16th World Congress in Computational Mechanics (WCCM)
21-26 July 2024, Vancouver, Canada
A. Korobenko, M. Laforest, S. Prudhomme, R. Vaziri (Eds)

ON THE ITERATIVE SOLUTION OF SADDLE POINT PROBLEMS USING A SYMMETRIC POSITIVE DEFINITE PRECONDITIONER

PHILIPPE DEVLOO[1], **GIOVANE AVANCINI**[2] **AND MARINA MENEGHEL**[3]

[1] State University of Campinas, Faculty of Civil Engineering
Av. Albert Einstein, 901 - Cidade Universitária, Campinas - SP, 13083-852
phil@unicamp.br

[2] State University of Campinas, Faculty of Civil Engineering
Av. Albert Einstein, 901 - Cidade Universitária, Campinas - SP, 13083-852
giovanea@unicamp.br

[3] State University of Campinas, Faculty of Civil Engineering
Av. Albert Einstein, 901 - Cidade Universitária, Campinas - SP, 13083-852
m240534@dac.unicamp.br

Key words: Iterative method, saddle-point problem, positive-definite preconditioner, H(div) approximation

Abstract. Saddle point problems frequently appear in many mathematical and engineering applications. Most systems of partial differential equations with constraints give rise to saddle point linear systems. Typical examples include mixed finite element formulations to solve fluid flows and/or elasticity problems under full incompressibility. The inversion of saddle point problems is challenging due to inherent numerical instability in the direct inversion methods. Many direct and iterative methods have been proposed to overcome this challenges, such as the Schur complement and the Uzawa's method. In the context of mixed finite element for incompressible flows using stable H(div)-L2 spaces for velocity and pressure, we propose an iterative method that can effectively solve a saddle point problem iteratively by summing a small compressibility to the original matrix. The preconditioning matrix is symmetric positive-definite, which allows the usage of Cholesky decomposition and/or CG-like iterative solvers to compute the incremental solution for the velocities unknowns. A procedure to compute the average pressure of each element of the incompressible problem is developed using the unbalanced fluxes caused by the compressibility perturbation. The average is updated during the iterative process as a function of the velocity increment at each iteration.

1 INTRODUCTION

Saddle point problems frequently appear in many mathematical and engineering applications. Most sets of partial differential equations with constraints give rise to saddle point linear systems. This is the case, for instance, when one uses mixed finite element formulations to solve incompressible problems such as Darcy flows in a porous media, Navier-Stokes flows and elasticity [1, 2, 3].

The numerical solution of saddle point problems is one of the most challenging in numerical analysis for many reasons. This family of matrices often shows poor spectral properties, and indefiniteness due to the null diagonal block related to the constraint equations. The balance between the approximation space and restraint space is delicate and the inversion of the algebraic system of equations can become unstable if terms with large differences are present [4].

On the other hand, numerical inversion of symmetric positive-definite systems are inherently stable: numerical perturbations introduced by round-off will necessarily be attenuated [5].

A main concern when using iterative methods is to guarantee the convergence within an acceptable number of iterations. Usually, this condition is directly related to the matrix spectral properties. In practice, the strategies to solve a saddle-point problem are divided into two groups - in the first one are the methods that compute the fields in a staggered manner. The most common method in this category is the Uzawa's method [6] and its variations. In the second group are the methods that solve the fields simultaneously [7].

In the context of mixed finite element applied to Darcy equations using stable $H(\text{div})$-L^2 spaces for flux and pressure, we have been developing an iterative method that can effectively solve a saddle point problem by introducing a small compressibility to the original matrix allowing for the static condensation of pressures. The resulting matrix is symmetric positive-definite, which allows the usage of Cholesky decomposition or CG-like iterative solvers to compute the incremental solution for the velocities or displacement unknowns. The pressure correction is shown to be proportional to the unbalanced force caused by the compressibility perturbation, and can be explicitly updated during the iterative process once the state variable increment is obtained.

2 DARCY PROBLEM

Let Ω be an open domain with Lipschitz boundary $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$, where $\partial\Omega_D$ and $\partial\Omega_N$ stand for the Dirichlet and Neumann boundaries, respectively. The mixed form of Darcy equations consist of finding the flux $\boldsymbol{\sigma} \in H(\text{div}, \Omega)$ and the pressure $u \in L^2(\Omega)$ such that

$$\begin{aligned} \boldsymbol{\sigma} = -\mathcal{K}\nabla u, &\quad \text{in} \quad \Omega, \\ \nabla \cdot \boldsymbol{\sigma} = f, &\quad \text{in} \quad \Omega, \\ u = u_D, &\quad \text{on} \quad \partial\Omega_D, \\ \boldsymbol{\sigma} \cdot \boldsymbol{n} = g, &\quad \text{on} \quad \partial\Omega_N, \end{aligned} \tag{1}$$

where $\mathcal{K}$ is the permeability tensor, $f \in L^2(\Omega)$ is the source term, $u_D \in H^{1/2}(\Omega)$ is the Dirichlet boundary condition, $g \in L^2(\Omega)$ is the Neumann boundary condition and $\mathbf{n}$ is the outward normal vector to $\partial\Omega_N$.

2.1 Weak statement

Applying the standard Galerkin method to Eqs. (1), the weak form of the Darcy problem reads: find $\boldsymbol{\sigma} \in H(\text{div}, \Omega)$ and $u \in L^2(\Omega)$ such that for all $\mathbf{w}_{\boldsymbol{\sigma}} \in H(\text{div}, \Omega)$ and $w_u \in L^2(\Omega)$:

$$\begin{aligned} \int_\Omega \mathcal{K}^{-1}\boldsymbol{\sigma} \cdot \mathbf{w}_{\boldsymbol{\sigma}} \, d\Omega - \int_\Omega u \nabla \cdot \mathbf{w}_{\boldsymbol{\sigma}} \, d\Omega = - \int_{\partial\Omega_D} u_D [\![\mathbf{w}_{\boldsymbol{\sigma}}]\!] \, d\partial\Omega, \\ \int_\Omega \nabla \cdot \boldsymbol{\sigma} w_u \, d\Omega = \int_\Omega f w_u \, d\Omega, \end{aligned} \tag{2}$$

where $[\![\bullet]\!]$ refers to the jump operator.

2.2 Finite element discretization

The discretization of Eqs. (2) is performed using the Finite Element Method [8]. We employ an approximation space based on stable $H(\text{div})$-L^2 pair which is De Rham compatible [9, 10]. A conformal $\mathcal{T} = \{\Omega_e, e = 1, \cdots, n_e\}$ of Ω in n_e finite elements Ω_e is defined. Let $\gamma(h, k)$ be a discretization parameter that solely depends on the mesh size h and the polynomial degree k, then the finite element spaces are defined as follows:

$$\mathbf{V}^\gamma = \{\boldsymbol{\sigma} \in H(\text{div}, \Omega) : \boldsymbol{\sigma}|_{\Omega_e} \in \mathbf{V}(\Omega_e), \forall \Omega_e \in \mathcal{T}\}, \tag{3}$$

$$W^\gamma = \{u \in L^2(\Omega) : u|_{\Omega_e} \in W(\Omega_e), \forall \Omega_e \in \mathcal{T}\}. \tag{4}$$

Rewriting Eqs. (2) using the finite element spaces from Eqs. (3)-(4), the discretized problem thus reads: find $\boldsymbol{\sigma}^\gamma \in \mathbf{V}^\gamma$ and $u^\gamma \in W^\gamma$ such that for all $\mathbf{w}^\gamma_{\boldsymbol{\sigma}} \in \mathbf{V}^\gamma$ and $w^\gamma_u \in W^\gamma$:

$$\begin{aligned} \sum_{\Omega_e \in \mathcal{T}} \left(\int_{\Omega_e} \mathcal{K}^{-1}\boldsymbol{\sigma}^\gamma \cdot \mathbf{w}^\gamma_{\boldsymbol{\sigma}} \, d\Omega - \int_{\Omega_e} u^\gamma \nabla \cdot \mathbf{w}^\gamma_{\boldsymbol{\sigma}} \, d\Omega \right) = - \sum_{\Omega_e \in \partial\Omega_D} \int_{\partial\Omega_e} u_D [\![\mathbf{w}^\gamma_{\boldsymbol{\sigma}}]\!] \, d\partial\Omega, \\ \sum_{\Omega_e \in \mathcal{T}} \int_{\Omega_e} \nabla \cdot \boldsymbol{\sigma}^\gamma w^\gamma_u \, d\Omega = \sum_{\Omega_e \in \mathcal{T}} \int_{\Omega_e} f w^\gamma_u \, d\Omega. \end{aligned} \tag{5}$$

3 ITERATIVE SOLUTION USING A SYMMETRIC POSITIVE-DEFINITE PRECONDITIONER

Equations (5) can be expressed in matrix form as:

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\sigma}^\gamma \\ \mathbf{u}^\gamma \end{bmatrix} = \begin{bmatrix} \mathbf{f}^\gamma_{\boldsymbol{\sigma}} \\ \mathbf{f}^\gamma_u \end{bmatrix}, \tag{6}$$

where matrix $\mathbf{A}$ represents the flux contribution and matrix $\mathbf{B}$ is the divergent operator which plays the role of imposing a constraint to the solution. The right-hand side vectors $\mathbf{f}^\gamma_{\boldsymbol{\sigma}}$ and $\mathbf{f}^\gamma_u$ are the Dirichlet contribution and source term, respectively. This latter often is assumed to be zero.

It is worth a discussion about some properties of problem (6). Matrix $\mathbf{A}$ is symmetric positive-definite and contains contributions from facet and internal fluxes. The latter ones can be eliminated from the global system by using a static condensation procedure [11]. Using the element-wise divergence constant approximation space of [9] results in a single pressure unknown per element. Thus, the global system comprises only facet fluxes and elemental pressures contributions.

Assuming a solution vector $\mathbf{u} = \{\mathbf{u}_1, \mathbf{u}_2\}^T$ such that $\mathbf{B}^T.\mathbf{u}_1 = \mathbf{0}$ (i.e. $\mathbf{u}_1 \in Ker\left(\mathbf{B}^T\right)$), then:

$$\{\mathbf{u}_1, \mathbf{u}_2\} \cdot \mathbf{G} \cdot \begin{Bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{Bmatrix} = \mathbf{u}_1^T \cdot \mathbf{A} \cdot \mathbf{u}_1 > 0 \tag{7}$$

We propose a modified matrix

$$\tilde{\mathbf{G}} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & -\mathbf{C} \end{bmatrix} \tag{8}$$

where matrix $\mathbf{C}$ is symmetric positive-definite. Once $\mathbf{C}$ plays the role of adding an artificial compressibility to the system, the static condensation procedure can also be applied to eliminate the pressure unknowns:

$$\tilde{\mathbf{G}}/\mathbf{C} = \mathbf{A} + \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^T \tag{9}$$

where $\tilde{\mathbf{G}}/\mathbf{C}$ is also symmetric positive-definite and can be seen as the Schur complement of block $\mathbf{C}$ of matrix $\tilde{\mathbf{G}}$.

The proposed iterative method uses matrix $\tilde{\mathbf{G}}/\mathbf{C}$ as a preconditioner for matrix $\mathbf{G}$. Given an approximate solution $\mathbf{u}^k$, where k stands for the iteration counter, an updated solution $\mathbf{u}^{k+1}$ can be retrieved by doing: first, the residual $\mathbf{r}^k$ is computed as:

$$\mathbf{r}^k = \mathbf{f} - \mathbf{G}\mathbf{u}^k \tag{10}$$

with $\mathbf{f}$ standing for the external forces. Then, we compute the solution increment:

$$\Delta\mathbf{u}^k = \tilde{\mathbf{G}}^{-1}\mathbf{r}^k \tag{11}$$

finally yielding

$$\mathbf{u}^{k+1} = \mathbf{u}^k + \Delta\mathbf{u}^k \tag{12}$$

In engineering terms, the matrix of an incompressible problem is preconditioned by a matrix corresponding to a slightly compressible system. To demonstrate its convergence, we start from the basic description of the iterative method:

$$\Delta\mathbf{u}^k = \mathbf{u}^{k+1} - \mathbf{u}^k = \tilde{\mathbf{G}}^{-1}\left(\mathbf{f} - \mathbf{G}\mathbf{u}^k\right) \tag{13}$$

$$\Delta\mathbf{u}^{k+1} = \mathbf{u}^{k+2} - \mathbf{u}^{k+1} = \tilde{\mathbf{G}}^{-1}\left(\mathbf{f} - \mathbf{G}\mathbf{u}^{k+1}\right) \tag{14}$$

such that

$$\Delta\mathbf{u}^{k+1} - \Delta\mathbf{u}^k = \tilde{\mathbf{G}}^{-1}\mathbf{G}\Delta\mathbf{u}^k \tag{15}$$

or

$$\Delta\mathbf{u}^{k+1} = \left(\mathbf{I} - \tilde{\mathbf{G}}^{-1}\mathbf{G}\right)\Delta\mathbf{u}^k \tag{16}$$

which implies that the iterative method will converge if $\left(\mathbf{I} - \tilde{\mathbf{G}}^{-1}\mathbf{G}\right)$ is a contraction i.e. if its maximum eigenvalue is less than one. One notices that for $\tilde{\mathbf{G}} = \mathbf{G}$, no solution increment is obtained, so the method converges in a single iteration.

From [12], we can approximate the inverse of $\tilde{\mathbf{G}}$ as:

$$\tilde{\mathbf{G}}^{-1} = \mathbf{G}^{-1} - \mathbf{G}^{-1}\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & -\mathbf{C}\end{bmatrix}\mathbf{G}^{-1} + O(\mathbf{C}) \tag{17}$$

so neglecting higher order terms yeilds:

$$\left(\mathbf{I} - \tilde{\mathbf{G}}^{-1}\mathbf{G}\right) \simeq \mathbf{G}^{-1}\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & -\mathbf{C}\end{bmatrix} \tag{18}$$

demonstrating that the convergence of the iterative method can be controlled by the size of $\mathbf{C}$. For all the examples analyzed, we arbitrarily chose $\mathbf{C} = \alpha\mathbf{I}$, where α refers to the artificial compressibility parameter.

4 NUMERICAL RESULTS

In this section, a tridimensional problem is used to assess the performance and stability of the proposed iterative method. In all analyses, convergence is reached when the Euclidian norm of the residual is less than 10^{-9}, and the maximum number of iterations is set to 50. The solver step is performed with 8 threads on a workstation with an Intel(R) Xeon(R) Gold 6130 2.10GHz CPU.

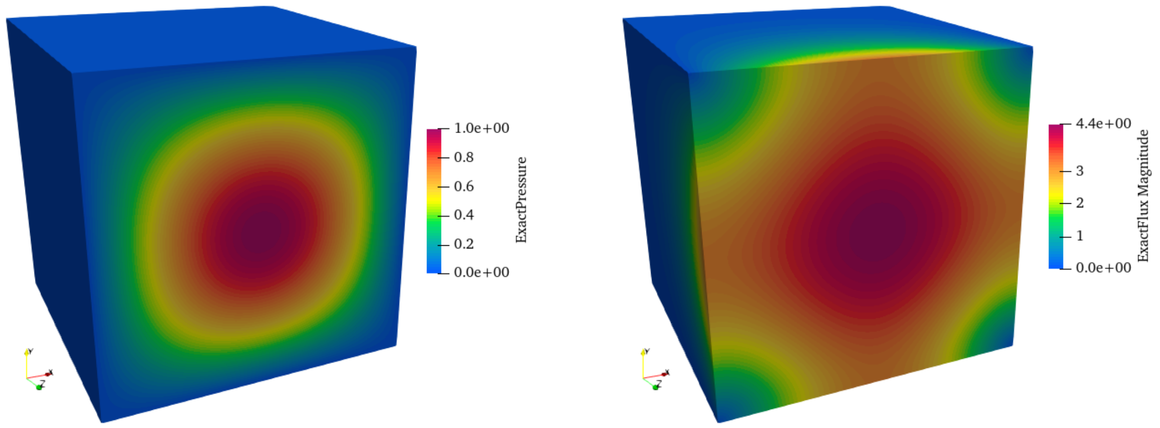
4.1 3D Darcy problem

The computational domain consists on a unit cube $\Omega = (0,1)\times(0,1)\times(0,1)$ and a uniform mesh of hexahedral elements with $k = 2$ and characteristic size $h_e = 1/n$, where $n = n_x = n_y = n_z$ is the number of elements used in each dimension, starting from $n = 5$ up to $n = 60$. The permeability tensor is assumed to be isotropic and unitary, so $\mathcal{K} = \mathbf{I}_3$. The following harmonic solution is adopted for the pressure field:

$$u(x,y,z) = \frac{\sin(\pi x)\sin(\pi y)\sinh(\sqrt{2}\pi z)}{\sinh(\sqrt{2}\pi)}. \tag{19}$$

It is straightforward to derive the corresponding flux field from Eq. (19). Figure 1 shows the analytic solutions over the domain. Dirichlet boundary conditions are imposed on all the facets.

Firstly, we analyse the convergence of the iterative method as a function of the compressibility parameter α. Figure 2 shows the number of iterations required to reach convergence for different values of α. It is evident that α plays a key role on the method efficiency, as the higher it gets, the more iterations are required to reach convergence. As the value decreases, we observe a faster convergence as the preconditioner system approximates to the original matrix. However, for excessive small values of α, the method shows a loss in the precision. This behaviour is expected as the lower the α, the problem tends to the original indefinite saddle-point matrix where Cholesky decomposition likely will fail to find a solution. The optimal value of α



(a)

(b)

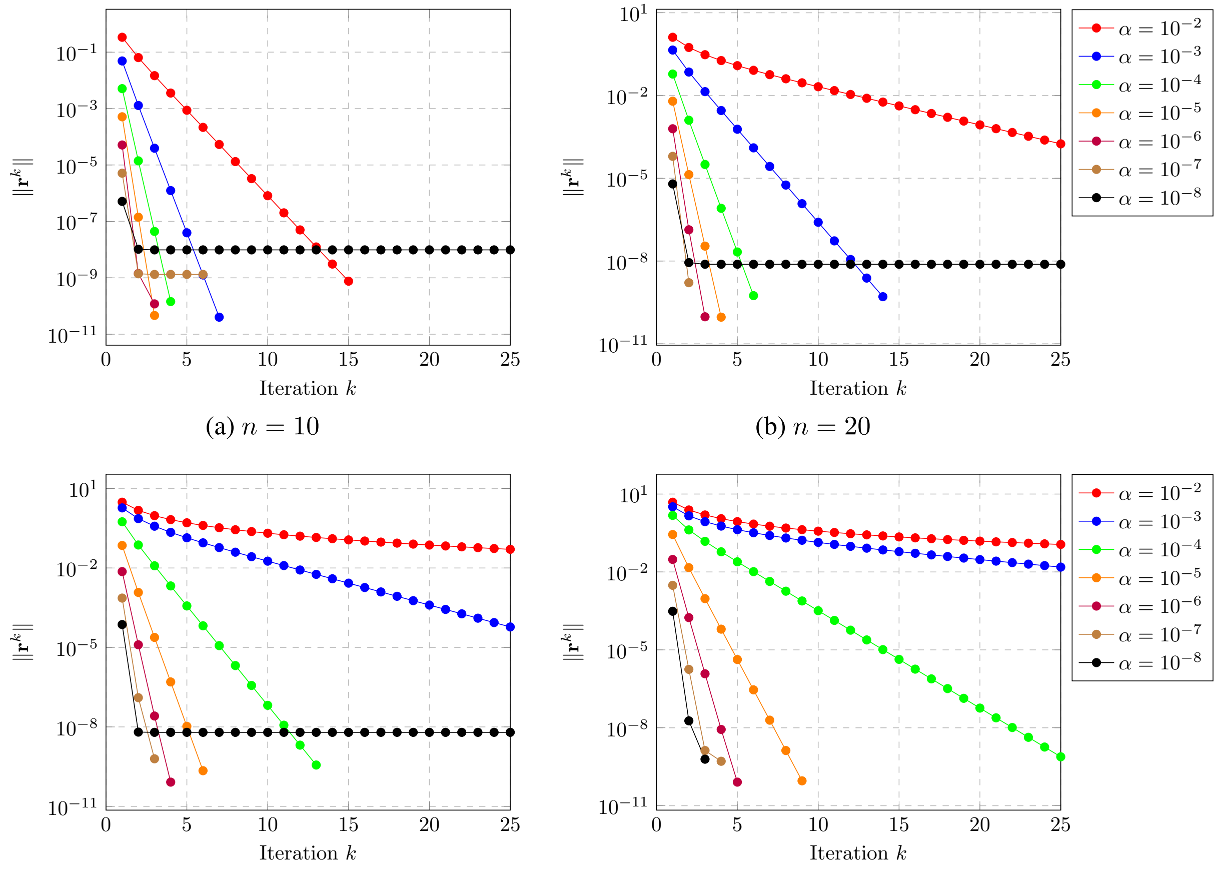
Figure 1: 3D Darcy problem - analytic solution for the pressure (at left) and flux magnitude (at right).

also varies with the mesh size. The more refined the mesh, the smaller is the magnitude of the elemental contributions in the global matrix, so the compressibility parameter must be adjusted accordingly.

The computational cost of the proposed iterative method is also investigated and displayed in Table 1. For these analyses, we use the Cholesky implementation of PARDISO [13] to build the positive-definite preconditioner for the iterative solver, while the solution of the original saddle-point problem is performed with the LDLT decomposition from PARDISO as well. It shall be highlited that matrix $\tilde{\mathbf{G}}/\mathbf{C}$ only needs to be decomposed once within the proposed iterative scheme. Thus, the total time is computed as the sum of the time required to build the preconditioner plus the time spent during each iteration solve.

Table 1: 3D Darcy problem - Computational cost of the proposed iterative method as a function of the compressibility parameter α.

Method	Time (ms)					
	$n = 5$	$n = 10$	$n = 20$	$n = 40$	$n = 50$	$n = 60$
$\alpha = 10^{-2}$	133	756	20 605	764 405	-	-
$\alpha = 10^{-3}$	106	541	7 000	279 441	-	-
$\alpha = 10^{-4}$	97	458	4 869	145 215	521 451	1 552 290
$\alpha = 10^{-5}$	100	441	4 308	127 344	460 492	1 335,251
$\alpha = 10^{-6}$	95	432	4 065	115 296	425 015	1 281 101
$\alpha = 10^{-7}$	137	507	4 055	112 178	416 897	1 267 023
$\alpha = 10^{-8}$	-	-	-	-	-	1 253 520
Direct	80	379	4 150	150 839	540 429	1 757 863



(a) $n = 10$

(b) $n = 20$

(c) $n = 40$

(d) $n = 60$

Figure 2: 3D Darcy problem - Number of iterations required to reach convergence as a function of the compressibility parameter α.

It can be noticed that, for an optimum value of α, the iterative method starts to be more efficient than the direct solver for $n \geq 20$. For the finest mesh, a speedup of up to 30 % can be achieved for $\alpha = 10^-8$. However, further studies are needed to determine the optimal value of α as a function of the mesh size. Figure 3 shows the time spent during the solution step using the proposed method with the optimum α as a percentage of the time consumed by the direct solver.

5 CONCLUSIONS

In this work, an iterative method was proposed based on the introduction of a small compressibility to solve saddle point problems using a symmetric positive-definite preconditioner. This scheme not only reduces the number of unkonwns in the global system but also allows the usage of optimized solvers such as Cholesky decomposition and CG-like methods to compute the fluxes increment. The use of an iterative solver to also decompose the preconditioner matrix

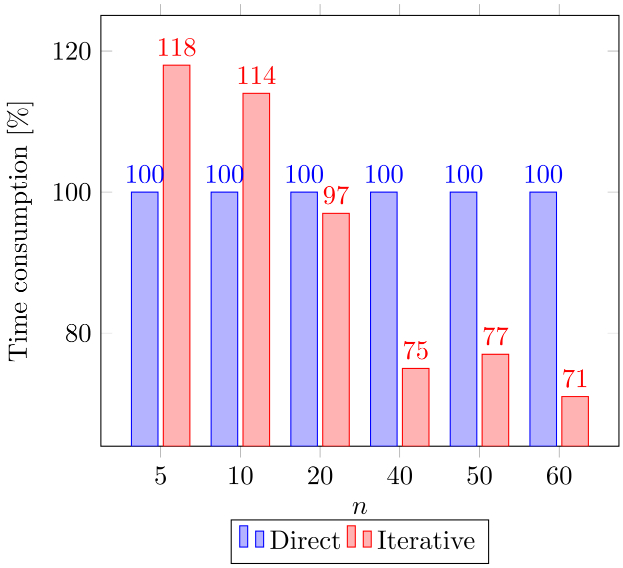


Figure 3: 3D Darcy problem - Time consumption obtained using the optimum value of α compared to the direct method for different levels of refinement.

shall be studied in future works.

The numerical results demonstrated that as the compressibility parameter increases, the number of iterations required to reach convergence also increases. However, for smaller values of α, the method converges in less than 3 iterations. Also, an excessive small value of α can lead to numerical instability, as the system becomes ill-conditioned. Therefore, a criterion on how to determine the optimal value of α a priori must be investigated in the future.

As the number of degrees of freedom increases, the proposed method demonstrated to be more efficient than the direct solver. Choosing the apropriate α, the computational cost was reduced by up to 30 % for the most refined mesh.

Acknowledgements. We gratefully acknowledge the support of EPIC - Energy Production Innovation Center through FAPESP/Equinor (Grant 2023/06981/-5), Total Energies Brazil through FUNCAMP (Process 76042-23) and the Brazilian National Council for Scientific and Technological Development (grants 305823/2017-5 and 309597/2021-8). We also acknowledge the support of ANP (Brazil's National Oil, Natural Gas and Biofuels Agency).

REFERENCES

[1] F. Brezzi and M. Fortin, *Mixed and hybrid finite element methods*, vol. 15. Springer Science & Business Media, 2012.

[2] O. Duran, P. R. Devloo, S. M. Gomes, and F. Valentin, "A multiscale hybrid method for darcy's problems using mixed finite element local solvers," *Computer methods in applied*

mechanics and engineering, vol. 354, pp. 213–244, 2019.

[3] U. Brink and E. Stein, “On some mixed finite element methods for incompressible and nearly incompressible finite elasticity,” *Computational Mechanics*, vol. 19, no. 1, pp. 105–119, 1996.

[4] R. E. Bank, B. D. Welfert, and H. Yserentant, “A class of iterative methods for solving saddle point problems,” *Numerische Mathematik*, vol. 56, no. 7, pp. 645–666, 1989.

[5] J. Rohn, “Positive definiteness and stability of interval matrices,” *SIAM Journal on Matrix Analysis and Applications*, vol. 15, no. 1, pp. 175–184, 1994.

[6] H. Uzawa, “Iterative methods for concave programming,” *Studies in linear and nonlinear programming*, vol. 6, pp. 154–165, 1958.

[7] M. Benzi, G. H. Golub, and J. Liesen, “Numerical solution of saddle point problems,” *Acta numerica*, vol. 14, pp. 1–137, 2005.

[8] E. Becker, G. Carey, and J. Oden, *Finite Elements: An introduction*. No. v. 1 in ACM monograph series, Prentice-Hall, 1981.

[9] P. R. Devloo, J. W. Fernandes, S. M. Gomes, F. T. Orlandini, and N. Shauer, “An efficient construction of divergence-free spaces in the context of exact finite element de rham sequences,” *Computer Methods in Applied Mechanics and Engineering*, vol. 402, p. 115476, 2022.

[10] D. De Siqueira, P. R. Devloo, and S. M. Gomes, “A new procedure for the construction of hierarchical high order hdiv and hcurl finite element spaces,” *Journal of Computational and Applied Mathematics*, vol. 240, pp. 204–214, 2013.

[11] R. J. Guyan, “Reduction of stiffness and mass matrices,” *AIAA journal*, vol. 3, no. 2, pp. 380–380, 1965.

[12] M. E. Gurtin, *An introduction to continuum mechanics*. Academic press, 1982.

[13] O. Schenk, K. Gärtner, and W. Fichtner, “Efficient sparse lu factorization with left-right looking strategy on shared memory multiprocessors,” *BIT Numerical Mathematics*, vol. 40, pp. 158–176, 2000.