Numerische Mathematik (2022) 150:257–298
https://doi.org/10.1007/s00211-021-01267-8

Numerische Mathematik

Preconditioning techniques for the coupled Stokes–Darcy problem: spectral and field-of-values analysis

Fatemeh Panjeh Ali Beik[1] · **Michele Benzi**[2]

Received: 3 May 2021 / Revised: 8 October 2021 / Accepted: 16 December 2021 /
Published online: 20 January 2022

Abstract
We study the performance of some preconditioning techniques for a class of block three-by-three linear systems of equations arising from finite element discretizations of the coupled Stokes–Darcy flow problem. In particular, we investigate preconditioning techniques including block preconditioners, constraint preconditioners, and augmented Lagrangian-based ones. Spectral and field-of-value analyses are established for the exact versions of these preconditioners. The result of numerical experiments are reported to illustrate the performance of inexact variants of the various preconditioners used with flexible GMRES in the solution of a 3D test problem with large jumps in the permeability.

Mathematics Subject Classification 65F10

1 Introduction

The coupled Stokes–Darcy model describes the interaction between free flow and porous media flow. It is a fundamental problem in several fields [14]. In one subregion of the flow domain Ω a free-flowing fluid is governed by the (Navier–)Stokes equations; in another subregion, the fluid follows Darcy's Law. The equations are coupled by conditions across the interface between the two subregions. In this paper we will only consider the case of stationary problems and Stokes flow.

Let Ω be a computational domain partitioned into two non-overlapping subdomains Ω_1 and Ω_2, separated by an interface Γ_{12}. We assume that the flow in Ω_1 is governed

✉ Michele Benzi
michele.benzi@sns.it

Fatemeh Panjeh Ali Beik
f.beik@vru.ac.ir; beik.fatemeh@gmail.com

[1] Department of Mathematics, Vali-e-Asr University of Rafsanjan, P.O. Box 518, Rafsanjan, Iran

[2] Scuola Normale Superiore, Piazza dei Cavalieri, 7, 56126 Pisa, Italy

by the stationary Stokes equations:

$$\begin{aligned} -\nabla \cdot (2\nu D(\mathbf{u}_1) - p_1 \mathbf{I}) &= \mathbf{f}_1 \quad \text{in} \quad \Omega_1, \\ \nabla \cdot \mathbf{u}_1 &= 0 \quad \text{in} \quad \Omega_1, \\ \mathbf{u}_1 &= 0 \quad \text{on} \quad \Gamma_1 = \partial\Omega_1 \cap \partial\Omega. \end{aligned}$$

Here $\nu > 0$ represents the kinematic viscosity, $\mathbf{u}_1$ and p_1 denote the velocity and pressure in Ω_1, $\mathbf{f}_1$ is an external force acting on the fluid, $\mathbf{I}$ is the identity matrix, and

$$D(\mathbf{u}_1) = \frac{1}{2}\left(\nabla \mathbf{u}_1 + \nabla \mathbf{u}_1^T\right)$$

is the rate of strain tensor. We also assume that the boundary $\Gamma_2 = \partial\Omega \cap \partial\Omega_2$ of the porous medium is partitioned into disjoint Neumann and Dirichlet parts Γ_{2N} and Γ_{2D}, with Γ_{2D} having positive measure. The flow in Ω_2 is governed by Darcy's Law:

$$\begin{aligned} -\nabla \cdot \mathbf{K} \nabla p_2 &= f_2 \quad \text{in} \quad \Omega_2, \\ p_2 &= g_D \quad \text{on} \quad \Gamma_{2D}, \\ \mathbf{K} \nabla p_2 \cdot \mathbf{n}_2 &= g_N \quad \text{on} \quad \Gamma_{2N}. \end{aligned}$$

Here p_2 represents the Darcy pressure in Ω_2, and the symmetric positive definite (SPD) matrix $\mathbf{K}$ represents the hydraulic conductivity in the porous medium. The Darcy velocity can be obtained from the pressure using

$$\mathbf{u}_2 = -\mathbf{K} \nabla p_2 \quad \text{in} \quad \Omega_2.$$

The coupling between the two flows comes from the following interface conditions on the internal boundary Γ_{12}. Let $\mathbf{n}_{12}$ and $\mathbf{t}_{12}$ denote the unit normal vector directed from Ω_1 to Ω_2 and the unit tangent vector to the interface. Then we impose

$$\begin{aligned} \mathbf{u}_1 \cdot \mathbf{n}_{12} &= -\mathbf{K} \nabla p_2 \cdot \mathbf{n}_{12}, \\ (-2\nu D(\mathbf{u}_1)\, \mathbf{n}_{12} + p_1 \mathbf{n}_{12}) \cdot \mathbf{n}_{12} &= p_2, \\ \mathbf{u}_1 \cdot \mathbf{t}_{12} &= -2\nu G(D(\mathbf{u}_1)\, \mathbf{n}_{12}) \cdot \mathbf{t}_{12}. \end{aligned}$$

The first two conditions enforce mass conservation and the balance of normal forces across the interface; the third condition represents the Beavers–Joseph–Saffman (BJS) law, in which G is an experimentally determined constant. Let

$$\mathbf{X} = \{\mathbf{v}_1 \in (H^1(\Omega_1))^2 \,|\, \mathbf{v}_1 = \mathbf{0} \ \text{ on } \ \Gamma_1\}, \quad Q_1 = L^2(\Omega_1)$$

be the Stokes velocity and pressure spaces and let

$$Q_2 = \{q_2 \in H^1(\Omega_2) \,|\, q_2 = 0 \ \text{ in } \ \Gamma_{2D}\}$$

be the Darcy pressure space. The weak formulation of the coupled Stokes–Darcy problem is:
find $\mathbf{u}_1 \in \mathbf{X}$, $p_1 \in Q_1$ and $p_2 \in Q_2$ such that

$$a(\mathbf{u}_1, p_2; \mathbf{v}_1, q_2) + b(\mathbf{v}_1, p_1) = \mathbf{f}(\mathbf{v}_1, q_2) \quad \forall \mathbf{v}_1 \in \mathbf{X}, \ \forall q_2 \in Q_2,$$
$$b(\mathbf{u}_1, q_1) = 0 \ \forall q_1 \in Q_1.$$

Here

$$a(\mathbf{u}_1, p_2; \mathbf{v}_1, q_2) = a_{\Omega_1}(\mathbf{u}_1, \mathbf{v}_1) + a_{\Omega_2}(p_2, q_2) + a_{\Gamma_{12}}(\mathbf{u}_1, p_2; \mathbf{v}_1, q_2)$$

where

$$a_{\Omega_1}(\mathbf{u}_1, \mathbf{v}_1) = 2\nu \int_{\Omega_1} D(\mathbf{u}_1) : D(\mathbf{v}_1) + \frac{1}{G} \int_{\Gamma_{12}} (\mathbf{u}_1 \cdot \mathbf{t}_{12})(\mathbf{v}_1 \cdot \mathbf{t}_{12}),$$
$$a_{\Omega_2}(p_2, q_2) = \int_{\Omega_2} \mathbf{K} \nabla p_2 \cdot \nabla q_2,$$
$$a_{\Gamma_{12}}(\mathbf{u}_1, p_2; \mathbf{v}_1, q_2) = \int_{\Gamma_{12}} (p_2 \mathbf{v}_1 - q_2 \mathbf{u}_1) \cdot \mathbf{n}_{12}.$$

Also,

$$b(\mathbf{u}_1, q_1) = -\int_{\Omega_1} (\nabla \cdot \mathbf{u}_1) q_1,$$

and

$$\mathbf{f}(\mathbf{v}_1, q_2) = \int_{\Omega_1} \mathbf{f}_1 \cdot \mathbf{v}_1 + \int_{\Omega_2} f_2 q_2 + \int_{\Gamma_{2N}} g_N q_2.$$

The well-posedness of the weak formulation is a consequence of Brezzi-Fortin theory (see, e.g., [11]). The weak form is discretized using conforming finite elements spaces $\mathbf{X}^h \subset \mathbf{X}$, $Q_1^h \subset Q_1$ satisfying the inf-sup condition for the Stokes velocity and pressure, such as the MINI and Taylor–Hood elements. For the Darcy pressure a space of piecewise continuous polynomials $Q_2^h \subset Q_2$ is used (linear in 2D, quadratic in 3D).

The discrete form of the weak formulation can be cast as a block linear system of the form

$$\mathcal{A}u = \begin{bmatrix} A_{\Omega_2} & A_{\Gamma_{12}}^T & 0 \\ -A_{\Gamma_{12}} & A_{\Omega_1} & B^T \\ 0 & B & 0 \end{bmatrix} \begin{bmatrix} \hat{p}_2 \\ \hat{\mathbf{u}}_1 \\ \hat{p}_1 \end{bmatrix} = \begin{bmatrix} \hat{f}_{2,h} \\ \hat{\mathbf{f}}_{1,h} \\ \hat{g}_h \end{bmatrix} = b$$

where A_{Ω_2}, A_{Ω_1}, $A_{\Gamma_{12}}$ are the matrices of the discrete bilinear forms and B is the discrete divergence. Under our assumptions A_{Ω_2} and A_{Ω_1} are SPD and B has full row rank; we refer the reader to [12,13] for further details.

We now introduce a slight change of notation and rewrite the previous linear system of equations in the following form:

$$\mathcal{A}u = \begin{bmatrix} A_{11} & A_{12} & 0 \\ A_{21} & A_{22} & B^T \\ 0 & B & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = b, \tag{1}$$

where A_{11} and A_{22} are both SPD, $A_{21} := -A_{12}^T$ and B has full row rank. We observe in passing that this system is an example of a double saddle point problem, and that similarly structured systems arise in a number of applications, see [5]. In particular we note that using the following similarity transformation,

$$\mathcal{A}_1 = \begin{bmatrix} A_{22} & B^T & -A_{12}^T \\ B & 0 & 0 \\ A_{12} & 0 & A_{11} \end{bmatrix} = X^T \mathcal{A} X, \tag{2}$$

where

$$X = \begin{bmatrix} 0 & 0 & I \\ I & 0 & 0 \\ 0 & I & 0 \end{bmatrix},$$

we can immediately conclude (under our assumptions) the invertibility of $\mathcal{A}$ from [5, Proposition 2.1].

The iterative solution of the discrete coupled Stokes–Darcy equations has attracted considerable attention in recent years. Here we limit ourselves to discussing solution algorithms based on preconditioned Krylov subspace methods. In [13], the following two constraint-type preconditioners were proposed for accelerating the convergence of Krylov subspace methods:

$$\mathcal{P}_{con_D} = \begin{bmatrix} A_{11} & 0 & 0 \\ 0 & A_{22} & B^T \\ 0 & B & 0 \end{bmatrix} \quad \text{and} \quad \mathcal{P}_{con_T} = \begin{bmatrix} A_{11} & 0 & 0 \\ A_{21} & A_{22} & B^T \\ 0 & B & 0 \end{bmatrix}. \tag{3}$$

Also Cai et al. [12] proposed the following block triangular preconditioner,

$$\mathcal{P}_{T_1,\rho} := \mathcal{P}_{T_1}(\rho) = \begin{bmatrix} A_{11} & 0 & 0 \\ 0 & A_{22} & 0 \\ 0 & B & -\rho M_p \end{bmatrix}, \tag{4}$$

where M_p is the mass matrix associated with the Stokes pressure space, and $\rho > 0$ is a user-defined parameter. For 2D problems, the preconditioner $\mathcal{P}_{con_T}$ outperforms the other preconditioners in terms of both iterations and CPU-time when used with GMRES. Exact variants of the preconditioners result in h-independent rates of convergence, as predicted by the theory. For 3D problems, only the inexact variants of these

preconditioners are feasible. Of the inexact variants, $\mathcal{P}_{T_1,\rho}$ with suitable ρ requires the least CPU time, according to [13].

The constraint preconditioners and $\mathcal{P}_{T_1,\rho}$ are *norm equivalent* to $\mathcal{A}$ in (1) under certain conditions; see [12,13]. On the other hand, the *Field-of-Values (FOV) equivalence* of constraint preconditioners with $\mathcal{A}$ was proved in [13]. It is well-known that if a preconditioner is norm equivalent to the coefficient matrix of a linear system of equations, the spectra of the preconditioned system remain uniformly bounded and bounded away from zero as the mesh size $h \to 0$, see [23] for more details. Here, we directly determine some bounds for the eigenvalues of the preconditioned matrices associated with $\mathcal{P}_{con_D}$, $\mathcal{P}_{con_T}$ and $\mathcal{P}_{T_1,\rho}$. In particular, we show that the eigenvalues of the preconditioned matrix corresponding to $\mathcal{P}_{con_T}$ are nicely clustered under certain conditions.

In the present work, we also consider the following block triangular preconditioner:

$$\mathcal{P}_r = \begin{bmatrix} A_{11} & A_{12} & 0 \\ 0 & A_{22} + rB^T Q^{-1} B & B^T \\ 0 & 0 & -\frac{1}{r} Q \end{bmatrix} \tag{5}$$

applied to the *augmented* linear system of equations $\bar{\mathcal{A}} u = \bar{b}$, where

$$\bar{\mathcal{A}} = \begin{bmatrix} A_{11} & A_{12} & 0 \\ A_{21} & A_{22} + rB^T Q^{-1} B & B^T \\ 0 & B & 0 \end{bmatrix}, \tag{6}$$

and $\bar{b} = (b_1; b_2 + rB^T Q^{-1} b_3; b_3)$, with Q being an arbitrary SPD matrix and $r > 0$ a user-defined parameter. Evidently, the linear system of equations $\bar{\mathcal{A}} x = \bar{b}$ is equivalent to $\mathcal{A} u = b$. This approach is motivated by the success of the use of grad-div stabilization and augmented Lagrangian techniques for solving saddle point problems.

The remainder of this paper is organized as follows. In Sect. 2 we derive lower and upper bounds for the eigenvalues of the preconditioned matrices corresponding to all of the above-mentioned preconditioners. In Sect. 3, we establish FOV-type bounds for the preconditioned system associated with the preconditioner of type (5). Some numerical experiments are reported in Sect. 4 to compare the performance of preconditioners, in particular in the presence of inexact solves. Brief conclusive remarks are given in Sect. 5.

Notations. We use "i" for the imaginary unit. The notation $\sigma(A)$ is used for the spectrum of a square matrix A. When all eigenvalues of A are real and positive, we use $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ to respectively represent the minimum and maximum eigenvalues of A. The notation $\rho(A)$ stands for the spectral radius of A. When A is symmetric positive (semi)definite, we write $A \succ 0$ ($A \succcurlyeq 0$). Furthermore, for two given matrices A and B, by $A \succ B$ ($A \succcurlyeq B$) we mean $A - B \succ 0$ ($A - B \succcurlyeq 0$). For $H \succ 0$, the corresponding vector norm is given by

$$\langle v, v \rangle_H = \langle Hv, v \rangle = v^T Hv = \|v\|_H^2 ,$$

whose induced matrix norm is defined by

$$\|M\|_H = \max_{0 \neq v} \frac{\|Mv\|_H}{\|v\|_H}.$$

For given vectors x, y and z of dimensions n, m and p, $(x; y; z)$ will denote a column vector of dimension $n + m + p$.

2 Spectral analysis

This section is devoted to deriving lower and upper bounds for the eigenvalues of preconditioned matrices $\mathcal{P}_{con_T}^{-1}\mathcal{A}$, $\mathcal{P}_{con_D}^{-1}\mathcal{A}$, $\mathcal{P}_{T_{1,\rho}}^{-1}\mathcal{A}$ and $\mathcal{P}_r^{-1}\bar{\mathcal{A}}$. To this end, first, we recall the following basic lemma which is an immediate consequence of Weyl's Theorem, see [20, Theorem 4.3.1].

Lemma 2.1 *Let A and B be two Hermitian matrices. Then,*

$$\lambda_{\max}(A + B) \leq \lambda_{\max}(A) + \lambda_{\max}(B),$$
$$\lambda_{\min}(A + B) \geq \lambda_{\min}(A) + \lambda_{\min}(B).$$

The following two results provide information on the eigenvalues of the constraint-preconditioned matrices.

Theorem 2.2 *Suppose that $\mathcal{P}_{con_D}$ is defined by (3). The eigenvalues of $\mathcal{P}_{con_D}^{-1}\mathcal{A}$ are either equal to unity, or complex numbers of the form $1 \pm \mathrm{i}\sqrt{\xi}$, where*

$$0 < \xi \leq \frac{\lambda_{\max}(A_{12}^T A_{11}^{-1} A_{12})}{\lambda_{\min}(A_{22})}.$$

Proof Let $\lambda \in \sigma(\mathcal{P}_{con_D}^{-1}\mathcal{A})$ with the corresponding eigenvector $(x; y; z)$. Therefore, we have

$$A_{11}x + A_{12}y = \lambda A_{11}x \tag{7}$$
$$A_{21}x + A_{22}y + B^T z = \lambda(A_{22}y + B^T z) \tag{8}$$
$$By = \lambda By. \tag{9}$$

Notice that $\lambda = 1$ is a possible eigenvalue of $\mathcal{P}_{con_D}^{-1}\mathcal{A}$ with corresponding eigenvector $(0; y; z)$ where $y \in \mathrm{Ker}(A_{12})$ and z is an arbitrary vector such that y and z are not both zero.

It is also immediate to observe that $\lambda \neq 1$ implies $y \neq 0$. Otherwise, in view of (7), we have $x = 0$ when $y = 0$. In this case, Eq. (8) results in $B^T z = 0$, which is equivalent to $z = 0$ since B^T has full column rank. Hence, we get $(x; y; z) = (0; 0; 0)$, which is impossible since $(x; y; z)$ is an eigenvector.

In the rest of proof, we assume that $\lambda \neq 1$, which implies $By = 0$ by (9). Now, we can compute vector x from (7) as follows:

$$x = \frac{1}{\lambda - 1} A_{11}^{-1} A_{12} y.$$

Multiplying both sides of Eq. (8) by y^* and recalling that $A_{21} = -A_{12}^T$, we get

$$(\lambda - 1)^2 = -\frac{y^* A_{12}^T A_{11}^{-1} A_{12} y}{y^* A_{22} y},$$

which is equivalent to

$$\lambda = 1 \pm \mathrm{i} \sqrt{\frac{y^* A_{12}^T A_{11}^{-1} A_{12} y}{y^* A_{22} y}},$$

from which the assertion follows immediately. □

Theorem 2.3 *Let $\mathcal{P}_{con_T}$ be defined by (3). Then*

$$\sigma(\mathcal{P}_{con_T}^{-1} \mathcal{A}) \subset \left[1, 1 + \frac{\lambda_{\max}(A_{12}^T A_{11}^{-1} A_{12})}{\lambda_{\min}(A_{22})}\right].$$

Proof Let λ be an arbitrary eigenvalue of $\mathcal{P}_{con_T}^{-1} \mathcal{A}$ with the corresponding eigenvector $(x; y; z)$. Therefore, we have,

$$\begin{aligned} A_{11}x + A_{12}y &= \lambda A_{11}x && (10)\\ A_{21}x + A_{22}y + B^T z &= \lambda(A_{21}x + A_{22}y + B^T z) && (11)\\ By &= \lambda By. && (12) \end{aligned}$$

For $y \in \mathrm{Ker}(A_{12})$, we observe that $1 \in \sigma(\mathcal{P}_{con_T}^{-1} \mathcal{A})$ with the corresponding eigenvector $(x; y; z)$ where x and z are arbitrary vectors with at least one of them nonzero when $y = 0$. In the rest of proof, we assume that $\lambda \neq 1$. From (10), we obtain

$$x = \frac{1}{\lambda - 1} A_{11}^{-1} A_{12} y.$$

Notice that from the fact that B^T has full column rank, similar to the proof of Theorem 2.2, one can conclude that $y \neq 0$. In addition, Eq. (12) shows that $By = 0$ when $\lambda \neq 1$. Now, we first substitute x from the above expression into (11) and then multiply both sides by $(\lambda - 1)y^*$, which yields

$$y^* A_{21} A_{11}^{-1} A_{12} y + (\lambda - 1) y^* A_{22} y = \lambda y^* A_{21} A_{11}^{-1} A_{12} y + \lambda(\lambda - 1) y^* A_{22} y.$$

Using $A_{21} = -A_{12}^T$, we obtain the following quadratic equation:

$$\lambda^2 - (2 + \hat{\gamma})\lambda + (1 + \hat{\gamma}) = 0, \tag{13}$$

where

$$\hat{\gamma} = \frac{y^* \left(A_{12}^T A_{11}^{-1} A_{12} \right) y}{y^* A_{22} y}. \tag{14}$$

The roots of (13) are given by $\lambda_1 = 1$ and $\lambda_2 = 1 + \hat{\gamma}$. This shows that the eigenvalue $\lambda \in \sigma(\mathcal{P}_{con_T}^{-1}\mathcal{A})$, not being equal to one, can be written in the following form:

$$1 + \frac{y^* \left(A_{12}^T A_{11}^{-1} A_{12} \right) y}{y^* A_{22} y},$$

which completes the proof. □

Remark 2.4 If $A_{22} \succcurlyeq A_{12}^T A_{11}^{-1} A_{12}$, then from the proof Theorem 2.3 we can conclude that $\lambda \in [1, 2]$ for $\lambda \in \sigma(\mathcal{P}_{con_T}^{-1}\mathcal{A})$. Similarly, from the proof of Theorem 2.2, for $\lambda \in \sigma(\mathcal{P}_{con_D}^{-1}\mathcal{A})$, we can deduce that $|\mathrm{Im}(\lambda)| \leq 1$ when $A_{22} \succcurlyeq A_{12}^T A_{11}^{-1} A_{12}$.

We have been able to verify numerically that for linear systems of the form (1) arising from the finite element discretization of coupled Stokes–Darcy flow, the condition $A_{22} \succ A_{12}^T A_{11}^{-1} A_{12}$ in the above remark is indeed satisfied. In fact, we observed that the condition $A_{22} \succ A_{12}^T A_{11}^{-1} A_{12} + \sigma B^T M_p^{-1} B$ (with $0 < \sigma \leq 2$) holds true for problems of small or moderate size, and numerical tests suggest that it may hold for larger problems as well.

We now turn to the spectral analysis of the block triangular preconditioner $\mathcal{P}_{T_1,\rho}$. Based on numerical experiments, Cai et al. [12] pointed out that the performance of $\mathcal{P}_{T_1,\rho}$ is not very sensitive to the scaling factor ρ, particularly when it belongs to interval [0.6, 1.05]; see [12, Table 2]. Moreover, it is numerically observed that the spectrum $\mathcal{P}_{T_1,\rho}^{-1}\mathcal{A}$ lies in a semi-annulus which does not include zero and is entirely contained in the right half-plane; see [12, Figure 2]. In what follows, we show that the experimentally observed eigenvalue distribution of $\mathcal{P}_{T_1,\rho}^{-1}\mathcal{A}$ can be theoretically proven for certain values of ρ. To do so, first, we recall a theorem established by Kakeya [21] in 1912.

Theorem 2.5 *If $p(z) = \sum_{j=0}^{n} a_j z^j$ is a polynomial of degree n with real and positive coefficients, then all the zeros of p lie in the annulus $R_1 \leq |z| \leq R_2$ where $R_1 = \min_{0 \leq j \leq n-1} a_j/a_{j+1}$ and $R_2 = \max_{0 \leq j \leq n-1} a_j/a_{j+1}$.*

Notice that if a monotonicity assumption holds for the coefficients of the polynomial p in the above theorem, i.e., $0 \leq a_0 \leq a_1 \leq \cdots \leq a_n$, then all zeros of p are strictly less than unity in modulus. The latter results is the well-known Eneström–Kakeya Theorem [1].

Proposition 2.6 *Consider the preconditioned matrix $\mathcal{P}_{T_1,\rho}^{-1}\mathcal{A}$ where $\mathcal{P}_{T_1,\rho}$ is defined by (4), with $\rho > 0$. If $A_{22} \succ A_{12}^T A_{11}^{-1} A_{12} + \rho^{-1} B^T M_p^{-1} B$ and $\lambda \in \sigma(\mathcal{P}_{T_1,\rho}^{-1}\mathcal{A})$, then either $\lambda = 1$, or $\xi \leq |\lambda - 1| \leq 1$ and $|\lambda| > \tau$, for some positive constants $\xi < 1$ and τ.*

Proof Let λ be an arbitrary eigenvalue of $\mathcal{P}_{T_1,\rho}^{-1}\mathcal{A}$ with the corresponding eigenvector $(x; y; z)$. As a result, we have

$$\begin{aligned} A_{11}x + A_{12}y &= \lambda A_{11}x && (15)\\ A_{21}x + A_{22}y + B^T z &= \lambda A_{22}y && (16)\\ By &= \lambda By - \rho\lambda M_p z. && (17) \end{aligned}$$

Note that $1 \in \sigma(\mathcal{P}_{T_1,\rho}^{-1}\mathcal{A})$ with the corresponding eigenvector $(0; y; 0)$ for $0 \neq y \in \mathrm{Ker}(A_{12})$. From now on, we assume that $\lambda \neq 1$. From (15) and (17), we obtain $x = (\lambda - 1)^{-1} A_{11}^{-1} A_{12} y$ and $z = \rho^{-1}\lambda^{-1}(\lambda - 1) M_p^{-1} By$, respectively. Notice that y is nonzero, otherwise x and z are both zero which is in contradiction with the fact that $(x; y; z)$ is an eigenvector. In the sequel, without loss of generality, we assume that $y^* y = 1$. Noting that $A_{21} = -A_{12}^T$, we substitute x and z in (16) which yields

$$(1 - \lambda)^{-1} p + (1 - \lambda) q + \lambda^{-1}(\lambda - 1) r = 0,$$

where,

$$p = y^* A_{12}^T A_{11}^{-1} A_{12} y, \quad q = y^* A_{22} y \quad \text{and} \quad r = \rho^{-1} y^* B^T M_p^{-1} B y.$$

Multiplying both sides of the preceding relation by $\lambda(1 - \lambda)$, we derive

$$\lambda(\lambda - 1)^2 - \frac{r}{q}(\lambda - 1)^2 + \frac{p}{q}\lambda = 0.$$

For simplicity, we set $t = \lambda - 1$ and rewrite the previous relation as follows:

$$t^2(t + 1) - \frac{r}{q} t^2 + \frac{p}{q}(t + 1) = 0. \tag{18}$$

It is easy to check that that if $\lambda \neq 1$, then $y \in \mathrm{Ker}(B) \cap \mathrm{Ker}(A_{12})$ implies that y is a zero vector, which is impossible. As a result, p and r cannot be both zero. When $p = 0$, we readily obtain

$$\lambda = \frac{r}{q} \geq \min \left\{ \frac{\rho^{-1} y^* B^T M_p^{-1} B y}{\lambda_{\max}(A_{22})} \,\middle|\, y \notin \mathrm{Ker}(B) \right\}.$$

Notice that if $r = 0$ then either $t = \pm \mathrm{i}\frac{p}{q}$ or $t + 1 = 0$. Since $\lambda \neq 0$, in this case, we only conclude that $\lambda = 1 \pm \mathrm{i}\frac{p}{q}$ and $\varphi_1 < \frac{p}{q} < 1$ with

$$\varphi_1 = \min \left\{ \frac{y^* A_{12}^T A_{11}^{-1} A_{12} y}{\lambda_{\max}(A_{22})} \middle| \, y \notin \mathrm{Ker}(A_{12}) \right\} > 0.$$

Next, we assume $r, p \neq 0$ (i.e., $y \notin \mathrm{Ker}(B) \cup \mathrm{Ker}(A_{12})$) and rewrite (18) as follows:

$$t^3 + \left(1 - \frac{r}{q}\right) t^2 + \frac{p}{q} t + \frac{p}{q} = 0.$$

By the assumption $q > r + p$, hence

$$\frac{p}{q} \leq \left(1 - \frac{r}{q}\right) < 1.$$

Therefore, from Theorem 2.5, we conclude that $\xi \leq |t| = |\lambda - 1| \leq 1$ by setting $\xi = \min\{\varphi_1, \varphi_2\}$ where

$$\varphi_2 = \frac{\lambda_{\min}(A_{22} - \rho^{-1} B^T M_p^{-1} B)}{\lambda_{\max}(A_{22})} > 0,$$

keeping in mind that $\frac{p}{q} \leq \frac{p}{q-r}$. Now, from Eq. (18), we observe that

$$\begin{aligned} |\lambda| &= |t + 1| \\ &\geq \frac{r|t|^2}{q|t|^2 + p} \geq \frac{r\xi^2}{q + p} \\ &\geq \frac{\tau}{\lambda_{\max}(A_{22} + A_{12}^T A_{11}^{-1} A_{12})}, \end{aligned}$$

where $\tau = \min \left\{ \xi^2 \rho^{-1} y^* B^T M_p^{-1} B y \middle| \, y \notin \mathrm{Ker}(B) \right\} > 0.$ □

Theorem 2.7 *Suppose that $\mathcal{P}_r$ and $\bar{\mathcal{A}}$ are respectively defined by (5) and (6). The eigenvalues of $\mathcal{P}_r^{-1}\bar{\mathcal{A}}$ are all real and positive. More precisely, we have*

$$\sigma(\mathcal{P}_r^{-1}\bar{\mathcal{A}}) \subseteq \left[\theta, 2 + \frac{\lambda_{\max}(A_{12}^T A_{11}^{-1} A_{12})}{\lambda_{\min}(A_{22})}\right],$$

where

$$\theta = \frac{\zeta}{2 + \lambda_{\max}(A_{12}^T A_{11}^{-1} A_{12})/\lambda_{\min}(A_{22})},$$

with

$$\zeta = \min \left\{ \frac{r y^* B^T Q^{-1} B y}{y^* A_{22} y + r y^* B^T Q^{-1} B y} \middle| \, y \notin \mathrm{Ker}(B) \right\}.$$

Proof For ease of notation, we set $\bar{A}_{22} = A_{22} + r B^T Q^{-1} B$. Let $\lambda \in \sigma(\mathcal{P}_r^{-1} \bar{\mathcal{A}})$ be an arbitrary eigenvalue with the corresponding eigenvector $(x; y; z)$. Therefore, we have

$$A_{11} x + A_{12} y = \lambda (A_{11} x + A_{12} y) \tag{19}$$

$$A_{21} x + \bar{A}_{22} y + B^T z = \lambda (\bar{A}_{22} y + B^T z) \tag{20}$$

$$B y = -\frac{\lambda}{r} Q z \tag{21}$$

Notice that for $y \notin \mathrm{Ker}(B)$, we have that $\lambda = 1$ is an eigenvalue of $\mathcal{P}_r^{-1} \mathcal{A}$ with the corresponding eigenvector $(0; y; -r Q^{-1} B y)$. Also, $\lambda = 1$ is obviously an eigenvalue of $\mathcal{P}_r^{-1} \mathcal{A}$ associated with eigenvector $(0; y; 0)$ when $0 \neq y \in \mathrm{Ker}(B)$.

In the rest of proof, we assume that $\lambda \neq 1$. From Eqs. (19) and (21), we respectively derive

$$x = -A_{11}^{-1} A_{12} y \quad \text{and} \quad z = -\frac{r}{\lambda} Q^{-1} B y.$$

The preceding two relations for x and z make it clear that y cannot be the zero vector. Without loss of generality, we may assume that $\|y\|_2 = 1$.

If $0 \neq y \in \mathrm{Ker}(B)$, we deduce that $z = 0$. It follows that

$$\lambda = 1 + \frac{y^* A_{12}^T A_{11}^{-1} A_{12} y}{y^* \bar{A}_{22} y} \leq 1 + \frac{\lambda_{\max}(A_{12}^T A_{11}^{-1} A_{12})}{\lambda_{\min}(A_{22})},$$

recalling that $A_{21} = -A_{12}^T$. Now, we discuss the case that $y \notin \mathrm{Ker}(B)$. Substituting vectors x and z into (20) and performing straightforward computations, we obtain

$$y^* A_{12}^T A_{11}^{-1} A_{12} y + (1 - \lambda) y^* \bar{A}_{22} y + r \left(1 - \frac{1}{\lambda}\right) y^* B^T Q^{-1} B y = 0.$$

Multiplying both sides of the above relation by $-\lambda$, we obtain the following quadratic equation:

$$\lambda^2 - \gamma \lambda + \eta = 0, \tag{22}$$

where

$$\gamma = 1 + \frac{y^* \left(A_{12}^T A_{11}^{-1} A_{12} + r B^T Q^{-1} B\right) y}{y^* \bar{A}_{22} y} \quad \text{and} \quad \eta = \frac{r y^* B^T Q^{-1} B y}{y^* \bar{A}_{22} y}.$$

Evidently, $\gamma = 1 + \tilde{\gamma} + \eta$ with

$$\tilde{\gamma} = \frac{y^* \left(A_{12}^T A_{11}^{-1} A_{12} \right) y}{y^* \bar{A}_{22} y}.$$

Observing that $\gamma^2 - 4\eta \geq (1+\eta)^2 - 4\eta \geq 0$, we have that all of the eigenvalues of $\mathcal{P}_r^{-1}\bar{\mathcal{A}}$ are real. Moreover, if λ_1 and λ_2 are the roots of (22) then

$$\lambda_1 \lambda_2 = \eta \quad \text{and} \quad \lambda_1 + \lambda_2 = \gamma.$$

Hence, recalling that $\bar{A}_{22} = A_{22} + r B^T Q^{-1} B$, we easily obtain

$$\lambda_1 \lambda_2 \to 1 \quad \text{and} \quad (\lambda_1 + \lambda_2) \to 2,$$

as $r \to \infty$, i.e., all eigenvalues satisfying (22) tend to 1 for $r \to \infty$. From (22), we have

$$\lambda_1 = \frac{\gamma - \sqrt{\gamma^2 - 4\eta}}{2} \quad \text{and} \quad \lambda_2 = \frac{\gamma + \sqrt{\gamma^2 - 4\eta}}{2}.$$

It is not difficult to verify that

$$\lambda_1 = \frac{2\eta}{\gamma + \sqrt{\gamma^2 - 4\eta}} \geq \frac{\eta}{\gamma}. \tag{23}$$

Consequently, in view of the facts that $\lambda_{\min}(B^T Q^{-1} B) = 0$, $\eta \leq 1$ and $0 < \zeta \leq \eta$ for $y \notin \mathrm{Ker}(B)$, we conclude that

$$\begin{aligned} \lambda_1 \geq \frac{\eta}{\gamma} &\geq \frac{\zeta}{2 + y^*(A_{12}^T A_{11}^{-1} A_{12}) y / y^* \bar{A}_{22} y} \\ &\geq \frac{\zeta}{2 + \lambda_{\max}(A_{12}^T A_{11}^{-1} A_{12}) / \lambda_{\min}(A_{22})}. \end{aligned}$$

Evidently, since $\eta \leq 1$, we have

$$\lambda_2 \leq \gamma \leq 2 + \frac{y^* A_{12}^T A_{11}^{-1} A_{12} y}{y^* A_{22} y},$$

which completes the proof. □

Remark 2.8 Notice that when $A_{22} \succcurlyeq A_{12}^T A_{11}^{-1} A_{12}$, then from the proof of Theorem 2.7, we obtain that $\sigma(\mathcal{P}_r^{-1}\bar{\mathcal{A}})$ lies in the interval $[1, 2]$ in the limit as $r \to \infty$. Moreover, "most" of the eigenvalues of the preconditioned matrix are either equal to 1, or tend to 1 as $r \to \infty$.

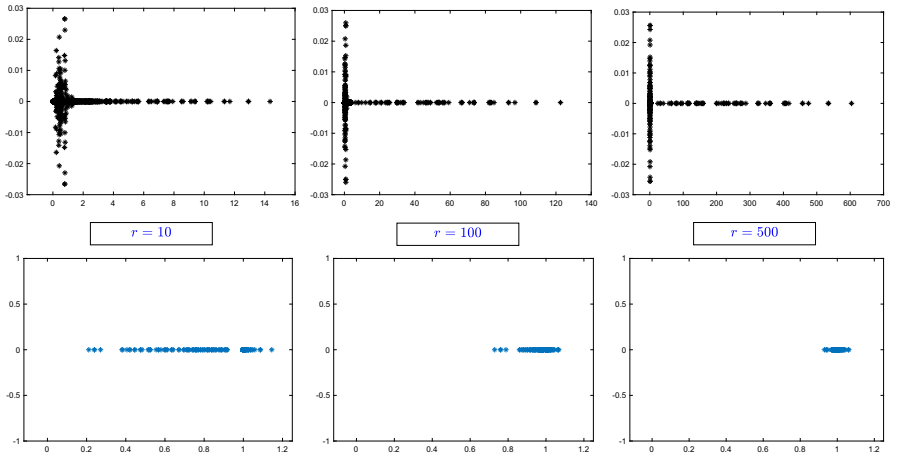

Fig. 1 Eigenvalue distributions of $\bar{\mathcal{A}}$ (top) versus that of the preconditioned matrix $\mathcal{P}_r^{-1}\bar{\mathcal{A}}$ (bottom) for different values of r, with $Q = \text{diag}(M_p)$ for a 3D coupled Stokes–Darcy problem with 1695 degrees of freedom

The eigenvalue distribution in Theorem 2.7 and Remark 2.8 is illustrated in Fig. 1. A full description of the test problem and of its finite element discretization can be found in Sect. 4.

3 Field-of-values analysis

In the previous section, we established eigenvalue bounds for the preconditioned matrices. A clustered spectrum (away from 0) often results in rapid convergence, particularly when the preconditioned matrix is close to normal; see [6] for more details. However, the situation is more complicated when the problem is far from normal. Indeed, the eigenvalues may not describe the convergence of nonsymmetric matrix iterations like GMRES; see [19]. The notions of norm equivalence and of field-of-values equivalence often provide the theoretical framework needed to establish optimality of a class of preconditioners for Krylov methods like GMRES [2,4,9,15,16,22,23]; see also [7] for a recent overview.

Here we derive FOV-type bounds for the preconditioned matrix associated with preconditioner $\mathcal{P}_r$. For the constraint preconditioners $\mathcal{P}_{con_D}$ and $\mathcal{P}_{con_T}$, this equivalence has been established by Chidyagwai et al. [13].

3.1 Basic concepts

In this subsection we briefly overview the required background for establishing FOV-type bounds, see [13,23] for more details.

We begin by reviewing the notion of spectral equivalence for families of SPD matrices [3]. Recall that two families of SPD matrices $\{A_n\}$ and $\{B_n\}$ (parametrized by their dimension n) are said to be *spectrally equivalent* if there exist n-independent

constants α and β with

$$0 < \alpha \leq \lambda_i(B_n^{-1}A_n) \leq \beta, \quad \forall i.$$

Equivalently, $\{A_n\}$ and $\{B_n\}$ are spectrally equivalent if the spectral condition number $\kappa(B_n^{-1}A_n)$ is uniformly bounded with respect to n. Yet another equivalent condition is that the *generalized Rayleigh quotients* associated with A_n and B_n are uniformly bounded:

$$0 < \alpha \leq \frac{\langle A_n x, x\rangle}{\langle B_n x, x\rangle} \leq \beta, \quad \forall x \neq 0.$$

Note that this is an equivalence relation between families of matrices.

Next, we recall the concepts of H-norm-equivalence and H-field-of-value-equivalence, where H corresponds to a given SPD matrix. For simplicity, in the following we drop the subscript n but it should always be kept in mind that matrices representing discretizations always depend on the dimension n (which in turn depends on the mesh size h). Similarly, with a slight abuse of language we will talk of equivalence of matrices rather than of families of matrices.

Definition 3.1 Two nonsingular matrices $M, N \in \mathbb{R}^{n\times n}$ are H-norm-equivalent, $M \sim_H N$, if there exist positive constants α_0 and β_0 independent of n such that the following holds for all nonzero $x \in \mathbb{R}^n$:

$$\alpha_0 \leq \frac{\|Mx\|_H}{\|Nx\|_H} \leq \beta_0.$$

Equivalently, $M \sim_H N$ is equivalent to

$$\left\|MN^{-1}\right\|_H \leq \beta_0,$$
$$\left\|NM^{-1}\right\|_H \leq \alpha_0^{-1}.$$

Definition 3.2 Let H be an $n \times n$ symmetric positive definite matrix and let $A \in \mathbb{R}^{n\times n}$. The H-field of values of A is the set

$$\mathcal{F}_H(A) := \{z \in \mathbb{C} \,|\, z = \langle Ax, x\rangle_H,\ \|x\|_H = 1\}.$$

Definition 3.3 Two nonsingular matrices $M, N \in \mathbb{R}^{n\times n}$ are H-field-of-values-equivalent, $M \approx_H N$, if there exist positive constants α_0 and β_0 independent of n such that the following holds for all nonzero $x \in \mathbb{R}^n$:

$$\alpha_0 \leq \frac{\left\langle MN^{-1}x, x\right\rangle_H}{\langle x, x\rangle_H} \quad \text{and} \quad \left\|MN^{-1}\right\|_H \leq \beta_0. \tag{24}$$

This definition implies that if $M \approx_H N$, the H-field of values of MN^{-1} lies in the right half-plane and is both bounded and bounded away from 0 uniformly in n.

Remark 3.4 If M and N are symmetric positive definite and $H = I_n$, then $M \approx_H N$ reduces to spectral equivalence.

Proposition 3.5 *Let M and N be two symmetric nonsingular matrices such that $M \approx_H N$ where H is a given symmetric positive definite matrix. Then $M^{-1} \approx_{H^{-1}} N^{-1}$.*

Proof The result follows from some algebraic computations. Note that M and N are both symmetric and relations (24) hold for some constant α_0 and β_0.

Let v be an arbitrary nonzero vector. Setting $y = H^{-1}v$, we have

$$\frac{\langle M^{-1}Nv, v\rangle_{H^{-1}}}{\langle v, v\rangle_{H^{-1}}} = \frac{\langle M^{-1}NHy, y\rangle}{\langle Hy, y\rangle} = \frac{\langle y, NM^{-1}y\rangle_H}{\langle y, y\rangle_H}.$$

Now setting $y = MN^{-1}z$ in the above relation, we find that

$$\frac{\langle M^{-1}Nv, v\rangle_{H^{-1}}}{\langle v, v\rangle_{H^{-1}}} = \frac{\langle MN^{-1}z, z\rangle_H}{\|MN^{-1}z\|_H^2} = \frac{\langle MN^{-1}z, z\rangle_H / \|z\|_H^2}{\|MN^{-1}z\|_H^2 / \|z\|_H^2} \geq \frac{\alpha_0}{\beta_0^2}.$$

To complete the proof, we need to show that there exists $\hat{\beta}_0$ such that

$$\left\|M^{-1}N\right\|_{H^{-1}} \leq \hat{\beta}_0.$$

For an arbitrary nonzero vector v, we have that

$$\frac{\|v\|_{H^{-1}}}{\|M^{-1}Nv\|_{H^{-1}}} = \frac{\|v\|_{H^{-1}}\|M^{-1}Nv\|_{H^{-1}}}{\|M^{-1}Nv\|_{H^{-1}}^2} \geq \frac{\langle M^{-1}Nv, v\rangle_{H^{-1}}}{\langle M^{-1}Nv, M^{-1}Nv\rangle_{H^{-1}}}.$$

Consequently, setting $v = N^{-1}My$, we derive

$$\frac{\|v\|_{H^{-1}}}{\|M^{-1}Nv\|_{H^{-1}}} \geq \frac{\langle MN^{-1}H^{-1}y, y\rangle}{\langle H^{-1}y, y\rangle}.$$

Now, we set $y = Hx$ which together with the above relation implies that

$$\frac{\|v\|_{H^{-1}}}{\|M^{-1}Nv\|_{H^{-1}}} \geq \frac{\langle MN^{-1}x, x\rangle_H}{\langle x, x\rangle_H} \geq \alpha_0.$$

Defining $\hat{\beta}_0 = \alpha_0^{-1}$, we obtain

$$\frac{\left\|M^{-1}Nv\right\|_{H^{-1}}}{\|v\|_{H^{-1}}} \leq \hat{\beta}_0,$$

which completes the proof. □

Remark 3.6 With a similar argument used in the proof of the above proposition, one can verify that $M \approx_H N$ implies $N \approx_H M$ for any symmetric nonsingular matrices M, N and symmetric positive definite matrix H with appropriate dimensions.

We also recall the following useful definitions and properties.

Definition 3.7 Let $M \in \mathbb{R}^{m\times n}$ and let $H_1 \in \mathbb{R}^{n\times n}$, $H_2 \in \mathbb{R}^{m\times m}$ be two symmetric positive definite matrices, then

$$\|M\|_{H_1,H_2} = \max_{v\in\mathbb{R}^n\setminus\{0\}} \frac{\|Mv\|_{H_2}}{\|v\|_{H_1}}.$$

Note that when $H_1 = H_2 = H$ and $m = n$, the above definition reduces to the following standard matrix norm,

$$\|M\|_{H,H} = \|M\|_H.$$

Also, it can be seen that

$$\|H_2^{-1/2} M H_1^{-1/2}\|_2 = \|M\|_{H_1,H_2^{-1}} = \left\|MH_1^{-1}\right\|_{H_1^{-1},H_2^{-1}} = \left\|H_2^{-1}M\right\|_{H_1,H_2}.$$

It can be shown that when $H_1 = H_2 = H$, from the above relation, we have

$$\left\|H^{-1/2}MH^{-1/2}\right\|_2 = \left\|MH^{-1}\right\|_{H^{-1}} = \left\|H^{-1}M\right\|_H.$$

We further observe that

$$\|MN\|_{H_3,H_1} \leq \|N\|_{H_3,H_2}\|M\|_{H_2,H_1},$$

where H_3 is a given arbitrary SPD matrix of the appropriate size.

Henceforth we assume that the matrix $\mathcal{A} \in \mathbb{R}^{n\times n}$ satisfies the following stability conditions [12,13]:

$$\max_{w\in\mathbb{R}^n\setminus\{0\}} \max_{v\in\mathbb{R}^n\setminus\{0\}} \frac{w^T\mathcal{A}v}{\|w\|_H\|v\|_H} \leq c_1, \tag{25a}$$

$$\min_{w\in\mathbb{R}^n\setminus\{0\}} \max_{v\in\mathbb{R}^n\setminus\{0\}} \frac{w^T\mathcal{A}v}{\|w\|_H\|v\|_H} \geq c_2, \tag{25b}$$

where c_1 and c_2 are positive constants independent of n, and the matrix H is SPD. For the coupled Stokes–Darcy problem H is a block diagonal matrix with diagonal blocks $H_1 \in \mathbb{R}^{n_1 \times n_1}$ and $H_2 \in \mathbb{R}^{n_2 \times n_2}$ (with $n_1 + n_2 = n$) given by

$$H_1 = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}, \quad H_2 = M_p, \tag{26}$$

where M_p denotes the mass matrix for the Stokes pressure space; see [13] for more details.

Similar to [13], we also need the following results from [23] which can be obtained making use of the stability conditions (25).

Lemma 3.8 *Let* (25) *hold, then* $H \sim_{H^{-1}} \mathcal{A}$ *and* $H^{-1} \sim_H \mathcal{A}^{-1}$, *and in particular*

$$\|H^{-1}\mathcal{A}\|_H = \|\mathcal{A}H^{-1}\|_{H^{-1}} \leq c_1,$$
$$\|\mathcal{A}^{-1}H\|_H = \|H\mathcal{A}^{-1}\|_{H^{-1}} \leq c_2^{-1}.$$

Lemma 3.9 *Let* (25) *hold and assume that* $\mathcal{P} \sim_{H^{-1}} H$, *then*

$$\mathcal{P} \sim_{H^{-1}} \mathcal{A} \quad \text{and} \quad \mathcal{P}^{-1} \sim_H \mathcal{A}^{-1}.$$

Lemma 3.10 *Let* (25) *hold, then* $\|A\|_{H_1, H_1^{-1}} \leq c_1$, $\|C\|_{H_1, H_2^{-1}} \leq c_1$, *where* $C = [0\ B]$ *and*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}.$$

Lemma 3.11 *Let* (25) *hold. If there exists* c_3 *independent of* n_1 *such that*

$$\min_{w \in \mathbb{R}^{n_1} \setminus \{0\}} \max_{v \in \mathbb{R}^{n_1} \setminus \{0\}} \frac{w^T A v}{\|w\|_{H_1} \|v\|_{H_1}} \geq c_3, \tag{27}$$

then $S = CA^{-1}C^T$ *satisfies* $S \sim_{H_2^{-1}} H_2$ *and* $H_2^{-1} \sim_{H_2} S^{-1}$, *where* C *and* A *are defined as in Lemma* 3.10. *Hence, there exists* c_4 *independent of* n_1 *such that* $\|S^{-1}\|_{H_2^{-1}, H_2} \leq c_4$.

Lemma 3.12 [13, Lemma 3.8] $\|M\|_{H_1, H_2^{-1}} = \|M^T\|_{H_2, H_1^{-1}}$

3.2 FOV-type bounds

The following proposition is established in [17] for symmetric matrices. In [18, Poposition 2.1], it is pointed out that the result remains true for nonsymmetric matrices as well.

Proposition 3.13 *Suppose that A is a general $n \times n$ matrix, C is full row rank $p \times n$ matrix ($p \leq n$), and W is a $p \times p$ matrix. Define,*

$$\mathcal{K}(W) = \begin{bmatrix} A + C^T W C & C^T \\ C & 0 \end{bmatrix}. \tag{28}$$

If $\mathcal{K} := \mathcal{K}(0)$ is nonsingular, then $\mathcal{K}(W)$ is a nonsingular matrix for any nonzero W and

$$\mathcal{K}^{-1}(W) = \mathcal{K}^{-1} - \begin{bmatrix} 0 & 0 \\ 0 & W \end{bmatrix}.$$

Let us consider the following partitioning for $\bar{\mathcal{A}}$ and write the matrix in the form of (28),

$$\bar{\mathcal{A}} = \left[\begin{array}{cc|c} A_{11} & A_{12} & 0 \\ A_{21} & A_{22} + rB^T Q^{-1} B & B^T \\ \hline 0 & B & 0 \end{array}\right] := \left[\begin{array}{c|c} A + rC^T Q^{-1} C & C^T \\ \hline C & 0 \end{array}\right].$$

As a result, from Proposition 3.13, we have

$$\bar{\mathcal{A}}^{-1} = \mathcal{A}^{-1} - \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & rQ^{-1} \end{bmatrix}. \tag{29}$$

The above discussion allows one to find sufficient conditions which ensure that stability conditions similar to (25) hold for $\bar{\mathcal{A}}$. In particular, it turns out that stability conditions for $\bar{\mathcal{A}}$ can be deduced from (25) by suitable choices of Q including the case that $Q = M_p$. To this end, we recall the following observation, which is a consequence of [23, Lemma 2.1] and Lemma 3.12.

Remark 3.14 Let the matrices H and $\mathcal{A}$ be defined as before. Then

$$\|\mathcal{A}\|_{H,H^{-1}} = \max_{w \in \mathbb{R}^n \setminus \{0\}} \max_{v \in \mathbb{R}^n \setminus \{0\}} \frac{w^T \mathcal{A} v}{\|w\|_H \|v\|_H}, \tag{30}$$

$$\left\|\mathcal{A}^{-1}\right\|^{-1}_{H^{-1},H} = \min_{w \in \mathbb{R}^n \setminus \{0\}} \max_{v \in \mathbb{R}^n \setminus \{0\}} \frac{w^T \mathcal{A} v}{\|w\|_H \|v\|_H}. \tag{31}$$

In view of the above remark, we need to establish that $\|\bar{\mathcal{A}}\|_{H,H^{-1}}$ and $\|\bar{\mathcal{A}}^{-1}\|_{H^{-1},H}$ are bounded from above in order to show that $\bar{\mathcal{A}}$ satisfies stability conditions similar to (25). Notice that (25) together with (29) imply that

$$\begin{aligned} \|\bar{\mathcal{A}}^{-1}\|_{H^{-1},H} &\leq \|\mathcal{A}^{-1}\|_{H^{-1},H} + r\|H_2^{1/2} Q^{-1} H_2^{1/2}\|_2 \\ &\leq c_2^{-1} + r\lambda_{\max}(M_p Q^{-1}). \end{aligned} \tag{32}$$

On the other hand, we have that

$$\begin{aligned}\|\bar{\mathcal{A}}\|_{H,H^{-1}} &\le \|\mathcal{A}\|_{H,H^{-1}} + r\|H_1^{-1/2}C^T Q^{-1} C H_1^{-1/2}\|_2 \\ &\le c_1 + r\|H_1^{-1/2}C^T H_2^{-1/2} H_2^{1/2} Q^{-1} H_2^{1/2} H_2^{-1/2} C H_1^{-1/2}\|_2 \\ &\le c_1 + r\|H_1^{-1/2}C^T H_2^{-1/2}\|_2 \|H_2^{1/2} Q^{-1} H_2^{1/2}\|_2 \|H_2^{-1/2} C H_1^{-1/2}\|_2 \\ &= c_1 + r\|C^T\|_{H_2,H_1^{-1}} \|Q^{-1}\|_{H_2^{-1},H_2} \|C\|_{H_1,H_2^{-1}}.\end{aligned}$$

It is known that $\|C^T\|_{H_2,H_1^{-1}} = \|C\|_{H_1,H_2^{-1}}$. Therefore, by Lemma 3.10, the following inequality holds when the first condition in (25) is satisfied:

$$\begin{aligned}\|\bar{\mathcal{A}}\|_{H,H^{-1}} &\le c_1 + r c_1^2 \|Q^{-1}\|_{H_2^{-1},H_2} \\ &= c_1 + r c_1^2 \lambda_{\max}(M_p Q^{-1}).\end{aligned} \tag{33}$$

From the above discussions, it can be observed that if we set $Q = M_p$ then Eqs. (32) and (33) reduce to the following inequalities, respectively:

$$\|\bar{\mathcal{A}}^{-1}\|_{H^{-1},H} \le c_2^{-1} + r,$$

and

$$\|\bar{\mathcal{A}}\|_{H,H^{-1}} \le c_1 + r c_1^2$$

In order to deal with the augmented system, the following lemma provides a useful expression for the Schur complement. The lemma is an immediate consequence of Proposition 3.13, see [10, Lemma 4.1] for more details.

Lemma 3.15 *Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{m\times n}$ $(m \le n)$. Let $\gamma \in \mathbb{R}$, and suppose that A, $A + \gamma B^T W^{-1} B$, $BA^{-1}B^T$, and $B(A + \gamma B^T W^{-1} B)^{-1} B^T$ are all nonsingular matrices. Then*

$$[B(A + \gamma B^T W^{-1} B)^{-1} B^T]^{-1} = (BA^{-1}B^T)^{-1} + \gamma W^{-1}. \tag{34}$$

We denote the negative Schur complement associated with $\bar{\mathcal{A}}$ by $\bar{S}$. By the above lemma, it is immediate to see that

$$\bar{S}^{-1} = [C(A + rC^T Q^{-1} C)^{-1} C^T]^{-1} = S^{-1} + rQ^{-1}.$$

Hence, we have

$$\begin{aligned}\left\|\bar{S}^{-1}\right\|_{H_2^{-1},H_2} &\le \left\|S^{-1}\right\|_{H_2^{-1},H_2} + r\left\|Q^{-1}\right\|_{H_2^{-1},H_2} \\ &= \left\|S^{-1}\right\|_{H_2^{-1},H_2} + r\lambda_{\max}(M_p Q^{-1}).\end{aligned}$$

Assuming that the assumption of Lemma 3.11 holds, we readily obtain

$$\left\| \bar{S}^{-1} \right\|_{H_2^{-1}, H_2} \leq c_4 + r\lambda_{\max}(M_p Q^{-1}). \tag{35}$$

For simplicity we set $\bar{A} = A + rC^T Q^{-1} C$, $S_A = A_{22} - A_{21}A_{11}^{-1}A_{12}$, and $S_{\bar{A}} = S_A + rB^T Q^{-1} B$. It is well known (see, e.g., [8, page 18]) that

$$\bar{A}^{-1} = \begin{bmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}S_{\bar{A}}^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}S_{\bar{A}}^{-1} \\ -S_{\bar{A}}^{-1}A_{21}A_{11}^{-1} & S_{\bar{A}}^{-1} \end{bmatrix},$$

in which $A_{21} = -A_{12}^T$. Straightforward computations show that

$$S_{\bar{A}} = S_A^{1/2}\left(I + rS_A^{-1/2}B^T Q^{-1} B S_A^{-1/2}\right) S_A^{1/2}.$$

Hence, considering the singularity of $S_A^{-1/2}B^T Q^{-1} B S_A^{-1/2}$, one can conclude that

$$\begin{aligned}
\left\| S_{\bar{A}}^{-1} \right\|_{A_{22}^{-1}, A_{22}} &= \left\| A_{22}^{1/2} S_{\bar{A}}^{-1} A_{22}^{1/2} \right\|_2 \\
&= \lambda_{\max}\left(A_{22}^{1/2} S_A^{-1/2}(I + rS_A^{-1/2}B^T Q^{-1} B S_A^{-1/2})^{-1} S_A^{-1/2} A_{22}^{1/2}\right) \\
&= \lambda_{\max}\left(S_A^{-1/2} A_{22} S_A^{-1/2}(I + rS_A^{-1/2}B^T Q^{-1} B S_A^{-1/2})^{-1}\right) \\
&\leq \lambda_{\max}\left(S_A^{-1/2} A_{22} S_A^{-1/2}\right) \lambda_{\max}\left((I + rS_A^{-1/2}B^T Q^{-1} B S_A^{-1/2})^{-1}\right) \\
&= \lambda_{\max}\left(A_{22}^{1/2} S_A^{-1} A_{22}^{1/2}\right) / \left(1 + r\lambda_{\min}(S_A^{-1/2}B^T Q^{-1} B S_A^{-1/2})\right) \\
&= \left\| S_A^{-1} \right\|_{A_{22}^{-1}, A_{22}}.
\end{aligned} \tag{36}$$

Let H_1 be defined by (26). Under the assumptions of Lemma 3.11, we now show that

$$\min_{w \in \mathbb{R}^n \setminus \{0\}} \max_{v \in \mathbb{R}^n \setminus \{0\}} \frac{w^T \bar{A} v}{\|w\|_{H_1} \|v\|_{H_1}} \geq \bar{c}_3$$

for some $\bar{c}_3$. The assumption (27) is equivalent to $\left\| A^{-1} \right\|_{H_1^{-1}, H_1} \leq c_3^{-1}$; see [23, Lemma 2.1]. In order to show that $\left\| \bar{A}^{-1} \right\|_{H_1^{-1}, H_1}$ is bounded from above by a constant, we need to show that the norm of each four blocks of $H_1^{1/2} \bar{A}^{-1} H_1^{1/2}$ is bounded from above. To this end, we first note that

$$\begin{aligned}\left\|S_A^{-1}\right\|_{A_{22}^{-1},A_{22}} &= \left\|A_{22}^{1/2}S_A^{-1}A_{22}^{1/2}\right\|_2\\ &= \lambda_{\max}\left(A_{22}^{1/2}S_A^{-1}A_{22}^{1/2}\right)\\ &= \lambda_{\max}\left(\left(I + A_{22}^{-1/2}A_{12}^T A_{11}^{-1}A_{12}A_{22}^{-1/2}\right)^{-1}\right) < 1,\end{aligned}$$

hence $\left\|S_{\bar{A}}^{-1}\right\|_{A_{22}^{-1},A_{22}} < 1$ from (36). By Lemma 3.10 we have $\|A\|_{H_1,H_1^{-1}} \le c_1$. On the other hand, we have that $\|A\|^2_{H_1,H_1^{-1}} = 1 + \|A_{21}\|^2_{A_{11},A_{22}^{-1}}$ as $A_{21} = -A_{12}^T$ and H_1 is a block diagonal matrix with blocks A_{11} and A_{22}. This ensures that $\|A_{21}\|_{A_{11},A_{22}^{-1}} = \|A_{12}\|_{A_{22},A_{11}^{-1}} \le c_1$. For the (1, 1) block of $\bar{A}^{-1}$, we get

$$\begin{aligned}&\left\|A_{11}^{-1} + A_{11}^{-1}A_{12}S_{\bar{A}}^{-1}A_{21}A_{11}^{-1}\right\|_{A_{11}^{-1},A_{11}}\\ &= \left\|I + A_{11}^{-1/2}A_{12}S_{\bar{A}}^{-1}A_{21}A_{11}^{-1/2}\right\|_2\\ &\le 1 + \left\|A_{11}^{-1/2}A_{12}A_{22}^{-1/2}A_{22}^{1/2}S_{\bar{A}}^{-1}A_{22}^{1/2}A_{22}^{-1/2}A_{21}A_{11}^{-1/2}\right\|_2\\ &\le 1 + \|A_{12}\|_{A_{22},A_{11}^{-1}}\left\|S_{\bar{A}}^{-1}\right\|_{A_{22}^{-1},A_{22}}\|A_{21}\|_{A_{11},A_{22}^{-1}}.\end{aligned}$$

Now, Eq. (36) implies that

$$\begin{aligned}\left\|A_{11}^{-1} + A_{11}^{-1}A_{12}S_{\bar{A}}^{-1}A_{21}A_{11}^{-1}\right\|_{A_{11}^{-1},A_{11}} &\le 1 + \|A_{12}\|_{A_{22},A_{11}^{-1}}\left\|S_A^{-1}\right\|_{A_{22}^{-1},A_{22}}\|A_{21}\|_{A_{11},A_{22}^{-1}}\\ &\le 1 + c_1^2. \qquad (37)\end{aligned}$$

It is not difficult to verify that

$$\left\|A_{11}^{-1}A_{12}S_{\bar{A}}^{-1}\right\|_{A_{22}^{-1},A_{11}} = \left\|S_{\bar{A}}^{-1}A_{21}A_{11}^{-1}\right\|_{A_{11}^{-1},A_{22}} \le c_1.$$

Therefore, the above relation together with the fact that $\left\|S_{\bar{A}}^{-1}\right\|_{A_{22}^{-1},A_{22}} < 1$ and inequality (37) show that $\left\|\bar{A}^{-1}\right\|_{H_1^{-1},H_1} \le \bar{c}_3^{-1}$ for some constant $\bar{c}_3$, i.e., $\bar{c}_3^{-1} = 1+(1+c_1)^2$. Let us assume that there exists a constant $\bar{\eta}$ such that

$$r\lambda_{\max}(M_p Q^{-1}) \le \bar{\eta}. \qquad (38)$$

Then, by virtue of the above observation, Eqs. (32), (33) and (35), we can find constants $\bar{c}_1$, $\bar{c}_2$ and $\bar{c}_4$ such that

$$\max_{w\in\mathbb{R}^n\setminus\{0\}}\max_{v\in\mathbb{R}^n\setminus\{0\}}\frac{w^T\bar{\mathcal{A}}v}{\|w\|_H\|v\|_H} \le \bar{c}_1, \qquad (39a)$$

$$\min_{w\in\mathbb{R}^n\setminus\{0\}}\max_{v\in\mathbb{R}^n\setminus\{0\}}\frac{w^T\bar{\mathcal{A}}v}{\|w\|_H\|v\|_H}\geq\bar{c}_2, \tag{39b}$$

and $\|\bar{S}^{-1}\|_{H_2^{-1},H_2}\leq\bar{c}_4$ provided the stability conditions (25) hold for $\mathcal{A}$. As pointed earlier, if $Q=M_p$ then $\lambda_{\max}(M_pQ^{-1})=1$ and we get $\bar{\eta}=r$. In practice, the matrix M_p can be efficiently approximated by its main diagonal. Therefore, we suggest choosing Q as the main diagonal of M_p in numerical experiments. In the sequel, we assume that Q is chosen such that there exits an $\bar{\eta}$ for which (38) holds. Consequently, similar to Lemma 3.11, the following lemma can be stated.

Lemma 3.16 *Let the assumptions of Lemma 3.11 hold and suppose there exists $\bar{\eta}>0$ such that (38) is satisfied independent of n. Then $\bar{S}\sim_{H_2^{-1}}H_2$ and $H_2^{-1}\sim_{H_2}\bar{S}^{-1}$. Hence, there exists $\bar{c}_4$ independent of n such that $\|\bar{S}^{-1}\|_{H_2^{-1},H_2}\leq\bar{c}_4$, where $\bar{S}=C\bar{A}^{-1}C^T$.*

Consider again the matrix $\bar{\mathcal{A}}$ in the following form:

$$\bar{\mathcal{A}}=\begin{bmatrix}\bar{A} & C^T\\ C & 0\end{bmatrix}.$$

We note that for $r=0$, $\bar{A}=A$ and $\bar{\mathcal{A}}$ reduces to

$$\mathcal{A}=\begin{bmatrix}A & C^T\\ C & 0\end{bmatrix}.$$

The constraint preconditioners can be written in the following form

$$\mathcal{P}_{con}=\begin{bmatrix}P_{con} & C^T\\ C & 0\end{bmatrix},$$

and it is shown that $\mathcal{P}_{con}$ is H-norm equivalent (and consequently H-f.o.v equivalent) to the operator $\mathcal{A}$ where

$$H=\begin{bmatrix}H_1 & 0\\ 0 & H_2\end{bmatrix}, \tag{40}$$

with H_1 and H_2 being symmetric positive definite given by (26) under suitable conditions, see [13].

Now consider the following preconditioner:

$$\mathcal{P}_r=\begin{bmatrix}A_{11} & A_{12} & 0\\ 0 & A_{22}+rB^TQ^{-1}B & B^T\\ 0 & 0 & -\frac{1}{r}Q\end{bmatrix},$$

where Q is symmetric positive definite and $r>0$ is given. Here, we comment that Q can be taken to be any SPD matrix such that (38) holds for a constant $\bar{\eta}>0$.

For simplicity, we rewrite $\mathcal{P}_r$ as follows:

$$\mathcal{P}_r = \begin{bmatrix} \bar{P} & C^T \\ 0 & -\frac{1}{r}Q \end{bmatrix} \tag{41}$$

with the obvious definition of $\bar{P}$. Now we establish a proposition which will be useful in proving norm-equivalence of the preconditioner $\mathcal{P}_r$ to $\bar{\mathcal{A}}$. To this end, we first need to recall the following well known fact.

Theorem 3.17 [20, Theorem 7.7.3] *Let A and B be two $n \times n$ real symmetric matrices such that A is positive definite and B is positive semidefinite. Then $A \succcurlyeq B$ if and only if $\rho(A^{-1}B) \leq 1$, and $A \succ B$ if and only if $\rho(A^{-1}B) < 1$.*

Proposition 3.18 *Under the assumptions of Lemma 3.16, if*

$$0 < r \leq \frac{1}{\rho(Q^{-1}\bar{S})}, \tag{42}$$

then there there exists $\bar{c}_4$ independent of n such that $\|rQ^{-1}\|_{H_2^{-1},H_2} \leq \bar{c}_4$ where $\bar{S} = C\bar{A}^{-1}C^T$.

Proof First note that Theorem 3.17 implies

$$\bar{S} - \frac{1}{r}Q \preccurlyeq 0.$$

Since H_2 is a symmetric positive definite matrix, we have $H_2^{-1/2}(\bar{S} - \frac{1}{r}Q)H_2^{-1/2} \preccurlyeq 0$. As a result, for any nonzero unit vector x, we have

$$\left\langle H_2^{-1/2}\bar{S}H_2^{-1/2}x, x\right\rangle \leq \left\langle r^{-1}H_2^{-1/2}QH_2^{-1/2}x, x\right\rangle,$$

which is equivalent to

$$\begin{aligned} \left\langle r^{-1}H_2^{-1/2}QH_2^{-1/2}x, x\right\rangle^{-1} &\leq \left\langle H_2^{-1/2}\bar{S}H_2^{-1/2}x, x\right\rangle^{-1} \\ &\leq \left(\min_{\|x\|_2=1}\left\langle H_2^{-1/2}\bar{S}H_2^{-1/2}x, x\right\rangle\right)^{-1} \\ &= \left\|H_2^{1/2}\bar{S}^{-1}H_2^{1/2}\right\|_2 \\ &= \left\|\bar{S}^{-1}\right\|_{H_2^{-1},H_2}. \end{aligned}$$

Lemma 3.16 ensures that there exists $\bar{c}_4$ such that $\|\bar{S}^{-1}\|_{H_2^{-1},H_2} \leq \bar{c}_4$. Hence, from the preceding relation, we deduce that

$$\left\langle r^{-1}H_2^{-1/2}QH_2^{-1/2}x, x\right\rangle^{-1} \leq \bar{c}_4,$$

for any nonzero vector x with $\|x\|_2 = 1$. □

Proposition 3.19 *Let the stability conditions* (25) *and the assumptions of Proposition* 3.18 *hold. Suppose that H_1 and $\bar{P}$ are defined as in* (26) *and* (41)*, respectively. If $P \sim_{H_1^{-1}} H_1$ then $\bar{P} \sim_{H_1^{-1}} H_1$, where P is obtained from $\bar{P}$ by setting $r = 0$.*

Proof By the assumptions, there exist α_0 and β_0 such that

$$\left\| P H_1^{-1} \right\|_{H_1^{-1}} \leq \beta_0,$$
$$\left\| H_1 P^{-1} \right\|_{H_1^{-1}} \leq \alpha_0^{-1}.$$

It is obvious that

$$\bar{P} = P + \begin{bmatrix} 0 & 0 \\ 0 & r B^T Q^{-1} B \end{bmatrix}.$$

Consequently, we have

$$\begin{aligned} \left\| \bar{P} H_1^{-1} \right\|_{H_1^{-1}} &= \left\| H_1^{-1/2} \bar{P} H_1^{-1/2} \right\|_2 \\ &\leq \left\| H_1^{-1/2} P H_1^{-1/2} \right\|_2 + r \left\| A_{22}^{-1/2} B^T Q^{-1} B A_{22}^{-1/2} \right\|_2 \\ &\leq \left\| P H_1^{-1} \right\|_{H_1^{-1}} + \left\| A_{22}^{-1/2} B^T H_2^{-1/2} \right\|_2 \left\| r H_2^{1/2} Q^{-1} H_2^{1/2} \right\|_2 \left\| H_2^{-1/2} B A_{22}^{-1/2} \right\|_2 \\ &\leq \beta_0 + \left\| C^T \right\|_{H_2, H_1^{-1}} \left\| r Q^{-1} \right\|_{H_2^{-1}, H_2} \|C\|_{H_1, H_2^{-1}}. \end{aligned} \tag{43}$$

Now from Lemma 3.10 and Proposition 3.18, we find that for $\bar{\beta}_0 = \beta_0 + c_1^2 \bar{c}_4$, we have

$$\left\| \bar{P} H_1^{-1} \right\|_{H_1^{-1}} \leq \bar{\beta}_0.$$

Note that

$$\left\| H_1 \bar{P}^{-1} \right\|_{H_1^{-1}} = \left\| H_1^{1/2} \bar{P}^{-1} H_1^{1/2} \right\|_2,$$

and

$$\begin{aligned} H_1^{1/2} \bar{P}^{-1} H_1^{1/2} &= \begin{bmatrix} I & -A_{11}^{-1/2} A_{12} \bar{A}_{22}^{-1} A_{22}^{1/2} \\ 0 & A_{22}^{1/2} \bar{A}_{22}^{-1} A_{22}^{1/2} \end{bmatrix} \\ &= \begin{bmatrix} I & -A_{11}^{-1/2} A_{12} A_{22}^{-1/2} A_{22}^{1/2} \bar{A}_{22}^{-1} A_{22}^{1/2} \\ 0 & A_{22}^{1/2} \bar{A}_{22}^{-1} A_{22}^{1/2} \end{bmatrix}, \end{aligned}$$

where $\bar{A}_{22} = A_{22} + r B^T Q^{-1} B$. Since $\left\| A_{22}^{1/2} \bar{A}_{22}^{-1} A_{22}^{1/2} \right\|_2 = 1$, we have that

$$\left\| H_1^{1/2} \bar{P}^{-1} H_1^{1/2} \right\|_2 \leq 2 + \|A_{21}\|_{A_{11},A_{22}^{-1}} \|A_{12}\|_{A_{22},A_{11}^{-1}}.$$

Recalling that $\|A_{21}\|_{A_{11},A_{22}^{-1}} = \|A_{12}\|_{A_{22},A_{11}^{-1}} \leq c_1$, from the preceding inequality we deduce that $\left\| H_1 \bar{P}^{-1} \right\|_{H_1^{-1}} \leq \bar{\alpha}_0^{-1}$ for $\bar{\alpha}_0^{-1} = 2 + c_1^2$, which completes the proof. □

The proof of the next theorem follows from a similar argument used in [13, Theorem 3.9], where $P_{con} \sim_{H_1^{-1}} H_1$ was an assumption. In view of the previous proposition, the assumption $\bar{P} \sim_{H_1^{-1}} H_1$ in the theorem is a consequence of the fact that $P \sim_{H_1^{-1}} H_1$, where P is the block upper triangular part of A. On the other hand, the matrix P_{con} in the constraint preconditioners $\mathcal{P}_{con_D}$ and $\mathcal{P}_{con_T}$ is, respectively, the block diagonal and block lower triangular part of A. However, considering Eq. (26), one can see that if P_{con} and P are the block lower or the block upper triangular part of A, then $P_{con} \sim_{H_1^{-1}} H_1$ and $P \sim_{H_1^{-1}} H_1$ can be deduced from $\|A_{21}\|_{A_{11},A_{22}^{-1}} = \|A_{12}\|_{A_{22},A_{11}^{-1}} \leq c_1$. More precisely, it turns out that

$$\|P_{con} H_1^{-1}\|_{H_1^{-1}} = \|H_1 P_{con}^{-1}\|_{H_1^{-1}} \leq 2 + \|A_{21}\|_{A_{11},A_{22}^{-1}}$$

and

$$\|P H_1^{-1}\|_{H_1^{-1}} = \|H_1 P^{-1}\|_{H_1^{-1}} \leq 2 + \|A_{12}\|_{A_{22},A_{11}^{-1}}.$$

We comment that if P_{con} is the block diagonal part of A, then

$$\|P_{con} H_1^{-1}\|_{H_1^{-1}} = \|H_1 P_{con}^{-1}\|_{H_1^{-1}} = 1.$$

Therefore, in the analysis of [13, Section 3], there is no need to require that $P_{con} \sim_{H_1^{-1}} H_1$ for establishing FOV bounds (independent of the mesh-width) when the preconditioner is applied "exactly", i.e., when direct methods are used for the block solves.

Theorem 3.20 *Let H and $\mathcal{P}_r$ be defined as in (40) and (41), respectively. In addition to the hypotheses of Proposition 3.18, assume that $r > 0$ is such that*

$$H_2 - \frac{1}{r} Q \succcurlyeq 0. \tag{44}$$

If $\bar{P} \sim_{H_1^{-1}} H_1$, then $\mathcal{P}_r \sim_{H^{-1}} \bar{\mathcal{A}}$ and $\mathcal{P}_r^{-1} \sim_H \bar{\mathcal{A}}^{-1}$.

Proof When the stability conditions (39) hold, considering Lemma 3.9, we only need to show that $\mathcal{P}_r \sim_{H^{-1}} H$. To this end we need to derive upper bounds (independent of n) for $\|\mathcal{P}_r H^{-1}\|_{H^{-1}} = \|H^{-1/2} \mathcal{P}_r H^{-1/2}\|_2$ and $\|H \mathcal{P}_r^{-1}\|_{H^{-1}} = \|H^{1/2} \mathcal{P}_r^{-1} H^{1/2}\|_2$.

The assumption $\bar{P} \sim_{H_1^{-1}} H_1$ implies that there exist positive constants α_1 and β_1 such that $\|\bar{P}H_1^{-1}\|_{H_1^{-1}} \leq \beta_1$ and $\|H_1\bar{P}^{-1}\|_{H_1^{-1}} \leq \alpha_1^{-1}$. Evidently, we have

$$H^{-1/2}\mathcal{P}_r H^{-1/2} = \begin{bmatrix} H_1^{-1/2}\bar{P}H_1^{-1/2} & H_1^{-1/2}C^T H_2^{-1/2} \\ 0 & -\frac{1}{r}H_2^{-1/2}QH_2^{-1/2} \end{bmatrix}.$$

Notice that from (44), we get

$$\frac{1}{r}\|H_2^{-1/2}QH_2^{-1/2}\|_2 \leq 1.$$

By Lemmas 3.10 and 3.12 , we have

$$\begin{aligned} \|H^{-1/2}\mathcal{P}_r H^{-1/2}\|_2 &\leq \|H_1^{-1/2}\bar{P}H_1^{-1/2}\|_2 + \|H_1^{-1/2}C^T H_2^{-1/2}\|_2 + \frac{1}{r}\|H_2^{-1/2}QH_2^{-1/2}\|_2 \\ &\leq \|\bar{P}H_1^{-1}\|_{H_1^{-1}} + \|C^T\|_{H_2,H_1^{-1}} + 1 \\ &\leq \beta_1 + c_1 + 1. \end{aligned}$$

Furthermore, we have

$$\mathcal{P}_r^{-1} = \begin{bmatrix} \bar{P}^{-1} & r\bar{P}^{-1}C^T Q^{-1} \\ 0 & -rQ^{-1} \end{bmatrix}.$$

Consequently, we have

$$H^{1/2}\mathcal{P}_r^{-1}H^{1/2} = \begin{bmatrix} H_1^{1/2}\bar{P}^{-1}H_1^{1/2} & rH_1^{1/2}\bar{P}C^T Q^{-1}H_2^{1/2} \\ 0 & -rH_2^{1/2}Q^{-1}H_2^{1/2} \end{bmatrix}.$$

From the assumption, $\|H_1\bar{P}^{-1}\|_{H_1^{-1}} = \|H_1^{1/2}\bar{P}^{-1}H_1^{1/2}\|_2 \leq \alpha_1^{-1}$. In addition, Proposition 3.18 ensures that $\|rQ^{-1}\|_{H_2^{-1},H_2} = \|rH_2^{1/2}Q^{-1}H_2^{1/2}\|_2 \leq \bar{c}_4$. Observing that

$$rH_1^{1/2}\bar{P}C^T Q^{-1}H_2^{1/2} = H_1^{1/2}\bar{P}H_1^{1/2}H_1^{-1/2}C^T H_2^{-1/2}H_2^{1/2}(rQ^{-1})H_2^{1/2},$$

we obtain

$$\begin{aligned} \left\|rH_1^{1/2}\bar{P}C^T Q^{-1}H_2^{1/2}\right\|_2 &\leq \left\|H_1^{1/2}\bar{P}H_1^{1/2}\right\|_2 \left\|H_1^{-1/2}C^T H_2^{-1/2}\right\|_2 \left\|H_2^{1/2}(rQ^{-1})H_2^{1/2}\right\|_2 \\ &\leq \alpha_1^{-1}\left\|C^T\right\|_{H_2,H_1^{-1}} \left\|rQ^{-1}\right\|_{H_2^{-1},H_2} \\ &\leq \alpha_1^{-1}c_1\bar{c}_4. \end{aligned}$$

Therefore, it is immediate to see that

$$\left\|H^{1/2}\mathcal{P}_r^{-1}H^{1/2}\right\|_2 \leq \alpha_1^{-1} + \alpha_1^{-1}c_1\bar{c}_4 + \bar{c}_4,$$

which completes the proof. □

Remark 3.21 By Theorem 3.17, assumption (44) is equivalent to setting the following lower bound for r:

$$r \geq \lambda_{\max}(H_2^{-1}Q).$$

Noting that

$$\frac{\lambda_{\max}(Q)}{\lambda_{\min}(H_2)} \geq \lambda_{\max}(H_2^{-1}Q),$$

we deduce that the condition (44) holds for $r \geq \lambda_{\max}(Q)/\lambda_{\min}(H_2)$.

We are now in a position to establish the main result of this section.

Theorem 3.22 *Let $\bar{\mathcal{A}}$, $\mathcal{P}_r$ be defined as before. In addition to the hypotheses of Proposition 3.18, suppose that the condition (44) is satisfied and there exists a constant $\nu > 0$ such that for any nonzero $y \in \mathbb{R}^{n_2}$ the following inequality holds:*

$$\nu \leq \frac{\left\langle \tilde{S}_r Q^{-1} y, y \right\rangle_{H_2^{-1}}}{\langle y, y \rangle_{H_2^{-1}}}, \tag{45}$$

where $\tilde{S}_r = C\bar{P}^{-1}C^T = C(P + rC^T Q^{-1}C)^{-1}C^T$. If $r > 1$ and $\bar{A} \approx_{H_1^{-1}} \bar{P}$, then there exists $\rho_0 > 0$ such that $\bar{\mathcal{A}} \approx_{H^{-1}} \mathcal{P}_r$ for all $r \geq \rho_0$ provided

$$\|\bar{A}\bar{P}^{-1} - I\|_{H_1^{-1}} \leq r^{-1}.$$

Proof The assumption $\bar{A} \approx_{H_1^{-1}} \bar{P}$ implies $\bar{A} \sim_{H_1^{-1}} \bar{P}$. On the other hand, we have $\bar{A} \sim_{H_1^{-1}} H_1$ in view of stability conditions (39). As a result, we can deduce that $\bar{P} \sim_{H_1^{-1}} H_1$. Therefore, by the previous theorem, $\|\bar{\mathcal{A}}\mathcal{P}_r^{-1}\|_{H^{-1}}$ is bounded from above.

Let $x = (x_1; x_2)$ be given. To complete the proof, in the sequel, we show that there exists a positive constant τ such that

$$\begin{aligned} x^T H^{-1}\bar{\mathcal{A}}\mathcal{P}_r^{-1}x &\geq \tau x^T H^{-1}x \\ &= \tau(x_1^T H_1^{-1}x_1 + x_2^T H_2^{-1}x_2) \\ &= \tau(\|x_1\|^2_{H_1^{-1}} + \|x_2\|^2_{H_2^{-1}}). \end{aligned}$$

Next, observe that

$$H^{-1}\bar{\mathcal{A}}\mathcal{P}_r^{-1} = \begin{bmatrix} H_1^{-1}\bar{A}\bar{P}^{-1} & rH_1^{-1}(\bar{A}\bar{P}^{-1} - I)C^T Q^{-1} \\ H_2^{-1}C\bar{P}^{-1} & rH_2^{-1}C\bar{P}^{-1}C^T Q^{-1} \end{bmatrix}.$$

From $\bar{A} \approx_{H_1^{-1}} \bar{P}$ we know that there exists a positive constat α_0 such that

$$x_1^T H_1^{-1} \bar{A} \bar{P}^{-1} x_1 \geq \alpha_0 \|x_1\|^2_{H_1^{-1}}. \tag{46}$$

By some simple computations, we have

$$\begin{aligned}
&\left| r x_1^T H_1^{-1} (\bar{A}\bar{P}^{-1} - I) C^T Q^{-1} x_2 \right| \\
&= \left| r x_1^T H_1^{-1/2} H_1^{-1/2} (\bar{A}\bar{P}^{-1} - I) C^T Q^{-1} H_2^{1/2} H_2^{-1/2} x_2 \right| \\
&\leq \left\| r (\bar{A}\bar{P}^{-1} - I) C^T Q^{-1} \right\|_{H_2^{-1}, H_1^{-1}} \|x_1\|_{H_1^{-1}} \|x_2\|_{H_2^{-1}} \\
&\leq \left\| r C^T Q^{-1} \right\|_{H_2^{-1}, H_1^{-1}} \left\| \bar{A}\bar{P}^{-1} - I \right\|_{H_1^{-1}} \|x_1\|_{H_1^{-1}} \|x_2\|_{H_2^{-1}} \\
&\leq \left\| r Q^{-1} \right\|_{H_2^{-1}, H_2} \left\| C^T \right\|_{H_2, H_1^{-1}} \left\| \bar{A}\bar{P}^{-1} - I \right\|_{H_1^{-1}} \|x_1\|_{H_1^{-1}} \|x_2\|_{H_2^{-1}} \\
&\leq \bar{c}_4 c_1 \left\| \bar{A}\bar{P}^{-1} - I \right\|_{H_1^{-1}} \|x_1\|_{H_1^{-1}} \|x_2\|_{H_2^{-1}}.
\end{aligned} \tag{47}$$

By making use of $\left\| \bar{A}\bar{P}^{-1} - I \right\|_{H_1^{-1}} \leq r^{-1}$, it can be observed that

$$\left| r x_1^T H_1^{-1} (\bar{A}\bar{P}^{-1} - I) C^T Q^{-1} x_2 \right| \leq \bar{c}_4 c_1 r^{-1} \|x_1\|_{H_1^{-1}} \|x_2\|_{H_2^{-1}}. \tag{48}$$

The fact that $\bar{P} \sim_{H_1^{-1}} H_1$ ensures the existence of a positive constant α_1 such that

$$\left\| H_1 \bar{P}^{-1} \right\|_{H_1^{-1}} \leq \alpha_1^{-1}.$$

Hence, we obtain

$$\begin{aligned}
\left| x_2^T H_2^{-1} C \bar{P}^{-1} x_1 \right| &= \left| x_2^T H_2^{-1/2} H_2^{-1/2} C H_1^{-1/2} H_1^{1/2} \bar{P}^{-1} H_1^{1/2} H_1^{-1/2} x_1 \right| \\
&\leq \left\| H_2^{-1/2} C H_1^{-1/2} \right\|_2 \left\| H_1^{1/2} \bar{P}^{-1} H_1^{1/2} \right\|_2 \|x_1\|_{H_1^{-1}} \|x_2\|_{H_2^{-1}} \\
&= \|C\|_{H_1, H_2^{-1}} \left\| H_1 \bar{P}^{-1} \right\|_{H_1^{-1}} \|x_1\|_{H_1^{-1}} \|x_2\|_{H_2^{-1}} \\
&\leq c_1 \alpha_1^{-1} \|x_1\|_{H_1^{-1}} \|x_2\|_{H_2^{-1}}.
\end{aligned} \tag{49}$$

Evidently, by the assumption (45), we have

$$x_2^T (r H_2^{-1} C \bar{P}^{-1} C^T Q^{-1}) x_2 = \|x_2\|^2_{H_2^{-1}} \frac{x_2^T (r H_2^{-1} C \bar{P}^{-1} C^T Q^{-1}) x_2}{x_2^T H_2^{-1} x_2} \geq r\nu \|x_2\|^2_{H_2^{-1}}, \tag{50}$$

From Eqs. (46)–(50), we derive the following bound

$$x^T H^{-1}\bar{\mathcal{A}}\mathcal{P}_r^{-1}x \geq \alpha_0 \|x_1\|^2_{H_1^{-1}} - (\bar{c}_4 c_1 r^{-1} + c_1\alpha_1^{-1})\|x_1\|_{H_1^{-1}}\|x_2\|_{H_2^{-1}} + r\nu \|x_2\|^2_{H_2^{-1}}.$$

Since $r > 1$, we have

$$x^T H^{-1}\bar{\mathcal{A}}\mathcal{P}_r^{-1}x \geq \alpha_0 \|x_1\|^2_{H_1^{-1}} - \gamma\|x_1\|_{H_1^{-1}}\|x_2\|_{H_2^{-1}} + r\nu \|x_2\|^2_{H_2^{-1}}$$

where $\gamma = c_1(\bar{c}_4 + \alpha_1^{-1})$. If we set

$$\rho_0 = \max\left\{1, \frac{\gamma^2}{2\alpha_0\nu} + \frac{\alpha_0}{2\nu}\right\}.$$

then it holds that

$$\begin{aligned} x^T H^{-1}\bar{\mathcal{A}}\mathcal{P}_r^{-1}x &\geq \frac{\alpha_0}{2}\left(\|x_1\|^2_{H_1^{-1}} + \|x_2\|^2_{H_2^{-1}}\right) \\ &\quad + \frac{\alpha_0}{2}\|x_1\|^2_{H_1^{-1}} - \gamma\|x_1\|_{H_1^{-1}}\|x_2\|_{H_2^{-1}} + \frac{\gamma^2}{2\alpha_0}\|x_2\|^2_{H_2^{-1}} \\ &= \frac{\alpha_0}{2}\left(\|x_1\|^2_{H_1^{-1}} + \|x_2\|^2_{H_2^{-1}}\right) + \left(\frac{\sqrt{\alpha_0}}{\sqrt{2}}\|x_1\|_{H_1^{-1}} - \frac{\gamma}{\sqrt{2\alpha_0}}\|x_2\|_{H_2^{-1}}\right)^2 \\ &\geq \frac{\alpha_0}{2}\left(\|x_1\|^2_{H_1^{-1}} + \|x_2\|^2_{H_2^{-1}}\right) \end{aligned}$$

for $r \geq \rho_0$. Hence,

$$\frac{x^T H^{-1}\bar{\mathcal{A}}\mathcal{P}_r^{-1}x}{x^T H^{-1}x} \geq \frac{\alpha_0}{2},$$

therefore we can take $\tau = \frac{\alpha_0}{2}$ and the proof is complete. □

Remark 3.23 The assumption that $r > 1$ in the inequality $\|\bar{A}\bar{P}^{-1} - I\|_{H_1^{-1}} \leq r^{-1}$ can be relaxed if we assume that there exists a constant c_5 such that $\|Q^{-1}\|_{H_2^{-1},H_2} \leq c_5$. Then, there is no need to set the assumption $r \geq 1$ in the statement of Theorem 3.22. Indeed, in view of Eq. (47), Eq. (48) can be replaced by

$$\left|r x_1^T H_1^{-1}(\bar{A}\bar{P}^{-1} - I)C^T Q^{-1}x_2\right| \leq c_1 c_5 \|x_1\|_{H_1^{-1}}\|x_2\|_{H_2^{-1}},$$

since $r\left\|\bar{A}\bar{P}^{-1} - I\right\|_{H_1^{-1}} \leq 1$. Thus, in the proof of previous theorem ρ_0 can be taken to be

$$\rho_0 = \frac{\gamma^2}{2\alpha_0\nu} + \frac{\alpha_0}{2\nu}$$

with $\gamma = c_1(c_5 + \alpha_1^{-1})$, while the rest of the proof remains unchanged after removing the restriction $r \geq 1$.

For the sake of generality, we add the following remark showing that the above result can be stated under weaker conditions.

Remark 3.24 In Proposition 3.18, the assumption (42) was used to deduce that $\|rQ^{-1}\|_{H_2^{-1},H_2}$ is bounded above by a constant. On the other hand, the condition (44) ensures that

$$\|r^{-1}H_2^{-1/2}QH_2^{-1/2}\| \leq 1.$$

Following the preceding discussion, the assumption (44) can be relaxed by setting the condition that

$$\|r^{-1}H_2^{-1/2}QH_2^{-1/2}\| = \|r^{-1}QH_2^{-1}\|_{H_2^{-1}}$$

is bounded above by a constant. Now, in view of the following equality

$$\|rQ^{-1}\|_{H_2^{-1},H_2} = \|H_2(r^{-1}Q)^{-1}\|_{H_2^{-1}},$$

one can relax the assumptions (42) and (44). To this end, we need to choose r and Q such that $\frac{1}{r}Q \sim_{H_2^{-1}} H_2$.

We have checked numerically that for linear systems of the form (1) arising from the finite element discretization of coupled Stokes–Darcy flow, the condition $\lambda_{\max}(A_{22}^{-1}A_{12}^TA_{11}^{-1}A_{12}) \leq 0.5$ holds true for problems of small or moderate size. As seen in Theorem 3.22, it is assumed that $\bar{A} \approx_{H^{-1}} \bar{P}$. Note that for $r = 0$ this condition reduces to $A \approx_{H_1^{-1}} P$ which is similar to the assumption in [13, Theorem 3.10]. The following proposition establishes sufficient conditions under which $\bar{A} \approx_{H_1^{-1}} \bar{P}$.

Proposition 3.25 *Let $\bar{A}$ and $\bar{P}$ be defined by*

$$\bar{A} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & \bar{A}_{22} \end{bmatrix} \quad \text{and} \quad \bar{P} = \begin{bmatrix} A_{11} & A_{12} \\ 0 & \bar{A}_{22} \end{bmatrix},$$

where $A_{21} = -A_{12}^T$ and H_1 is given by (26). If the stability conditions (39) hold, then there exists $\beta_0 > 0$ such that

$$\left\| \bar{A}\bar{P}^{-1} \right\|_{H_1^{-1}} \leq \beta_0.$$

Furthermore, assume that $A_{22} \succcurlyeq A_{12}^TA_{11}^{-1}A_{12}$ and $\lambda_M := \lambda_{\max}(A_{22}^{-1}A_{12}^TA_{11}^{-1}A_{12}) \leq 3/4$. If the following relation holds:

$$\frac{1}{2} - \lambda_M \frac{r\lambda_{\max}(A_{22}^{-1}B^TQ^{-1}B)}{1 + r\lambda_{\max}(A_{22}^{-1}B^TQ^{-1}B)} \geq 0, \tag{51}$$

then $\bar{A} \approx_{H_1^{-1}} \bar{P}$.

Proof It is straightforward to check that

$$\bar{A}\bar{P}^{-1} = \begin{bmatrix} I & 0 \\ A_{21}A_{11}^{-1} & I + A_{12}^T A_{11}^{-1} A_{12} \bar{A}_{22}^{-1} \end{bmatrix},$$

having in mind that $A_{12}^T = -A_{21}$. Moreover, we have that

$$\begin{aligned} \left\| \bar{A}\bar{P}^{-1} \right\|_{H_1^{-1}} &= \left\| H_1^{-1/2} \bar{A}\bar{P}^{-1} H_1^{1/2} \right\|_2 \\ &\leq 2 + c_1 + c_1^2 =: \beta_0, \end{aligned}$$

where we made use of

$$\begin{aligned} &H_1^{-1/2} \bar{A}\bar{P}^{-1} H_1^{1/2} \\ &= \begin{bmatrix} I & 0 \\ A_{22}^{-1/2} A_{21} A_{11}^{-1/2} & I + A_{22}^{-1/2} A_{12}^T A_{11}^{-1/2} A_{11}^{-1/2} A_{12} A_{22}^{-1/2} A_{22}^{1/2} \bar{A}_{22}^{-1} A_{22}^{1/2} \end{bmatrix}, \end{aligned} \tag{52}$$

$\left\| A_{22}^{1/2} \bar{A}_{22}^{-1} A_{22}^{1/2} \right\|_2 = 1$ and $\|A_{12}^T\|_{A_{11}, A_{22}^{-1}} = \|A_{12}\|_{A_{22}, A_{11}^{-1}} \leq c_1$.

To prove the assertion, we need to show that there exists α_0 such that

$$\alpha_0 \leq \frac{\left\langle \bar{A}\bar{P}^{-1} x, x \right\rangle_{H_1^{-1}}}{\langle x, x \rangle_{H_1^{-1}}}.$$

To this end, we first show that the assumption (51) guarantees that the matrix $\mathcal{F} = \frac{1}{2} I + A_{22}^{-1/2} \hat{S} \bar{A}_{22}^{-1} A_{22}^{1/2}$ (where we set $\hat{S} := A_{12}^T A_{11}^{-1} A_{12}$ for notational simplicity) is positive semi-definite, in the sense that the quadratic form $\langle \mathcal{F} z, z \rangle$ is nonnegative for any real vector z. We comment that $\mathcal{F}$ is symmetric positive definite for $r = 0$ since in this case $\bar{A}_{22} = A_{22}$.

By the Sherman-Morrison-Woodbury matrix identity we have

$$\bar{A}_{22}^{-1} = A_{22}^{-1} - r A_{22}^{-1} B^T Q^{-1/2} (I + r Q^{-1/2} B A_{22}^{-1} B^T Q^{-1/2})^{-1} Q^{-1/2} B A_{22}^{-1},$$

hence we can write

$$A_{22}^{-1/2} \hat{S} \bar{A}_{22}^{-1} A_{22}^{1/2} = A_{22}^{-1/2} \hat{S} A_{22}^{-1/2} - A_{22}^{-1/2} \hat{S} A_{22}^{-1/2} E,$$

where E is a symmetric positive semi-definite matrix given by

$$E = r A_{22}^{-1/2} B^T Q^{-1/2} (I + r Q^{-1/2} B A_{22}^{-1} B^T Q^{-1/2})^{-1} Q^{-1/2} B A_{22}^{-1/2}.$$

Next, we observe that the nonzero eigenvalues of E are the same as those of the matrix

$$\tilde{E} = r Q^{-1/2} B A_{22}^{-1} B^T Q^{-1/2} (I + r Q^{-1/2} B A_{22}^{-1} B^T Q^{-1/2})^{-1}.$$

and that the spectrum of $\tilde{E}$ is given by

$$\sigma(\tilde{E}) = \left\{ \frac{r\lambda}{1+r\lambda} \;\middle|\; \lambda \in \sigma(Q^{-1/2} B A_{22}^{-1} B^T Q^{-1/2}) \right\}.$$

Notice that the function $g_r(x) = \frac{rx}{1+rx}$ is monotonically increasing for $x, r > 0$. Hence, we have

$$\begin{aligned} \|E\|_2 &= \lambda_{\max}(E) = \lambda_{\max}(\tilde{E}) \\ &= \frac{r\lambda_{\max}(Q^{-1/2} B A_{22}^{-1} B^T Q^{-1/2})}{1 + r\lambda_{\max}(Q^{-1/2} B A_{22}^{-1} B^T Q^{-1/2})} \\ &= \frac{r\lambda_{\max}(A_{22}^{-1} B^T Q^{-1} B)}{1 + r\lambda_{\max}(A_{22}^{-1} B^T Q^{-1} B)}. \end{aligned}$$

Let z be an arbitrary real vector, then, using the above relation, we have

$$\begin{aligned} \langle \mathcal{F} z, z \rangle &= \frac{1}{2} \langle z, z \rangle + \left\langle A_{22}^{-1/2} \hat{S} A_{22}^{-1/2} z, z \right\rangle - \left\langle A_{22}^{-1/2} \hat{S} A_{22}^{-1/2} E z, z \right\rangle \\ &\geq \left(\frac{1}{2} - \left\| A_{22}^{-1/2} \hat{S} A_{22}^{-1/2} \right\|_2 \|E\|_2 \right) \|z\|_2^2 \\ &= \left(\frac{1}{2} - \lambda_M \frac{r\lambda_{\max}(A_{22}^{-1} B^T Q^{-1} B)}{1 + r\lambda_{\max}(A_{22}^{-1} B^T Q^{-1} B)} \right) \|z\|_2^2 \geq 0, \end{aligned}$$

as claimed. Considering (52), we can rewrite $H_1^{-1/2} \bar{A} \bar{P}^{-1} H_1^{1/2}$ as follows:

$$H_1^{-1/2} \bar{A} \bar{P}^{-1} H_1^{1/2} = \begin{bmatrix} I & 0 \\ A_{22}^{-1/2} A_{21} A_{11}^{-1/2} & \frac{1}{2} I + \mathcal{F} \end{bmatrix}.$$

By assumption, we have

$$\begin{aligned} \|A_{22}^{-1/2} A_{21} A_{11}^{-1/2}\|_2^2 &= \rho(A_{11}^{-1/2} A_{21}^T A_{22}^{-1} A_{21} A_{11}^{-1/2}) \\ &= \rho(A_{22}^{-1} A_{21} A_{11}^{-1} A_{21}^T) = \rho(A_{22}^{-1} A_{12}^T A_{11}^{-1} A_{12}) \leq 3/4. \end{aligned}$$

Now let $x = (x_1; x_2)$ be an arbitrary nonzero vector and $w = H_1^{-1/2} x$ (with block partitioning $w = (w_1; w_2)$). Using the Cauchy–Schwarz inequality and some straightforward computations, we obtain

$$\begin{aligned} \left\langle \bar{A} \bar{P}^{-1} x, x \right\rangle_{H_1^{-1}} &= \left\langle H_1^{-1/2} \bar{A} \bar{P}^{-1} H_1^{1/2} w, w \right\rangle \\ &= \langle w_1, w_1 \rangle + \left\langle A_{22}^{-1/2} A_{21} A_{11}^{-1/2} w_1, w_2 \right\rangle + \frac{1}{2} \langle w_2, w_2 \rangle + \langle \mathcal{F} w_2, w_2 \rangle \\ &\geq \langle w_1, w_1 \rangle - \|A_{22}^{-1/2} A_{21} A_{11}^{-1/2}\|_2 \|w_1\|_2 \|w_2\|_2 + \frac{1}{2} \langle w_2, w_2 \rangle \end{aligned}$$

$$\begin{aligned}
&\geq \frac{1}{4}\langle w_1, w_1\rangle + \frac{3}{4}\langle w_1, w_1\rangle - \frac{\sqrt{3}}{2}\|w_1\|_2\|w_2\|_2 + \frac{1}{2}\langle w_2, w_2\rangle \\
&= \frac{1}{4}(\langle w_1, w_1\rangle + \langle w_2, w_2\rangle) + \left(\frac{\sqrt{3}}{2}\|w_1\|_2 - \frac{1}{2}\|w_2\|_2\right)^2 \\
&\geq \frac{1}{4}(\langle w_1, w_1\rangle + \langle w_2, w_2\rangle) = \frac{1}{4}\langle x, x\rangle_{H_1^{-1}}.
\end{aligned}$$

Setting $\alpha_0 = \frac{1}{4}$, the proof is complete.

□

We end this section with the following comments on assumption (45) in Theorem 3.22.

Remark 3.26 Assume that $\tilde{S}_0 \approx_{H_2^{-1}} Q$, where $\tilde{S}_0 = CP^{-1}C^T = BA_{22}^{-1}B^T$. As a result of Remark 3.6, there exists a constant $\tilde{\gamma}$ (independent of n) such that $\left\| Q\tilde{S}_0^{-1} \right\|_{H_2^{-1}} \leq \tilde{\gamma}$. The assumption $\tilde{S}_0 \approx_{H_2^{-1}} Q$ implies that there exists $\nu_0 > 0$ such that

$$\nu_0 \leq \frac{\left\langle \tilde{S}_0 Q^{-1} y, y \right\rangle_{H_2^{-1}}}{\langle y, y\rangle_{H_2^{-1}}}, \tag{53}$$

for any nonzero vector y. Next, we show that we can find $\nu > 0$ such that (45) holds for $r < \nu_0^{-1}$. Note that using Lemma 3.15, we obtain

$$\begin{aligned}
Q\tilde{S}_r^{-1} &= Q(C\bar{P}^{-1}C^T)^{-1} = Q(C(P + rC^TQ^{-1}C)^{-1}C^T)^{-1} \\
&= Q((CP^{-1}C^T)^{-1} + rQ^{-1}) = Q\tilde{S}_0^{-1} + rI.
\end{aligned}$$

Let y be an arbitrary nonzero vector and set $y = Q\tilde{S}_r^{-1}w$. Now, using a similar argument to the one in [9, Page 781], we get

$$\begin{aligned}
\frac{\left\langle \tilde{S}_r Q^{-1} y, y \right\rangle_{H_2^{-1}}}{\langle y, y\rangle_{H_2^{-1}}} &= \frac{\left\langle w, Q\tilde{S}_r^{-1}w \right\rangle_{H_2^{-1}}}{\left\| Q\tilde{S}_r^{-1}w \right\|_{H_2^{-1}}^2} \\
&\geq \frac{\left\langle w, Q\tilde{S}_0^{-1}w \right\rangle_{H_2^{-1}} + r\langle w, w\rangle_{H_2^{-1}}}{2(\|Q\tilde{S}_0^{-1}w\|_{H_2^{-1}}^2 + r^2\|w\|_{H_2^{-1}}^2)} \\
&= \frac{k + rt}{2(1 + r^2t)},
\end{aligned}$$

where $t = \dfrac{\langle w,w\rangle_{H_2^{-1}}}{\left\|Q\tilde{S}_0^{-1}w\right\|^2_{H_2^{-1}}}$ and $k = \dfrac{\left\langle w, Q\tilde{S}_0^{-1}w\right\rangle_{H_2^{-1}}}{\left\|Q\tilde{S}_0^{-1}w\right\|^2_{H_2^{-1}}}$. Notice that setting $z = Q\tilde{S}_0^{-1}w$, using the assumption (53), we have

$$k = \frac{\left\langle \tilde{S}_0 Q^{-1}z, z\right\rangle_{H_2^{-1}}}{\|z\|^2_{H_2^{-1}}} \geq \nu_0.$$

Let $f(x) = \frac{\nu_0 + rx}{2(1+r^2x)}$. The function $f(x)$ is monotonically increasing for $0 < r < \nu_0^{-1}$. Using the fact that $t \geq \tilde{\gamma}^{-2}$, it follows that

$$\frac{\left\langle \tilde{S}_r Q^{-1}y, y\right\rangle_{H_2^{-1}}}{\langle y, y\rangle_{H_2^{-1}}} \geq \nu,$$

where $\nu = \frac{\nu_0 + r\tilde{\gamma}^{-2}}{2(1+r^2\tilde{\gamma}^{-2})}$. Finally, we observe that in [23, Theorem 3.8], for the case $r = 0$, it is assumed that $Q^{-1} \approx_{H_2} \tilde{S}_0^{-1}$. Here we point out that Proposition 3.5 shows that $\tilde{S}_0 \approx_{H_2^{-1}} Q$ is a consequence of $\tilde{S}_0^{-1} \approx_{H_2} Q^{-1}$ (which is equivalent to $Q^{-1} \approx_{H_2} \tilde{S}_0^{-1}$ by Remark 3.6).

4 Numerical experiments

In practice, all the preconditioners considered so far must be applied inexactly, especially when solving 3D problems. Whether the mesh-independent behavior is retained or not by the inexact variants is not clear a priori; as we will see, the choice of inexact solver may impact some preconditioners more than others. In this section we illustrate the performance of inexact variants of the block preconditioners using a test problem, taken from [13, Subsection 5.3], which corresponds to a 3D coupled flow problem in a cube $\Omega = \Omega_1 \cup \Omega_2$ with $\Omega_1 = [0,2] \times [0,2] \times [1,2]$ and $\Omega_2 = [0,2] \times [0,2] \times [0,1]$. The porous medium Ω_2 contains an embedded impermeable cube $[0.75, 1.25] \times [0.75, 1.25] \times [0, 0.50]$. The hydraulic conductivities of the porous medium and embedded impermeable enclosure are $\kappa_1\mathbf{I}$ an $\kappa_2\mathbf{I}$, respectively, with $\kappa_1 = 1$ and $\kappa_2 = 10^{-10}$. The kinematic viscosity is set to $\nu = 1.0$. On the horizontal part of $\Gamma_1 = \partial\Omega_1 \cap \partial\Omega$ we prescribe $\mathbf{u}_1 = (0,0,-1)^T$ at $z = 2$ and the no-slip condition on the lateral sides of Γ_1. We prescribe homogeneous Dirichlet boundary conditions on $\Gamma_2 = \partial\Omega \cap \partial\Omega_2$ ($z = 0$) and homogeneous Neumann conditions on the rest of the boundary of the porous medium. The large jump in the hydraulic conductivity in the porous medium region makes this problem challenging.

We report the performances of several preconditioner variants in conjunction with FGMRES [24]. The initial guess is taken to be the zero vector and the iterations are stopped once $\|\mathcal{A}u_k - b\|_2 \leq 10^{-7}\|b\|_2$ (or $\|\bar{\mathcal{A}}u_k - \bar{b}\|_2 \leq 10^{-7}\|\bar{b}\|_2$ for the augmented Lagrangian variants) where u_k is the obtained k-th approximate solution. In addition,

we have used right-hand sides corresponding to random solution vectors and averaged results over 10 test runs. At each iteration of FGMRES, we need to solve at least two SPD linear systems as subtasks. To this end we applied two different approaches, discussed in the following two subsections.

All computations were carried out on a computer with an Intel Core i7-10750H CPU @ 2.60GHz processor and 16.0GB RAM using MATLAB.R2020b.

4.1 Implementation based on IC-CG

First we present the results of experiments in which, inside FGMRES, the SPD subsystems were solved inexactly by the preconditioned conjugate gradient (PCG) method using loose tolerances. More precisely, the inner PCG solver for linear systems with coefficient matrix A_{11} (A_{22} and $A_{22} + rB^TQ^{-1}B$) was terminated when the relative residual norm was below 10^{-1} (respectively, 10^{-2}) or when the maximum number of 5 (respectively, 25) iterations was reached. In the implementation of the preconditiner $\mathcal{P}_{T_1,\rho}$, the inverse of M_p was applied inexactly using PCG with a relative residual tolerance of 10^{-2} and a maximum number of 20 iterations. The preconditioner for PCG are incomplete Cholesky factorizations constructed using the MATLAB function "`ichol(., opts)`" where `opts.type ='ict'` with drop tolerances between 10^{-4} and 10^{-2}. The FGMRES iteration count is reported in the tables under "Iter". Under "Iter$_{pcg_i}$" ("Iter$_{cg_i}$") we further report the total number of inner PCG (or CG) iterations performed for solving the linear systems corresponding to block (i, i) of the preconditioner, where $i = 1, 2$. For more details, in the "Appendix" we summarize the implementation of preconditioners $\mathcal{P}_{con_D}$, $\mathcal{P}_{con_T}$ and $\mathcal{P}_{T_1,\rho}$ in Algorithms 1-3.

For the linear system corresponding to $A_{22} + rB^TQ^{-1}B$, we distinguish between two approaches:

- Approach I. The matrix $A_{22} + rB^TQ^{-1}B$ is not formed explicitly and the CG method is used without preconditioning with a relative residual tolerance of 10^{-3} and a maximum allowed number of 25 iterations.
- Approach II. The matrix $A_{22} + rB^TQ^{-1}B$ is formed explicitly, and PCG with incomplete Cholesky preconditionign was used. We note that while we could successfully compute the "ichol" factor without diagonal shifts for the two smallest problem sizes, adding the shift 0.01 was found to be necessary for larger sizes. We further note that with this approach we can use larger values of r, leading to faster FGMRES convergence.

In Tables 1 and 2 , we report the performance $\mathcal{P}_r$ for Approaches I and II. From the results presented, we can see that even when implemented inexactly, the augmented Lagrangian-based preconditioner $\mathcal{P}_r$ results in convergence rates of FGMRES that are essentially mesh-independent, as predicted by our theoretical analysis. As for the number of inner PCG iterations, we observe some differences in the results obtained with Approaches I and II. In the case of Approach I we see an increase in the total number of inner PCG iterations as the mesh is refined, reflecting the known fact that the CG method, with or without incomplete Cholesky preconditioning, is not mesh-independent in general. With Approach II this increase is not observed, however, the total timings are much higher and still scale superlinearly with the number

Table 1 Results for FGMRES in conjunction with preconditioner $\mathcal{P}_r$, Approach I

Size	$r = 2$				$r = 5$			
	FGMRES		Inner iterations		FGMRES		Inner iterations	
	Iter	CPU time	Iter_{pcg_1}	Iter_{cg_2}	Iter	CPU time	Iter_{pcg_1}	Iter_{cg_2}
1695	22	0.0521	70	432	15	0.0404	52	337
10809	20	0.7018	79	525	15	0.5293	63	387
76653	20	7.4126	93	570	15	5.7076	71	390
576213	23	71.439	110	595	21	63.847	98	536

Table 2 Results for FGMRES in conjunction with preconditioner $\mathcal{P}_r$, Approach II

Size	$r = 5$				$r = 10$			
	FGMRES		Inner iterations		FGMRES		Inner iterations	
	Iter	CPU time	Iter_{pcg_1}	Iter_{pcg_2}	Iter	CPU time	Iter_{pcg_1}	Iter_{pcg_2}
1695	16	0.0864	57	150	13	0.0467	43	80
10809	15	1.3697	63	121	12	0.8918	49	69
76653	13	11.966	64	95	11	8.2454	50	56
576213	14	189.43	64	103	11	155.603	49	58

Table 3 Results for FGMRES in conjunction with preconditioners $\mathcal{P}_{con_D}$ and $\mathcal{P}_{con_T}$

Size	$\mathcal{P}_{con_D}$				$\mathcal{P}_{con_T}$			
	FGMRES		Inner iterations		FGMRES		Inner iterations	
	Iter	CPU time	Iter_{pcg_1}	Iter_{pcg_2}	Iter	CPU time	Iter_{pcg_1}	Iter_{pcg_2}
1695	21	0.1295	77	606	18	0.1001	68	462
10809	20	1.2693	97	697	19	1.2694	89	619
76653	29	18.417	109	975	26	16.920	103	860
576213	61	319.99	167	805	72	380.37	189	974

of degrees of freedom. This is due to the fact that explicitly forming the augmented matrix $A_{22} + rB^T Q^{-1} B$ and computing its incomplete Cholesky factorization leads to a considerably less sparse matrix and superlinear growth in the fill-in in the incomplete factors, and thus to more expensive PCG iterations. We conclude that with $\mathcal{P}_r$, Approach I is to be preferred to Approach II.

In Table 3, we report the results corresponding to the constraint preconditioners. Our numerical tests illustrate that in the inexact implementation of these preconditioners using IC-CG for the inner SPD linear solves, the outer iteration counts grow each time the mesh size is halved. Hence, the mesh-independence of the outer FGMRES iteration is lost when the preconditioner is applied inexactly using incomplete Cholesky as the preconditioner for the inner PCG iterations. We add that in our implementation of these preconditioners, we approximated $BA_{22}^{-1}B^T$ by M_p and solved the corresponding linear system using PCG with a maximum number of 25 iterations and relative residual tolerance tolerance 10^{-2} and with incomplete Cholesky precon-

Table 4 Results for FGMRES in conjunction with preconditioners $\mathcal{P}_{T_1,0.6}$ and $\tilde{\mathcal{P}}_{T_1,0.6}$ (Case I)

Size	$\mathcal{P}_{T_1,0.6}$				$\tilde{\mathcal{P}}_{T_1,0.6}$			
	FGMRES		Inner iterations		FGMRES		Inner iterations	
	Iter	CPU time	Iter_{pcg_1}	Iter_{pcg_2}	Iter	CPU time	Iter_{pcg_1}	Iter_{pcg_2}
1695	21	0.1482	69	437	31	0.2066	97	680
10809	21	2.2306	57	418	37	3.9359	127	836
76653	21	22.175	63	466	38	26.219	117	511
576213	22	214.43	101	516	37	317.45	167	753

ditioning where opts.droptol" was set to 10^{-2}. We also tried approximating $BA_{22}^{-1}B^T$ by the diagonal of M_p but the results were generally worse and we do not report them here.

Next, we consider inexact variants of the following block triangular preconditioners,

$$\mathcal{P}_{T_1,\rho} := \mathcal{P}_{T_1}(\rho) = \begin{bmatrix} A_{11} & 0 & 0 \\ 0 & A_{22} & 0 \\ 0 & B & -\rho M_p \end{bmatrix}, \tag{54}$$

and

$$\tilde{\mathcal{P}}_{T_1,\rho} := \tilde{\mathcal{P}}_{T_1}(\rho) = \begin{bmatrix} A_{11} & 0 & 0 \\ 0 & A_{22} & 0 \\ 0 & B & -\rho\ \text{diag}(M_p) \end{bmatrix}.$$

In Tables 4 and 5 , we present results for $\rho = 0.6$. In [12], it was experimentally observed that the performance of $\mathcal{P}_{T_1,\rho}$ is not sensitive to ρ when $\rho \in [0.6, 1.05]$. However, based on our experimental results, we found the optimum value $\rho = 0.6$. For more details, we report the results for two different cases (referred as Cases I and II) in Tables 4 and 5 by setting opts.droptol" to be 10^{-4} and 10^{-2}, respectively. Similar to $\mathcal{P}_r$, in Table 4, it is seen that for $\mathcal{P}_{T_1,0.6}$, the outer iteration count for FGMRES remains essentially constant as the grid is refined, in agreement with our analysis for the exact case. Although the results in Table 5 indicate a better performance of the preconditioner for the first three problem sizes, the number of outer iterations increases drastically for the largest problem size.

From these results we see that replacing the inexact solves involving the mass matrix M_p with a simple diagonal scaling involving the diagonal of M_p leads to a degradation of the rate of convergence. We found that in terms of CPU time, this degradation more than offsets the savings obtained by using simple diagonal scalings in place of solves of linear systems involving M_p.

Overall, when the subsystems associated with the block preconditioners are solved using (P)CG with incomplete Cholesky factorization, the fastest solution times are achieved with the inexact variant of the augmented Lagrangian preconditionr $\mathcal{P}_r$ using what we called "Approach I". For the largest size problems, this approach is about

Table 5 Results for FGMRES in conjunction with preconditioners $\mathcal{P}_{T_1,0.6}$ and $\tilde{\mathcal{P}}_{T_1,0.6}$ (Case II)

Size	$\mathcal{P}_{T_1,0.6}$				$\tilde{\mathcal{P}}_{T_1,0.6}$			
	FGMRES		Inner iterations		FGMRES		Inner iterations	
	Iter	CPU time	Iter_{pcg_1}	Iter_{pcg_2}	Iter	CPU time	Iter_{pcg_1}	Iter_{pcg_2}
1695	21	0.0724	85	429	35	0.1232	131	765
10809	22	0.8061	105	467	36	1.2528	167	780
76653	28	9.0421	100	523	39	11.849	142	918
576213	80	231.15	177	633	109	300.31	191	1446

3.3 times faster, in terms of CPU time, than the block triangular preconditioner from [12], which is in turn far more efficient than the inexact variants of the constraint preconditioners. Furthermore, the construction cost of the incomplete Cholesky factorizations used with these preconditioners is negligible. The CPU time scaling of all these methods with respect to the mesh size is, however, superlinear in the number of unknowns due to the use of IC preconditioning in conjunction with the CG method to perform the inner iterations.

4.2 Implementation based on ARMS preconditioner

In an attempt to have better scalability of the number of inner iterations with respect to mesh refinements, as an alternative to using IC-CG, we performed some experiments with an algebraic multilevel solver for approximately solving the subsystems associated with the block preconditioners. We chose the MATLAB implementation of the ARMS preconditioner [25], which can be downloaded from https://www-users.cs.umn.edu/~saad/software/.

Since the ARMS preconditioner is not SPD, for inexact solves involving sub-blocks we use it with GMRES (with relative residual tolerance 0.1 and a maximum number of iterations equal to 20) in conjunction with ARMS. With this approach, all tested preconditioners (including constraint preconditioners) appear robust, displaying mesh-independent convergence of the outer FGMRES iteration, and faster convergence of the inner iterations. The obtained numerical results are shown in Tables 6 for the block triangular, constraint preconditioners, and the augmented Lagrangian-based preconditioner. To implement the preconditioner $\mathcal{P}_r$, the subsystem corresponding to sub-block (1, 1) is solved by GMRES in conjunction with the ARMS preconditioner. For the subsystem associated with $A_{22} + rB^T Q^{-1} B$, forming the ARMS preconditioner is not practically feasible for the larger problem sizes. Therefore, the matrix $A_{22} + rB^T Q^{-1} B$ is not formed explicitly and the corresponding system is solved by the preconditioned GMRES where the ARMS preconditioner for A_{22} is used. As before, we report under "Iter" the number of (outer) FGMRES iterations. Under "Iter_i" we report the total number of inner iterations performed for solving the linear systems corresponding to block (i, i) of the preconditioner where $i = 1, 2$. To obtain results in Table 6, PCG with tolerance 10^{-2} and a maximum of 20 iterations was used for solv-

Table 6 Results for constraint preconditioners, $\mathcal{P}_{T_1,0.6}$ and $\mathcal{P}_r$; Inner solvers: GMRES in conjunction with ARMS preconditioners for sub-blocks (1, 1) and (2, 2)

Size	Pre	FGMRES		Inner iterations		
		Iter	CPU time	Iter_1	Iter_2	Iter_3
1695	$\mathcal{P}_{T_1,0.6}$	20	0.0689	123	100	33
	$\mathcal{P}_{con_D}$	20	0.1170	125	186	35
	$\mathcal{P}_{con_T}$	15	0.0811	91	135	25
	$\mathcal{P}_{15}$	11	0.0684	51	101	–
	$\mathcal{P}_7$	13	0.0705	80	131	–
	$\mathcal{P}_5$	15	0.0912	92	150	–
10809	$\mathcal{P}_{T_1,0.6}$	20	0.7176	164	148	31
	$\mathcal{P}_{con_D}$	18	1.1279	150	259	31
	$\mathcal{P}_{con_T}$	14	0.8637	114	194	22
	$\mathcal{P}_{15}$	12	0.5631	71	117	–
	$\mathcal{P}_7$	13	0.6443	95	131	–
	$\mathcal{P}_5$	15	0.7134	94	146	–
76653	$\mathcal{P}_{T_1,0.6}$	20	8.3387	190	209	25
	$\mathcal{P}_{con_D}$	16	12.136	160	310	26
	$\mathcal{P}_{con_T}$	13	9.2989	130	240	19
	$\mathcal{P}_{15}$	13	5.5103	92	130	–
	$\mathcal{P}_7$	13	5.5541	102	131	–
	$\mathcal{P}_5$	14	6.1664	111	141	–
576213	$\mathcal{P}_{T_1,0.6}$	20	82.834	186	200	20
	$\mathcal{P}_{con_D}$	14	100.91	141	272	17
	$\mathcal{P}_{con_T}$	12	88.234	120	230	12
	$\mathcal{P}_{15}$	23	91.508	160	230	–
	$\mathcal{P}_7$	18	75.246	130	180	–
	$\mathcal{P}_5$	16	63.587	126	156	–

ing linear systems associated with M_p. The corresponding total number of iterations are given under "Iter_{M_p}".

Iteration times were also found to exhibit better (though not perfect) scalability than in the experiments described in the previous subsection. The construction costs for ARMS, however, appear to be prohibitive, at least in the MATLAB implementation, completely off-setting any gains in performance. In particular, for the largest problem sizes it takes hours to compute the ARMS preconditioners.

We can see from these experiments that for all preconditioners tested, both the number of outer FGMRES iterations and (for large enough problem sizes) the total number of inner preconditioned GMRES and CG iterations remain almost constant, with outer iteration counts even improving for smaller mesh sizes. As mentioned, however, this improved scaling behavior comes at the price of much higher preconditioner construction costs. The reported solution times show that in conjunction with ARMS, the augmented Lagrangian-based preconditioner $\mathcal{P}_r$ is both efficient and fairly robust with respect to the parameter r, and outperforms all other preconditioners for large enough problem sizes. However, it does not outperform the implementation of $\mathcal{P}_r$

based on IC-CG for the inner solves. Given the enormous set-up costs associated with ARMS, we conclude that its use does not bring about any actual advantage in terms of times to solution, at least when working in MATLAB.

In conclusion, the results of our experiments indicate that among all preconditioner variants we tested, the inexact variant of $\mathcal{P}_r$ with IC-CG inner solves (Approach I) is, by a large margin, the fastest solver in terms of total solution times.

5 Conclusions

In this paper we have provided a theoretical analysis of several types of block preconditioners for the discrete Stokes–Darcy problem. Both eigenvalue bounds and FOV-equivalence have been considered, completing the analyses given in [12] and in [13]. Our analysis extends previous results and explains the experimentally observed mesh-independence of the exact variants of the block preconditioner based on the augmented Lagrangian approach.

Numerical experiments show that inexact variants of these block preconditioners may or may not retain mesh-independence, depending on the solver used for the inexact solves. All preconditioners show near mesh-independence when a multilevel algebraic solver (ARMS) is used for the inexact solves, but this preconditionr is found to have exceedingly high construction costs. When cheaper inner solvers based on incomplete Cholesky-preconditioned CG are used, the fastest total solution times are achieved by the augmented Lagrangian-type preconditioner. It is possible, of course, that better results may be achieved with different multilevel solvers.

Future work should consider the development of similar preconditioners for the coupled Navier–Stokes–Darcy model.

Acknowledgements The authors would like to thank Scott Ladenheim for providing the test problem. Thanks also to three anonymous referees for their careful reading of the manuscript and helpful suggestions.

Appendix

For the sake of clarity, we summarize the required steps for implementation of the constraint preconditioners and $\mathcal{P}_{T_1,\rho}$ inside FGMRES in the following algorithms. We recall that in the numerical experiments the matrix $BA_{22}^{-1}B^T$ was replaced by M_p.

Algorithm 1: Computation of $(w_1; w_2; w_3) = \mathcal{P}_{con_D}^{-1}(w_1; w_2; w_3)$.

Step 1. Solve $A_{11}w_1 = r_1$ for w_1;
Step 2. Solve $A_{22}z = r_2$ for z;
Step 3. Solve $BA_{22}^{-1}B^T w_3 = Bz - r_3$ for w_3;
Step 4. Solve $A_{22}v = B^T w_3$ for v;
Step 5. Set $w_2 = z - v$.

Algorithm 2: Computation of $(w_1; w_2; w_3) = \mathcal{P}_{con_T}^{-1}(r_1; r_2; r_3)$.

Step 1. Solve $A_{11}w_1 = r_1$ for w_1;
Step 2. Solve $A_{22}z = r_2 - A_{21}w_1$ for z;
Step 3. Solve $BA_{22}^{-1}B^T w_3 = Bz - r_3$ for w_3;
Step 4. Solve $A_{22}v = B^T w_3$ for v;
Step 5. Set $w_2 = z - v$.

Algorithm 3: Computation of $(w_1; w_2; w_3) = \mathcal{P}_{T_1,\rho}^{-1}(r_1; r_2; r_3)$.

Step 1. Solve $A_{11}w_1 = r_1$ for w_1;
Step 2. Solve $A_{22}w_2 = r_2$ for w_2;
Step 3. Solve $M_p w_3 = -\rho^{-1}(r_3 - Bw_2)$ for w_3.

References

1. Anderson, N., Saff, S.H., Varga, R.S.: On the Eneström-Kakeya theorem and its sharpness. Linear Algebra Appl. **28**, 5–16 (1979)
2. Aulisa, E., Bornia, G., Howle, V., Ke, G.: Field-of-values analysis of preconditioned linearized Rayleigh-Bénard convection problems. J. Comput. Appl. Math. **369**, art. 112582 (2020)
3. Axelsson, O., Barker, V.A.: Finite Element Solution of Boundary Value Problems: Theory and Computation, SIAM Classics in Applied Mathematics 35. Society for Industrial and Applied Mathematics, Philadelphia (2001)
4. Beckermann, B., Goreinov, S.A., Tyrtyshnikov, E.E.: Some remarks on the Elman estimate for GMRES. SIAM J. Matrix Anal. Appl. **27**, 72–778 (2005)
5. Beik, F.P.A., Benzi, M.: Iterative methods for double saddle point systems. SIAM J. Matrix Anal. Appl. **39**, 902–921 (2018)
6. Benzi, M.: Preconditioning techniques for large linear systems: a survey. J. Comput. Phys. **182**, 418–477 (2002)
7. Benzi, M.: Some uses of the field of values in numerical analysis. Boll. Unione Matematica Italiana **14**, 159–177 (2021)
8. Benzi, M., Golub, G.H., Liesen, J.: Numerical solution of saddle point problems. Acta Numer. **14**, 1–137 (2005)
9. Benzi, M., Olshanskii, M.A.: Field-of-values convergence analysis of augmented Lagrangian preconditioners for the linearized Navier–Stokes problem. SIAM J. Numer. Anal. **49**, 770–788 (2011)
10. Benzi, M., Olshanskii, M.A.: An augmented Lagrangian-based approach to the Oseen problem. SIAM J. Sci. Comput. **28**, 2095–2113 (2006)
11. Boffi, D., Brezzi, F., Fortin, M.: Mixed Finite Element Methods and Applications, Springer Series in Computational Mathematics vol. 44. Springer, Berlin (2013)
12. Cai, M., Mu, M., Xu, J.: Preconditioning techniques for a mixed Stokes/Darcy model in porous media applications. J. Comput. Appl. Math. **233**, 346–355 (2009)

13. Chidyagwai, P., Ladenheim, S., Szyld, D.B.: Constraint preconditioning for the coupled Stokes–Darcy system. SIAM J. Sci. Comput. **38**, A668–A690 (2016)
14. Discacciati, M., Quarteroni, A.: Navier–Stokes/Darcy coupling: modeling, analysis and numerical approximation. Rev. Mat. Complut. **22**, 315–426 (2009)
15. Eisenstat, S.C., Elman, H.C., Schultz, M.H.: Variational iterative methods for nonsymmetric systems of linear equations. SIAM J. Numer. Anal. **20**, 345–357 (1983)
16. Elman, H.C.: Iterative Methods for Sparse Nonsymmetric Systems of Linear Equations. Ph.D. Thesis, Yale University, Department of Computer Science (1982)
17. Fletcher, R.: An ideal penalty function for constrained optimization. In: Nonlinear Programming, 2, pp. 121–163. Academic Press, New York (1974)
18. Golub, G.H., Greif, C.: On solving block-structured indefinite linear systems. SIAM J. Sci. Comput. **24**, 2076–2092 (2003)
19. Greenbaum, A., Pták, V., Strakoš, Z.: Any nonincreasing convergence curve is possible for GMRES. SIAM J. Matrix Anal. Appl. **17**, 465–469 (1996)
20. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, Cambridge (1985)
21. Kakeya, S.: On the limits of roots of an algebraic equation with positive coefficient. Tôhoku Math. J. First Series **2**, 140–142 (1912)
22. Klawonn, A., Starke, K.: Block preconditioners for nonsymmetric saddle point problems. Numer. Math. **81**, 577–594 (1999)
23. Loghin, D., Wathen, A.J.: Analysis of preconditioners for saddle-point problems. SIAM J. Sci. Comput. **25**, 2029–2049 (2004)
24. Saad, Y.: A flexible inner-outer preconditioned GMRES algorithm. SIAM J. Sci. Comput. **14**, 461–469 (1993)
25. Saad, Y., Suchomel, B.: ARMS: an algebraic recursive multilevel solver for general sparse linear systems. Numer. Linear Algebra Appl. **9**, 359–378 (2002)