}_\varrho) = \sigma(\tilde{\mathbf{P}} \tilde{\mathbf{F}} \tilde{\mathbf{P}} | \text{Im} \tilde{\mathbf{P}}) \cup \{\varrho\},$$

and

$$\min\{\bar{\lambda}_{\min}(\tilde{\mathbf{P}} \tilde{\mathbf{F}} \tilde{\mathbf{P}}), \varrho\} \leq \lambda_i(\tilde{\mathbf{H}}_\varrho) \leq \max\{\|\tilde{\mathbf{F}}\|, \varrho\}. \tag{8.1}$$

We shall look for the bounds on $\tilde{\mathbf{P}} \tilde{\mathbf{F}} \tilde{\mathbf{P}}$ in two steps.

8.1 Reducing the problem to subdomain boundaries

Notice that $\tilde{\mathbf{B}}$ has zero columns in the positions corresponding to the variables in the interiors of Ω_i . To enhance this observation, let us decompose the set of indices into two sets corresponding to the subdomains boundary and interior nodes \mathcal{B} and \mathcal{I} , respectively. Then, it is easy to check that

$$[\tilde{\mathbf{K}}^+]_{\mathcal{B}\mathcal{B}} = \tilde{\mathbf{S}}^+, \quad \tilde{\mathbf{S}} = \tilde{\mathbf{K}}_{\mathcal{B}\mathcal{B}} - \tilde{\mathbf{K}}_{\mathcal{B}\mathcal{I}}\tilde{\mathbf{K}}_{\mathcal{I}\mathcal{I}}^{-1}\tilde{\mathbf{K}}_{\mathcal{I}\mathcal{B}}.$$

The matrix $\tilde{\mathbf{S}}$ is called the Schur complement of $\tilde{\mathbf{K}}$ with respect to the block interior variables. A similar formula holds for the clusters, i.e.,

$$[\tilde{\mathbf{K}}_i^+]_{\mathcal{B}_i\mathcal{B}_i} = \tilde{\mathbf{S}}_i^+, \quad \tilde{\mathbf{S}}_i = \tilde{\mathbf{K}}_{\mathcal{B}_i\mathcal{B}_i} - \tilde{\mathbf{K}}_{\mathcal{B}_i\mathcal{I}_i}\tilde{\mathbf{K}}_{\mathcal{I}_i\mathcal{I}_i}^{-1}\tilde{\mathbf{K}}_{\mathcal{I}_i\mathcal{B}_i},$$

where \mathcal{B}_i and \mathcal{I}_i denote the sets of indices of the global variables on the boundaries and in the interiors of the subdomains assembled into the cluster, respectively. Notice that the multiplication by \mathbf{Z} does not change the variables associated with the interior of subdomains as

$$\mathbf{Z} = \begin{bmatrix} \mathbf{I}_{\mathcal{I}\mathcal{I}} & \mathbf{O} \\ \mathbf{O} & \tilde{\mathbf{Z}} \end{bmatrix},$$

where $\tilde{\mathbf{Z}}$ maps the global variables to the feasible ones. It follows that

$$[\tilde{\mathbf{K}}^+]_{\mathcal{B}\mathcal{B}} = \tilde{\mathbf{S}}^+ = (\tilde{\mathbf{Z}}^T \tilde{\mathbf{S}} \tilde{\mathbf{Z}})^+, \quad \mathbf{S} = \text{diag}(\mathbf{S}_1, \dots, \mathbf{S}_s). \tag{8.2}$$

Thus, we can define the H-TFETI-DP dual problem by the matrices

$$\tilde{\mathbf{B}}_{*\mathcal{B}} \quad \text{and} \quad \tilde{\mathbf{R}}_{\mathcal{B}*},$$

obtained by deleting the entries corresponding to the interior variables to get

$$\tilde{\mathbf{F}} = \tilde{\mathbf{B}}\tilde{\mathbf{K}} + \tilde{\mathbf{B}}^T = \tilde{\mathbf{B}}_{*\mathcal{B}}\tilde{\mathbf{S}}^+ + \tilde{\mathbf{B}}_{*\mathcal{B}}^T. \tag{8.3}$$

The following lemma shows that $\tilde{\mathbf{R}}_{\mathcal{B}*}$ is the the kernel of \mathbf{S} .

Lemma 1 *Let $\text{Im}\tilde{\mathbf{R}}$ be the kernel of $\tilde{\mathbf{K}}$. Then,*

$$\text{Im}\tilde{\mathbf{R}}_{\mathcal{B}*} = \text{Ker}\mathbf{S}. \tag{8.4}$$

Proof By the assumption

$$\mathbf{K}_{\mathcal{I}\mathcal{I}}\mathbf{R}_{\mathcal{I}*} + \mathbf{K}_{\mathcal{I}\mathcal{B}}\mathbf{R}_{\mathcal{B}*} = \mathbf{O}, \quad \text{i.e.,} \quad \mathbf{K}_{\mathcal{I}\mathcal{B}}\mathbf{R}_{\mathcal{B}*} = -\mathbf{K}_{\mathcal{I}\mathcal{I}}\mathbf{R}_{\mathcal{I}*}.$$

It follows that

$$\mathbf{S}\mathbf{R}_{\mathcal{B}*} = \left(\mathbf{K}_{\mathcal{B}\mathcal{B}} - \mathbf{K}_{\mathcal{B}\mathcal{I}}\mathbf{K}_{\mathcal{I}\mathcal{I}}^{-1}\mathbf{K}_{\mathcal{I}\mathcal{B}}\right)\mathbf{R}_{\mathcal{B}*} = \mathbf{K}_{\mathcal{B}\mathcal{B}}\mathbf{R}_{\mathcal{B}*} + \mathbf{K}_{\mathcal{B}\mathcal{I}}\mathbf{R}_{\mathcal{I}*} = \mathbf{O}. \quad \square$$

In what follows, we consider only the objects defined on the boundaries of subdomains and simplify the notation by omitting the specification of related index sets, e.g., we use $\tilde{\mathbf{R}}$ and $\tilde{\mathbf{B}}$ to denote $\tilde{\mathbf{R}}_{\mathcal{B}*}$ and $\tilde{\mathbf{R}}_{\mathcal{B}*}$.

8.2 The spectrum of Schur complements of subdomains and clusters

The bounds on the spectrum of $\widetilde{\mathbf{PFP}}$ can be estimated by those of the Schur complements.

Lemma 2 *Let there be constants $0 < c < C$ such that for each $\lambda \in \mathbb{R}^m$*

$$c\|\lambda\|^2 \leq \|\widetilde{\mathbf{B}}^T \lambda\|^2 \leq C\|\lambda\|^2. \tag{8.5}$$

Then, there are constants $0 < c_1 < C_1$ that are independent of h and H_s such that

$$c_1 \left(\max_i \{\|\widetilde{\mathbf{S}}_i\|\} \right)^{-1} \leq \bar{\lambda}_{\min}(\widetilde{\mathbf{PFP}}) \leq \|\widetilde{\mathbf{PFP}}\| \leq C_1 \left(\min_i \{\bar{\lambda}_{\min}(\widetilde{\mathbf{S}}_i)\} \right)^{-1}. \tag{8.6}$$

Proof The proof of this lemma is rather trivial; it uses only the observations that if $\lambda \in \text{Im}\widetilde{\mathbf{P}}$, then $\widetilde{\mathbf{G}}\lambda = \widetilde{\mathbf{R}}^T \widetilde{\mathbf{B}}^T \lambda = \mathbf{o}$, i.e., $\widetilde{\mathbf{B}}^T \lambda$ is orthogonal to the kernel of $\widetilde{\mathbf{S}}$, so that $\widetilde{\mathbf{B}}^T \lambda \in \text{Im}\widetilde{\mathbf{S}}$, and that the nonzero eigenvalues of $\widetilde{\mathbf{S}}$ are reciprocal to the corresponding eigenvalues of $\widetilde{\mathbf{S}}^+$. \square

Lemma 2 reduced the problem to find bounds on the spectrum of $\widetilde{\mathbf{H}}$ to the problem to find bounds on the spectrum of $\widetilde{\mathbf{S}}_i$. Some bounds were proved recently (see [20]):

Theorem 1 *For each integer $m > 1$, let $\widetilde{\mathbf{S}}$ denote the Schur complement of the cluster with the side-length H_c comprising $m \times m$ square subdomains of the side-length $H_s = H_c/m$ discretized by the regular grid with the step-length h and interconnected by the edge averages. Let $\bar{\lambda}_{\min}(\mathbf{S})$ denote the smallest nonzero eigenvalue of*

$$\mathbf{S} = \text{diag}(\mathbf{S}_1, \dots, \mathbf{S}_s),$$

where \mathbf{S}_i denote the Schur complements of the subdomain stiffness matrices \mathbf{K}_i , $i = 1, \dots, s = m^2$, with respect to the interior variables. Then,

$$\|\mathbf{S}\| = \lambda_{\max}(\mathbf{S}) \geq \lambda_{\max}(\widetilde{\mathbf{S}}), \tag{8.7}$$

$$\bar{\lambda}_{\min}(\mathbf{S}) \geq \bar{\lambda}_{\min}(\widetilde{\mathbf{S}}) \geq \frac{2n_e}{n_s} \bar{\lambda}_{\min}(\mathbf{S}_i) \sin^2\left(\frac{\pi}{2m}\right) \approx \frac{1}{2} \bar{\lambda}_{\min}(\mathbf{S}_i) \left(\frac{\pi}{2m}\right)^2. \tag{8.8}$$

8.3 $H_s - h$ bounds on \mathbf{S}_i

The following lemma gives bounds on the Schur complements \mathbf{S}_i in terms of h and H_s ,

Lemma 3 *Let \mathbf{S}_i denote the Schur complement of subdomain Ω_i with the sidelength H_s that is discretized with the steplength h . Then, there are constants $0 < c < C$, independent of h and H_s , such that*

$$c \frac{h}{H_s} \leq \bar{\lambda}_{\min}(\mathbf{S}_i) \leq \|\mathbf{S}_i\| \leq C, \quad i = 1, \dots, s, \tag{8.9}$$

and there are constants $0 < c_1 < C_1$, independent of h and H_s , such that

$$c_1 \frac{h}{H_s m^2} \leq \min_i \{\bar{\lambda}_{\min}(\widetilde{\mathbf{S}}_i)\} \leq \max_i \{\|\widetilde{\mathbf{S}}_i\|\} \leq C_1. \tag{8.10}$$

Proof See, e.g., Brenner [2] or Pechstain [38, Lemma 1.59]). \square

Now, we can formulate the main result, which is at the core of the proof of optimality of the presented algorithms.

Proposition 1 *Let Ω be decomposed into s square subdomains Ω_i with the sidelength H_s and discretized with the parameter h as in Sections 3 and 4. Let the subdomains be interconnected by the edge averages into $c = s/m^2$ square clusters with the sidelength $H_c = mH_s$, each cluster comprising $m \times m$ subdomains. Let $\varrho > 0$ and let there be constants $0 < c < C$ such that for each $\lambda \in \mathbb{R}^m$*

$$c\|\lambda\|^2 \leq \|\tilde{\mathbf{B}}^T \lambda\|^2 \leq C\|\lambda\|^2. \quad (8.11)$$

Then, there are constants $0 < c_1 < C_1$ such that

$$c_1 \leq \bar{\lambda}_{\min}(\tilde{\mathbf{P}}\tilde{\mathbf{F}}\tilde{\mathbf{P}}) \leq \|\tilde{\mathbf{P}}\tilde{\mathbf{F}}\tilde{\mathbf{P}}\| \leq C_1 \frac{H_c m}{h}. \quad (8.12)$$

Proof Use Lemma 2, and Lemma 3, in particular the inequalities (8.10). \square

9 Optimal solvers to bound and equality constrained problems

Here, we shall present two algorithms that can be combined to solve approximately a class of problems (7.4) in a uniformly bounded number of matrix-vector multiplications. We shall formulate the optimality results in the next section. For simplicity, we formulate the algorithms without the stopping criteria, which are presented separately.

9.1 SMALBE-M

The first algorithm is the semi-monotonic augmented Lagrangian method called SMALBE-M [18, Chapter 9]. It generates the approximations for the Lagrange multipliers for equality constraints in the outer loop using the active set based algorithm for bound constrained auxiliary minimization problems in the inner loop.

SMALBE-M is a variant of the algorithm proposed by Conn, Gould, and Toint [5] for identifying stationary points of more general problems. Its early modification called SMALBE (semi-monotonic augmented Lagrangians for bound and equality constrained problems) by Dostál, Friedlander and Santos [15] and Dostál, Friedlander and Santos [14] was later shown to be in a sense optimal for the solution of a class of problems with uniformly bounded spectrum [8]. Dostál and Horák used SMALBE-M to develop scalable FETI based algorithms for variational inequalities [11].

A unique feature of SMALBE-M is the adaptive precision control of auxiliary problems that guarantees the increase of Lagrangian that is sufficient for the optimality results. The algorithm was implemented in PERMON [39] and ESPRESO [22] software. Recent improvement with adaptive reorthogonalization is described in [13].

If we introduce a new Lagrange multiplier vector μ to enforce the equality constraints, the corresponding augmented Lagrangian for problem (7.4) can be written as

$$L(\lambda, \mu, \rho) = \frac{1}{2} \lambda^T \tilde{\mathbf{H}}_\rho \lambda - \lambda^T \tilde{\mathbf{P}} \mathbf{d} + \mu^T \tilde{\mathbf{G}} \lambda,$$

where $\tilde{\mathbf{H}}_\rho$, $\tilde{\mathbf{P}}$, and $\tilde{\mathbf{G}}$ are defined by (7.5) and (7.6). The gradient of $L(\lambda, \mu, \rho)$ is given by

$$\tilde{\mathbf{g}}(\lambda, \mu, \rho) = \tilde{\mathbf{H}}_\rho \lambda - \tilde{\mathbf{P}} \mathbf{d} + \tilde{\mathbf{G}}^T \mu.$$

Let \mathcal{I} denote the set of the indices of the inequality constrained entries of λ , $\lambda_{\mathcal{I}} \geq -\tilde{\lambda}_{\mathcal{I}}$. The *projected gradient*

$$\tilde{\mathbf{g}}^P = \tilde{\mathbf{g}}^P(\lambda, \mu, \rho)$$

of L at λ is given componentwise by

$$\tilde{g}_i^P = \begin{cases} \tilde{g}_i & \text{for } \lambda_i > -\tilde{\lambda}_i \text{ or } i \notin \mathcal{I}, \\ \tilde{g}_i^- & \text{for } \lambda_i = -\tilde{\lambda}_i \text{ and } i \in \mathcal{I}, \end{cases}$$

where $\tilde{g}_i^- = \min\{\tilde{g}_i, 0\}$. It can be verified directly that $(\bar{\lambda}, \bar{\mu}, \rho)$ solves problem (7.4) if and only if

$$\tilde{\mathbf{g}}^P(\bar{\lambda}, \bar{\mu}, \rho) = \mathbf{0}.$$

The above condition is a quantitative refinement of the Karush–Kuhn–Tucker conditions [10, Section 6.2.1].

The algorithm that implements the outer loop reads as follows.

Algorithm 1 SMALBE-M (semi-monotonic augmented Lagrangians for bound and equality constraints).

Step 0. {Initialization.}

Choose $\eta, \rho, M_0 > 0, 0 < \beta < 1, \lambda^0 \in \mathbb{R}^m$

for $k = 0, 1, 2, \dots$

Step 1. {Inner iteration with adaptive precision control.}

Find λ^k such that $\lambda_{\mathcal{I}}^k \geq -\tilde{\lambda}_{\mathcal{I}}$ and

$$\|\mathbf{g}^P(\lambda^k, \mu^k, \rho)\| \leq \min\{M_k \|\tilde{\mathbf{G}} \mathbf{x}^k\|, \eta\} \tag{9.1}$$

Step 2. {Updating the Lagrange multipliers.}

$$\mu^{k+1} = \mu^k + \tilde{\mathbf{G}} \mu^k \tag{9.2}$$

Step 3. {Update M if the increase of the Lagrangian is not sufficient.}

if $k > \bar{k}$ and $L(\lambda^k, \mu^k, 0) < L(\lambda^{k-1}, \mu^{k-1}, \rho)$

$$M_{k+1} = \beta M_k$$

else

$$M_{k+1} = M_k$$

end for

Step 1 can be implemented by any algorithm for minimization of the augmented Lagrangian L with respect to λ subject to $\lambda_{\mathcal{I}} \geq -\tilde{\lambda}_{\mathcal{I}}$ which guarantees convergence of

the projected gradient to zero. To get a bound on the number of matrix multiplication that are necessary to get $\bar{\lambda} = \lambda^k$ which satisfies

$$\|\mathbf{g}^P(\lambda^k, \mu^k, \varrho)\| \leq \varepsilon \quad \text{and} \quad \|\tilde{\mathbf{G}}\lambda^k\| \leq \varepsilon, \tag{9.3}$$

it is necessary to carry out Step 1 in a uniformly bounded number of matrix–vector multiplications, i.e., to solve the problem

$$\min L(\lambda, \mu, \rho) \quad \text{subject to} \quad \lambda_{\mathcal{I}} \geq -\tilde{\lambda}_{\mathcal{I}} \tag{9.4}$$

with the rate of convergence in terms of bounds on the spectrum of the Hessian matrix of L .

9.2 MPRGP

Step 1 of SMALBE-M requires an approximate solution of the convex bound constrained QP problem. Here, we implement Step 1 of SMALBE-M by MPRGP (modified proportioning with reduced gradient projections) (see Dostál and Schöberl [12] and Dostál [7], [10, Chap. 5], and [18, Chap. 8]). To describe it, let us recall that the unique solution $\bar{\lambda} = \bar{\lambda}(\mu, \rho)$ of (9.4) satisfies the Karush-Kuhn-Tucker conditions

$$\mathbf{g}^P(\bar{\lambda}, \mu, \rho) = \mathbf{o}. \tag{9.5}$$

Let $\mathcal{A}(\lambda)$ and $\mathcal{F}(\lambda)$ denote the *active set* and *free set* of the indices of λ , respectively, i.e.,

$$\mathcal{A}(\lambda) = \{i \in \mathcal{I} : \lambda_i = -\tilde{\lambda}_i\} \quad \text{and} \quad \mathcal{F}(\lambda) = \{i : \lambda_i > -\tilde{\lambda}_i \text{ or } i \notin \mathcal{I}\}.$$

To enable an alternative reference to the KKT conditions, let us define the *free gradient* $\varphi(\lambda)$ and the *chopped gradient* $\beta(\lambda)$ by

$$\varphi_i(\lambda) = \begin{cases} g_i(\lambda) & \text{for } i \in \mathcal{F}(\lambda) \\ 0 & \text{for } i \in \mathcal{A}(\lambda) \end{cases} \quad \text{and} \quad \beta_i(\lambda) = \begin{cases} 0 & \text{for } i \in \mathcal{F}(\lambda) \\ g_i^-(\lambda) & \text{for } i \in \mathcal{A}(\lambda) \end{cases}$$

so that the KKT conditions are satisfied if and only if the *projected gradient* $\mathbf{g}^P(\lambda) = \varphi(\lambda) + \beta(\lambda)$ is equal to zero. We call λ *feasible* if $\lambda_i \geq -\tilde{\lambda}_i$ for $i \in \mathcal{I}$. The projector Π to the set of feasible vectors is defined for any λ by

$$\Pi(\lambda)_i = \max\{\lambda_i, -\tilde{\lambda}_i\} \quad \text{for } i \in \mathcal{I}, \quad \Pi(\lambda)_i = \lambda_i \quad \text{for } i \notin \mathcal{I}.$$

Recall that $\tilde{\mathbf{H}}_\varrho$ is the Hessian of L with respect to λ . The *expansion step* is defined by

$$\lambda^{k+1} = \Pi(\lambda^k - \bar{\alpha}\varphi(\lambda^k)) \tag{9.6}$$

with the steplength $\bar{\alpha} \in (0, 2\|\tilde{\mathbf{H}}_\varrho\|^{-1})$ (see [9], recommended $\bar{\alpha} = 1.90\|\tilde{\mathbf{H}}_\varrho\|^{-1}$). This step can expand the current active set. To describe it without Π , let $\tilde{\varphi}(\lambda)$ be the *reduced free gradient* for any feasible λ , with entries

$$\tilde{\varphi}_i = \tilde{\varphi}_i(\lambda) = \min\{\lambda_i/\bar{\alpha}, \varphi_i\} \quad \text{for } i \in \mathcal{I}, \quad \tilde{\varphi}_i = \varphi_i \quad \text{for } i \in \mathcal{F}(\lambda)$$

such that

$$\Pi(\lambda - \bar{\alpha}\varphi(\lambda)) = \lambda - \bar{\alpha}\tilde{\varphi}(\lambda). \tag{9.7}$$

If the inequality

$$\|\beta(\lambda^k)\|^2 \leq \Gamma^2 \tilde{\varphi}(\lambda^k)^T \varphi(\lambda^k) \tag{9.8}$$

holds, then we call the iterate λ^k *strictly proportional*. The test (9.8) is used to decide which component of the projected gradient $g^P(\lambda^k)$ will be reduced in the next step.

The *proportioning step* is defined by

$$\lambda^{k+1} = \lambda^k - \alpha_{cg} \beta(\lambda^k).$$

The steplength α_{cg} is chosen to minimize $L(\lambda^k - \alpha \beta(\lambda^k), \mu^k, \rho_k)$ with respect to α , i.e.,

$$\alpha_{cg} = \frac{\beta(\lambda^k)^T \tilde{\mathbf{g}}(\lambda^k)}{\beta(\lambda^k)^T \tilde{\mathbf{H}}_\rho \beta(\lambda^k)}.$$

The purpose of the proportioning step is to remove indexes from the active set.

The *conjugate gradient step* is defined by

$$\lambda^{k+1} = \lambda^k - \alpha_{cg} \mathbf{p}^k, \tag{9.9}$$

where \mathbf{p}^k is the conjugate gradient direction [1] which is defined recurrently. The recurrence starts (or restarts) with $\mathbf{p}^k = \varphi(\lambda^k)$ whenever λ^k is generated by the expansion step or the proportioning step. If \mathbf{p}^k is known, then \mathbf{p}^{k+1} is given by the formulae [1]

$$\mathbf{p}^{k+1} = \varphi(\lambda^k) - \gamma \mathbf{p}^k, \quad \gamma = \frac{\varphi(\lambda^k)^T \tilde{\mathbf{H}}_\rho \mathbf{p}^k}{(\mathbf{p}^k)^T \tilde{\mathbf{H}}_\rho \mathbf{p}^k}. \tag{9.10}$$

The conjugate gradient steps are used to carry out the minimization in the face $\mathcal{W}_\mathcal{J} = \{\lambda : \lambda_i = 0 \text{ for } i \in \mathcal{J}\}$ given by $\mathcal{J} = \mathcal{A}(\lambda^s)$ efficiently. The algorithm that we use may now be described as follows.

Algorithm 2 MPRGP (Modified proportioning with reduced gradient projections).

Let λ^0 be an n -vector such that $\lambda_i \geq -\tilde{\lambda}_i$ for $i \in \mathcal{I}$, $\bar{\alpha} \in (0, \|\tilde{\mathbf{H}}_\rho\|^{-1}]$, and $\Gamma > 0$ be given. For $k \geq 0$ and λ^k known, choose λ^{k+1} by the following rules:

- Step 1. {Solution found}
 - if $\mathbf{g}^P(\lambda^k) = \mathbf{o}$ then $\lambda^{k+1} = \lambda^k$.
- Step 2. {Explore current face or expand active set}
 - if λ^k is strictly proportional and $\mathbf{g}^P(\lambda^k) \neq \mathbf{o}$
 - try to generate λ^{k+1} by the conjugate gradient step.
 - if $\lambda_i^{k+1} \geq 0$ for $i \in \mathcal{I}$, then accept it
 - else generate λ^{k+1} by the expansion step.
- Step 3. {Reduce active set}
 - if λ^k is not strictly proportional
 - define λ^{k+1} by proportioning.

The MPRGP algorithm has a linear rate of convergence in terms of the bounds on the spectrum of the Hessian $\tilde{\mathbf{H}}_\rho$ of L [12]. The norm of projected gradient converges to zero with qualitatively the same rate of convergence. More about the properties and implementation of SMALBE and MPRGP algorithms may be found in the books [10, Chap. 5, 6] and [18, Chaps. 8,9].

10 Optimality of H-TFETI-DP

To plug these observations into optimality analysis, notice that the discretized problem (7.4) is fully specified by the regularization parameter $\varrho > 0$ and parameters H_c , m , and h ,

$$0 < H_c = mH_s \leq 1/2, \quad m \geq 2, \quad h \leq 1/8, \quad 1/H_c, 1/H_s, H_s/h \in \mathbb{N}.$$

More formally, let us denote by \mathcal{D} the set of all triples $d = (H_s, m, h)$ that define some discretized problem, so the smallest discretized H-TFETI-DP problem is characterized by the triple $d = (1/2, 2, 1/8)$. For any $D \geq 2$, let us define

$$\mathcal{D}_D = \left\{ d \in \mathcal{D} : \frac{H_c m}{h} \leq D \right\}.$$

Theorem 2 *Let $D \geq 2$ and let each $(H_s, m, h) \in \mathcal{D}_D$ specifies a problem (7.4) with $\tilde{\mathbf{B}}$ which satisfies (7.1) and $\varrho \approx \|\tilde{\mathbf{F}}\|$.*

Then, there are constants $c, C > 0$ independent of (H_c, m, h) such that

$$c \leq \lambda_{\min}(\tilde{\mathbf{H}}_\varrho) \leq \|\tilde{\mathbf{H}}_\varrho\| \leq C \frac{mH_c}{h} \leq CD. \quad (10.1)$$

Proof Combine the assumptions $\varrho \approx \|\tilde{\mathbf{F}}\|$ and $(H_s, m, h) \in \mathcal{D}_D$ with Proposition 1. \square

To show that Algorithm 1 with the inner loop implemented by Algorithm 2 is optimal for the solution of the class of problems (7.4), let us consider a class of problems defined by $d \in \mathcal{D}_D$ and ϱ_d , $D \geq 2$ and $\varrho_d \approx \|\tilde{\mathbf{F}}\|$. For any $d \in \mathcal{D}$, we shall define

$$\begin{aligned} \mathbf{A}_d &= \tilde{\mathbf{H}}_{\varrho_d}, & \mathbf{b}_d &= \tilde{\mathbf{P}}\mathbf{d} \\ \mathbf{C}_d &= \tilde{\mathbf{G}}, & \ell_{d,\mathcal{I}} &= -\tilde{\lambda}_{\mathcal{I}} \text{ and } \ell_{d,\mathcal{E}} = -\infty \end{aligned}$$

by the vectors and matrices generated with the parameters H_c , h , and m , so that the problem (7.4) is equivalent to the problem

$$\text{minimize } \Theta_d(\boldsymbol{\lambda}_d) \text{ s.t. } \mathbf{C}_d \boldsymbol{\lambda}_d = \mathbf{o} \text{ and } \boldsymbol{\lambda}_d \geq \ell_d \quad (10.2)$$

with

$$\Theta_d(\boldsymbol{\lambda}) = \frac{1}{2} \boldsymbol{\lambda}^T \mathbf{A}_d \boldsymbol{\lambda} - \mathbf{b}_d^T \boldsymbol{\lambda}.$$

Using these definitions, $\tilde{\mathbf{G}}\tilde{\mathbf{G}}^T = \mathbf{I}$, and assuming $\tilde{\boldsymbol{\lambda}} \geq \mathbf{o}$, we obtain

$$\|\mathbf{C}_d\| = 1 \quad \text{and} \quad \|\ell_d^+\| = 0. \quad (10.3)$$

Moreover, using (8.1) and Theorem 2, we get that there are positive constants a_{\min} and a_{\max} such that

$$a_{\min} \leq \lambda_{\min}(\mathbf{A}_d) \leq \lambda_{\max}(\mathbf{A}_d) \leq a_{\max} \quad (10.4)$$

for any $d \in \mathcal{D}_D$. It follows that the assumptions of [18, Theorem 9.4](i.e., the inequalities (10.3) and (10.4)) are satisfied for any discretization specified by parameters $d \in \mathcal{D}_D$, $D \geq 2$ and we have the following result:

Theorem 3 *Let $D \geq 2$ denote a given constant, $d \in \mathcal{D}_D$, and let $\{\lambda_d^k\}$, $\{\mu_d^k\}$, and $\{M_d^k\}$ be generated by Algorithm 1 (SMALBE-M) for the solution of problem (10.2) arising from the discretization and decomposition of problem (2.1) with parameters H_c , h , and m with*

$$\|\mathbf{b}_d\| \geq \eta_d > 0, \quad 1 > \beta > 0, \quad M_0^l = M_0 > 0, \quad \varrho_d \approx \|\mathbf{A}_d\|, \quad \text{and} \quad \mu_d^0 = \mathbf{o}.$$

Let Step 1 of algorithm 9.1 (SMALBE-M) be implemented by Algorithm 2 (MPRGP) with the parameters

$$\Gamma > 0 [\Gamma \approx 1] \quad \text{and} \quad \bar{\alpha} \in (0, 2a_{\max}^{-1})$$

to generate iterates $\lambda_d^{k,0}, \lambda_d^{k,1}, \dots, \lambda_d^{k,l} = \lambda_d^k$ for the solution of (9.1) starting from $\lambda_d^{k,0} = \lambda_d^{k-1}$ with $\lambda_d^{-1} = \mathbf{o}$, where $l = l_{d,k}$ is the first index satisfying

$$\|\mathbf{g}^P(\lambda_d^{k,l}, \mu_d^k, \rho_d)\| \leq M_k^d \|\mathbf{C}_d \lambda_d^{k,l}\| \tag{10.5}$$

or

$$\|\mathbf{g}^P(\lambda_d^{k,l}, \mu_d^k, \rho_d)\| \leq \varepsilon \|\mathbf{b}_d\| \min\{1, M_k^{-1}\}. \tag{10.6}$$

Then, Algorithm 1 generates an approximate solution λ^k which satisfies

$$\|\mathbf{g}^P(\lambda_d^k, \mu_d^k, \rho_d)\| \leq \varepsilon \|\mathbf{b}_d\| \quad \text{and} \quad \|\mathbf{C}_d \lambda_d^k\| \leq \varepsilon \|\mathbf{b}_d\| \tag{10.7}$$

at $O(1)$ matrix-vector multiplications by the Hessian of the augmented Lagrangian L_d for (10.2).

11 Numerical experiments

We implemented H-TFETI-DP into the PERMON package [39] developed at the Department of Applied Mathematics of the Technical University of Ostrava and the Institute of Geonics AS CR Ostrava. We carried out the computations on the Salomon cluster, which consists of 1008 compute nodes; each node contains 24 core Intel Xeon E5-2680v3 processors with 128 GB RAM, interconnected by 7D Enhanced hypercube InfiniBand. We carried out some numerical experiments to compare the performance of H-TFETI-DP and TFETI and to get information about the effect of clustering. The experiments were limited to the semicoercive problem. In all experiments, we use the relative precision stopping criterion with $\varepsilon = 1e^{-4}$.

11.1 Effect of clustering

To understand the effect of clustering, we decomposed the domain of each membrane into 1024 square subdomains discretized by 100×100 degrees of freedom each. The subdomains were interconnected into $m \times m$ clusters, $m \in \{1, 2, 4, 8\}$, with $m = 1$ corresponding to the standard TFETI method. The (primal) dimension of the resulting

Table 1 Performance of H-TFETI-DP with $m \times m$ clusters

| Cores | Clusters | m | SMALBE-M iterations | Matrix \times vector | Time (s) |
|-------|----------|-----|---------------------|------------------------|----------|
| 32 | 32 | 8 | 12 | 218 | 142.87 |
| 128 | 128 | 4 | 16 | 186 | 37.79 |
| 512 | 512 | 2 | 25 | 252 | 24.30 |
| 2048 | 2048 | 1 | 52 | 243 | 74.51 |

discretized problem was 20,480,000, 3169 inequalities enforced the nonpenetration. We allocated each cluster one computational core.

The results in Table 1 indicate a positive effect of clustering on the rate of convergence of the outer loop. At least a partial explanation provided Jarošová et al. [28], who proved that we could interpret the interconnecting of subdomains into clusters as the preconditioning by conjugate projector [6]. The clustering reduces the number of equality constraints m^2 -times, so it is not surprising that the number of the outer iterations decreases with m .

11.2 Scalability

The purpose of the second set of numerical experiments was to demonstrate the performance of the algorithms and numerical scalability supported by Theorem 3. We used

$$H_c/h = mH_s/h = 255, \quad m \in \{1, 2, 4, 8\}$$

to decompose the domain into 338, 648, 968, and 1250 clusters discretized by 22,151,168 to 81,920,000 nodal variables. The numbers of the most expensive operations, the multiplication by the matrix **PF**, are depicted in Fig. 5. We can see that the

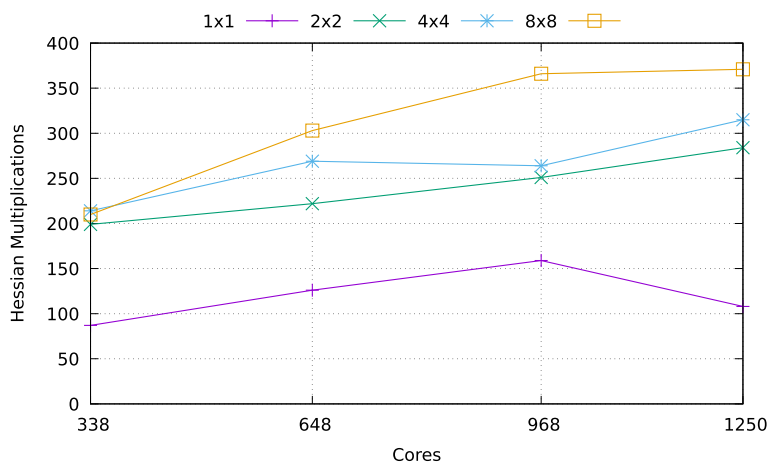
**Fig. 5** Matrix–vector multiplications for $m = 1, 2, 4, 8$

Table 2 Variational inequality with 4×4 clusters and $H_s/h = 99$

| Primal dimension | Dual dimension | SMALBE-M clusters | SMALBE-M iterations | Matrix \times vector | Time (s) |
|------------------|----------------|-------------------|---------------------|------------------------|----------|
| 169,280,000 | 1,670,305 | 1058 | 36 | 350 | 42.95 |
| 327,680,000 | 3,243,199 | 2048 | 50 | 416 | 81.08 |
| 677,120,000 | 6,685,639 | 4032 | 70 | 549 | 276.83 |

number of matrix-vector multiplications increases very mildly in agreement with the theory. Notice that the expansion step requires two matrix–vector multiplications.

11.3 Clusters with fine grid

To see the performance of the algorithm on larger problems, we fixed the number of nodes on the edge of subdomains to 100 and carried out the computation with relatively large 4×4 clusters. The results are in Table 2.

We can see that the number of iterates of the H-TFETI-DP algorithm without preconditioning increases still rather slowly in agreement with the theory. The number of iterations can be affected by large subdomains and a very fine grid that generates many nodes on the contact interface that touch the support. The performance of the algorithms can be improved by using a recently proposed adaptive reorthogonalization [13]. It is also possible to improve the performance of MPRGP [34].

11.4 3D Elastic cube on rigid support

Our final benchmark indicates the importance of the small coarse grid. We resolved the obstacle problem defined by a clumped elastic cube over the sinus-shaped obstacle as in Fig. 6, loaded down by its weight, decomposed into $4 \times 4 \times 4$ clusters, $H_s/h = 14$. We used the ESPRESO [22] implementation of H-TFETI-DP for contact problems developed in a National Supercomputer Center IT4Innovations of VSB-TU

Fig. 6 Elastic body on rigid support

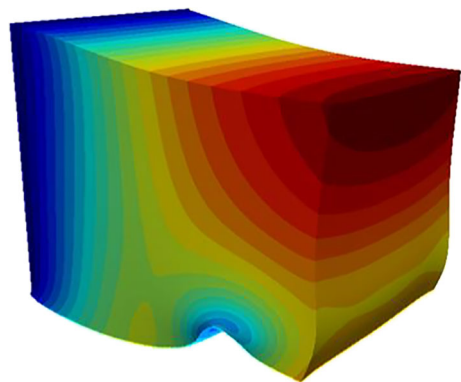


Table 3 Elastic body on the sinus-shaped support, $m=4$, $H_s/h = 14$

| Clusters | Subdomains | Unknowns $\times 10^6$ | H-TFETI-DP (iter/s) | TFETI (iter/s) |
|----------|------------|------------------------|------------------------|-------------------|
| 64 | 4,096 | 13 | 169/23.9 | 117/24.9 |
| 512 | 72,900 | 99 | 208/30.2 | 152/115.1 |
| 1,000 | 656,100 | 193 | 206/42.6 | 173/279.9 |

Ostrava. We can see in Table 3 that TFETI needs a much smaller number of iterations, but H-TFETI-DP is still faster due to 64-times smaller coarse space and better exploitation of the node-core memory organization. In general, if we use $m \times m \times m$ clusters, the hybrid strategy reduces the dimension and the cost of the coarse problem by m^3 and m^6 , respectively.

12 Comments and conclusions

We have used recently established bounds on the regular condition number of the Schur complements of floating clusters arising from the interconnecting of square subdomains by edge averages [17] to develop a theoretically supported massively parallel algorithm for the solution of variational inequalities. The performance of the algorithms was demonstrated by solving an academic benchmark and a 3D obstacle problem discretized by hundreds of millions of nodal variables. In particular, the results show that joining the subdomains into $m \times m$ clusters increases only slowly the number of iterations necessary to achieve a prescribed relative precision and reduces m^4 -times (m^6 for 3D problems) the cost of preparation of the coarse problem. The theoretical results and numerical experiments indicate that unpreconditioned H-TFETI-DP with large clusters can be a competitive computational engine for solving huge systems of variational inequalities discretized by sufficiently regular structured grids. Using the reorthogonalization-based preconditioning [17], we can achieve the same performance for the problems with variable coefficients provided they are constant on the clusters. The methods of proofs can be used to get similar results for liner elasticity and more general grids, particularly those obtained by the deformation of a structured grid.

Funding This work was supported by The Ministry of Education, Youth, and Sports from the National Programme of Sustainability (NPS II) project “IT4Innovations excellence in science - LQ1602” and by the IT4Innovations infrastructure which is supported from the Large Infrastructures for Research, Experimental Development and Innovations project “e-INFRA CZ– LM2018140.” The second and third authors were supported by the grants 19-11441S of the Czech Science Foundation (GACR), SGS SP2021/103 of the VŠB – Technical University of Ostrava and “RRC/10/2019” Support for Science and Research in the Moravia–Silesia Region 2.

Data availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Conflict of interest The authors no competing interests.


References

1. Axelsson, O.: Iterative Solution Methods. Cambridge University Press, Cambridge (1994)
2. Brenner, S.C.: The condition number of the Schur complement. *Numer. Math.* **83**, 187–203 (1999)
3. Brzobohatý, T., Dostál, Z., Kozubek, T., Kovář, P., Markopoulos, A.: Cholesky decomposition with fixing nodes to the stable computation of a generalized inverse of the stiffness matrix of a floating structure. *Int. J. Numer. Methods Eng.* **88**(5), 493–509 (2011)
4. Brzobohatý, T., Jarošová, M., Kozubek, T., Menšík, M., Markopoulos, A.: The hybrid total FETI method. In: Proceedings of the Third International Conference on Parallel, Distributed, Grid, and Cloud Computing for Engineering Civil-Comp (2013)
5. Conn, A.R., Gould, N.I.M., Toint, P.L.: A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds. *SIAM J. Numer. Anal.* **28**, 545–572 (1991)
6. Dostál, Z.: Conjugate gradient method with preconditioning by projector. *Int. J. Comput. Math.* **23**, 315–324 (1988)
7. Dostál, Z.: A proportioning based algorithm with rate of convergence for bound constrained quadratic programming. *Numer. Algorithms* **34**(2–4), 293–302 (2003)
8. Dostál, Z.: An optimal algorithm for bound and equality constrained quadratic programming problems with bounded spectrum. *Computing* **78**, 311–328 (2006)
9. Dostál, Z.: On the decrease of a quadratic function along the projected–gradient path. *ETNA* **31**, 25–59 (2008)
10. Dostál, Z.: Optimal quadratic programming algorithms, with applications to variational inequalities, 1st edn. Springer, New York (2009)
11. Dostál, Z., Horák, D.: Theoretically supported scalable FETI for numerical solution of variational inequalities. *SIAM J. Numer. Anal.* **45**(2), 500–513 (2007)
12. Dostál, Z., Schöberl, J.: Minimizing quadratic functions subject to bound constraints with the rate of convergence and finite termination. *Comput. Opt. Appl.* **30**(1), 23–44 (2005)
13. Dostál, Z., Vlach, O.: An accelerated augmented Lagrangian algorithm with adaptive orthogonalization strategy for bound and equality constrained quadratic programming and its application to large-scale contact problems of elasticity. *J. Comput. Appl. Math.* **394**(1), 113565 (2021)
14. Dostál, Z., Gomes, F.A.M., Santos, S.A.: Duality based domain decomposition with natural coarse space for variational inequalities. *J. Comput. Appl. Math.* **126**(1–2), 397–415 (2000)
15. Dostál, Z., Friedlander, A., Santos, S.A.: Augmented Lagrangians with adaptive precision control for quadratic programming with simple bounds and equality constraints. *SIAM J. Optim.* **13**(4), 1120–1140 (2003)
16. Dostál, Z., Horák, D., Kučera, R.: Total FETI—an easier implementable variant of the FETI method for numerical solution of elliptic PDE. *Commun. Numer. Methods Eng.* **22**, 1155–1162 (2006)
17. Dostál, Z., Kozubek, T., Vlach, O.: Reorthogonalization based stiffness preconditioning in FETI algorithms with applications to variational inequalities. *Numer. Lin. Algebra Appl.* **22**(6), 987–998 (2015)
18. Dostál, Z., Kozubek, T., Sadowská, M., Vondrák, V.: Scalable Algorithms for Contact Problems AMM, vol. 36. Springer, New York (2016)
19. Dostál, Z., Horák, D., Sojka, R.: On the efficient reconstruction of displacements in FETI methods for contact problems. *Adv. Electr. Electron. Eng.* **15**, 237–241 (2017)
20. Dostál, Z., Horák, D., Brzobohatý, T., Vodstrčil, P.: Bounds on the spectra of Schur complements of large H-TFETI clusters for 2D Laplacian and applications. *Numer. Lin. Algebra Appl.* <https://doi.org/10.1002/nla.2344>
21. Dostál, Z., Brzobohatý, T., Vlach, O.: Schur complement spectral bounds for large hybrid FETI-DP clusters and huge three-dimensional scalar problems. *J. Numer. Math.* (2021)
22. ESPRESO—Highly Parallel Framework for Engineering Applications. <http://numbox.it4i.cz>
23. Farhat, C., Roux, F.-X.: A method of finite element tearing and interconnecting and its parallel solution algorithm. *Int. J. Numer. Methods Eng.* **32**, 1205–1227 (1991)

24. Farhat, C., Roux, F.-X.: An unconventional domain decomposition method for an efficient parallel solution of large-scale finite element systems. *SIAM J. Sci. Comput.* **13**, 379–396 (1992)
25. Farhat, C., Mandel, J., Roux, F.-X.: Optimal convergence properties of the FETI domain decomposition method. *Comput. Methods Appl. Mech. Eng.* **115**, 365–385 (1994)
26. Farhat, C., Lesoinne, M., Pierson, K.: A scalable dual-primal domain decomposition method. *Numer. Lin. Algebra Appl.* **7**(7–8), 687–714 (2000)
27. Hlaváček, I., Haslinger, J., Nečas, J., Lovíšek, J.: *Solution of variational inequalities in mechanics*. Springer, Berlin (1988). *Topics in Matrix Analysis*. Cambridge University Press, Cambridge (1991)
28. Jarošová, M., Klawonn, A., Rheinbach, O.: Projector preconditioning and transformation of basis in FETI-DP algorithms for contact problem. *Math. Comput. Simul.* **82**(10), 1894–1907 (2012)
29. Klawonn, A., Rheinbach, O.: A parallel implementation of dual-primal FETI methods for three dimensional linear elasticity using a transformation of basis. *SIAM J. Sci. Comput.* **28**(5), 1886–1906 (2006)
30. Klawonn, A., Rheinbach, O.: A hybrid approach to 3-level FETI. *Proc. Appl. Math. Mech.* **90**(1), 10841–10843 (2008)
31. Klawonn, A., Rheinbach, O.: Highly scalable parallel domain decomposition methods with an application to biomechanics. *Z. Angew. Math. Mech.* **90**(1), 5–32 (2010)
32. Klawonn, A., Widlund, O.: Dual-primal FETI method for linear elasticity. *Commun. Pure Appl. Anal.* **LIX**, 1523–1572 (2006)
33. Klawonn, A., Lanser, M., Rheinbach, O.: Toward extremally scalable nonlinear domain decomposition methods for elliptic partial differential equations. *SIAM J. Sci. Comput.* (37) **6**, C667–C696 (2015)
34. Kružík, J., Horák, D., Čermák, M., Pospíšil, L., Pecha, M.: Active set expansion strategies in MPRGP algorithm. *Adv. Eng. Softw.* **149**(1028), 95 (2020)
35. Lee, J.: Domain decomposition methods for auxiliary linear problems of an elliptic variational inequality. In: Bank, R., et al (eds.) *Domain Decomposition Methods in Science and Engineering XX*, Lecture Notes in Computational Science and Engineering, vol. 91, pp. 319–326 (2013)
36. Lee, J.: Two domain decomposition methods for auxiliary linear problems for a multibody variational inequality. *SIAM J. Sci. Comput.* **35**(3), 1350–1375 (2013)
37. Li, J., Widlund, O.B.: FETI-DP, BDDC, and block Cholesky method. *Int. J. Num. Methods Eng.* **66**, 250–271 (2006)
38. Pechstein, C.: *Finite and boundary element tearing and interconnecting solvers for multiscale problems*. Springer, Heidelberg (2013)
39. PERMON—Parallel, efficient, robust, modular, object-oriented numerical software toolbox. <http://permon.vsb.cz/>
40. Toselli, A., Widlund, O.B.: *Domain decomposition methods—algorithms and theory* Springer Series on Computational Mathematics, vol. 34. Springer, Berlin (2005)
41. Vodstrčil, P., Bouchala, J., Jarošová, M., Dostál, Z.: On conditioning of Schur complements of h-TFETI clusters for 2D problems governed by Laplacian. *Appl. Math.* **62**(6), 699–718 (2017)

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Zdeněk Dostál¹  · David Horák² · Jakub Kružík² · Tomáš Brzobohatý³ · Oldřich Vlach¹

David Horák
david.horak@vsb.cz

Jakub Kružík
jakub.kruzik@vsb.cz

Tomáš Brzobohatý
tomas.brzobohaty@vsb.cz

Oldřich Vlach
oldrich.vlach2@vsb.cz

¹ Department of Applied Mathematics and IT4Innovations-National Supercomputing Center, VSB-Technical University of Ostrava, 17. listopadu 15, 70833, Ostrava, Czech Republic

² Department of Applied Mathematics, VSB-TU Ostrava and Institute of Geonics ASCR, Ostrava, Czech Republic

³ IT4Innovations-National Supercomputing Center, VSB-TU Ostrava, Ostrava, Czech Republic