Field-of-Values analysis of preconditioned linearized Rayleigh-Bénard convection problems

Eugenio Aulisa[a], Giorgio Bornia[a,*], Victoria Howle[a], Guoyi Ke[b]

[a] *Department of Mathematics and Statistics, Texas Tech University, Lubbock, TX, USA*
[b] *Department of Mathematics and Physical Sciences, Louisiana State University at Alexandria, Alexandria, LA, USA*

Abstract

In this paper we use the notion of field-of-values (FOV) equivalence of matrices to study a class of block-triangular preconditioners for the fixed-point linearization of the Rayleigh-Bénard convection problem discretized with inf-sup stable finite element spaces. First, sufficient conditions on the nondimensional parameters of the problem are determined in order to establish the FOV-equivalence between the system matrix and the preconditioners. Four upper triangular block preconditioners belonging to the general proposed class are then considered. Numerical experiments show that the Generalized Minimal Residual (GMRES) convergence is robust with respect to the mesh size for these preconditioned systems. We also compare the performance of the different preconditioners in terms of computational time.

Keywords: Rayleigh-Bénard convection, block preconditioning, incompressible flows, FOV-equivalence

1. Introduction

In this paper we analyze block preconditioners for the numerical solution of the Rayleigh-Bénard convection problem. Our analysis is driven by the notion of Field-Of-Values(FOV)-equivalence of matrices [15, 5, 23, 19, 2, 1]. For the sake of completeness, we first present the problem in its dimensional form and we obtain a nondimensional version (clearly, other nondimensionalizations are also possible). Under the Oberbeck-Boussinesq approximation, the Rayleigh-Bénard convection equations read

$$\begin{cases} \rho(\boldsymbol{u}\cdot\nabla)\boldsymbol{u} - \mu\Delta\boldsymbol{u} + \nabla p = \rho\beta(T - T_b)\boldsymbol{g} + \boldsymbol{f}, \\ -\nabla\cdot\boldsymbol{u} = 0, \\ \rho c_p(\boldsymbol{u}\cdot\nabla)T - k\Delta T = g, \end{cases} \tag{1}$$

*Corresponding author
Email address: giorgio.bornia@ttu.edu (Giorgio Bornia)

posed on some subset $\Omega \subset \mathbb{R}^d$ for $d = 2, 3$. Here, $\boldsymbol{g}$ is the gravity vector, ρ, μ, β, c_p, k denote fluid density, dynamic viscosity, thermal expansion coefficient, specific heat at constant pressure, and thermal conductivity, respectively. Also, T_b is a reference temperature. The unknowns are the velocity $\boldsymbol{u}$, the temperature T and the piezometric head $p = \bar{p} - \rho \boldsymbol{g} \cdot \boldsymbol{x}$, where $\bar{p}$ is the thermodynamic pressure and $\boldsymbol{x}$ is the position. The terms $\boldsymbol{f}$ and g are nonhomogeneous terms that can be physically interpreted as momentum and energy sources or that can take into account nonhomogeneous boundary conditions. We refer to a typical set of boundary conditions for heated enclosed flow problems given by

$$T = T_D \ \text{ on } \Gamma_D\,, \quad \nabla T \cdot \boldsymbol{n} = 0 \ \text{ on } \Gamma_N\,, \quad \boldsymbol{u} = \boldsymbol{0} \ \text{ on } \Gamma\,, \tag{2}$$

where $\Gamma = \Gamma_D \cup \Gamma_N$ and Γ_D has positive measure. A nondimensional form is obtained as follows. Denote with L_r a reference value for length. Then, we choose reference values for velocity and piezometric head as

$$U_r = \frac{\mu}{\rho L_r}\,, \quad p_r = \rho U_r^2\,,$$

and we define the nondimensional unknowns

$$\widetilde{\boldsymbol{u}} = \frac{\boldsymbol{u}}{U_r}\,, \quad \widetilde{p} = \frac{p}{p_r}\,, \quad \widetilde{T} = \frac{T - T_b}{\bar{\Theta}}\,,$$

where $\bar{\Theta}$ is a reference temperature difference. We also define the Rayleigh and Prandtl numbers

$$Ra = \frac{\rho^2 c_p \|\boldsymbol{g}\| \beta \bar{\Theta} L^3}{\mu k}\,, \quad Pr = \frac{\mu}{\rho k}.$$

If the momentum and energy balances are scaled with respect to the corresponding reference diffusion terms, the equations in nondimensional form which will be considered from now on read (with abuse of notation, we drop the tilde sign for the nondimensional quantities)

$$\begin{cases} (\boldsymbol{u} \cdot \nabla)\boldsymbol{u} - \Delta \boldsymbol{u} + \nabla p = \dfrac{Ra}{Pr} \widehat{\boldsymbol{g}} T + \boldsymbol{f}, \\ -\nabla \cdot \boldsymbol{u} = 0, \\ (\boldsymbol{u} \cdot \nabla) T - \dfrac{1}{Pr} \Delta T = g. \end{cases} \tag{3}$$

Here, the vector $\widehat{\boldsymbol{g}} = \boldsymbol{g}/\|\boldsymbol{g}\|$ is the unit vector along the gravity direction. Among the applications of Rayleigh-Bénard convection, we recall boiling water nuclear reactors (BWR), multiphase flows and atmospheric flows [24, 22, 21]. For more details, see [16, 18, 10, 14].

After linearizing the system and discretizing it using a finite element approximation, we obtain a system matrix of the type

$$J = \begin{bmatrix} F & B^{\intercal} & M_1 \\ B & 0 & 0 \\ 0 & 0 & K \end{bmatrix}. \tag{4}$$

As $n \to \infty$, the system of $J \in \mathbb{R}^{n\times n}$ becomes a large and sparse matrix, so a preconditioner needs to be applied to the linearized system. Moreover, one wishes to design a preconditioner that has a number of iterations independent of the matrix size n. To this end, in this work we intend to construct such preconditioners for the Rayleigh-Bénard convection problem, and we do so by starting from a result from [20] according to which the speed of convergence of GMRES is independent of the mesh size if the preconditioner is FOV-equivalent to the system matrix J. The results in [20] concern a block system arising from discretizations of the Navier-Stokes equations. Here, we intend to extend their analysis to the case of the Rayleigh-Bénard problem.

The FOV analysis has been addressed in various works as a tool for the convergence study of preconditioned Krylov subspace methods. In [23] bounds on the rate of convergence for such methods are presented, based on the smallest real part of the field-of-values of the coefficient matrix and its inverse. Here, finite element discretizations of nonsymmetric elliptic BVPs are considered and convergence bounds are given as a function of the mesh size for preconditioners of hierarchical basis or multilevel type. Block triangular preconditioners for the GMRES method applied to nonsymmetric saddle point problems are analyzed in [19]. Here, a result on the rate of convergence for GMRES is provided in terms of certain quantities that depend on the choice of an appropriate matrix norm. Previous versions of such convergence results were obtained by [9] and [6]. In [2] field-of-values estimates are provided for the analysis of preconditioners of augmented Lagrangian type for the Oseen equations. Indefinite preconditioners for the coupled Stokes-Darcy system are studied using field-of-values analysis in [4].

The paper is organized as follows. In Section 2 we provide some preliminary definitions and results needed for the subsequent analysis. Block preconditioners for the Picard linearization of the Rayleigh-Bénard convection problem are described in Section 6. and their norm- and FOV-equivalence to the linearization matrix are proved in Section 6.3. Numerical results showing mesh-independent convergence are reported in Section 7.

2. Finite element approximation of the fixed-point linearized Rayleigh-Bénard problem

Let Ω be a subset of $\mathbb{R}^d$ with Lipschitz-continuous boundary Γ. We denote as $L^2(\Omega)$ the space of square integrable functions with respect to the Lebesgue measure in $\mathbb{R}^d$ and we define $H^1(\Omega) = \{v \in L^2(\Omega) : \nabla v \in L^2(\Omega)\}$. Boldface notations $\boldsymbol{L}^2(\Omega)$ and $\boldsymbol{H}^1(\Omega)$ are used for vector-valued functions. The notations $(\cdot,\cdot)$ and $\|\cdot\|$ denote the standard inner product and induced norm either in $L^2(\Omega)$ or $\boldsymbol{L}^2(\Omega)$ depending on the context. The space of zero-mean L^2 functions is

$$L^2_0(\Omega) = \{p \in L^2(\Omega) \ : \ \int_\Omega p \, d\boldsymbol{x} = 0\} \tag{5}$$

and, given any subset $\Gamma_s \subset \Gamma$, we denote the space $H^1_{\Gamma_s}(\Omega)$ as

$$H^1_{\Gamma_s}(\Omega) = \{u \in H^1(\Omega) \mid \gamma_{\Gamma_s} u = 0\}\,, \tag{6}$$

where γ_{Γ_s} is the trace operator on Γ_s. When $\Gamma_s \equiv \Gamma$, we write $H^1_0(\Omega) = H^1_\Gamma(\Omega)$.

Some properties that are used in the following analysis are reported here. We have the antisymmetry properties [13]

$$((\boldsymbol{a}\cdot\nabla)\boldsymbol{u},\boldsymbol{u}) = 0 \quad \forall \boldsymbol{a} \in \boldsymbol{H}^1(\Omega), \nabla\cdot\boldsymbol{a} = 0 \text{ weakly}, \boldsymbol{a}\cdot\boldsymbol{n}|_\Gamma = 0, \quad \forall \boldsymbol{u} \in \boldsymbol{H}^1(\Omega)\,, \tag{7}$$

$$((\boldsymbol{a}\cdot\nabla)T, T) = 0 \quad \forall \boldsymbol{a} \in \boldsymbol{H}^1(\Omega), \nabla\cdot\boldsymbol{a} = 0 \text{ weakly}, \boldsymbol{a}|_\Gamma = \boldsymbol{0}, \quad \forall T \in H^1(\Omega)\,. \tag{8}$$

Coercivity properties also hold. If $\Gamma_s \subseteq \partial\Omega$ has nonzero measure, then there exists $C_p > 0$ such that

$$\begin{aligned} \|\boldsymbol{u}\|_{L^2} &\le C_p \|\nabla\boldsymbol{u}\|_{L^2} \quad \forall \boldsymbol{u} \in \boldsymbol{H}^1_{\Gamma_s}(\Omega)\,, \\ \|T\|_{L^2} &\le C_p \|\nabla T\|_{L^2} \quad \forall T \in H^1_{\Gamma_s}(\Omega)\,. \end{aligned} \tag{9}$$

For the divergence operator, we may use Hölder's inequality to get

$$\|\nabla\cdot\boldsymbol{u}\|_{L^2} \le \sqrt{3}\|\nabla\boldsymbol{u}\|_{L^2}\,. \tag{10}$$

A weak form of the Rayleigh-Bénard problem for the unknowns $(\boldsymbol{u}, p, T) \in \boldsymbol{H}^1_0(\Omega) \times L^2_0(\Omega) \times H^1_{\Gamma_D}(\Omega)$ reads

$$\begin{aligned} (\nabla\boldsymbol{u},\nabla\boldsymbol{v}) + ((\boldsymbol{u}\cdot\nabla)\boldsymbol{u},\boldsymbol{v}) - (p,\nabla\cdot\boldsymbol{v}) - \frac{Ra}{Pr}(\hat{g}T,\boldsymbol{v}) - (\boldsymbol{f},\boldsymbol{v}) = 0 \quad \forall \boldsymbol{v} \in \boldsymbol{H}^1_0(\Omega)\,, \\ (q,\nabla\cdot\boldsymbol{u}) = 0 \quad \forall q \in L^2_0(\Omega)\,, \\ \frac{1}{Pr}(\nabla T,\nabla r) + ((\boldsymbol{u}\cdot\nabla)T, r) - (g,r) = 0 \quad \forall r \in H^1_{\Gamma_D}(\Omega)\,. \end{aligned}$$

Let us consider conforming finite element approximations on a quasi-uniform triangulation of Ω for the variables $\boldsymbol{u}, p, T$, respectively:

$$\boldsymbol{\Phi}_h \subset \boldsymbol{H}^1_0(\Omega)\,, \quad \Psi_h \subset L^2_0(\Omega)\,, \quad \Theta_h \subset H^1_{\Gamma_D}(\Omega)\,. \tag{11}$$

The spaces $\boldsymbol{\Phi}_h$ and Ψ_h are required to satisfy the inf-sup condition

$$\inf_{p\in\Psi_h} \sup_{\boldsymbol{v}\in\boldsymbol{\Phi}_h} \frac{(p,\nabla\cdot\boldsymbol{u})}{\|\nabla\boldsymbol{v}\|\|p\|} \ge \beta\,. \tag{12}$$

A classical choice that satisfies (12) is given by the Taylor-Hood pair of continuous piecewise-biquadratic (or triquadratic) polynomials for $\boldsymbol{\Phi}_h$ and continuous piecewise-linear polynomials for Ψ_h, i.e.

$$\begin{aligned} \boldsymbol{\Phi}_h &= \Phi_h^n \cap \boldsymbol{H}^1_0(\Omega)\,, \quad \text{with } \Phi_h = \{\phi \in \mathcal{C}^0(\overline{\mathcal{O}}_h) \,:\, \phi|_\kappa \in Q_2(\kappa) \quad \forall\, \kappa \in \mathcal{O}_h\}\,, \\ \Psi_h &= \widetilde{\Psi}_h \cap L^2_0(\Omega)\,, \quad \text{with } \widetilde{\Psi}_h = \{\psi \in \mathcal{C}^0(\overline{\mathcal{O}}_h) \,:\, \psi|_\kappa \in P_1(\kappa) \quad \forall\, \kappa \in \mathcal{O}_h\}\,. \end{aligned}$$

Also, we choose $\Theta_h = \Phi_h \cap H^1_{\Gamma_D}(\Omega)$. For compactness, we denote $\boldsymbol{V}_h := \boldsymbol{\Phi}_h \times \Psi_h \times \Theta_h$.

Thanks to the Poincaré inequalities (9), we may define an inner product $(\cdot,\cdot)_a$ on $\boldsymbol{V}_h \times \boldsymbol{V}_h$ with induced norm $\|\cdot\|_a$ on $\boldsymbol{V}_h$ as

$$((\boldsymbol{u}_h, p_h, T_h), (\boldsymbol{v}_h, q_h, r_h))_a := (\nabla \boldsymbol{u}_h, \nabla \boldsymbol{v}_h) + \frac{1}{Pr}(\nabla T_h, \nabla r_h) + (p_h, q_h)\,, \quad (13)$$

$$\|(\boldsymbol{u}_h, p_h, T_h)\|_a := \sqrt{((\boldsymbol{u}_h, p_h, T_h), (\boldsymbol{u}_h, p_h, T_h))_a} \quad (14)$$

Let $\boldsymbol{a}_h \in \boldsymbol{\Phi}_h$, $\nabla \cdot \boldsymbol{a}_h = 0$ weakly, be the velocity at the previous iteration in the Picard method. If we define a bilinear form on $\boldsymbol{V}_h \times \boldsymbol{V}_h$ given by

$$\begin{aligned} a((\boldsymbol{u}_h, p_h, T_h), (\boldsymbol{v}_h, q_h, r_h)) &= (\nabla \boldsymbol{u}_h, \nabla \boldsymbol{v}_h) + ((\boldsymbol{a}_h \cdot \nabla)\boldsymbol{u}_h, \boldsymbol{v}_h) - (p_h, \nabla \cdot \boldsymbol{v}_h) \\ &\quad - \frac{Ra}{Pr}(\hat{g} T_h, \boldsymbol{v}_h) + (q_h, \nabla \cdot \boldsymbol{u}_h) + \frac{1}{Pr}(\nabla T_h, \nabla r_h) + ((\boldsymbol{a}_h \cdot \nabla) T_h, r_h)\,, \quad (15) \end{aligned}$$

then the finite element problem consists in finding $(\boldsymbol{u}_h, p_h, T_h) \in \boldsymbol{V}_h$ such that

$$a((\boldsymbol{u}_h, p_h, T_h), (\boldsymbol{v}_h, q_h, r_h)) = (\boldsymbol{f}, \boldsymbol{v}_h) + (g, r_h) \quad \forall (\boldsymbol{v}_h, q_h, r_h) \in \boldsymbol{V}_h\,. \quad (16)$$

2.1. Sup-sup and inf-sup properties

The following lemma identifies a sufficient condition involving Ra and Pr for the bilinear form $a(\cdot,\cdot)$ to satisfy certain sup-sup and inf-sup properties. In turn, these properties are sufficient conditions for a convergence result on the preconditioned GMRES algorithm that will be considered later.

Lemma 1. *Let the product space $\boldsymbol{V}_h$ be such that* (11) *and* (12) *hold. Then, there exists a positive constant C_2 independent of the mesh size such that*

$$\sup_{(\boldsymbol{u}_h, p_h, T_h) \in \boldsymbol{V}_h} \sup_{(\boldsymbol{v}_h, q_h, r_h) \in \boldsymbol{V}_h} \frac{a((\boldsymbol{u}_h, p_h, T_h), (\boldsymbol{v}_h, q_h, r_h))}{\|(\boldsymbol{u}_h, p_h, T_h)\|_a \|(\boldsymbol{v}_h, q_h, r_h)\|_a} \leq C_2\,. \quad (17)$$

Moreover, if Pr and Ra satisfy

$$\frac{Ra}{\sqrt{Pr}} < 2C_p^2 \quad (18)$$

where C_p is the constant in Poincaré's inequalities (9)*, there exists a positive constant C_1 independent of the mesh size such that*

$$\inf_{(\boldsymbol{u}_h, p_h, T_h) \in \boldsymbol{V}_h} \sup_{(\boldsymbol{v}_h, q_h, r_h) \in \boldsymbol{V}_h} \frac{a((\boldsymbol{u}_h, p_h, T_h), (\boldsymbol{v}_h, q_h, r_h))}{\|(\boldsymbol{u}_h, p_h, T_h)\|_a \|(\boldsymbol{v}_h, q_h, r_h)\|_a} \geq C_1\,. \quad (19)$$

Proof. We begin with condition (17). Using the Poincaré inequalities, the continuity of the divergence operator (10), the Cauchy-Schwarz inequality in $\mathbb{R}^2$

and $\mathbb{R}^3$, and Young's inequality, we have

$$
\begin{aligned}
&a((\boldsymbol{u}_h, p_h, T_h), (\boldsymbol{v}_h, q_h, r_h)) \\
&\le \|\nabla \boldsymbol{u}_h\| \|\nabla \boldsymbol{v}_h\| + \|\boldsymbol{a}_h\|_{L^\infty} \|\nabla \boldsymbol{u}_h\| \|\nabla \boldsymbol{v}_h\| + \|p_h\| \|\nabla \cdot \boldsymbol{v}_h\| \\
&+ \|q_h\| \|\nabla \cdot \boldsymbol{u}_h\| + \frac{Ra}{Pr} \|T_h\| \|\boldsymbol{v}_h\| + \frac{1}{Pr} \|\nabla T_h\| \|\nabla r_h\| + \|\boldsymbol{a}_h\|_{L^\infty} \|\nabla T_h\| \|r_h\| \\
&\le \|\nabla \boldsymbol{u}_h\| \|\nabla \boldsymbol{v}_h\| + \|\boldsymbol{a}_h\|_{L^\infty} \|\nabla \boldsymbol{u}_h\| \|\nabla \boldsymbol{v}_h\| + \sqrt{3} \|p_h\| \|\nabla \boldsymbol{v}_h\| + \sqrt{3} \|q_h\| \|\nabla \boldsymbol{u}_h\| \\
&+ \frac{Ra}{Pr} C_p^2 \|\nabla T_h\| \|\nabla \boldsymbol{v}_h\| + \frac{1}{Pr} \|\nabla T_h\| \|\nabla r_h\| + C_p \|\boldsymbol{a}_h\|_{L^\infty} \|\nabla T_h\| \|\nabla r_h\| \\
&\le (1 + \|\boldsymbol{a}_h\|_{L^\infty}) \|\nabla \boldsymbol{u}_h\| \|\nabla \boldsymbol{v}_h\| + \sqrt{3} \|p_h\| \|\nabla \boldsymbol{v}_h\| \\
&+ \sqrt{3} \|q_h\| \|\nabla \boldsymbol{u}_h\| + \frac{Ra}{Pr} C_p^2 \|\nabla T_h\| \|\nabla \boldsymbol{v}_h\| + (\frac{1}{Pr} + C_p \|\boldsymbol{a}_h\|_{L^\infty}) \|\nabla T_h\| \|\nabla r_h\| \\
&\le (1 + \|\boldsymbol{a}_h\|_{L^\infty}) \|\nabla \boldsymbol{u}_h\| \|\nabla \boldsymbol{v}_h\| + \sqrt{3} \|p_h\| \|\nabla \boldsymbol{v}_h\| \\
&+ \sqrt{3} \|q_h\| \|\nabla \boldsymbol{u}_h\| + \frac{Ra}{Pr} C_p^2 \|\nabla T_h\| \|\nabla \boldsymbol{v}_h\| + (\frac{1}{Pr} + C_p \|\boldsymbol{a}_h\|_{L^\infty}) \|\nabla T_h\| \|\nabla r_h\| \\
&\le \|\nabla \boldsymbol{u}_h\| \left((1 + \|\boldsymbol{a}_h\|_{L^\infty}) \|\nabla \boldsymbol{v}_h\| + \sqrt{3} \|q_h\| \right) + \left(\sqrt{3} \|p_h\| + \frac{Ra}{Pr} C_p^2 \|\nabla T_h\| \right) \|\nabla \boldsymbol{v}_h\| \\
&+ \left((\frac{1}{Pr} + C_p \|\boldsymbol{a}_h\|_{L^\infty}) \|\nabla T_h\| \right) \|\nabla r_h\| \\
&\le \sqrt{\|\nabla \boldsymbol{u}_h\|^2 + \left(\sqrt{3} \|p_h\| + \frac{Ra}{Pr} C_p^2 \|\nabla T_h\| \right)^2 + \left((\frac{1}{Pr} + C_p \|\boldsymbol{a}_h\|_{L^\infty}) \|\nabla T_h\| \right)^2} \\
&\sqrt{\left((1 + \|\boldsymbol{a}_h\|_{L^\infty}) \|\nabla \boldsymbol{v}_h\| + \sqrt{3} \|q_h\| \right)^2 + \|\nabla \boldsymbol{v}_h\|^2 + \|\nabla r_h\|^2} \\
&\le \sqrt{\|\nabla \boldsymbol{u}_h\|^2 + 2 \left(3 \|p_h\|^2 + (\frac{Ra}{Pr} C_p^2)^2 \|\nabla T_h\|^2 \right) + (\frac{1}{Pr} + C_p \|\boldsymbol{a}_h\|_{L^\infty})^2 \|\nabla T_h\|^2} \\
&\sqrt{2 \left((1 + \|\boldsymbol{a}_h\|_{L^\infty})^2 \|\nabla \boldsymbol{v}_h\|^2 + 3 \|q_h\|^2 \right) + \|\nabla \boldsymbol{v}_h\|^2 + \|\nabla r_h\|^2} \\
&\le \sqrt{\|\nabla \boldsymbol{u}_h\|^2 + 6 \|p_h\|^2 + 2 (\frac{Ra}{Pr} C_p^2)^2 \|\nabla T_h\|^2 + (\frac{1}{Pr} + C_p \|\boldsymbol{a}_h\|_{L^\infty})^2 \|\nabla T_h\|^2} \\
&\sqrt{2 (1 + \|\boldsymbol{a}_h\|_{L^\infty})^2 \|\nabla \boldsymbol{v}_h\|^2 + 6 \|q_h\|^2 + \|\nabla \boldsymbol{v}_h\|^2 + \|\nabla r_h\|^2} \\
&\le \sqrt{\|\nabla \boldsymbol{u}_h\|^2 + 6 \|p_h\|^2 + \left(2 (\frac{Ra}{Pr} C_p^2)^2 + (\frac{1}{Pr} + C_p \|\boldsymbol{a}_h\|_{L^\infty})^2 \right) \|\nabla T_h\|^2} \\
&\sqrt{\left(2 (1 + \|\boldsymbol{a}_h\|_{L^\infty})^2 + 1 \right) \|\nabla \boldsymbol{v}_h\|^2 + 6 \|q_h\|^2 + \|\nabla r_h\|^2}
\end{aligned}
$$

where the last two factors are equivalent to the norm $\|\cdot\|_a$.

Inf-sup condition. We move along the lines of [12]. Let $(\boldsymbol{u}_h, p_h, T_h) \in \boldsymbol{\Phi}_h \times \Psi_h \times \Theta_h$. We first need to construct some $\overline{(\boldsymbol{v}_h, q_h, r_h)}$ such that

$$
a((\boldsymbol{u}_h, p_h, T_h), \overline{(\boldsymbol{v}_h, q_h, r_h)}) \ge \beta \|\overline{(\boldsymbol{v}_h, q_h, r_h)}\|^2 . \tag{20}
$$

We construct $\overline{(\boldsymbol{v}_h, q_h, r_h)}$ by first looking at $a((\boldsymbol{u}_h, p_h, T_h), (\boldsymbol{u}_h, p_h, T_h))$. Us-

ing the skew symmetry properties (7) and Young's inequality, we have

$$\begin{aligned} a((\boldsymbol{u}_h, p_h, T_h), (\boldsymbol{u}_h, p_h, T_h)) &= (\nabla \boldsymbol{u}_h, \nabla \boldsymbol{u}_h) - \frac{Ra}{Pr}(\hat{g} T_h, \boldsymbol{u}_h) + \frac{1}{Pr}(\nabla T_h, \nabla T_h) \\ &\geq \|\nabla \boldsymbol{u}_h\|^2 + \frac{1}{Pr}\|\nabla T_h\|^2 - \frac{Ra}{Pr}(\hat{g} T_h, \boldsymbol{u}_h) \\ &\geq \|\nabla \boldsymbol{u}_h\|^2 \left(1 - \frac{Ra}{Pr}\frac{1}{C_p^2}\epsilon\right) + \|\nabla T_h\|^2 \left(\frac{1}{Pr} - \frac{Ra}{Pr}\frac{1}{C_p^2}\frac{1}{4\epsilon}\right). \end{aligned}$$

Choose $\epsilon > 0$ such that

$$\Xi := \left(1 - \frac{Ra}{Pr}\frac{1}{C_p^2}\epsilon\right) > 0\,, \quad \Psi := \left(\frac{1}{Pr} - \frac{Ra}{Pr}\frac{1}{C_p^2}\frac{1}{4\epsilon}\right) > 0\,. \tag{21}$$

This is possible due to the hypothesis (18). Next, due to the inf-sup condition (12), there exists a function $\boldsymbol{z}_h$ such that

$$\frac{(p_h, \nabla \cdot \boldsymbol{z}_h)}{\|\nabla \boldsymbol{z}_h\| \|p_h\|} \geq \beta\,. \tag{22}$$

This function can be chosen without loss of generality such that $\|\nabla \boldsymbol{z}_h\| = \|p_h\|$. Then,

$$\begin{aligned} &a((\boldsymbol{u}_h, p_h, T_h), (-\boldsymbol{z}_h, 0, 0)) = \\ &\quad - (\nabla \boldsymbol{u}_h, \nabla \boldsymbol{z}_h) - ((\boldsymbol{a}_h \cdot \nabla)\boldsymbol{u}_h, \boldsymbol{z}_h) + (p_h, \nabla \cdot \boldsymbol{z}_h) + \frac{Ra}{Pr}(\hat{g} T_h, \boldsymbol{z}_h) \\ &\quad\quad = (p_h, \nabla \cdot \boldsymbol{z}_h) + \frac{Ra}{Pr}(\hat{g} T_h, \boldsymbol{z}_h) - (\nabla \boldsymbol{u}_h, \nabla \boldsymbol{z}_h) + ((\boldsymbol{a}_h \cdot \nabla)\boldsymbol{z}_h, \boldsymbol{u}_h)\,. \end{aligned}$$

Define the quantity

$$\Upsilon = -\frac{Ra}{Pr}(\hat{g} T_h, \boldsymbol{z}_h) + (\nabla \boldsymbol{u}_h, \nabla \boldsymbol{z}_h) - ((\boldsymbol{a}_h \cdot \nabla)\boldsymbol{z}_h, \boldsymbol{u}_h)\,.$$

Then

$$\begin{aligned} \Upsilon &\leq \|\nabla \boldsymbol{u}_h\| \|p_h\| + C_p \|\boldsymbol{a}_h\|_{L^\infty} \|\nabla \boldsymbol{u}_h\| \|p_h\| + \frac{Ra}{Pr}\|T_h\| \|\boldsymbol{z}_h\| \\ &\leq \|\nabla \boldsymbol{u}_h\| \|p_h\| + C_p \|\boldsymbol{a}_h\|_{L^\infty} \|\nabla \boldsymbol{u}_h\| \|p_h\| + \frac{Ra}{Pr} C_p \|T_h\| \|\nabla \boldsymbol{z}_h\| \\ &\leq \|\nabla \boldsymbol{u}_h\| \|p_h\| + C_p \|\boldsymbol{a}_h\|_{L^\infty} \|\nabla \boldsymbol{u}_h\| \|p_h\| + \frac{Ra}{Pr} C_p \|T_h\| \|p_h\| \\ &= \left(\|\nabla \boldsymbol{u}_h\|(1 + C_p \|\boldsymbol{a}_h\|_{L^\infty}) + C_p \frac{Ra}{Pr}\|T_h\|\right) \|p_h\| \\ &\leq \left(\|\nabla \boldsymbol{u}_h\|(1 + C_p \|\boldsymbol{a}_h\|_{L^\infty}) + C_p^2 \frac{Ra}{Pr}\|\nabla T_h\|\right) \|p_h\|\,. \end{aligned}$$

If we set

$$\zeta_a = \max\left\{1 + C_p \|\boldsymbol{a}_h\|_{L^\infty}, C_p^2 \frac{Ra}{Pr}\right\}$$

we obtain

$$\Upsilon \leq \zeta_a \left(\|\nabla \boldsymbol{u}_h\| + \|\nabla T_h\| \right) \|p_h\| ,$$

so that

$$a((\boldsymbol{u}_h, p_h, T_h), (-\boldsymbol{z}_h, 0, 0)) \geq (p_h, \nabla \cdot \boldsymbol{z}_h) - \Upsilon \geq \beta \|p_h\|^2 - \zeta_a \left(\|\nabla \boldsymbol{u}_h\| + \|\nabla T_h\| \right) \|p_h\| .$$

Define $G = \|\nabla \boldsymbol{u}_h\| + \|\nabla T_h\|$, $B = \|p_h\|$. Then Young's inequality gives

$$\frac{\zeta_a G B}{\sqrt{\beta}} \sqrt{\beta} \leq \frac{1}{2} \beta B^2 + \frac{1}{2} \frac{\zeta_a^2 G^2}{\beta} .$$

Hence,

$$a((\boldsymbol{u}_h, p_h, T_h), (-\boldsymbol{z}_h, 0, 0)) \geq \frac{1}{2} \beta \|p_h\|^2 - \frac{\zeta_a^2}{\beta} \left(\|\nabla \boldsymbol{u}_h\|^2 + \|\nabla T_h\|^2 \right) .$$

Now, we can construct

$$\overline{(\boldsymbol{v}_h, q_h, r_h)} = (\boldsymbol{u}_h, p_h, T_h) + \rho_a (-\boldsymbol{z}_h, 0, 0)$$

with ρ_a to be determined. Clearly,

$$\begin{aligned}
& a((\boldsymbol{u}_h, p_h, T_h), \overline{(\boldsymbol{v}_h, q_h, r_h)}) \\
& \geq (\min\{\Xi, \Psi\} - \rho_a \frac{\zeta_a^2}{\beta}) \left(\|\nabla \boldsymbol{u}_h\|^2 + \|\nabla T_h\|^2 \right) + \rho_a \frac{1}{2} \frac{\beta}{\sigma_a} \sigma_a \|p_h\|^2 \\
& \geq \min \left\{ \min\{\Xi, \Psi\} - \rho_a \frac{\zeta_a^2}{\beta} , \rho_a \frac{1}{2} \frac{\beta}{\sigma_a} \right\} \left(\|\nabla \boldsymbol{u}_h\|^2 + \|\nabla T_h\|^2 + \sigma_a \|p_h\|^2 \right) \\
& = \min \left\{ \min\{\Xi, \Psi\} - \rho_a \frac{\zeta_a^2}{\beta} , \rho_a \frac{1}{2} \frac{\beta}{\sigma_a} \right\} \|(\boldsymbol{u}_h, p_h, T_h)\|_{\sigma_a}^2 \\
& \geq \min \left\{ \min\{\Xi, \Psi\} - \rho_a \frac{\zeta_a^2}{\beta} , \rho_a \frac{1}{2} \frac{\beta}{\sigma_a} \right\} C_{Pr} \|(\boldsymbol{u}_h, p_h, T_h)\|_a^2
\end{aligned}$$

where C_{Pr} is a constant that depends on Pr, and for $\sigma_a > 0$ we defined the norm

$$\|(\boldsymbol{u}_h, p_h, T_h)\|_{\sigma_a} := \sqrt{(\nabla \boldsymbol{u}_h, \nabla \boldsymbol{u}_h) + (\nabla T_h, \nabla T_h) + \sigma_a (p_h, p_h)} .$$

which is equivalent to $\|\cdot\|_a$. We may choose the constants $\rho_a > 0$ and $\sigma_a > 0$ such that

$$\begin{aligned}
\sigma_a &= 2\beta\rho_a \\
\rho_a \frac{\zeta_a^2}{\beta} &< \min\{\Xi, \Psi\} .
\end{aligned}$$

Moreover,

$$\|(-\boldsymbol{z}_h, 0, 0)\|_a^2 = \|\nabla \boldsymbol{z}_h\|^2 = \|p_h\|^2 \leq \|(\boldsymbol{u}_h, p_h, T_h)\|_a^2$$

Therefore, using the arithmetic-geometric mean inequality,

$$\begin{aligned}\|\overline{(\boldsymbol{v}_h, q_h, r_h)}\|_a^2 &\le 2\left(\|(\boldsymbol{u}_h, p_h, T_h)\|_a^2 + \rho_a^2\|(-\boldsymbol{z}_h, 0, 0)\|_a^2\right)\\ &\le 2(1+\rho_a^2)\|(\boldsymbol{u}_h, p_h, T_h)\|_a^2\end{aligned}$$

Hence

$$\sup_{(\boldsymbol{v}_h, q_h, r_h)} \frac{a((\boldsymbol{u}_h, p_h, T_h), (\boldsymbol{v}_h, q_h, r_h))}{\|(\boldsymbol{u}_h, p_h, T_h)\|_a \|(\boldsymbol{v}_h, q_h, r_h)\|_a} \ge \frac{a((\boldsymbol{u}_h, p_h, T_h), \overline{(\boldsymbol{v}_h, q_h, r_h)})}{\|(\boldsymbol{u}_h, p_h, T_h)\|_a \|\overline{(\boldsymbol{v}_h, q_h, r_h)}\|_a} \ge C\,.$$

Taking the infimum over $(\boldsymbol{u}_h, p_h, T_h)$ completes the proof. □

3. Matrix description of the FE problem

If we denote

$$\boldsymbol{\Phi}_h = \operatorname{span}\{\boldsymbol{\phi}_i\}_{i=1}^{n_1}, \quad \Psi_h = \operatorname{span}\{\psi_i\}_{i=1}^{n_2}, \quad \Theta_h = \operatorname{span}\{\theta_i\}_{i=1}^{n_3},$$

we define the matrices

$$(F)_{ij} = (\nabla\boldsymbol{\phi}_j, \nabla\boldsymbol{\phi}_i) + ((\boldsymbol{a}_h\cdot\nabla)\boldsymbol{\phi}_j, \boldsymbol{\phi}_i)\,, (B)_{ij} = (\nabla\cdot\boldsymbol{\phi}_j, \psi_i)\,, \tag{23}$$

$$(M_1)_{ij} = -\frac{Ra}{Pr}(\hat{g}\theta_j, \boldsymbol{\phi}_i)\,, \quad K_{ij} = \frac{1}{Pr}(\nabla\theta_j, \nabla\theta_i) + ((\boldsymbol{a}_h\cdot\nabla)\theta_j, \theta_i)\,. \tag{24}$$

Notice that F and K are not symmetric. We also define the vectors $\widetilde{\boldsymbol{f}}$ and $\widetilde{\boldsymbol{g}}$ as

$$(\widetilde{\boldsymbol{f}})_i = (\boldsymbol{f}, \boldsymbol{\phi}_i)\,, \quad (\widetilde{\boldsymbol{g}})_i = (g, \theta_i)\,. \tag{25}$$

Then, the Picard-linearized finite element Rayleigh-Bénard problem (16) is equivalent to the matrix problem of finding $[\boldsymbol{u}^\intercal, p^\intercal, T^\intercal]^\intercal \in \mathbb{R}^{n_1+n_2+n_3}$ such that

$$\begin{bmatrix} F & B^\intercal & M_1 \\ B & 0 & 0 \\ 0 & 0 & K \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ p \\ T \end{bmatrix} = \begin{bmatrix} \widetilde{\boldsymbol{f}} \\ \boldsymbol{0} \\ \widetilde{\boldsymbol{g}} \end{bmatrix}. \tag{26}$$

We denote the matrix in (26) as J. In view of the following definition of preconditioners for the system (26), we denote the Schur complement of F with respect to the Navier-Stokes block as $S = -BF^{-1}B^\intercal$, and we set

$$(A_p)_{ij} = (\nabla\psi_j, \nabla\psi_i)\,, \quad (F_p)_{ij} = ((\boldsymbol{a}_h\cdot\nabla)\psi_j, \nabla\psi_i) + (\nabla\psi_j, \nabla\psi_i)\,. \tag{27}$$

The blocks A_p and F_p are the Laplace operator and the convection-diffusion operator in the space for p. With the symmetric positive-definite matrices

$$(A_{\boldsymbol{u}})_{ij} = (\nabla\boldsymbol{\phi}_j, \nabla\boldsymbol{\phi}_i) = \frac{F+F^\intercal}{2},\ (M_p)_{ij} = (\psi_j, \psi_i),\ (A_T)_{ij} = (\nabla\theta_j, \nabla\theta_i) = \frac{K+K^\intercal}{2},$$

we may define the matrix $H \in \mathbb{R}^{n\times n}$ as

$$H = \begin{bmatrix} H_1 & 0 & 0 \\ 0 & H_2 & 0 \\ 0 & 0 & H_3 \end{bmatrix} = \begin{bmatrix} A_{\boldsymbol{u}} & 0 & 0 \\ 0 & M_p & 0 \\ 0 & 0 & A_T \end{bmatrix}, \tag{28}$$

as well as the weighted scalar product and norm

$$< \boldsymbol{u}, \boldsymbol{v} >_H = \boldsymbol{v}^\intercal H \boldsymbol{u}\,, \|\boldsymbol{u}\|_H = \sqrt{< \boldsymbol{u}, \boldsymbol{u} >_H}\,. \tag{29}$$

We remind that the square root matrices $H_i^{1/2}$ are well-defined since H_i are symmetric positive-definite. Notice that, if $(\boldsymbol{u}_h, p_h, T_h)$ and $(\boldsymbol{v}_h, q_h, r_h)$ have column vectors of finite element degrees of freedom denoted by $[\boldsymbol{u}^\intercal, p^\intercal, T^\intercal]^\intercal$ and $[\boldsymbol{v}^\intercal, q^\intercal, r^\intercal]^\intercal$ respectively, then by definition

$$\|(\boldsymbol{u}_h, p_h, T_h)\|_a = \|[\boldsymbol{u}^\intercal, p^\intercal, T^\intercal]^\intercal\|_H\,, \tag{30}$$

$$a((\boldsymbol{u}_h, p_h, T_h), (\boldsymbol{v}_h, q_h, r_h)) = [\boldsymbol{v}^\intercal, q^\intercal, r^\intercal] J [\boldsymbol{u}^\intercal, p^\intercal, T^\intercal]^\intercal\,. \tag{31}$$

We also provide a definition of matrix norm. Given two symmetric and positive-definite matrices $H_a \in \mathbb{R}^{n\times n}$ and $H_b \in \mathbb{R}^{m\times m}$, we define for any $Q \in \mathbb{R}^{m\times n}$ the matrix norm

$$\|Q\|_{H_a,H_b} = \max_{\boldsymbol{w}\in\mathbb{R}^n\setminus\{\boldsymbol{0}\}} \frac{\|Q\boldsymbol{w}\|_{H_b}}{\|\boldsymbol{w}\|_{H_a}}. \tag{32}$$

When $m = n$ and $H_a = H_b = \overline{H}$, we denote $\|Q\|_{\overline{H},\overline{H}} = \|Q\|_{\overline{H}}$. No confusion should arise from the context with respect to the similar notation of the vector norm (29). We recall some results about this matrix norm from [20] that are used several times in this work.

Lemma 2. *Let H_1, H_2 and H_3 be symmetric positive definite matrices. Given $R \in \mathbb{R}^{n_1\times n_2}$, $Q \in \mathbb{R}^{n_2\times n_3}$, the following hold:*

$$\|RQ\|_{H_3,H_1} \le \|Q\|_{H_3,H_2} \|R\|_{H_2,H_1}\,, \tag{33}$$

$$\left\|H_2^{-1/2} Q H_1^{-1/2}\right\|_{l_2} = \|Q\|_{H_1,H_2^{-1}} = \left\|QH_1^{-1}\right\|_{H_1^{-1},H_2^{-1}} = \left\|H_2^{-1}Q\right\|_{H_1,H_2}\,. \tag{34}$$

Lemma 3. *Let $Q \in \mathbb{R}^{m\times n}$ have full rank and let $H_a \in \mathbb{R}^{n\times n}$ and $H_b \in \mathbb{R}^{m\times m}$ be two symmetric and positive definite matrices. Then*

$$\|Q\|_{H_a,H_b^{-1}} = \max_{\boldsymbol{v}\in\mathbb{R}^n\setminus\{\boldsymbol{0}\}} \max_{\boldsymbol{w}\in\mathbb{R}^m\setminus\{\boldsymbol{0}\}} \frac{\boldsymbol{w}^\intercal Q \boldsymbol{v}}{\|\boldsymbol{v}\|_{H_a} \|\boldsymbol{w}\|_{H_b}}, \tag{35}$$

$$\min_{\boldsymbol{v}\notin\ker(Q)} \frac{\|Q\boldsymbol{v}\|_{H_b^{-1}}}{\|\boldsymbol{v}\|_{H_a}} = \min_{\boldsymbol{v}\notin\ker(Q)} \max_{\boldsymbol{w}\in\mathbb{R}^m\setminus\{\boldsymbol{0}\}} \frac{\boldsymbol{w}^\intercal Q \boldsymbol{v}}{\|\boldsymbol{v}\|_{H_a} \|\boldsymbol{w}\|_{H_b}}. \tag{36}$$

If $n = m$,

$$\left\|Q^{-1}\right\|^{-1}_{H_b^{-1},H_a} = \min_{\boldsymbol{v}\in\mathbb{R}^n\setminus\{\boldsymbol{0}\}} \max_{\boldsymbol{w}\in\mathbb{R}^n\setminus\{\boldsymbol{0}\}} \frac{\boldsymbol{w}^\intercal Q \boldsymbol{v}}{\|\boldsymbol{v}\|_{H_a} \|\boldsymbol{w}\|_{H_b}}. \tag{37}$$

4. A GMRES convergence result using FOV-equivalence

Here, we discuss certain notions of equivalence between matrices as introduced in [20]. These definitions lead to a fundamental sufficient condition for GMRES convergence. We begin with the concept of H-norm equivalence.

Definition 1 (H-norm equivalence)**.** *Nonsingular matrices $R, Q \in \mathbb{R}^{n\times n}$ are said to be H-norm equivalent if there exist constants $\gamma, \Gamma > 0$ independent of n such that for all $\boldsymbol{x} \in \mathbb{R}^n \backslash \{\mathbf{0}\}$*

$$\gamma \leq \frac{\|R\boldsymbol{x}\|_H}{\|Q\boldsymbol{x}\|_H} \leq \Gamma.$$

We write

$$R \sim_H Q.$$

We now prove the following result.

Lemma 4. *$R \sim_H Q$ if and only if $\left\|RQ^{-1}\right\|_H \leq \Gamma$ and $\left\|QR^{-1}\right\|_H \leq \gamma^{-1}$.*

Proof. Let $R \sim_H Q$, then by Definition 1, we have $\gamma \leq \frac{\|R\boldsymbol{x}\|_H}{\|Q\boldsymbol{x}\|_H} \leq \Gamma$. Let $\boldsymbol{v} = Q\boldsymbol{x}$, then $\frac{\left\|RQ^{-1}\boldsymbol{v}\right\|_H}{\|\boldsymbol{v}\|_H} \leq \Gamma$. Using equation (32), inequality $\left\|RQ^{-1}\right\|_H \leq \Gamma$ follows. Similarly, we have $\frac{\|Q\boldsymbol{x}\|_H}{\|R\boldsymbol{x}\|_H} \leq \gamma^{-1}$. Let $\boldsymbol{v} = R\boldsymbol{x}$, then $\frac{\left\|QR^{-1}\boldsymbol{x}\right\|_H}{\|\boldsymbol{x}\|_H} \leq \gamma^{-1}$. Using equation (32) again, inequality $\left\|QR^{-1}\right\|_H \leq \gamma^{-1}$ follows. The reverse of the proof is straightforward. □

We now provide the definition of Field-of-values equivalence.

Definition 2 (FOV equivalence)**.** *Nonsingular matrices $R, Q \in \mathbb{R}^{n\times n}$ are said to be FOV-equivalent if there exist constants $\gamma, \Gamma > 0$ independent of n such that for all $\boldsymbol{x} \in \mathbb{R}^n \backslash \{\mathbf{0}\}$,*

$$\gamma \leq \frac{< \boldsymbol{x}, RQ^{-1}\boldsymbol{x} >_H}{< \boldsymbol{x}, \boldsymbol{x} >_H}, \quad \frac{\left\|RQ^{-1}\boldsymbol{x}\right\|_H}{\|\boldsymbol{x}\|_H} \leq \Gamma.$$

We write

$$R \approx_H Q.$$

Notice that the FOV-equivalence definition is given for a certain SPD matrix H and is a stronger statement than H-norm equivalence: if $R \approx_H Q$, then $R \sim_H Q$. Also, while H-norm equivalence is an equivalence relation on the set of nonsingular matrices, the term "FOV equivalence" is somewhat misleading since this relation is not symmetric.

4.1. General convergence result for GMRES

Here we state the main convergence result that is used in this work, see [20]. For the sake of clarity, we state it for the case of right preconditioning which is addressed in this work, although a left preconditioning version also holds. Notice that the theorem relies on a notion of FOV-equivalence between the system matrix and its preconditioner.

Theorem 1 (Generalized Minimum Residual (GMRES))**.** *If*

$$R \approx_{H^{-1}} Q, \tag{38}$$

then the GMRES algorithm applied to R with right preconditioner Q converges with respect to $< \cdot, \cdot >_{H^{-1}}$ in a number of iterations independent of n. Moreover, the residuals satisfy

$$\frac{\left\|r^k\right\|_{H^{-1}}}{\|r^0\|_{H^{-1}}} \leq \left(1 - \frac{\gamma^2}{\Gamma^2}\right)^{k/2},$$

where γ and Γ are the constants in Definition 2.

5. Linear algebra results implied by FE results

Here we briefly state how the properties in the finite element spaces transfer to the matrix blocks. We notice that some properties depend on the assumption (18), while other results on single matrix blocks have general validity. From the strong coercivity properties 9, we have the corresponding strong and weak coercivity results for F and K.

Theorem 2. *There exists constants $\eta, \xi > 0$ such that*

$$\boldsymbol{v}^\intercal F \boldsymbol{v} \geq \eta \|\boldsymbol{v}\|_{H_1}^2 , \quad \boldsymbol{v}^\intercal K \boldsymbol{v} \geq \xi \|\boldsymbol{v}\|_{H_3}^2 . \tag{39}$$

Hence, there exist constants $C_3 > 0$, $C_4 > 0$ such that

$$\begin{aligned} \min_{\boldsymbol{w} \in \mathbb{R}^{n_1} \setminus \{\boldsymbol{0}\}} \max_{\boldsymbol{v} \in \mathbb{R}^{n_1} \setminus \{\boldsymbol{0}\}} \frac{\boldsymbol{w}^\intercal F \boldsymbol{v}}{\|\boldsymbol{w}\|_{H_1} \|\boldsymbol{v}\|_{H_1}} &\geq C_3, \\ \min_{\boldsymbol{w} \in \mathbb{R}^{n_3} \setminus \{\boldsymbol{0}\}} \max_{\boldsymbol{v} \in \mathbb{R}^{n_3} \setminus \{\boldsymbol{0}\}} \frac{\boldsymbol{w}^\intercal K \boldsymbol{v}}{\|\boldsymbol{w}\|_{H_3} \|\boldsymbol{v}\|_{H_3}} &\geq C_4 . \end{aligned} \tag{40}$$

Theorem 3. *Condition* (17) *is equivalent to*

$$\max_{\boldsymbol{w} \in \mathbb{R}^{n} \setminus \{\boldsymbol{0}\}} \max_{\boldsymbol{v} \in \mathbb{R}^{n} \setminus \{\boldsymbol{0}\}} \frac{\boldsymbol{w}^\intercal J \boldsymbol{v}}{\|\boldsymbol{w}\|_{H} \|\boldsymbol{v}\|_{H}} \leq C_1 . \tag{41}$$

Moreover, if (18) *holds, condition* (19) *is equivalent to*

$$\min_{\boldsymbol{w} \in \mathbb{R}^{n} \setminus \{\boldsymbol{0}\}} \max_{\boldsymbol{v} \in \mathbb{R}^{n} \setminus \{\boldsymbol{0}\}} \frac{\boldsymbol{w}^\intercal J \boldsymbol{v}}{\|\boldsymbol{w}\|_{H} \|\boldsymbol{v}\|_{H}} \geq C_2 . \tag{42}$$

Proof. It is an immediate consequence of Theorem 1 with the given definitions of the norms and of the bilinear forms, see (30). □

Using the notion of H-norm equivalence, we have the following results involving the linearization matrix J in (26) and the matrix H defining the vector norm (29).

Lemma 5. *Let H as in* (28) *and let* (18) *hold. Then*

$$H \sim_{H^{-1}} J, \quad H^{-1} \sim_H J^{-1}. \tag{43}$$

In particular,

$$\left\|H^{-1}J\right\|_H = \left\|JH^{-1}\right\|_{H^{-1}} \leq C_1, \tag{44}$$

$$\left\|J^{-1}H\right\|_H = \left\|HJ^{-1}\right\|_{H^{-1}} \leq C_2^{-1}. \tag{45}$$

Moreover, if $P \in \mathbb{R}^{n\times n}$ satisfies $P \sim_{H^{-1}} H$, then

$$P \sim_{H^{-1}} J, \quad P^{-1} \sim_H J^{-1}. \tag{46}$$

Proof. Since (41) and (42) hold with the given assumptions due to Theorem (3), the proof of (44) and (45) is the same as in Lemma 2.7 in [20]. Then, (46) is a result of the transitivity of norm equivalence and (43). □

From the results on the global matrix J, we may also determine the following results which characterize its blocks.

Lemma 6. *Let H as in* (28)*. Then*

$$\|F\|_{H_1H_1^{-1}} \leq C_1, \quad \|B\|_{H_1,H_2^{-1}} \leq C_1, \quad \|B^\intercal\|_{H_2,H_1^{-1}} \leq C_1,$$
$$\|M_1\|_{H_3,H_1^{-1}} \leq C_1, \quad \|K\|_{H_3,H_3^{-1}} \leq C_1\,.$$

Proof. Condition (17) holds true and it implies (41) which in turn implies (44). Equation (34) with $H_a = H_b = H$ and $Q = J$ yields

$$\|H^{-1/2}JH^{-1/2}\|_{l_2} = \|H^{-1}J\|_H \leq C_1,$$

from which it follows that the l_2-norm of each block of

$$H^{-1/2}JH^{-1/2}$$
$$= \begin{bmatrix} H_1^{-1/2} & 0 & 0 \\ 0 & H_2^{-1/2} & 0 \\ 0 & 0 & H_3^{-1/2} \end{bmatrix} \begin{bmatrix} F & B^\intercal & M_1 \\ B & 0 & 0 \\ 0 & 0 & K \end{bmatrix} \begin{bmatrix} H_1^{-1/2} & 0 & 0 \\ 0 & H_2^{-1/2} & 0 \\ 0 & 0 & H_3^{-1/2} \end{bmatrix}$$

$$= \begin{bmatrix} H_1^{-1/2}FH_1^{-1/2} & H_1^{-1/2}B^\intercal H_2^{-1/2} & H_1^{-1/2}M_1H_3^{-1/2} \\ H_2^{-1/2}BH_1^{-1/2} & 0 & 0 \\ 0 & 0 & H_3^{-1/2}KH_3^{-1/2} \end{bmatrix}$$

is bounded by C_1. Finally the bounds for $\|F\|_{H_1,H_1^{-1}}$, $\|B\|_{H_1,H_2^{-1}}$, $\|B^\intercal\|_{H_2,H_1^{-1}}$, $\|M\|_{H_3,H_1^{-1}}$ and $\|K\|_{H_3,H_3^{-1}}$ follow using again equation (34). □

Norm-equivalence properties for matrix blocks are here given.

Lemma 7. *Let H as in* (28) *and let* (18) *hold. Then*

$$F \sim_{H_1^{-1}} H_1, \quad F^{-1} \sim_{H_1} H_1^{-1}, \tag{47}$$

$$-S \sim_{H_2^{-1}} H_2, \quad -S^{-1} \sim_{H_2} H_2^{-1}, \tag{48}$$

$$K \sim_{H_3^{-1}} H_3, \quad K^{-1} \sim_{H_3} H_3^{-1}. \tag{49}$$

Also

$$\left\|F^{-1}\right\|_{H_1^{-1},H_1}^{-1} \geq C_3, \tag{50}$$

$$\left\|K^{-1}\right\|_{H_3^{-1},H_3}^{-1} \geq C_4. \tag{51}$$

Furthermore, let $\widetilde{F}$ satisfy $\widetilde{F} \sim_{H_1^{-1}} H_1$, then $\widetilde{S} = -B\widetilde{F}^{-1}B^\intercal$ satisfies

$$-\widetilde{S} \sim_{H_2^{-1}} H_2, \quad -\widetilde{S}^{-1} \sim_{H_2} H_2^{-1}. \tag{52}$$

Proof. Eq. (41) holds for our problem, and (18) implies (42). Conditions (41)-(42), together with the weak coercivity conditions (40), allow us to fit within Lemma 3.2 in [20], by which the norm equivalences (47), (48) and (49) follow. The properties (50), (51) and (52) are also proven using Lemma 3.2 in [20]. □

FOV-equivalence properties between various matrix blocks are also needed.

Lemma 8. *Let H as in* (28) *and let* (18) *hold. Then*

$$F \approx_{H_1^{-1}} H_1,\ H_1^{-1} \approx_{H_1} F^{-1}\,, \tag{53}$$

$$-S \approx_{H_2^{-1}} H_2,\ H_2^{-1} \approx_{H_2} -S^{-1}\,, \tag{54}$$

$$K \approx_{H_3^{-1}} H_3,\ H_3^{-1} \approx_{H_3} K^{-1}\,, \tag{55}$$

Furthermore, let $\widetilde{F} \approx_{H_1^{-1}} H_1$. Then $\widetilde{S} = -B\widetilde{F}^{-1}B^\intercal$ satisfies

$$-\widetilde{S} \approx_{H_2^{-1}} H_2,\ H_2^{-1} \approx_{H_2} -\widetilde{S}^{-1}. \tag{56}$$

Proof. Again, Eq. (41) holds and (18) implies (42). Using conditions (41)-(42) together with the strong coercivity conditions (39), we may apply Lemma 3.4 in [20] and prove the above properties. □

6. A class of upper triangular block preconditioners

We assume right preconditioning, so that the preconditioned system reads

$$JP^{-1}\boldsymbol{x} = f\,, \quad P\boldsymbol{y} = \boldsymbol{x}\,. \tag{57}$$

All the preconditioners we consider in this work fall within the structure

$$P(\rho) = \begin{bmatrix} P_1 & B^\intercal & M_1 \\ 0 & \rho^{-1}P_2 & 0 \\ 0 & 0 & P_3 \end{bmatrix}, \tag{58}$$

where P_1, P_2, P_3 and $\rho \neq 0$. Notice that the preconditioned system matrix becomes

$$JP(\rho)^{-1} = \begin{bmatrix} FP_1^{-1} & \rho(I - FP_1^{-1})B^\intercal P_2^{-1} & (I - FP_1^{-1})M_1P_3^{-1} \\ BP_1^{-1} & \rho\widehat{S}P_2^{-1} & -BP_1^{-1}M_1P_3^{-1} \\ 0 & 0 & KP_3^{-1} \end{bmatrix}, \tag{59}$$

where we define the approximate Schur complement block $\widehat{S}$ as

$$\widehat{S} = -BP_1^{-1}B^\intercal. \tag{60}$$

Sufficient conditions on the blocks of P will be determined so that the matrix J in (4) is FOV-equivalent to $P(\rho)$ with respect to H^{-1} and Theorem 1 can be used. First, we begin with norm equivalence results.

6.1. A norm-equivalence result

In this section, we establish the norm equivalence of the preconditioner to the system matrix.

Theorem 4. *Let* (18) *hold and let* $P(\rho)$ *be as in* (58). *Assume*

$$P_1 \sim_{H_1^{-1}} H_1\,, \quad P_2 \sim_{H_2^{-1}} H_2\,, \quad P_3 \sim_{H_3^{-1}} H_3, \tag{61}$$

then for any $\rho \neq 0$ *we have* $P(\rho) \sim_{H^{-1}} J$ *and* $P(\rho)^{-1} \sim_H J^{-1}$.

Proof. We only prove $P(\rho) \sim_{H^{-1}} J$, as the second equivalence follows similarly. By Lemma 5, $H \sim_{H^{-1}} J$. Hence, by the transitivity of $\sim_{H^{-1}}$ we only need to show $P(\rho) \sim_{H^{-1}} H$. We want to show that

$$\|P(\rho)H^{-1}\|_{H^{-1}} \leq \Gamma_P \;\text{ and }\; \|HP(\rho)^{-1}\|_{H^{-1}} \leq \gamma_P,$$

for some constants Γ_P and γ_P.

Since $P_i \sim_{H_i^{-1}} H_i$, then by Lemma 4

$$\left\|H_i^{-1}P_i\right\|_{H_i} \leq \Gamma_i \;\text{ and }\; \left\|P_i^{-1}H_i\right\|_{H_i} \leq \gamma_i^{-1}, \text{ for } i = 1, 2, 3. \tag{62}$$

Since

$$\begin{aligned} &H^{-1/2}P(\rho)H^{-1/2} \\ &\quad = \begin{bmatrix} H_1^{-1/2} & 0 & 0 \\ 0 & H_2^{-1/2} & 0 \\ 0 & 0 & H_3^{-1/2} \end{bmatrix} \begin{bmatrix} P_1 & B^\intercal & M_1 \\ 0 & \rho^{-1}P_2 & 0 \\ 0 & 0 & P_3 \end{bmatrix} \begin{bmatrix} H_1^{-1/2} & 0 & 0 \\ 0 & H_2^{-1/2} & 0 \\ 0 & 0 & H_3^{-1/2} \end{bmatrix} \\ &\quad = \begin{bmatrix} H_1^{-1/2}P_1H_1^{-1/2} & H_1^{-1/2}B^\intercal H_2^{-1/2} & H_1^{-1/2}M_1H_3^{-1/2} \\ 0 & \rho^{-1}H_2^{-1/2}P_2H_2^{-1/2} & 0 \\ 0 & 0 & H_3^{-1/2}P_3H_3^{-1/2} \end{bmatrix}, \end{aligned}$$

the following inequalities are obtained using Lemma 6 and inequalities (62):

$$\begin{aligned}
\|P(\rho)H^{-1}\|_{H^{-1}} &= \left\|H^{-1/2}P(\rho)H^{-1/2}\right\|_{l_2} \\
&\leq \left\|H_1^{-1/2}P_1H_1^{-1/2}\right\|_{l_2} + \left\|H_1^{-1/2}B^\intercal H_2^{-1/2}\right\|_{l_2} + \left\|H_1^{-1/2}M_1H_3^{-1/2}\right\|_{l_2} \\
&\quad + \left\|H_2^{-1/2}\rho^{-1}P_2H_2^{-1/2}\right\|_{l_2} + \left\|H_3^{-1/2}P_3H_3^{-1/2}\right\|_{l_2} \\
&= \left\|H_1^{-1}P_1\right\|_{H_1} + \|B^\intercal\|_{H_2,H_1^{-1}} + \|M_1\|_{H_3,H_1^{-1}} \\
&\quad + \left\|H_2^{-1}\rho^{-1}P_2\right\|_{H_2} + \left\|H_3^{-1}P_3\right\|_{H_3} \\
&\leq \Gamma_1 + C_1 + C_1 + |\rho^{-1}|\Gamma_2 + \Gamma_3 \\
&= \Gamma_1 + |\rho^{-1}|\Gamma_2 + \Gamma_3 + 2C_1 \\
&= \Gamma_P.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&H^{1/2}P(\rho)^{-1}H^{1/2} \\
&= \begin{bmatrix} H_1^{1/2} & 0 & 0 \\ 0 & H_2^{1/2} & 0 \\ 0 & 0 & H_3^{1/2} \end{bmatrix} \begin{bmatrix} P_1^{-1} & -\rho P_1^{-1}B^\intercal P_2^{-1} & -P_1^{-1}M_1P_3^{-1} \\ 0 & \rho P_2^{-1} & 0 \\ 0 & 0 & P_3^{-1} \end{bmatrix} \begin{bmatrix} H_1^{1/2} & 0 & 0 \\ 0 & H_2^{1/2} & 0 \\ 0 & 0 & H_3^{1/2} \end{bmatrix} \\
&= \begin{bmatrix} H_1^{1/2}P_1^{-1}H_1^{1/2} & -\rho H_1^{1/2}P_1^{-1}B^\intercal P_2^{-1}H_2^{1/2} & -H_1^{1/2}P_1^{-1}M_1P_3^{-1}H_3^{1/2} \\ 0 & \rho H_2^{1/2}P_2^{-1}H_2^{1/2} & 0 \\ 0 & 0 & H_3^{1/2}P_3^{-1}H_3^{1/2} \end{bmatrix},
\end{aligned}$$

$$\begin{aligned}
\|HP(\rho)^{-1}\|_{H^{-1}} &= \left\|H^{1/2}P(\rho)^{-1}H^{1/2}\right\|_{l_2} \\
&\leq \left\|H_1^{1/2}P_1^{-1}H_1^{1/2}\right\|_{l_2} + \left\|\rho H_1^{1/2}P_1^{-1}B^\intercal P_2^{-1}H_2^{1/2}\right\|_{l_2} + \left\|H_1^{1/2}P_1^{-1}M_1P_3^{-1}H_3^{1/2}\right\|_{l_2} \\
&\quad + \left\|\rho H_2^{1/2}P_2^{-1}H_2^{1/2}\right\|_{l_2} + \left\|H_3^{1/2}P_3^{-1}H_3^{1/2}\right\|_{l_2} \\
&\leq \left\|P_1^{-1}H_1\right\|_{H_1} + \left\|\rho P_1^{-1}B^\intercal P_2^{-1}\right\|_{H_2^{-1},H_1} + \left\|P_1^{-1}M_1P_3^{-1}\right\|_{H_3^{-1},H_1} \\
&\quad + \left\|\rho P_2^{-1}H_2\right\|_{H_2} + \left\|P_3^{-1}H_3\right\|_{H_3} \\
&\leq \gamma_1^{-1} + |\rho|\left\|P_2^{-1}\right\|_{H_2^{-1},H_2}\|B^\intercal\|_{H_2,H_1^{-1}}\left\|P_1^{-1}\right\|_{H_1^{-1},H_1} \\
&\quad + \left\|P_3^{-1}\right\|_{H_3^{-1},H_3}\|M_1\|_{H_3,H_1^{-1}}\left\|P_1^{-1}\right\|_{H_1^{-1},H_1} + |\rho|\gamma_2^{-1} + \gamma_3^{-1} \\
&\leq \gamma_1^{-1} + |\rho|\gamma_2^{-1}C_1\gamma_1^{-1} + \gamma_3^{-1}C_1\gamma_1^{-1} + |\rho|\gamma_2^{-1} + \gamma_3^{-1} \\
&= \gamma_P.
\end{aligned}$$

□

6.2. FOV-equivalence results

We now address two theorems that identify different conditions that lead to the FOV-equivalence between the linearization matrix J and the preconditioners of type (58).

Theorem 5. *Let* (18) *hold. Let* $P(\rho)$ *be defined as in* (58) *with*

$$P_1 = F\,, \quad P_3 = K\,, \quad S \approx_{H_2^{-1}} P_2\,. \tag{63}$$

There exists $\rho_0 > 0$ *such that if* $\rho \geq \rho_0$ *then*

$$J \approx_{H^{-1}} P(\rho).$$

Proof. Since (18) holds, we may use Lemma 7 and $S \approx_{H_2^{-1}} P_2$ to get the norm equivalences for any $\rho \neq 0$

$$F \sim_{H_1^{-1}} H_1, \quad \rho^{-1} P_2 \sim_{H_2^{-1}} S \sim_{H_2^{-1}} H_2, \quad K \sim_{H_3^{-1}} H_3\,.$$

Using these in Theorem 4 leads to the equivalence $P(\rho) \sim_{H^{-1}} J$, which implies $\left\|JP(\rho)^{-1}\right\|_{H^{-1}} \leq \Gamma$. Hence, by Definition 2 we only need to find a constant $\gamma > 0$ such that for all $\boldsymbol{v} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$,

$$\boldsymbol{v}^\intercal H^{-1} J P(\rho)^{-1} \boldsymbol{v} \geq \gamma \left\|\boldsymbol{v}\right\|_{H^{-1}}.$$

Let $\boldsymbol{v}^\intercal = [\boldsymbol{x}^\intercal, \boldsymbol{y}^\intercal, \boldsymbol{z}^\intercal]$. Due to the hypothesis (63), (59) reduces to

$$JP(\rho)^{-1} = \begin{bmatrix} I & 0 & 0 \\ BF^{-1} & \rho S P_2^{-1} & -BF^{-1}MK^{-1} \\ 0 & 0 & I \end{bmatrix},$$

we need to show that

$$\begin{aligned} &\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \\ \boldsymbol{z} \end{bmatrix}^\intercal \begin{bmatrix} H_1^{-1} & 0 & 0 \\ H_2^{-1}BF^{-1} & \rho H_2^{-1} S P_2^{-1} & -H_2^{-1}BF^{-1}MK^{-1} \\ 0 & 0 & H_3^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \\ \boldsymbol{z} \end{bmatrix} \\ &\quad \geq \gamma \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \\ \boldsymbol{z} \end{bmatrix}^\intercal \begin{bmatrix} H_1^{-1} & 0 & 0 \\ 0 & H_2^{-1} & 0 \\ 0 & 0 & H_3^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \\ \boldsymbol{z} \end{bmatrix}. \end{aligned} \tag{64}$$

We start with lower bounds on the terms coming from the diagonal blocks in (64). Since $S \approx_{H_2^{-1}} P_2$, there exists a constant $\beta_1 > 0$ such that

$$\beta_1 \leq \frac{< \boldsymbol{x}, S P_2^{-1} \boldsymbol{x} >_{H_2^{-1}}}{< \boldsymbol{x}, \boldsymbol{x} >_{H_2^{-1}}}. \tag{65}$$

Using (65), we obtain

$$\boldsymbol{x}^\intercal H_1^{-1} \boldsymbol{x} = \left\|\boldsymbol{x}\right\|^2_{H_1^{-1}}, \quad \rho \boldsymbol{y}^\intercal H_2^{-1} S P_2^{-1} \boldsymbol{y} \geq \rho \beta_1 \left\|\boldsymbol{y}\right\|^2_{H_2^{-1}}, \quad \text{and} \quad \boldsymbol{z}^\intercal H_3^{-1} \boldsymbol{z} = \left\|\boldsymbol{z}\right\|^2_{H_3^{-1}}.$$

Concerning the off-diagonal blocks, we first get upper bounds in appropriate matrix norms. Lemma 6 and Lemma 7 give

$$\begin{aligned}\left\|H_2^{-1}BF^{-1}\right\|_{H_1^{-1},H_2} &\le \left\|F^{-1}\right\|_{H_1^{-1},H_1}\left\|H_2^{-1}B\right\|_{H_1,H_2}\\ &= \left\|F^{-1}\right\|_{H_1^{-1},H_1}\|B\|_{H_1,H_2^{-1}}\\ &\le C_1C_3^{-1},\end{aligned}$$

then by using Eq. (35) and Young's inequality

$$\left|\boldsymbol{y}^\intercal H_2^{-1}BF^{-1}\boldsymbol{x}\right| \le C_1C_3^{-1}\|\boldsymbol{x}\|_{H_1^{-1}}\|\boldsymbol{y}\|_{H_2^{-1}} \le \frac{1}{2}\|\boldsymbol{x}\|_{H_1^{-1}}^2 + \frac{\left(C_1C_3^{-1}\right)^2}{2}\|\boldsymbol{y}\|_{H_2^{-1}}^2.$$

Lemma 6 and Lemma 7 also give

$$\begin{aligned}&\left\|H_2^{-1}BF^{-1}M_1K^{-1}\right\|_{H_3^{-1},H_2}\\ &\quad\le \left\|K^{-1}\right\|_{H_3^{-1},H_3}\|M_1\|_{H_3,H_1^{-1}}\left\|F^{-1}\right\|_{H_1^{-1},H_1}\left\|H_2^{-1}B\right\|_{H_1,H_2}\\ &\quad= \left\|K^{-1}\right\|_{H_3^{-1},H_3}\|M_1\|_{H_3,H_1^{-1}}\left\|F^{-1}\right\|_{H_1^{-1},H_1}\|B\|_{H_1,H_2^{-1}}\\ &\quad\le C_4^{-1}C_1C_3^{-1}C_1\\ &\quad= C_1^2C_3^{-1}C_4^{-1},\end{aligned}$$

and again using Eq. (35) and Young's inequality

$$\begin{aligned}\left|\boldsymbol{y}^\intercal H_2^{-1}BF^{-1}M_1K^{-1}\boldsymbol{z}\right| &\le C_1^2C_3^{-1}C_4^{-1}\|\boldsymbol{y}\|_{H_2^{-1}}\|\boldsymbol{z}\|_{H_3^{-1}}\\ &\le \frac{1}{2}\|\boldsymbol{z}\|_{H_3^{-1}}^2 + \frac{\left(C_1^2C_3^{-1}C_4^{-1}\right)^2}{2}\|\boldsymbol{y}\|_{H_2^{-1}}^2.\end{aligned}$$

Hence, we have

$$\begin{aligned}&\boldsymbol{v}^\intercal H^{-1}JP(\rho)^{-1}\boldsymbol{v}\\ &\quad\ge \|\boldsymbol{x}\|_{H_1^{-1}}^2 + \rho\beta_1\|\boldsymbol{y}\|_{H_2^{-1}}^2 + \|\boldsymbol{z}\|_{H_3^{-1}}^2 - \left|\boldsymbol{y}^\intercal H_2^{-1}BF^{-1}\boldsymbol{x}\right| - \left|\boldsymbol{y}^\intercal H_2^{-1}BF^{-1}M_1K^{-1}\boldsymbol{z}\right|\\ &\quad\ge \frac{1}{2}\|\boldsymbol{x}\|_{H_1^{-1}}^2 + \left(\rho\beta_1 - \frac{\left(C_1C_3^{-1}\right)^2}{2} - \frac{\left(C_1^2C_3^{-1}C_4^{-1}\right)^2}{2}\right)\|\boldsymbol{y}\|_{H_2^{-1}}^2 + \frac{1}{2}\|\boldsymbol{z}\|_{H_3^{-1}}^2\\ &\quad\ge \frac{1}{2}(\|\boldsymbol{x}\|_{H_1^{-1}}^2 + \|\boldsymbol{y}\|_{H_2^{-1}}^2 + \|\boldsymbol{z}\|_{H_3^{-1}}^2) = \gamma\|\boldsymbol{v}\|_{H^{-1}},\end{aligned}$$

with $\gamma = 1/2$, provided that

$$\rho \ge \rho_0 = \frac{1 + (C_1C_3^{-1})^2 + (C_1^2C_3^{-1}C_4^{-1})^2}{2\beta_1}. \tag{66}$$

□

By a relaxation of the hypotheses in (63), we obtain the following result.

Theorem 6. *Let* (18) *hold. Let* $P(\rho)$ *as in* (58)*, chosen such that*

$$F \approx_{H_1^{-1}} P_1,\ K \approx_{H_3^{-1}} P_3,\ \widehat{S} \approx_{H_2^{-1}} P_2. \tag{67}$$

Then there exists $\rho_0 > 0$ *such that if* $\rho > \rho_0$ *and*

$$\left\|I - FP_1^{-1}\right\|_{H_1^{-1}} \leq \frac{1}{\rho} \tag{68}$$

then

$$J \approx_{H^{-1}} P(\rho),$$

Proof. Since we assume (67), there exist constants α_1, α_2, β_1, β_2, ζ_1 and ζ_2 such that for all $\boldsymbol{x} \in \mathbb{R}^{n_1}\backslash\{0\}$, $\boldsymbol{y} \in \mathbb{R}^{n_2}\backslash\{0\}$ and $\boldsymbol{z} \in \mathbb{R}^{n_3}\backslash\{0\}$,

$$\begin{aligned}
\alpha_1 &\leq \frac{\left\langle \boldsymbol{x}, FP_1^{-1}\boldsymbol{x}\right\rangle_{H_1^{-1}}}{<\boldsymbol{x},\boldsymbol{x}>_{H_1^{-1}}},\ \left\|FP_1^{-1}\right\|_{H_1^{-1}} \leq \alpha_2,\\
\beta_1 &\leq \frac{\left\langle \boldsymbol{y}, \widehat{S}P_2^{-1}\boldsymbol{y}\right\rangle_{H_2^{-1}}}{<\boldsymbol{y},\boldsymbol{y}>_{H_2^{-1}}},\ \left\|\widehat{S}P_2^{-1}\right\|_{H_2^{-1}} \leq \beta_2,\\
\zeta_1 &\leq \frac{\left\langle \boldsymbol{z}, KP_3^{-1}\boldsymbol{z}\right\rangle_{H_3^{-1}}}{<\boldsymbol{z},\boldsymbol{z}>_{H_3^{-1}}},\ \left\|KP_3^{-1}\right\|_{H_3^{-1}} \leq \zeta_2.
\end{aligned} \tag{69}$$

On the other hand, the hypotheses and Lemma 7 imply

$$P_1 \sim_{H_1^{-1}} F \sim_{H_1^{-1}} H_1\,, \tag{70}$$

$$P_3 \sim_{H_3^{-1}} K \sim_{H_3^{-1}} H_3\,. \tag{71}$$

Using again Lemma 7 with (70) and the hypotheses we also have $P_2 \sim_{H_2^{-1}} \widehat{S} \sim_{H_2^{-1}} H_2$. Then Theorem 4 yields $P(\rho) \sim_{H^{-1}} J$. Hence $\left\|JP(\rho)^{-1}\right\| \leq \Gamma$ for some $\Gamma > 0$. By Definition 2 we then only need to find a constant γ such that for all nonzero $\boldsymbol{v} \in \mathbb{R}^n$,

$$\frac{\left\langle \boldsymbol{v}, JP(\rho)^{-1}\boldsymbol{v}\right\rangle_{H^{-1}}}{<\boldsymbol{v},\boldsymbol{v}>_{H^{-1}}} \geq \gamma\,.$$

We need to establish the lower bound

$$\begin{aligned}
&\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\\ \boldsymbol{z}\end{bmatrix}^{\intercal}
\begin{bmatrix} H_1^{-1}FP_1^{-1} & \rho H_1^{-1}(I - FP_1^{-1})B^{\intercal}P_2^{-1} & H_1^{-1}(I - FP_1^{-1})M_1P_3^{-1}\\ H_2^{-1}BP_1^{-1} & \rho H_2^{-1}\widehat{S}P_2^{-1} & -H_2^{-1}BP_1^{-1}M_1P_3^{-1}\\ 0 & 0 & H_3^{-1}KP_3^{-1}\end{bmatrix}
\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\\ \boldsymbol{z}\end{bmatrix}\\
&\quad\geq \gamma \begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\\ \boldsymbol{z}\end{bmatrix}^{\intercal}
\begin{bmatrix} H_1^{-1} & 0 & 0\\ 0 & H_2^{-1} & 0\\ 0 & 0 & H_3^{-1}\end{bmatrix}
\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\\ \boldsymbol{z}\end{bmatrix}\\
&\quad= \gamma\left(\|\boldsymbol{x}\|^2_{H_1^{-1}} + \|\boldsymbol{y}\|^2_{H_2^{-1}} + \|\boldsymbol{z}\|^2_{H_3^{-1}}\right).
\end{aligned} \tag{72}$$

Concerning the diagonal blocks of $JP(\rho)^{-1}$ in (72), we get, using (69),

$$\boldsymbol{x}^\intercal H_1^{-1} F P_1^{-1} \boldsymbol{x} \geq \alpha_1 \|\boldsymbol{x}\|^2_{H_1^{-1}}, \quad \rho \boldsymbol{y}^\intercal H_2^{-1} \widehat{S} P_2^{-1} \boldsymbol{y} \geq \rho \beta_1 \|\boldsymbol{y}\|^2_{H_2^{-1}},$$

and

$$\boldsymbol{z}^\intercal H_3^{-1} K P_3^{-1} \boldsymbol{z} \geq \zeta_1 \|\boldsymbol{z}\|^2_{H_3^{-1}}.$$

For the off-diagonal terms, we first get upper bounds on the corresponding blocks in appropriate matrix norms. By an intermediate result in the proof of Theorem 3.8 in [20] we have

$$\left\|P_2^{-1}\right\|_{H_2^{-1}, H_2} \leq \beta_2 C_2^{-2} \alpha_1^{-1} C_1.$$

Then from Lemma 6 and (33) we get

$$\begin{aligned}
&\left\|H_1^{-1}(I - F P_1^{-1}) B^\intercal P_2^{-1}\right\|_{H_2^{-1}, H_1} \\
&\quad \leq \left\|P_2^{-1}\right\|_{H_2^{-1}, H_2} \|B^\intercal\|_{H_2, H_1^{-1}} \left\|H_1^{-1}(I - F P_1^{-1})\right\|_{H_1^{-1}, H_1} \\
&\quad \leq \beta_2 C_2^{-2} \alpha_1^{-1} C_1 C_1 \left\|I - F P^{-1}\right\|_{H_1^{-1}} \\
&\quad \leq \frac{1}{\rho} \beta_2 C_2^{-2} \alpha_1^{-1} C_1^2,
\end{aligned}$$

where we used the hypothesis (68). Thus using the norm characterization (35) we get

$$\left|\rho \boldsymbol{x}^\intercal H_1^{-1}(I - F P_1^{-1}) B^\intercal P_2^{-1} \boldsymbol{y}\right| \leq \beta_2 C_2^{-2} \alpha_1^{-1} C_1^2 \|\boldsymbol{x}\|_{H_1^{-1}} \|\boldsymbol{y}\|_{H_2^{-1}}.$$

Similarly using Lemma 6, Lemma 7 and Eq. (69) we find

$$
\begin{aligned}
&\left\|H_1^{-1}(I-FP_1^{-1})M_1P_3^{-1}\right\|_{H_3^{-1},H_1}\\
&\quad=\left\|H_1^{-1}(I-FP_1^{-1})M_1K^{-1}KP_3^{-1}\right\|_{H_3^{-1},H_1}\\
&\quad\le\left\|KP_3^{-1}\right\|_{H_3^{-1},H_3^{-1}}\left\|K^{-1}\right\|_{H_3^{-1},H_3}\|M_1\|_{H_3,H_1^{-1}}\left\|H_1^{-1}(I-FP_1^{-1})\right\|_{H_1^{-1},H_1}\\
&\quad\le\zeta_2C_4^{-1}C_1\left\|I-FP_1^{-1}\right\|_{H_1^{-1},H_1^{-1}}\\
&\quad\le\frac{\zeta_2C_4^{-1}C_1}{\rho},\\
&\left\|H_2^{-1}BP_1^{-1}\right\|_{H_1^{-1},H_2}\\
&\quad=\left\|H_2^{-1}BF^{-1}FP_1^{-1}\right\|_{H_1^{-1},H_2}\\
&\quad\le\left\|FP_1^{-1}\right\|_{H_1^{-1},H_1^{-1}}\left\|F^{-1}\right\|_{H_1^{-1},H_1}\left\|H_2^{-1}B\right\|_{H_1,H_2}\\
&\quad\le\alpha_2C_3^{-1}\|B\|_{H_1,H_2^{-1}}\\
&\quad\le\alpha_2C_3^{-1}C_1,\\
&\left\|H_2^{-1}BP_1^{-1}M_1P_3^{-1}\right\|_{H_3^{-1},H_2}\\
&\quad=\left\|H_2^{-1}BF^{-1}FP_1^{-1}M_1K^{-1}KP_3^{-1}\right\|_{H_3^{-1},H_2}\\
&\quad\le\left\|KP_3^{-1}\right\|_{H_3^{-1},H_3^{-1}}\left\|K^{-1}\right\|_{H_3^{-1},H_3}\|M_1\|_{H_3,H_1^{-1}}\left\|FP_1^{-1}\right\|_{H_1^{-1},H_1^{-1}}\\
&\qquad\left\|F^{-1}\right\|_{H_1^{-1},H_1}\left\|H_2^{-1}B\right\|_{H_1,H_2}\\
&\quad\le\zeta_2C_4^{-1}C_1\alpha_2C_3^{-1}\|B\|_{H_1,H_2^{-1}}\\
&\quad\le\zeta_2C_4^{-1}C_1^2\alpha_2C_3^{-1},
\end{aligned}
$$

and using again Eq. (35)

$$
\begin{aligned}
\left|\boldsymbol{x}^\intercal H_1^{-1}(I-FP_1^{-1})M_1P_3^{-1}\boldsymbol{z}\right|&\le\frac{\zeta_2C_4^{-1}C_1}{\rho}\|\boldsymbol{x}\|_{H_1^{-1}}\|\boldsymbol{z}\|_{H_3^{-1}},\\
\left|\boldsymbol{y}^\intercal H_2^{-1}BP_1^{-1}\boldsymbol{x}\right|&\le\alpha_2C_3^{-1}C_1\|\boldsymbol{y}\|_{H_2^{-1}}\|\boldsymbol{x}\|_{H_1^{-1}},\\
\left|\boldsymbol{y}^\intercal H_2^{-1}BP_1^{-1}M_1P_3^{-1}\boldsymbol{z}\right|&\le\zeta_2C_4^{-1}C_1^2\alpha_2C_3^{-1}\|\boldsymbol{y}\|_{H_2^{-1}}\|\boldsymbol{z}\|_{H_3^{-1}}.
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
&\boldsymbol{v}^\intercal H^{-1} JP(\rho)^{-1}\boldsymbol{v} \\
&\quad \geq \alpha_1 \|\boldsymbol{x}\|^2_{H_1^{-1}} + \rho\beta_1 \|\boldsymbol{y}\|^2_{H_2^{-1}} + \zeta_1 \|\boldsymbol{z}\|^2_{H_3^{-1}} \\
&\qquad - \beta_2 C_2^{-2}\alpha_1^{-1} C_1^2 \|\boldsymbol{x}\|_{H_1^{-1}} \|\boldsymbol{y}\|_{H_2^{-1}} - \frac{\zeta_2 C_4^{-1} C_1}{\rho} \|\boldsymbol{x}\|_{H_1^{-1}} \|\boldsymbol{z}\|_{H_3^{-1}} \\
&\qquad - \alpha_2 C_3^{-1} C_1 \|\boldsymbol{x}\|_{H_1^{-1}} \|\boldsymbol{y}\|_{H_2^{-1}} - \zeta_2 C_4^{-1} C_1^2 \alpha_2 C_3^{-1} \|\boldsymbol{y}\|_{H_2^{-1}} \|\boldsymbol{z}\|_{H_3^{-1}} \\
&\quad = \alpha_1 \|\boldsymbol{x}\|^2_{H_1^{-1}} + \rho\beta_1 \|\boldsymbol{y}\|^2_{H_2^{-1}} + \zeta_1 \|\boldsymbol{z}\|^2_{H_3^{-1}} \\
&\qquad - (\beta_2 C_2^{-2}\alpha_1^{-1} C_1^2 + \alpha_2 C_3^{-1} C_1) \|\boldsymbol{x}\|_{H_1^{-1}} \|\boldsymbol{y}\|_{H_2^{-1}} \\
&\qquad - \frac{\zeta_2 C_4^{-1} C_1}{\rho} \|\boldsymbol{x}\|_{H_1^{-1}} \|\boldsymbol{z}\|_{H_3^{-1}} - \zeta_2 C_4^{-1} C_1^2 \alpha_2 C_3^{-1} \|\boldsymbol{y}\|_{H_2^{-1}} \|\boldsymbol{z}\|_{H_3^{-1}} .
\end{aligned}
$$

Let $a = \beta_2 C_2^{-2}\alpha_1^{-1} C_1^2 + \alpha_2 C_3^{-1} C_1$, $b = \zeta_2 C_4^{-1} C_1$, $c = \zeta_2 C_4^{-1} C_1^2 \alpha_2 C_3^{-1}$. Then we use the following Young's inequalities

$$
\begin{aligned}
a \|\boldsymbol{x}\|_{H_1^{-1}} \|\boldsymbol{y}\|_{H_2^{-1}} &\leq \frac{\varepsilon_1\alpha_1}{2} \|\boldsymbol{x}\|^2_{H_1^{-1}} + \frac{a^2}{2\varepsilon_1\alpha_1} \|\boldsymbol{y}\|^2_{H_2^{-1}} , \\
\frac{b}{\rho} \|\boldsymbol{x}\|_{H_1^{-1}} \|\boldsymbol{z}\|_{H_3^{-1}} &\leq \frac{\varepsilon_2\alpha_1}{2} \|\boldsymbol{x}\|^2_{H_1^{-1}} + \frac{b^2}{2\varepsilon_2\alpha_1\rho^2} \|\boldsymbol{z}\|^2_{H_3^{-1}} , \\
c \|\boldsymbol{y}\|_{H_2^{-1}} \|\boldsymbol{z}\|_{H_3^{-1}} &\leq \frac{\varepsilon_3 c}{2} \|\boldsymbol{y}\|^2_{H_2^{-1}} + \frac{c}{2\varepsilon_3} \|\boldsymbol{z}\|^2_{H_3^{-1}} .
\end{aligned}
$$

Thus we obtain

$$
\begin{aligned}
\boldsymbol{v}^\intercal H^{-1} JP(\rho)^{-1}\boldsymbol{v} \geq \alpha_1(1 - \frac{\varepsilon_1}{2} - \frac{\varepsilon_2}{2}) \|\boldsymbol{x}\|^2_{H_1^{-1}} + (\rho\beta_1 - \frac{a^2}{2\varepsilon_1\alpha_1} - \frac{\varepsilon_3 c}{2}) \|\boldsymbol{y}\|^2_{H_2^{-1}} \\
+ (\zeta_1 - \frac{b^2}{2\varepsilon_2\alpha_1\rho^2} - \frac{c}{2\varepsilon_3}) \|\boldsymbol{z}\|^2_{H_3^{-1}} .
\end{aligned}
$$

Let

$$
\gamma = \alpha_1 \left(1 - \frac{\varepsilon_1}{2} - \frac{\varepsilon_2}{2}\right). \tag{73}
$$

We need to find further conditions on $\varepsilon_1, \varepsilon_2, \varepsilon_3, \rho > 0$ such that

$$
\gamma > 0, \quad \left(\rho\beta_1 - \frac{a^2}{2\varepsilon_1\alpha_1} - \frac{\varepsilon_3 c}{2}\right) > 0, \quad \left(\zeta_1 - \frac{b^2}{2\varepsilon_2\alpha_1\rho^2} - \frac{c}{2\varepsilon_3}\right) > 0. \tag{74}
$$

The first condition in (74) is satisfied with $\varepsilon_1 + \varepsilon_2 < 2$. The other two imply

$$
\rho > \frac{\gamma + \frac{a^2}{2\varepsilon_1\alpha_1} + \frac{\varepsilon_3 c}{2}}{\beta_1}, \quad \zeta_1 - \gamma - \frac{c}{2\varepsilon_3} > \frac{b^2}{2\varepsilon_2\alpha_1\rho^2} . \tag{75}
$$

The second inequality yields $\zeta_1 - \gamma > \frac{c}{2\varepsilon_3}$. The positivity of ε_3 requires $\zeta_1 \geq \gamma$. Since γ depends on $\varepsilon_1, \varepsilon_2$ by (73), we have to choose these two constants such that they also satisfy

$$
2(1 - \frac{\zeta_1}{\alpha_1}) \leq \varepsilon_1 + \varepsilon_2 < 2 . \tag{76}
$$

This choice is always possible since $1 - \frac{\zeta_1}{\alpha_1} < 1$. Hence, we can choose

$$\varepsilon_3 \geq \frac{c}{2}\frac{1}{\zeta_1 - \gamma}\,, \quad \rho > \rho_0 := \max\left(\frac{\gamma + \frac{a^2}{2\varepsilon_1\alpha_1} + \frac{\varepsilon_3 c}{2}}{\beta_1}, \frac{b}{\sqrt{2\varepsilon_2\alpha_1[\zeta_1 - \gamma - \frac{c}{2\varepsilon_3}]}}\right), \tag{77}$$

so that inequality (72) follows. □

6.3. Particular choices of preconditioners

We now propose four block triangular preconditioners having the structure of (58) given by

$$P_{R1}(\rho) = \begin{bmatrix} F & B^\intercal & M_1 \\ 0 & -\rho^{-1}A_pF_p^{-1}M_p & 0 \\ 0 & 0 & K \end{bmatrix}, \quad P_{R2}(\rho) = \begin{bmatrix} F & B^\intercal & M_1 \\ 0 & -\rho^{-1}M_p & 0 \\ 0 & 0 & K \end{bmatrix},$$

$$P_{R3}(\rho) = \begin{bmatrix} A_{\boldsymbol{u}} & B^\intercal & M_1 \\ 0 & -\rho^{-1}A_pF_p^{-1}M_p & 0 \\ 0 & 0 & A_T \end{bmatrix}, \text{ and } P_{R4}(\rho) = \begin{bmatrix} A_{\boldsymbol{u}} & B^\intercal & M_1 \\ 0 & -\rho^{-1}M_p & 0 \\ 0 & 0 & A_T \end{bmatrix}.$$

We refer to [17, 8, 7, 18] for more details on similar choices. The following two theorems show under what conditions these preconditioners are either norm-equivalent or FOV-equivalent to J.

Theorem 7. *Let* (18) *hold. Then for any* $\rho \neq 0$

$$J \sim_{H^{-1}} P_{R3}(\rho)\,, \quad J \sim_{H^{-1}} P_{R1}(\rho).$$

Proof. We fit $P_{R3}(\rho)$ in the general preconditioner structure (58) by letting

$$P_1 = A_{\boldsymbol{u}}\,, \quad P_2 = -A_pF_p^{-1}M_p\,, \quad P_3 = A_T. \tag{78}$$

We have $P_2 \sim_{H_2^{-1}} S$ (see [20], page 2046). Moreover, by Lemma 7 we have $S \sim_{H_2^{-1}} H_2$, so that $P_2 \sim_{H_2^{-1}} H_2$. Since $P_1 \equiv H_1$ and $P_3 \equiv H_3$, we may use Theorem 4 to get $J \sim_{H^{-1}} P_{R3}(\rho)$.

We fit $P_{R1}(\rho)$ in the general preconditioner structure (58) by letting

$$P_1 = F\,, \quad P_2 = -A_pF_p^{-1}M_p\,, \quad P_3 = K. \tag{79}$$

By Lemma 7 we obtain $P_1 \sim_{H_1^{-1}} H_1$ and $P_3 \sim_{H_3^{-1}} H_3$. Together with $P_2 \sim_{H_2^{-1}} H_2$ from above, we may use again Theorem 4 and have $J \sim_{H^{-1}} P_{R1}(\rho)$. □

Theorem 8. *Let* (18) *hold. Then there exists* $\rho_0 > 0$ *such that if* $\rho \geq \rho_0$ *then*

$$J \approx_{H^{-1}} P_{R2}(\rho)\,.$$

Also, there exists $\rho_1 > 0$*, such that if* $\rho > \rho_1$ *and* $\left\|I - FA_{\boldsymbol{u}}^{-1}\right\|_{H_1^{-1}} \leq \frac{1}{\rho}$ *we have*

$$J \approx_{H^{-1}} P_{R4}(\rho)\,.$$

Proof. Concerning $P_{R2}(\rho)$, let $P_1 = F$, $P_2 = -M_p = -H_2$ and $P_3 = K$ in (58). By Lemma 8 we have $P_1 \approx_{H_1^{-1}} H_1$ and $P_3 \approx_{H_3^{-1}} H_3$, and $-S \approx_{H_2^{-1}} H_2$, so $S \approx_{H_2^{-1}} P_2$. Using Theorem 5 we obtain $J \approx_{H^{-1}} P_{R2}(\rho)$.

Concerning $P_{R4}(\rho)$, let $P_1 = A_{\boldsymbol{u}}$, $P_2 = -M_p$, $P_3 = A_T$. Setting $\widetilde{S} = -BA_{\boldsymbol{u}}^{-1}B^\intercal$ in Lemma 8, we have $-\widetilde{S} \approx_{H_2^{-1}} H_2$ so that $\widetilde{S} \approx_{H_2^{-1}} P_2$. Using Theorem 6 we obtain $J \approx_{H^{-1}} P_{R4}(\rho)$.

□

Remark 6.1. *The GMRES convergence theorem 1 is based on the condition* (38) *of FOV-equivalence between the linearization matrix and the preconditioner. For the discretization framework and the class of block preconditioners considered here, this FOV equivalence condition as proved in Theorem 5 or 6 depends first of all on the fulfillment of inequality* (18)*, which bounds a function of the nondimensional parameters Ra and Pr in terms of the Poincaré constant C_p of the domain. The numerical determination of C_p is not a trivial task for general domains. Furthermore, these FOV equivalence conditions are valid with lower bounds ρ_0 on the parameter ρ, where ρ_0 cannot be in general computed nor estimated, see* (66) *or* (77)*.*

7. Numerical results

7.1. Description of the solvers

First, we describe the solvers and subsolvers used for the preconditioned systems. As pointed out earlier, the determination of a value of ρ that guarantees FOV-equivalence is not possible in general. We observed that the choice $\rho = 1$ for all preconditioners guarantees convergence of the following numerical tests. See also [3] for some numerical investigations on the value of ρ for a mixed Stokes-Darcy problem. The preconditioner P_{R1} requires solving four linear systems: one for the F block, two for the approximation of the inverse Schur complement $P_2^{-1} = -M_p^{-1}F_pA_p^{-1}$, and one for the K block. We use FGMRES right-preconditioned with an algebraic multigrid method (AMG) [11] for the F and K blocks and CG preconditioned with AMG for A_p and M_p. The A_p and M_p blocks do not depend on the solution and their inversion can be done only once outside the nonlinear loop. The preconditioner P_{R2} requires solving three linear systems: one for the F block, one for the M_p block, and one for the K block. We use the same subsolvers as in P_{R1} to solve for the F, M_p and K blocks. Again the M_p block does not depend on the solution and it inversion can be done only once outside the nonlinear loop.

In contrast to preconditioner P_{R1}, preconditioner P_{R3} requires to solve $A_{\boldsymbol{u}}$ and A_T blocks instead of F and K blocks. Since both $A_{\boldsymbol{u}}$ and A_T blocks are symmetric matrices, CG preconditioned with AMG can be applied to solve them. In this case all the blocks, $A_{\boldsymbol{u}}$, A_p, M_p and A_T, do not depend on the solution and their inversion can be done only once outside the nonlinear loop. The last preconditioner P_{R4} requires to solve $A_{\boldsymbol{u}}$, M_p and A_T blocks, which can all be solved by CG preconditioned with AMG. Also in this case all the blocks,

$A_{\boldsymbol{u}}$, M_p and A_T, do not depend on solution and their inversion can be done only once outside the nonlinear loop.

We set the threshold value for the scaled residual of each subsolver involved to be 10^{-4}. In the outer linear FGMRES solver for the preconditioned linearized systems, the convergence threshold of the scaled residual is set to be 10^{-8}. The stopping tolerance for the nonlinear iterations for the L_∞-norm of the absolute error between two successive solutions is set to 10^{-6}. We set a maximum number of 50 iterations.

Note that in preconditioners P_{R1} and P_{R2} the F and K blocks are inverted at any nonlinear iteration, while in the preconditioners P_{R3} and P_{R4} no inversion for any subblock is done inside the nonlinear loop. This yields an overhead in the computational time for P_{R1} and P_{R2}. On the other hand, we expect P_{R1} and P_{R2} to perform better in terms of number of GMRES iterations, since the F and K blocks are the same as in the linearization matrix J. All of these aspects will be taken into account in the following numerical examples.

7.2. Numerical experiments

A two-dimensional example is considered in the numerical study. The fluid domain is the unit square with no-slip boundary conditions, temperature $T = 1$ on the right side, temperature $T = 0$ on the left side, and zero heat flux on the remaining boundaries. At large Rayleigh numbers, this creates an instability leading to overturning cells. The unit square is partitioned into $N \times N$ squares, and each square is cut into two right triangles, so the total number of elements is $2N \times N$. We set $Pr = 1$ in the following numerical studies.

We observe numerically that the eigenvalues of the preconditioned system cluster around 1 with any choice of the preconditioners P_{R1}, P_{R2}, P_{R3} and P_{R4}. Figure 1 shows the eigenvalues for the systems with $N = 16$ and $Ra = 2\times10^3$, respectively. Figure 1 (a), (b), (c), (d) and (e) show the eigenvalues with preconditioners P_{R1}, P_{R2}, P_{R3}, P_{R4} and without preconditioner, respectively. The real part of the majority of the eigenvalues with preconditioners P_{R1} and P_{R2} is between 0 and 1, see Figure 1 (a) and (b). On the other hand, since $A_{\boldsymbol{u}}$ and A_T are just an approximation of F and K, the real part of more eigenvalues for both systems with preconditioners P_{R3} and P_{R4} is greater than 1, see Figure 1 (c) and (d). Then, we the expect preconditioners P_{R1} and P_{R2} to perform better than P_{R3} and P_{R4} in terms of total number of GMRES iterations.

In Tables 1 - 4 we report the number of nonlinear iterations, the average number of GMRES iterations per nonlinear iteration and the overall computational time for the four preconditioners, respectively. The number of total GMRES iterations is independent of the mesh size in all cases. We expected such results for P_{R2} and P_{R4}, since in Theorem 8 we proved FOV equivalence of these system with the matrix J. Nevertheless we get mesh independence also for preconditioners P_{R1} and P_{R3} which are just norm-equivalent to the matrix J. This suggests that the FOV-equivalence condition in Theorem 1 may be too strong, and there may be still room for improvement.

We consider two pairs of preconditioners P_{R1} versus P_{R2}, and P_{R3} versus P_{R4}. The difference in each pair is that the approximate Schur complement is

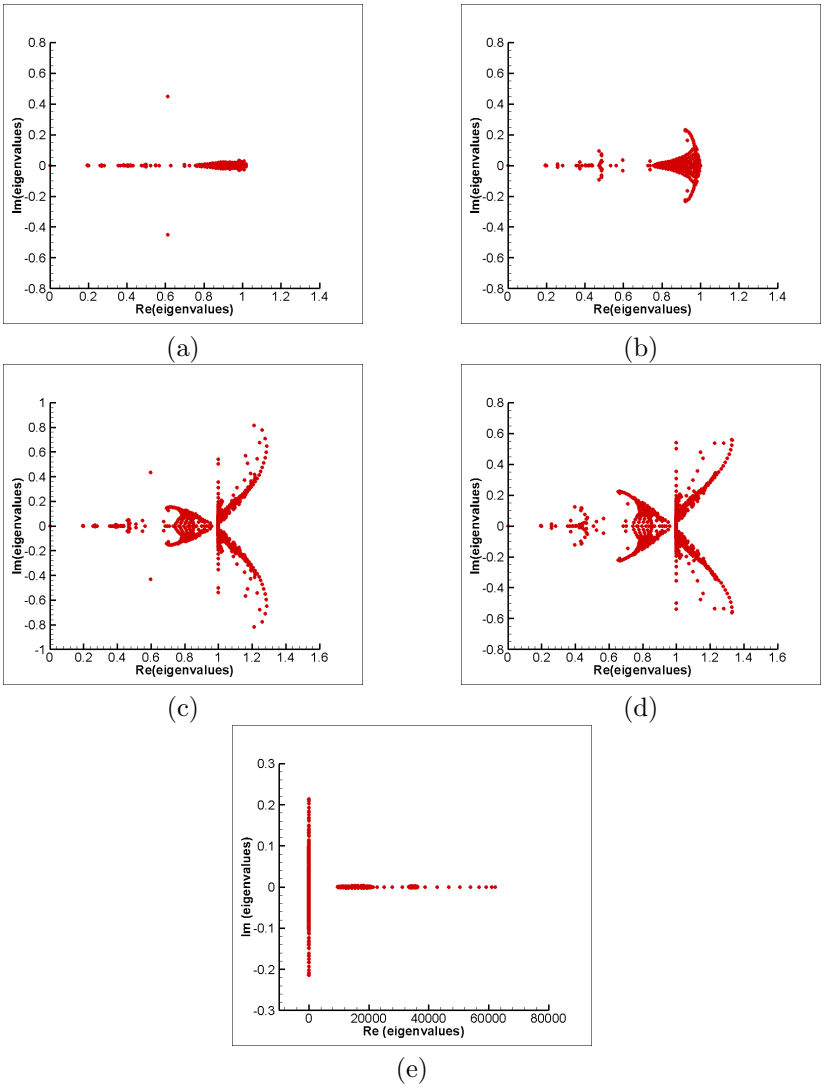

Figure 1: Eigenvalues for the Picard system for $N = 16$ with (a) preconditioner P_{R1}, (b) preconditioner P_{R2}, (c) preconditioner P_{R3}, (d) preconditioner P_{R4}, (e) no preconditioner.

changed from $-A_p F_p^{-1} M_p$ to $-M_p$. For each analyzed case and for each pair both the computational time and the number of linear iterations are comparable. At low Rayleigh number P_{R2} is slightly better than P_{R1} while at high Rayleigh number P_{R1} is the better one. In general P_{R4} performs slightly better that P_{R3}.

Ra	2×10^2		2×10^3	
N	Picard	FGMRES (Timing)	Picard	FGMRES (Timing)
32	3	27.7 (51.3s)	12	30.7 (225.2s)
64	3	29.3 (202.1s)	11	31.8 (837s)
128	2	31 (534.7s)	9	33.6 (2825s)

Table 1: Picard method with preconditioner P_{R1}

Ra	2×10^2		2×10^3	
N	Picard	FGMRES (Timing)	Picard	FGMRES (Timing)
32	3	28 (49.6s)	12	33.6 (235.3s)
64	3	29.3 (196.3s)	11	35.4 (898.5s)
128	2	29.5 (499.4s)	9	36.4 (2971s)

Table 2: Picard method with preconditioner P_{R2}

Ra	2×10^2		2×10^3	
N	Picard	FGMRES (Timing)	Picard	FGMRES (Timing)
32	3	31.7 (31.6s)	12	51 (189.7s)
64	3	33.7 (114s)	11	53.1 (635.5s)
128	2	34 (309.8s)	9	53.8 (2118s)

Table 3: Picard method with preconditioner P_{R3}

Ra	2×10^2		2×10^3	
N	Picard	FGMRES (Timing)	Picard	FGMRES (Timing)
32	3	31.7 (28.2s)	12	49.2 (165.9s)
64	3	33.3 (108.2s)	11	51.2 (572.4s)
128	2	33.5 (283.2s)	9	51.9 (1918s)

Table 4: Picard method with preconditioner P_{R4}

Next we consider other two pairs of preconditioners P_{R1} versus P_{R3}, and P_{R2} versus P_{R4}. In each pair P_1 changes from F to $A_{\boldsymbol{u}}$ and P_3 changes from K to A_T. The number of linear iterations increases from P_{R1} to P_{R3} and from P_{R2} to

P_{R4}. This is more evident at Rayleigh number 2×10^3, and it can be explained by looking at the eigenvalue distribution in the figures. For example, the eigenvalues in Figure 1(a) are closer to 1 than the ones in Figure 1(c). Consequently GMRES with preconditioner P_{R1} need fewer iterations to converge than the one with preconditioner P_{R3}. On the other hand, the preconditioners P_{R3} and P_{R4} perform much better than P_{R1} and P_{R2} in terms of computational time, and especially at high Rayleigh number. As already pointed out this substantial gain in the timing arises from the fact that both the $A_{\boldsymbol{u}}$ and A_T blocks are solution-independent and can be solved only once outside the nonlinear loop. Moreover since they are also SPD we can use a CG subsolver instead of GMRES which is used to invert the F and K in P_{R1} and P_{R2}.

Finally, in each simulation we observed that P_{R4} is the best preconditioner in terms of computational time.

8. Conclusions

We presented an analysis of block preconditioners for fixed-point linearizations of the Rayleigh-Bénard convection problem, discretized with inf-sup stable finite element spaces. In our analysis we considered either norm-equivalence or FOV-equivalence between the linearized systems and right preconditioners. Using these equivalences we proved that the total number of GMRES iterations is independent of the mesh size. Four different preconditioners were investigated. We showed that the eigenvalues of all preconditioned systems cluster around one. Preconditioner P_{R1} is the one whose clustering is the most effective, followed by preconditioners P_{R2}, P_{R3} and P_{R4}. In the numerical result session we confirmed our theoretical findings, showing that using each considered preconditioner the total number of GMRES iterations is independent of the mesh size. In accordance with the eigenvalue clustering we observed that P_{R1} requires the least number of iterations, followed by P_{R2}, P_{R3} and P_{R4}. However in terms of computational time preconditioners P_{R3} and P_{R4} work much better than P_{R1} and P_{R2}, since the computational time for each iteration of P_{R3} and P_{R4} is considerably less expensive than each iteration of P_{R1} and P_{R2}. Moreover among P_{R3} and P_{R4}, we found that preconditioner P_{R4} is the better one in terms of computational time.

9. Acknowledgments

This work was supported by the National Science Foundation grant DMS-1412796. The research of Dr. E. Aulisa was partially supported by the NSF grant DMS-1912902.

References

[1] Aulisa, E., Calandrini, S., Capodaglio, G., 2018. Fov-equivalent block triangular preconditioners for generalized saddle-point problems. Applied Mathematics Letters 75, 43–49.

[2] Benzi, M., Olshanskii, M. A., 2011. Field-of-values convergence analysis of augmented lagrangian preconditioners for the linearized navier-stokes problem. SIAM Journal on Numerical Analysis 49 (2), 770–788.

[3] Cai, M., Mu, M., Xu, J., 2009. Preconditioning techniques for a mixed stokes/darcy model in porous media applications. Journal of computational and applied mathematics 233 (2), 346–355.

[4] Chidyagwai, P., Ladenheim, S., Szyld, D. B., 2016. Constraint preconditioning for the coupled stokes–darcy system. SIAM Journal on Scientific Computing 38 (2), A668–A690.

[5] Eiermann, M., 1993. Fields of values and iterative methods. Linear Algebra and its Applications 180, 167–197.

[6] Eisenstat, S. C., Elman, H. C., Schultz, M. H., 1983. Variational iterative methods for nonsymmetric systems of linear equations. SIAM Journal on Numerical Analysis 20 (2), 345–357.

[7] Elman, H., Howle, V. E., Shadid, J., Shuttleworth, R., Tuminaro, R., 2006. Block preconditioners based on approximate commutators. SIAM Journal on Scientific Computing 27 (5), 1651–1668.

[8] Elman, H., Silvester, D., 1996. Fast nonsymmetric iterations and preconditioning for Navier-Stokes equations. SIAM Journal on Scientific Computing 17 (1), 33–46.

[9] Elman, H. C., 1982. Iterative methods for large, sparse, nonsymmetric systems of linear equations. Ph.D. thesis, Yale University New Haven, Conn.

[10] Ferziger, J. H., Peric, M., 1996. Computational methods for fluid dynamics. Springer, Heidelberg.

[11] Gee, M. W., Siefert, C. M., Hu, J. J., Tuminaro, R. S., Sala, M. G., 2006. ML 5.0 smoothed aggregation users guide. Tech. rep., Technical Report SAND2006-2649, Sandia National Laboratories.

[12] Gelhard, T., Lube, G., Olshanskii, M. A., Starcke, J.-H., 2005. Stabilized finite element schemes with lbb-stable elements for incompressible flows. Journal of Computational and Applied Mathematics 177 (2), 243–267.

[13] Girault, V., Raviart, P.-A., 1986. Finite element methods for Navier-Stokes equations: theory and algorithms. Springer-Verlag, Berlin Heidelberg New York Tokyo.

[14] Gresho, P. M., Sani, R. L., 1998. Incompressible flow and the finite element method. Volume one: Advection-diffusion and isothermal laminar flow. John Wiley and Sons, Ltd., West Sussex, England.

[15] Horn, R. A., Johnson, C. R., 2012. Matrix analysis. Cambridge university press.

[16] Howle, V. E., Kirby, R. C., 2012. Block preconditioners for finite element discretization of incompressible flow with thermal convection. Numerical Linear Algebra with Applications 19 (2), 427–440.

[17] Kay, D., Loghin, D., Wathen, A., 2002. A preconditioner for the steady-state Navier–Stokes equations. SIAM Journal on Scientific Computing 24 (1), 237–256.

[18] Ke, Guoyi and Aulisa, Eugenio and Bornia, Giorgio and Howle, Victoria, 2017. Block triangular preconditioners for linearization schemes of the Rayleigh-Bénard convection problem. Numerical Linear Algebra with Applications, e2096–n/ae2096 nla.2096.
URL http://dx.doi.org/10.1002/nla.2096

[19] Klawonn, A., Starke, G., 1999. Block triangular preconditioners for nonsymmetric saddle point problems: field-of-values analysis. Numerische Mathematik 81 (4), 577–594.

[20] Loghin, D., Wathen, A. J., 2004. Analysis of preconditioners for saddle-point problems. SIAM Journal on Scientific Computing 25 (6), 2029–2049.

[21] Lohse, D., Xia, K.-Q., 2010. Small-scale properties of turbulent Rayleigh-Bénard convection. Annual Review of Fluid Mechanics 42, 335–364.

[22] Oresta, P., Verzicco, R., Lohse, D., Prosperetti, A., 2009. Heat transfer mechanisms in bubbly Rayleigh-Bénard convection. Physical Review E 80 (2), 026304.

[23] Starke, G., 1997. Field-of-values analysis of preconditioned iterative methods for nonsymmetric elliptic problems. Numerische Mathematik 78 (1), 103–117.

[24] Tran, C. T., Kudinov, P., Dinh, T.-N., 2010. An approach to numerical simulation and analysis of molten corium coolability in a boiling water reactor lower head. Nuclear Engineering and Design 240 (9), 2148–2159.