Numerical Analysis and Scientific Computing
Preprint Seria

LU factorizations and ILU preconditioning for stabilized discretizations of incompressible Navier-Stokes equations

I. N. Konshin M. A. Olshanskii Yu. V. Vassilevski

Preprint #49

Department of Mathematics
University of Houston

May 2016

LU FACTORIZATIONS AND ILU PRECONDITIONING FOR STABILIZED DISCRETIZATIONS OF INCOMPRESSIBLE NAVIER-STOKES EQUATIONS *

IGOR N. KONSHIN†, MAXIM A. OLSHANSKII‡, AND YURI V. VASSILEVSKI§

Abstract. The paper studies numerical properties of LU and incomplete LU factorizations applied to the discrete linearized incompressible Navier-Stokes problem also known as the Oseen problem. A commonly used stabilized Petrov-Galerkin finite element method for the Oseen problem leads to the system of algebraic equations having a 2×2-block structure. While enforcing better stability of the finite element solution, the Petrov-Galerkin method perturbs the saddle-point structure of the matrix and may lead to less favourable algebraic properties of the system. The paper analyzes the stability of the LU factorization. This analysis quantifies the affect of the stabilization in terms of the perturbation made to a non-stabilized system. The further analysis shows how the perturbation depends on the particular finite element method, the choice of stabilization parameters, and flow problem parameters. The analysis of LU factorization and its stability further helps to understand the properties of threshold ILU factorization preconditioners for the system. Numerical experiments for a model problem of blood flow in a coronary artery illustrate the performance of the threshold ILU factorization as a preconditioner. The dependence of the preconditioner properties on the stabilization parameters of the finite element method is also studied numerically.

Key words. iterative methods, preconditioning, threshold ILU factorization, Navier–Stokes equations, finite element method, SUPG stabilization, haemodynamics

AMS subject classifications. 65F10, 65N22, 65F50.

1. Introduction. The paper addresses the question of developing fast algebraic solves for finite element discretizations of the linearized Navier-Stokes equations. The Navier-Stokes equations describe the motion of incompressible Newtonian fluids. For a bounded domain $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$, with boundary $\partial\Omega$, and time interval $[0, T]$, the equations read

$$\left\{\begin{aligned} \frac{\partial \mathbf{u}}{\partial t} - \nu\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p &= \mathbf{f} \ \text{ in } \Omega\times(0,T] \\ \operatorname{div}\mathbf{u} &= 0 \ \text{ in } \Omega\times[0,T] \\ \mathbf{u} = \mathbf{g} \ \text{ on } \Gamma_0\times[0,T], \quad -\nu(\nabla\mathbf{u})\cdot\mathbf{n} + p\mathbf{n} &= \mathbf{0} \ \text{ on } \Gamma_{\mathrm{N}}\times[0,T] \\ \mathbf{u}(\mathbf{x},0) &= \mathbf{u}_0(\mathbf{x}) \ \text{ in } \Omega. \end{aligned}\right. \tag{1.1}$$

The unknowns are the velocity vector field $\mathbf{u} = \mathbf{u}(\mathbf{x}, t)$ and the pressure field $p = p(\mathbf{x}, t)$. The volume forces $\mathbf{f}$, boundary and initial values $\mathbf{g}$ and $\mathbf{u}_0$ are given. Parameter ν is the kinematic viscosity; $\partial\Omega = \overline{\Gamma}_0 \cup \overline{\Gamma}_{\mathrm{N}}$ and $\Gamma_0 \neq \varnothing$. An important parameter of the flow is the dimensionless Reynolds number $\mathrm{Re} = \frac{UL}{\nu}$, where U and L are characteristic velocity and linear dimension. Solving (1.1) numerically is known to get harder for higher values of Re, in particular some special modelling of flow scales unresolved by the mesh may be needed. Implicit time discretization and linearization of the Navier–Stokes system (1.1) by Picard fixed-point iteration result in a sequence

*This work has been supported by Russian Science Foundation through the grant 14-31-00024.

†Institute of Numerical Mathematics, Institute of Nuclear Safety, Russian Academy of Sciences, Moscow; igor.konshin@gmail.com

‡Department of Mathematics, University of Houston; molshan@math.uh.edu

§Institute of Numerical Mathematics, Russian Academy of Sciences, Moscow Institute of Physics and Technology, Moscow; yuri.vassilevski@gmail.com

of (generalized) Oseen problems of the form

$$\begin{cases} \alpha\mathbf{u} - \nu\Delta\mathbf{u} + (\mathbf{w}\cdot\nabla)\mathbf{u} + \nabla p = \hat{\mathbf{f}} \ \text{ in } \Omega \\ \qquad\qquad\qquad\qquad \operatorname{div}\mathbf{u} = \hat{g} \ \text{ in } \Omega \\ \mathbf{u} = \mathbf{0} \ \text{ on } \Gamma_0, \quad -\nu(\nabla\mathbf{u})\cdot\mathbf{n} + p\mathbf{n} = \mathbf{0} \ \text{ on } \Gamma_{\mathrm{N}} \end{cases} \tag{1.2}$$

where $\mathbf{w}$ is a known velocity field from a previous iteration or time step and α is proportional to the reciprocal of the time step. Non-homogenous boundary conditions in the nonlinear problem are accounted in the right-hand side.

Finite element (FE) methods for (1.1) and (1.2) may suffer from different sources of instabilities. One is a possible incompatibility of pressure and velocity FE pairs. A remedy is a choice of FE spaces satisfying the inf-sup or LBB condition [13] or the use of pressure stabilizing techniques. A major source of instabilities stems from dominating inertia terms for large Reynolds numbers. There exist several variants of stabilized FE methods, which combine stability and accuracy, e.g. the streamline upwind Petrov-Galerkin (SUPG) method, the Galerkin/Least-squares, algebraic subgrid scale, and internal penalty techniques, see, e.g., [4, 7, 11, 27]. These methods simultaneously suppress spurious oscillations caused by both, dominating advection and non-LBB-stable FE spaces. The combination of LBB-stable velocity-pressure FE pairs with advection stabilization is also often used in practice and studied in the literature, see, e.g., [12, 36]. For numerical experiments and finite element analysis in this paper, we consider a variant of the SUPG method. Details of the method are given later in this paper.

A finite element spatial discretization of (1.2) results in large, sparse systems of the form

$$\begin{pmatrix} A & \widetilde{B}^T \\ B & -C \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}, \tag{1.3}$$

where u and p represent the discrete velocity and pressure, respectively, $A \in \mathbb{R}^{n\times n}$ is the discretization of the diffusion, convection, and time-dependent terms. The matrix A accounts also for certain stabilization terms. Matrices B and $\widetilde{B}^T \in \mathbb{R}^{n\times m}$ are (negative) discrete divergence and gradient. These matrices may also be perturbed due to stabilization. It is typical for the stabilized methods that $B \neq \widetilde{B}$, while for a plain Galerkin method these two matrices are the same. Matrix $C \in \mathbb{R}^{m\times m}$ results from possible pressure stabilization terms, and f and g contain forcing and boundary terms. For the LBB stable finite elements, no pressure stabilization is required and so $C = 0$ holds. If the LBB condition is not satisfied, the stabilization matrix $C \neq 0$ is typically symmetric and positive semidefinite. For $B = \widetilde{B}$ of the full rank and positive definite $A = A^T$ the solution to (1.3) is a saddle point. Otherwise, one often refers to (1.3) as a *generalized saddle point* system, see, e.g., [3].

Considerable work has been done in developing efficient preconditioners for Krylov subspace methods applied to system (1.3) with $\widetilde{B} = B$; see the comprehensive studies in [3, 9, 24] of the preconditioning exploiting the block structure of the system. A common approach is based on preconditioners for block A and pressure Schur complement matrix $S = BA^{-1}\widetilde{B}^T + C$, see [10, 25, 38] for recent developments. Well known block preconditioners are not completely robust with respect to variations of viscosity parameter, properties of advective velocity field $\mathbf{w}$, grid size and anisotropy ratio, and the domain geometry. The search of a more robust black-box type approach to solve algebraic system (1.3) stimulates an interest in developing preconditioners based on

incomplete factorizations. Clearly, computing a suitable incomplete LU factorizations of (1.3) is challenging and requires certain care for (at least) the following reasons. The matrix can be highly non-symmetric for higher Reynolds numbers flows; even in symmetric case the matrix is indefinite (both positive and negative eigenvalues occur in the spectrum); extra stabilization terms may break the positive definiteness of A and/or of the Schur complement. Nevertheless, a progress have been recently reported in developing incomplete LU preconditioners for saddle-point matrices and generalized saddle-point matrices. Thus the authors of [30,31] studied the signed incomplete Cholesky type preconditioners for symmetric saddle-point systems, corresponding to the Stokes problem. For the finite element discretization of the incompressible Navier–Stokes equations the authors of [8,37] developed ILU preconditioners, where the fill-in is allowed based on the connectivity of nodes rather than actual non-zeros in the matrix. The papers [32, 37] studied several reordering techniques for ILU factorization of (1.3) and found that some of the resulting preconditioners are competitive with the most advanced block preconditioners. Elementwise threshold incomplete LU factorizations for non-symmetric saddle point matrices were developed in [20]. In that paper, an extension of the Tismenetsky-Kaporin variant of ILU factorization for non-symmetric matrices is used as a preconditioner for the finite element discretizations of the Oseen equations. Numerical analysis and experiments with the (non-stabilized) Galerkin methods for the incompressible Navier-Stokes equations demonstrated the robustness and efficiency of this approach. An important advantage of preconditioners based on elementwise ILU decomposition is that they are straightforward to implement in standard finite element codes.

In the present paper we extend the method and analysis from [20] to the system of algebraic equations resulting from the *stabilized* formulations of the Navier-Stokes equations. Hence, we are interested in the numerically challenging case of higher Reynolds number flows. The effect of different stabilization techniques on the accuracy of finite element solutions is substantial and is well studied in the literature. However, not that much research has addressed the question of how the stabilization affects the algebraic properties of the discrete systems, see [9]. The present study intends to fill this gap. We analyze the stability of the (exact) LU factorization and numerical properties of a threshold ILU factorization for (1.3). One might expect that stabilization adds to the ellipticity of matrices and hence improves algebraic properties. This is certainly the situation in particular cases of scalar advection-diffusion equations and linear elements. However, for saddle-point problems and higher order elements the situation appears to be more delicate. In particular, algebraic stability may impose more restrictive bounds on the stabilization parameters than those satisfied by optimal parameters with respect to FE solution accuracy. We study the explicit dependence of algebraic properties of (1.3) on flow, stabilization and discretization parameters and show that larger values of the stabilization parameter may affect the algebraic stability. Therefore, for those fluid flow problems, which require SUPG stabilization, suitable parameters meet both restrictions: they are large enough to add necessary stability for the finite element solution, but not too large to guarantee stable factorizations of algebraic systems.

The remainder of the paper is organized as follows. In section 2 we give necessary details on the finite element method for the Oseen equations. Section 3 studies stability of the exact LU factorizations for (1.3). We derive the sufficient conditions for the existence and stability of the LU factorization without pivoting. These conditions and an estimate on the entries of the resulting LU factors are given in terms of the

properties of the (1,1)-block A, auxiliary Schur complement matrix $BA^{-1}B^T+C$, and the perturbation matrix $B-\widetilde{B}$. In section 4, we apply this analysis to system (1.3) arising from SUPG-stabilized FE discretization of the Oseen system. In section 5, we briefly discuss the implication of our analysis of LU factorization on the stability of a two-parameter Tismenetsky–Kaporin variant of the threshold ILU factorization for non-symmetric non-definite problems. This factorization is used in our numerical experiments. In section 6 we study the numerical performance of the method on the sequence of linear systems appearing in simulation of a blood flow in a right coronary artery. Section 7 collects conclusions and a few closing remarks.

2. FE method and SUPG stabilization. In this paper, we consider an inf-sup stable conforming FE method stabilized by the SUPG method. To formulate it, we first need the weak formulation of the Oseen problem. Let $\mathbf{V} := \{\mathbf{v} \in H^1(\Omega)^3 : \mathbf{v}|_{\Gamma_0} = \mathbf{0}\}$. Given $\mathbf{f} \in \mathbf{V}'$, the problem is to find $\mathbf{u} \in \mathbf{V}$ and $p \in L^2(\Omega)$ such that

$$\begin{aligned}
&\mathcal{L}(\mathbf{u},p;\mathbf{v},q) \;=\; (\mathbf{f},\mathbf{v})_* + (g,q) \qquad \forall\ \mathbf{v}\in\mathbf{V},\ q\in L^2(\Omega)\,,\\
&\mathcal{L}(\mathbf{u},p;\mathbf{v},q) := \alpha(\mathbf{u},\mathbf{v}) + \nu(\nabla\mathbf{u},\nabla\mathbf{v}) + ((\mathbf{w}\cdot\nabla)\,\mathbf{u},\mathbf{v}) - (p,\operatorname{div}\mathbf{v}) + (q,\operatorname{div}\mathbf{u})\,,
\end{aligned}$$

where $(\cdot,\cdot)$ denotes the $L^2(\Omega)$ inner product and $(\cdot,\cdot)_*$ is the duality paring for $\mathbf{V}'\times\mathbf{V}$.

We assume T_h to be a collection of tetrahedra which is a consistent subdivision of Ω satisfying the regularity condition

$$\max_{\tau\in T_h} \operatorname{diam}(\tau)/\rho(\tau) \le C_T, \tag{2.1}$$

where $\rho(\tau)$ is the diameter of the subscribed ball in the tetrahedron τ. A constant C_T measures the maximum anisotropy ratio for T_h. Further we denote $h_\tau = \operatorname{diam}(\tau)$, $h_{\min} = \min_{\tau\in T_h} h_\tau$. Given conforming FE spaces $\mathbb{V}_h \subset \mathbf{V}$ and $\mathbb{Q}_h \subset L^2(\Omega)$, the Galerkin FE discretization of (1.2) is based on the weak formulation: Find $\{\mathbf{u}_h, p_h\} \in \mathbb{V}_h \times \mathbb{Q}_h$ such that

$$\mathcal{L}(\mathbf{u}_h,p_h;\mathbf{v}_h,q_h) = (\mathbf{f},\mathbf{v}_h)_* + (g,q_h) \qquad \forall\,\mathbf{v}_h\in\mathbb{V}_h,\ q_h\in\mathbb{Q}_h\,. \tag{2.2}$$

In our experiments we shall use P2-P1 Taylor–Hood FE pair, which satisfies the LBB compatibility condition for $\mathbb{V}_h$ and $\mathbb{Q}_h$ [13] and hence ensures well-posedness and full approximation order for the FE linear problem.

A potential source of instabilities in (2.2) is the presence of dominating convection terms. This necessitates stabilization of the discrete system, if the mesh is not sufficiently fine to resolve all scales in the solution. We consider below one commonly used SUPG stabilization, while more details on the family of SUPG methods can be found in, e.g., [6,26,36]. Using (2.2) as the starting point, a weighted residual for the FE solution multiplied by an 'advection'-depended test function is added:

$$\begin{aligned}
\mathcal{L}(\mathbf{u}_h,p_h;\mathbf{v}_h,q_h) + \sum_{\tau\in T_h} \sigma_\tau(\alpha\mathbf{u}_h - \nu\Delta\mathbf{u}_h + \mathbf{w}\cdot\nabla\mathbf{u}_h + \nabla p_h - \mathbf{f}, \mathbf{w}\cdot\nabla\mathbf{v}_h)_\tau \\
= (\mathbf{f},\mathbf{v}_h) \qquad \forall\,\mathbf{v}_h\in\mathbb{V}_h,\ q_h\in\mathbb{Q}_h\,,
\end{aligned} \tag{2.3}$$

with $(f,g)_\tau := \int_\tau fg\,dx$. The second term in (2.3) is evaluated element-wise for each element $\tau \in T_h$. Parameters σ_τ are element- and problem-dependent. To define the parameters, we introduce mesh Reynolds numbers $\mathrm{Re}_\tau := \|\mathbf{w}\|_{L_\infty(\tau)} h_{\mathbf{w}}/\nu$ for all $\tau \in T_h$, where $h_{\mathbf{w}}$ is the diameter of τ in direction $\mathbf{w}$. Several recipes for the

particular choice of the stabilization parameters can be found in the literature. When we experiment with the stabilization, we set

$$\sigma_\tau = \begin{cases} \bar{\sigma}\dfrac{h_{\mathbf{w}}}{2\|\mathbf{w}\|_{L_\infty(\tau)}}\left(1-\dfrac{1}{\mathrm{Re}_\tau}\right), & \text{if } \mathrm{Re}_\tau > 1, \\ 0, & \text{if } \mathrm{Re}_\tau \le 1, \end{cases} \qquad \text{with } \ 0 < \bar{\sigma} < 1. \tag{2.4}$$

If one enumerates velocity unknowns first and pressure unknowns next, then the resulting discrete system has the 2×2-block form (1.3) with $C = 0$. The stabilization alters the (1,2)-block of the matrix making the latter not equal to the transpose of the (2,1)-block B. In this paper, we analyse factorizations for the matrix from (1.3) assuming that the perturbation of B^T in the (1,2)-block caused by (2.3) is relatively small due to the choice of σ_τ. The analysis and results of numerical experiments also show that the perturbation of A caused by (2.3) affects essentially the properties of LU and ILU decompositions.

We note that there was an intensive development of stabilized and multiscale finite element methods for fluid problems over last decade, see, for example, [7, 16] and references in more recent review papers [1, 4]. While these methods can be more accurate and less dissipative compared to (2.3), they add terms to the algebraic system of the same structure and similar algebraic properties as the SUPG method. The streamline diffusion stabilization as in (2.3) is a standard (and often the only available) option in many existing CFD software, so we decided to consider in the present studies this more classical approach as the particular example leading to the system (1.3).

3. Stability of LU factorization. The 2×2-block matrix from (1.3) is in general not sign definite and if $C = 0$, its diagonal has zero entries. An LU factorization of such matrices often requires pivoting (rows and columns permutations) for stability reasons. However, exploiting the block structure and the properties of blocks A and C, one readily verifies that the LU factorization

$$\mathcal{A} = \begin{pmatrix} A & \widetilde{B}^T \\ B & -C \end{pmatrix} = \begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix}\begin{pmatrix} U_{11} & U_{12} \\ 0 & -U_{22} \end{pmatrix} \tag{3.1}$$

with low (upper) triangle matrices L_{11}, L_{22} (U_{11}, U_{22}) exists without pivoting, once $\det(A) \neq 0$ and there exist LU factorizations for the (1,1)-block

$$A = L_{11}U_{11}$$

and the Schur complement matrix $\widetilde{S} := BA^{-1}\widetilde{B}^T + C$ is factorized as

$$\widetilde{S} = L_{22}U_{22}.$$

Decomposition (3.1) then holds with $U_{12} = L_{11}^{-1}\widetilde{B}^T$ and $L_{21} = BU_{11}^{-1}$.

Assume A is positive definite. Then the LU factorization of A exists without pivoting. Its numerical stability (the relative size of entries in factors L_{11} and U_{11}) may depend on how large is the skew-symmetric part of A comparing to the symmetric part. To make this statement more precise, we denote $A_{\mathrm{S}} = \frac{1}{2}(A + A^T)$, $A_{\mathrm{N}} = A - A_{\mathrm{S}}$ and let

$$C_A = \|A_{\mathrm{S}}^{-\frac{1}{2}} A_{\mathrm{N}} A_{\mathrm{S}}^{-\frac{1}{2}}\|.$$

Here and further, $\|\cdot\|$ and $\|\cdot\|_F$ denote the matrix spectral norm and the Frobenius norm, respectively, and $|M|$ denotes the matrix of absolute values of M-entries. The following bound on the size of elements of L_{11} and U_{11} holds (see eq.(3.2) in [20]):

$$\frac{\||L_{11}||U_{11}|\|_F}{\|A\|} \le n\,\left(1+C_A^2\right). \tag{3.2}$$

If $C \ge 0$, $\widetilde{B} = B$, and matrix B^T has the full column rank, then the positive definiteness of A implies that the Schur complement matrix is also positive definite. However, this is not the case for a general block $\widetilde{B} \ne B$. In the application studied in this paper, the (1,2)-block $\widetilde{B}^T$ is a *perturbation* of B^T. The analysis below shows that the positive definiteness of $\widetilde{S}$ and the stability of its LU factorization is guaranteed if the perturbation $E = \widetilde{B} - B$ is not too large. The size of the perturbation will enter our bounds as the parameter ε_E defined as

$$\varepsilon_E := \|A_{\rm S}^{-\frac{1}{2}} E^T\|.$$

For the ease of analysis we introduce further notations:

$$S = BA^{-1}B^T + C, \quad \widehat{A}_N = A_{\rm S}^{-\frac{1}{2}} A_{\rm N}\, A_{\rm S}^{-\frac{1}{2}}.$$

We shall repeatedly make use of the following identities:

$$\begin{aligned}
(A^{-1})_{\rm S} &= \frac{1}{2}\left(A^{-1}+A^{-T}\right) = A_{\rm S}^{-\frac{1}{2}}(I-\widehat{A}_N^2)^{-1}A_{\rm S}^{-\frac{1}{2}},\\
(A^{-1})_{\rm N} &= \frac{1}{2}\left(A^{-1}-A^{-T}\right) = A_{\rm S}^{-\frac{1}{2}}(I+\widehat{A}_N)^{-1}\widehat{A}_N(I-\widehat{A}_N)^{-1}A_{\rm S}^{-\frac{1}{2}}.
\end{aligned} \tag{3.3}$$

From the identities

$$\langle Sq,q\rangle = \langle Bv,q\rangle + \langle Cq,q\rangle = \langle v,B^Tq\rangle + \langle Cq,q\rangle = \langle Av,v\rangle + \langle Cq,q\rangle,$$

which are true for $q \in \mathbb{R}^m$ and $v := A^{-1}B^Tq \in \mathbb{R}^n$, we see that S is positive definite, if A is positive definite. For $\widetilde{S}$ we then compute:

$$\begin{aligned}
\langle \widetilde{S}q,q\rangle &= \langle Sq,q\rangle + \langle A^{-1}E^Tq, B^Tq\rangle\\
&= \langle Sq,q\rangle + \langle A_{\rm S}^{\frac{1}{2}}A^{-1}E^Tq, A_{\rm S}^{-\frac{1}{2}}B^Tq\rangle\\
&= \langle Sq,q\rangle + \langle A_{\rm S}^{\frac{1}{2}}A^{-1}E^Tq, (I-\widehat{A}_N)(I-\widehat{A}_N)^{-1}A_{\rm S}^{-\frac{1}{2}}B^Tq\rangle\\
&= \langle Sq,q\rangle + \langle \Big((I+\widehat{A}_N)A_{\rm S}^{\frac{1}{2}}A^{-1}A_{\rm S}^{\frac{1}{2}}\Big)\; A_{\rm S}^{-\frac{1}{2}}E^Tq, (I-\widehat{A}_N)^{-1}A_{\rm S}^{-\frac{1}{2}}B^Tq\rangle.
\end{aligned}$$

We employ identities (3.3) to get

$$\begin{aligned}
(I+\widehat{A}_N)A_{\rm S}^{\frac{1}{2}}A^{-1}A_{\rm S}^{\frac{1}{2}} &= (I+\widehat{A}_N)A_{\rm S}^{\frac{1}{2}}((A^{-1})_{\rm S}+(A^{-1})_{\rm N})A_{\rm S}^{\frac{1}{2}}\\
&= (I+\widehat{A}_N)((I-\widehat{A}_N^2)^{-1} + (I+\widehat{A}_N)^{-1}\widehat{A}_N(I-\widehat{A}_N)^{-1})\\
&= (I-\widehat{A}_N)^{-1} + \widehat{A}_N(I-\widehat{A}_N)^{-1}\\
&= (I+\widehat{A}_N)(I-\widehat{A}_N)^{-1}.
\end{aligned}$$

Noting $\|(I-\widehat{A}_N)^{-1}\| \le 1$ for a skew-symmetric $\widehat{A}_N$, we estimate

$$\begin{aligned}
\langle \widetilde{S}q, q\rangle &\ge \langle Sq, q\rangle - \|(I+\widehat{A}_N)(I-\widehat{A}_N)^{-1}\|\|A_{\mathrm{S}}^{-\frac{1}{2}}E^Tq\|\|(I-\widehat{A}_N)^{-1}A_{\mathrm{S}}^{-\frac{1}{2}}B^Tq\| \\
&\ge \langle Sq, q\rangle - \|(I+\widehat{A}_N)\|\|A_{\mathrm{S}}^{-\frac{1}{2}}E^T\|\|q\|\|(I-\widehat{A}_N)^{-1}A_{\mathrm{S}}^{-\frac{1}{2}}B^Tq\| \\
&\ge \langle Sq, q\rangle - (1+C_A)\varepsilon_E\|q\|\|(I-\widehat{A}_N)^{-1}A_{\mathrm{S}}^{-\frac{1}{2}}B^Tq\| \\
&= \langle Sq, q\rangle - (1+C_A)\varepsilon_E\|q\|\langle (I-\widehat{A}_N)^{-1}A_{\mathrm{S}}^{-\frac{1}{2}}B^Tq, (I-\widehat{A}_N)^{-1}A_{\mathrm{S}}^{-\frac{1}{2}}B^Tq\rangle^{\frac{1}{2}} \\
&= \langle Sq, q\rangle - (1+C_A)\varepsilon_E\|q\|\langle A_{\mathrm{S}}^{-\frac{1}{2}}B^Tq, (I+\widehat{A}_N)^{-1}(I-\widehat{A}_N)^{-1}A_{\mathrm{S}}^{-\frac{1}{2}}B^Tq\rangle^{\frac{1}{2}} \\
&= \langle Sq, q\rangle - (1+C_A)\varepsilon_E\|q\|\langle A_{\mathrm{S}}^{-\frac{1}{2}}B^Tq, (I-\widehat{A}_N^2)^{-1}A_{\mathrm{S}}^{-\frac{1}{2}}B^Tq\rangle^{\frac{1}{2}} \\
&= \langle Sq, q\rangle - (1+C_A)\varepsilon_E\|q\|\langle B^Tq, A_{\mathrm{S}}^{-\frac{1}{2}}(I-\widehat{A}_N^2)^{-1}A_{\mathrm{S}}^{-\frac{1}{2}}B^Tq\rangle^{\frac{1}{2}} \\
&= \langle Sq, q\rangle - (1+C_A)\varepsilon_E\|q\|\langle B(A^{-1})_{\mathrm{S}}B^Tq, q\rangle^{\frac{1}{2}} \\
&= \langle Sq, q\rangle - (1+C_A)\varepsilon_E\|q\|\langle BA^{-1}B^Tq, q\rangle^{\frac{1}{2}} \\
&= \langle Sq, q\rangle - (1+C_A)\varepsilon_E\|q\|\langle Sq, q\rangle^{\frac{1}{2}} \\
&\ge \left(1-(1+C_A)\varepsilon_E\lambda_{\min}^{-\frac{1}{2}}(S_{\mathrm{S}})\right)\langle Sq, q\rangle.
\end{aligned} \tag{3.4}$$

Hence, we conclude that $\widetilde{S}$ is positive definite if the perturbation matrix E is sufficiently small such that it holds

$$\kappa := (1+C_A)\varepsilon_E c_S^{-\frac{1}{2}} < 1 \tag{3.5}$$

where $c_S := \lambda_{\min}(S_{\mathrm{S}})$.

If $\widetilde{S}$ is positive definite, the factorization $\widetilde{S} = L_{22}U_{22}$ satisfies the stability bound similar to (3.2):

$$\frac{\||L_{22}||U_{22}|\|_F}{\|\widetilde{S}\|} \le m\,\left(1+\|\widetilde{S}_{\mathrm{S}}^{-\frac{1}{2}}\widetilde{S}_{\mathrm{N}}\widetilde{S}_{\mathrm{S}}^{-\frac{1}{2}}\|^2\right),$$

where $\widetilde{S}_{\mathrm{S}} = \frac{1}{2}(\widetilde{S}+\widetilde{S}^T)$, $\widetilde{S}_{\mathrm{N}} = \widetilde{S}-\widetilde{S}_{\mathrm{S}}$.

The quotients $C_A = \|A_{\mathrm{S}}^{-\frac{1}{2}}A_{\mathrm{N}}A_{\mathrm{S}}^{-\frac{1}{2}}\|$ and $\|\widetilde{S}_{\mathrm{S}}^{-\frac{1}{2}}\widetilde{S}_N\widetilde{S}_{\mathrm{S}}^{-\frac{1}{2}}\|$ are largely responsible for the stability of the LU factorization for (1.3). The following lemma shows the estimate of $\|\widetilde{S}_{\mathrm{S}}^{-\frac{1}{2}}\widetilde{S}_N\widetilde{S}_{\mathrm{S}}^{-\frac{1}{2}}\|$ in terms of C_A, ε_E and c_S.

LEMMA 3.1. *Let $A \in \mathbb{R}^{n\times n}$ be positive definite and* (3.5) *be satisfied, then it holds*

$$\|\widetilde{S}_{\mathrm{S}}^{-\frac{1}{2}}\widetilde{S}_N\widetilde{S}_{\mathrm{S}}^{-\frac{1}{2}}\| \le \frac{(1+\varepsilon_E c_S^{-\frac{1}{2}})C_A}{1-\kappa}. \tag{3.6}$$

Proof. Due to the skew-symmetry of $\widetilde{S}_{\mathrm{S}}^{-\frac{1}{2}}\widetilde{S}_N\widetilde{S}_{\mathrm{S}}^{-\frac{1}{2}}$ it holds $|\lambda| = |\mathrm{Im}(\lambda)|$ for $\lambda \in \mathrm{sp}(\widetilde{S}_{\mathrm{S}}^{-\frac{1}{2}}\widetilde{S}_N\widetilde{S}_{\mathrm{S}}^{-\frac{1}{2}})$, where we use $\mathrm{sp}(\cdot)$ to denote the spectrum. We apply Bendixson's theorem [33] to estimate

$$\begin{aligned}
\|\widetilde{S}_{\mathrm{S}}^{-\frac{1}{2}}\widetilde{S}_N\,\widetilde{S}_{\mathrm{S}}^{-\frac{1}{2}}\| &= \max\{|\lambda| \,:\, \lambda \in \mathrm{sp}(\widetilde{S}_{\mathrm{S}}^{-\frac{1}{2}}\widetilde{S}_N\widetilde{S}_{\mathrm{S}}^{-\frac{1}{2}})\} \\
&= \max\{|\mathrm{Im}(\lambda)| \,:\, \lambda \in \mathrm{sp}(\widetilde{S}_{\mathrm{S}}^{-\frac{1}{2}}\widetilde{S}_N\widetilde{S}_{\mathrm{S}}^{-\frac{1}{2}})\} \\
&\le \sup_{q\in\mathbb{C}^m}\frac{\left|\langle \widetilde{S}_N q, q\rangle\right|}{\langle \widetilde{S}_{\mathrm{S}}q, q\rangle}.
\end{aligned}$$

Thanks to (3.4) we estimate

$$\|\widetilde{S}_{\mathrm{S}}^{-\frac{1}{2}}\widetilde{S}_N\,\widetilde{S}_{\mathrm{S}}^{-\frac{1}{2}}\| \le \sup_{q\in\mathbb{C}^m}\frac{\left|\langle\widetilde{S}_N q,q\rangle\right|}{(1-\kappa)\langle S_{\mathrm{S}}q,q\rangle}\,. \tag{3.7}$$

Employing identities from (3.3), we can write

$$\begin{aligned}
S_{\mathrm{S}} &= BA_{\mathrm{S}}^{-\frac{1}{2}}(I-\widehat{A}_N^T)^{-1}(I-\widehat{A}_N)^{-1}A_{\mathrm{S}}^{-\frac{1}{2}}B^T + C,\\
\widetilde{S}_N &= BA_{\mathrm{S}}^{-\frac{1}{2}}(I-\widehat{A}_N^T)^{-1}\widehat{A}_N(I-\widehat{A}_N)^{-1}A_{\mathrm{S}}^{-\frac{1}{2}}\widetilde{B}^T.
\end{aligned}$$

With the help of the substitution $v_q = (I-\widehat{A}_N)^{-1}A_{\mathrm{S}}^{-\frac{1}{2}}B^Tq$ in the right-hand side of (3.7) and recalling that C is positive semidefinite, we obtain

$$\begin{aligned}
\|\widetilde{S}_{\mathrm{S}}^{-\frac{1}{2}}\widetilde{S}_N\,\widetilde{S}_{\mathrm{S}}^{-\frac{1}{2}}\| &\le \sup_{q\in\mathbb{C}^m}\frac{\left|\langle\widehat{A}_N v_q,v_q\rangle\right| + \left|\langle\widehat{A}_N(1-\widehat{A}_N)^{-1}A_{\mathrm{S}}^{-\frac{1}{2}}E^Tq,v_q\rangle\right|}{(1-\kappa)(\langle v_q,v_q\rangle+\langle Cq,q\rangle)}\\
&\le \sup_{q\in\mathbb{C}^m}\frac{\|\widehat{A}_N\|\|v_q\|^2 + \|\widehat{A}_N\|\varepsilon_E\|q\|\|v_q\|}{(1-\kappa)(\|v_q\|^2+\langle Cq,q\rangle)}\\
&\le \sup_{q\in\mathbb{C}^m}\frac{\|\widehat{A}_N\|\|v_q\|^2 + \|\widehat{A}_N\|\varepsilon_E\lambda_{\min}^{-\frac{1}{2}}(S_{\mathrm{S}})\langle S_{\mathrm{S}}q,q\rangle^{\frac{1}{2}}\|v_q\|}{(1-\kappa)(\|v_q\|^2+\langle Cq,q\rangle)}\\
&= \sup_{q\in\mathbb{C}^m}\frac{\|\widehat{A}_N\|\|v_q\|^2 + \|\widehat{A}_N\|\varepsilon_E\lambda_{\min}^{-\frac{1}{2}}(S_{\mathrm{S}})(\|v_q\|^2+\langle Cq,q\rangle)^{\frac{1}{2}}\|v_q\|}{(1-\kappa)(\|v_q\|^2+\langle Cq,q\rangle)}\\
&\le \frac{(1+\varepsilon_E c_S^{-\frac{1}{2}})\|\widehat{A}_N\|}{1-\kappa}.
\end{aligned}$$

□

To estimate the entries of U_{12} and L_{21} factors in (3.1) we repeat the arguments from [20] and arrive at the following bound

$$\frac{\|U_{12}\|_F+\|L_{21}\|_F}{\|U_{11}\|\|\widetilde{B}\|_F+\|L_{11}\|\|B\|_F} \le \frac{m(1+C_A)}{c_A}$$

with $c_A := \lambda_{\min}(A_{\mathrm{S}})$.

We summarize the results of this section in the following theorem.

THEOREM 3.2. *Assume matrix A is positive definite, C is positive semidefinite, and the inequality* (3.5) *holds with $\varepsilon_E = \|A_{\mathrm{S}}^{-\frac{1}{2}}(\widetilde{B}-B)^T\|$, $C_A = \|A_{\mathrm{S}}^{-\frac{1}{2}}A_{\mathrm{N}}A_{\mathrm{S}}^{-\frac{1}{2}}\|$, and $c_S = \lambda_{\min}(S_{\mathrm{S}})$, then the LU factrorization* (3.1) *exists without pivoting. The entries of the block factors satisfy the following bounds*

$$\begin{aligned}
\frac{\||L_{11}||U_{11}|\|_F}{\|A\|} &\le n\,\left(1+C_A^2\right),\\
\frac{\||L_{22}||U_{22}|\|_F}{\|\widetilde{S}\|} &\le m\,\left(1+\frac{(1+\varepsilon_E c_S^{-\frac{1}{2}})C_A}{1-\kappa}\right),\\
\frac{\|U_{12}\|_F+\|L_{21}\|_F}{\|U_{11}\|\|\widetilde{B}\|_F+\|L_{11}\|\|B\|_F} &\le \frac{m(1+C_A)}{c_A}
\end{aligned}$$

with κ from (3.5).

The above analysis indicates that the LU factorization for (1.3) exists if the (1,1) block A is positive definite and the perturbation of the (1,2)-block is sufficiently small. The stability bounds depend on the constant C_A which measures the ratio of skew-symmetry for A, the ellipticity constant c_A, the perturbation measure ε_E and the minimal eigenvalue of the symmetric part of the unperturbed Schur complement matrix S. In section 4 below, we estimate all these values for the discrete linearized Navier–Stokes system.

4. Properties of matrices A and $\widetilde{S}$. In this section we deduce the dependence of the critical constants c_A, C_A, ε_E and c_S from Theorem 3.2 on the problem and discretization parameters. This analysis relies on the SUPG-FE formulation from section 2. Recall that we assume an inf-sup finite element method, and so matrix C is zero. Let $\{\varphi_i\}_{1\le i\le n}$ and $\{\psi_j\}_{1\le j\le m}$ be bases of $\mathbb{V}_h$ and $\mathbb{Q}_h$, respectively. For arbitrary $v\in\mathbb{R}^n$ and corresponding $\mathbf{v}_h=\sum_{i=1}^n v_i\varphi_i$, one gets the following identity from the definition of matrix A:

$$\langle Av,v\rangle = \alpha\|\mathbf{v}_h\|^2+\nu\|\nabla\mathbf{v}_h\|^2+\sum_{\tau\in T_h}\sigma_\tau\|\mathbf{w}\cdot\nabla\mathbf{v}_h\|_\tau^2+\frac{1}{2}\int_{\Gamma_{\mathrm{N}}}(\mathbf{w}\cdot\mathbf{n})|\mathbf{v}_h|^2\,ds \\ +\frac{1}{2}\sum_{\tau\in T_h}((\operatorname{div}\mathbf{w})\mathbf{v}_h,\mathbf{v}_h)_\tau+\sum_{\tau\in T_h}\sigma_\tau(\alpha\mathbf{v}_h-\nu\Delta\mathbf{v}_h,\mathbf{w}\cdot\nabla\mathbf{v}_h)_\tau, \quad (4.1)$$

where $\mathbf{n}$ is the outward normal on Γ_{N}. We shall also need the velocity mass and stiffness matrices M and K: $M_{ij}=(\varphi_i,\varphi_j)$, $K_{ij}=(\nabla\varphi_i,\nabla\varphi_j)$ and the pressure mass matrix M_p: $(M_p)_{ij}=(\psi_i,\psi_j)$.

The first three terms on the right-hand side of (4.1) are positive and contribute to the ellipticity of the block A. However, the rest three terms are not necessarily sign definite and should be properly bounded. Although a modification of boundary conditions on Γ_{N} can be done to insure the resulting boundary integral is non-negative, see, e.g., [5], we shall use a FE trace inequality to estimate this term. We remark that this term disappears in the case of artificial outflow boundary conditions leading to Dirichlet conditions in (1.2) on the entire boundary [23, 29]. Next, $\mathbf{w}$ is typically a *finite element* velocity field, $\mathbf{w}\in\mathbb{V}_h$, satisfying only weak divergence free constraint $(\operatorname{div}\mathbf{w},q_h)=0\ \ \forall\ q_h\in\mathbb{Q}_h$. This weak divergence free equation does *not* imply $\operatorname{div}\mathbf{w}=0$ pointwise for most of stable FE pairs including P2-P1 elements. Therefore, the fifth term on the right-hand side of (4.1) should be controlled somehow. The last term in (4.1) is due to the SUPG stabilization. The ν-dependent part of it vanishes for P1 finite element velocities, but not for most of inf-sup stable disretization pressure-velocity pairs. Both analysis and numerical experiments below show that this term may significantly affect the properties of the matrix A, leading to unstable behavior of incomplete LU decomposition unless the stabilization parameters are chosen sufficiently small. We make the above statements more precise in Theorem 4.1. We need some preparation before we formulate the theorem.

First, recall the Sobolev trace inequality

$$\int_{\Gamma_{\mathrm{N}}}|v|^2\,ds\le C_0\|\nabla v\|^2\quad\forall\ v\in H^1(\Omega),\ v=0\text{ on }\partial\Omega\setminus\Gamma_{\mathrm{N}}. \quad (4.2)$$

For any tetrahedron $\tau\in T_h$ and arbitrary $\mathbf{v}_h\in\mathbb{V}_h$, the following FE trace and inverse

inequalities hold

$$\int_{\partial\tau} \mathbf{v}_h^2\, ds \le C_{\rm tr} h_\tau^{-1} \|\mathbf{v}_h\|_\tau^2,\ \ \|\nabla \mathbf{v}_h\|_\tau \le C_{\rm in} h_\tau^{-1} \|\mathbf{v}_h\|_\tau,\ \ \|\Delta \mathbf{v}_h\|_\tau \le \bar{C}_{\rm in} h_\tau^{-1} \|\nabla \mathbf{v}_h\|_\tau, \tag{4.3}$$

where the constants $C_{\rm tr}$, $C_{\rm in}$, $\bar{C}_{\rm in}$ depend only on the polynomial degree k and the shape regularity constant C_T from (2.1). In addition, denote by $C_{\rm f}$ the constant from the Friedrichs inequality:

$$\|\mathbf{v}_h\| \le C_{\rm f} \|\nabla \mathbf{v}_h\| \quad \forall\ \mathbf{v}_h \in \mathbb{V}_h, \tag{4.4}$$

and let $C_{\mathbf{w}} := \|(\mathbf{w}\cdot\mathbf{n})_-\|_{L^\infty(\Gamma_{\rm N})}$.

To avoid the repeated use of generic but unspecified constants, in the remainder of the paper the binary relation $x \lesssim y$ means that there is a constant c such that $x \le c\, y$, and c does not depend on the parameters which x and y may depend on, e.g., ν, α, mesh size, and properties of $\mathbf{w}$. Obviously, $x \gtrsim y$ is defined as $y \lesssim x$.

THEOREM 4.1. *Assume that* $\mathbf{w} \in L^\infty(\Omega)$*, problem and discretization parameters satisfy*

$$\begin{cases} C_{\mathbf{w}} C_{\rm tr} h_{\min}^{-1} \le \dfrac{\alpha}{4} \ \ \textit{or} \ \ C_{\mathbf{w}} C_0 \le \dfrac{\nu}{4}, \\ \|\operatorname{div}\mathbf{w}\|_{L^\infty(\Omega)} \le \dfrac{1}{4}\max\{\alpha, \nu C_f^{-1}\}, \\ \sigma_\tau \le \dfrac{1}{2}\left(\dfrac{h_\tau^2}{\nu \bar{C}_{\rm in}^2} + \dfrac{\alpha h_\tau^4}{\nu^2 \bar{C}_{\rm in}^2 C_{\rm in}^2}\right) \ \ \textit{and} \ \ \sigma_\tau \le \dfrac{h_\tau}{4\|\mathbf{w}\|_{L^\infty(\tau)} C_{\rm in}} \ \ \forall\ \tau \in T_h, \end{cases} \tag{4.5}$$

with constants defined in (4.2)–(4.4)*. Then the matrix A is positive definite and the constants c_A, C_A, c_S and ε_E can be estimated as follows:*

$$\begin{aligned} c_A &\ge \frac{1}{4}\lambda_{\min}(\alpha M + \nu K), \\ C_A &\lesssim 1 + \frac{\|\mathbf{w}\|_{L^\infty(\Omega)}}{\sqrt{\nu\alpha} + \nu + h_{\min}\alpha}, \\ c_S &\gtrsim \frac{\lambda_{\min}(M_p)}{(\nu + \alpha + \|\mathbf{w}\|_{L^\infty(\Omega)} + \|\operatorname{div}\mathbf{w}\|_{L^\infty(\Omega)})(1 + C_A^2)}, \\ \varepsilon_E &\le \Big(\frac{\bar{\sigma}}{2\nu}\lambda_{\max}(M_p)\Big)^{\frac{1}{2}}. \end{aligned} \tag{4.6}$$

Proof. Using the Cauchy inequality and (4.3), we bound the ν-dependent part of the last term in (4.1) as follows:

$$\begin{aligned} \left|\sum_{\tau\in T_h} \sigma_\tau \nu (\Delta\mathbf{v}_h, \mathbf{w}\cdot\nabla\mathbf{v}_h)_\tau\right| &\le \nu \left(\sum_{\tau\in T_h} \sigma_\tau \bar{C}_{\rm in}^2 h_\tau^{-2} \|\nabla\mathbf{v}_h\|_\tau^2\right)^{\frac{1}{2}} \left(\sum_{\tau\in T_h} \sigma_\tau \|\mathbf{w}\cdot\nabla\mathbf{v}_h\|_\tau^2\right)^{\frac{1}{2}} \\ &\le \frac{\nu^2}{2}\sum_{\tau\in T_h} \sigma_\tau \bar{C}_{\rm in}^2 h_\tau^{-2} \|\nabla\mathbf{v}_h\|_\tau^2 + \frac{1}{2}\sum_{\tau\in T_h} \sigma_\tau \|\mathbf{w}\cdot\nabla\mathbf{v}_h\|_\tau^2 \\ &\le \frac{\nu^2}{2}\bar{C}_{\rm in}^2 \sum_{\tau\in T_h} \sigma_\tau \frac{\nu\|\nabla\mathbf{v}_h\|_\tau^2 + \alpha\|\mathbf{v}_h\|_\tau^2}{\nu h_\tau^2 + C_{\rm in}^{-2}\alpha h_\tau^4} + \frac{1}{2}\sum_{\tau\in T_h} \sigma_\tau \|\mathbf{w}\cdot\nabla\mathbf{v}_h\|_\tau^2 \\ &\le \frac{1}{2}\sum_{\tau\in T_h} \frac{\nu^2 \sigma_\tau \bar{C}_{\rm in}^2 C_{\rm in}^2}{\nu h_\tau^2 C_{\rm in}^2 + \alpha h_\tau^4}\left(\nu\|\nabla\mathbf{v}_h\|_\tau^2 + \alpha\|\mathbf{v}_h\|_\tau^2\right) + \frac{1}{2}\sum_{\tau\in T_h} \sigma_\tau \|\mathbf{w}\cdot\nabla\mathbf{v}_h\|_\tau^2. \end{aligned} \tag{4.7}$$

The first term in the second line of (4.7) is bounded due to $\min\{\frac{a}{c};\frac{b}{d}\} \le \frac{a+b}{c+d}$ for $a,b,c,d>0$. Using similar arguments we bound the α-dependent part of the last term in (4.1):

$$\begin{aligned}\left|\sum_{\tau\in T_h}\sigma_\tau\alpha(\mathbf{v}_h,\mathbf{w}\cdot\nabla\mathbf{v}_h)_\tau\right| &\le \sum_{\tau\in T_h}\alpha\sigma_\tau\|\mathbf{w}\|_{L^\infty(\tau)}\|\mathbf{v}_h\|_\tau\|\nabla\mathbf{v}_h\|_\tau\\ &\le \sum_{\tau\in T_h}\alpha\sigma_\tau\|\mathbf{w}\|_{L^\infty(\tau)}C_{\rm in}h_\tau^{-1}\|\mathbf{v}_h\|_\tau^2.\end{aligned}\tag{4.8}$$

Applying (4.2), (4.7), and (4.8) in (4.1), we deduce

$$\begin{aligned}\langle Av,v\rangle \ge\ & \sum_{\tau\in T_h}\left(1-\frac{\nu^2\sigma_\tau\bar C_{\rm in}^2C_{\rm in}^2}{2(\nu h_\tau^2C_{\rm in}^2+\alpha h_\tau^4)}-\frac{\sigma_\tau\|\mathbf{w}\|_{L^\infty(\tau)}C_{\rm in}}{h_\tau}\right)\left(\nu\|\nabla\mathbf{v}_h\|_\tau^2+\alpha\|\mathbf{v}_h\|_\tau^2\right)\\ &+\frac12\sum_{\tau\in T_h}\sigma_\tau\|\mathbf{w}\cdot\nabla\mathbf{v}_h\|_\tau^2-\frac{C_{\mathbf{w}}}{2}\int_{\Gamma_{\rm N}}|\mathbf{v}_h|^2\,ds-\frac12\|\operatorname{div}\mathbf{w}\|_{L^\infty(\tau)}\|\mathbf{v}_h\|^2\\ \ge\ & \sum_{\tau\in T_h}\left(1-\frac{\nu^2\sigma_\tau\bar C_{\rm in}^2C_{\rm in}^2}{2(\nu h_\tau^2C_{\rm in}^2+\alpha h_\tau^4)}-\frac{\sigma_\tau\|\mathbf{w}\|_{L^\infty(\tau)}C_{\rm in}}{h_\tau}\right)\left(\nu\|\nabla\mathbf{v}_h\|_\tau^2+\alpha\|\mathbf{v}_h\|_\tau^2\right)\\ &-\frac{C_{\mathbf{w}}}{2}\min\{C_0\|\nabla\mathbf{v}_h\|^2,C_{\rm tr}h_{\min}^{-1}\|\mathbf{v}_h\|^2\}+\frac12\sum_{\tau\in T_h}\sigma_\tau\|\mathbf{w}\cdot\nabla\mathbf{v}_h\|_\tau^2-\frac12\|\operatorname{div}\mathbf{w}\|_{L^\infty(\tau)}\|\mathbf{v}_h\|^2.\end{aligned}\tag{4.9}$$

To ensure the right-hand side is positive, we assume conditions (4.5) on problem parameters and coefficients. Employing conditions (4.5) in (4.9), we deduce

$$\begin{aligned}\langle Av,v\rangle &\ge \frac14\left(\alpha\|\mathbf{v}_h\|^2+\nu\|\nabla\mathbf{v}_h\|_\tau^2+\sum_{\tau\in T_h}\sigma_\tau\|\mathbf{w}\cdot\nabla\mathbf{v}_h\|_\tau^2\right)\\ &\ge \frac14\left(\alpha\langle Mv,v\rangle+\nu\langle Kv,v\rangle\right)\quad\forall\ v\in\mathbb{R}^n,\end{aligned}\tag{4.10}$$

therefore, $c_A\ge\frac14\lambda_{\min}(\alpha M+\nu K)$. Further, we estimate

$$\begin{aligned}C_A:=\|A_{\rm S}^{-\frac12}A_{\rm N}A_{\rm S}^{-\frac12}\| &= \max\{|\lambda|\,:\,\lambda\in\operatorname{sp}(A_{\rm S}^{-\frac12}A_{\rm N}A_{\rm S}^{-\frac12})\}\\ &= \max\{|\lambda|\,:\,\lambda\in\operatorname{sp}(A_{\rm S}^{-1}A_{\rm N})\}\\ &\le \|A_{\rm S}^{-1}A_{\rm N}\|_*,\end{aligned}\tag{4.11}$$

and for $\|\cdot\|_*$ we choose a matrix norm induced by the vector norm $\langle(\alpha M+\nu K)\cdot,\cdot\rangle^{\frac12}$. For a given $v\in\mathbb{R}^n$ and $u=A_{\rm S}^{-1}A_{\rm N}\,v$ consider their finite element counterparts $\mathbf{v}_h,\mathbf{u}_h\in\mathbb{V}_h$. Then $A_{\rm S}u=A_{\rm N}\,v$ can be written in a finite element form as

$$\begin{aligned}&\nu(\nabla\mathbf{u}_h,\nabla\boldsymbol{\varphi}_h)+\alpha(\mathbf{u}_h,\boldsymbol{\varphi}_h)+\frac12\int_{\Gamma_{\rm N}}(\mathbf{w}\cdot\mathbf{n})\mathbf{u}_h\cdot\boldsymbol{\varphi}_h\,ds+\sum_{\tau\in T_h}\sigma_\tau(\mathbf{w}\cdot\nabla\mathbf{u}_h,\mathbf{w}\cdot\nabla\boldsymbol{\varphi}_h)_\tau\\ &+\frac12\sum_{\tau\in T_h}((\operatorname{div}\mathbf{w})\mathbf{u}_h,\boldsymbol{\varphi}_h)_\tau+\frac12\sum_{\tau\in T_h}\sigma_\tau[(\alpha\mathbf{u}_h-\nu\Delta\mathbf{u}_h,\mathbf{w}\cdot\nabla\boldsymbol{\varphi}_h)_\tau+(\alpha\boldsymbol{\varphi}_h-\nu\Delta\boldsymbol{\varphi}_h,\mathbf{w}\cdot\nabla\mathbf{u}_h)_\tau]\\ &\qquad=\frac12\sum_{\tau\in T_h}(1+\alpha\sigma_\tau)[(\mathbf{w}\cdot\nabla\mathbf{v}_h,\boldsymbol{\varphi}_h)_\tau-(\mathbf{w}\cdot\nabla\boldsymbol{\varphi}_h,\mathbf{v}_h)_\tau]\\ &\qquad-\frac12\sum_{\tau\in T_h}\sigma_\tau\nu[(\Delta\mathbf{v}_h,\mathbf{w}\cdot\nabla\boldsymbol{\varphi}_h)_\tau-(\Delta\boldsymbol{\varphi}_h,\mathbf{w}\cdot\nabla\mathbf{v}_h)_\tau]\quad\forall\boldsymbol{\varphi}_h\in\mathbb{V}_h.\end{aligned}\tag{4.12}$$

We set $\boldsymbol{\varphi}_h = \mathbf{u}_h$. For the left-hand side of (4.12) the lower bound (4.10) holds. To estimate the right-hand side, we apply the Cauchy–Schwarz inequality, the second restriction on σ_τ from (4.5) and finite element inverse inequality:

$$
\begin{aligned}
&\sum_{\tau\in T_h}(1+\alpha\sigma_\tau)[(\mathbf{w}\cdot\nabla\mathbf{v}_h,\mathbf{u}_h)_\tau-(\mathbf{w}\cdot\nabla\mathbf{u}_h,\mathbf{v}_h)_\tau]\\
&\le\sum_{\tau\in T_h}(1+\frac{\alpha h_\tau}{\|\mathbf{w}\|_{L^\infty(\tau)}C_{\mathrm{in}}})[(\mathbf{w}\cdot\nabla\mathbf{v}_h,\mathbf{u}_h)_\tau-(\mathbf{w}\cdot\nabla\mathbf{u}_h,\mathbf{v}_h)_\tau]\\
&\le\|\mathbf{w}\|_{L^\infty(\Omega)}(\|\nabla\mathbf{v}_h\|\|\mathbf{u}_h\|+\|\nabla\mathbf{u}_h\|\|\mathbf{v}_h\|)+\sum_{\tau\in T_h}\frac{\alpha h_\tau}{C_{\mathrm{in}}}(\|\nabla\mathbf{v}_h\|_\tau\|\mathbf{u}_h\|_\tau+\|\nabla\mathbf{u}_h\|_\tau\|\mathbf{v}_h\|_\tau)\\
&\le\|\mathbf{w}\|_{L^\infty(\Omega)}(\|\nabla\mathbf{v}_h\|\|\mathbf{u}_h\|+\|\nabla\mathbf{u}_h\|\|\mathbf{v}_h\|)+\sum_{\tau\in T_h}2\alpha\|\mathbf{v}_h\|_\tau\|\mathbf{u}_h\|_\tau\\
&\le\|\mathbf{w}\|_{L^\infty(\Omega)}(\|\nabla\mathbf{v}_h\|\|\mathbf{u}_h\|+\|\nabla\mathbf{u}_h\|\|\mathbf{v}_h\|)+32\alpha\|\mathbf{v}_h\|^2+\frac{\alpha}{32}\|\mathbf{u}_h\|^2.
\end{aligned}
\tag{4.13}
$$

Further we estimate terms on the right-hand side by employing Young's, Friedrichs, and finite element inverse inequalities. Thus, the product $\|\mathbf{u}_h\|\|\nabla\mathbf{v}_h\|$ one can estimate in three different ways:

$$
\begin{aligned}
\|\mathbf{w}\|_{L^\infty(\Omega)}\|\mathbf{u}_h\|\|\nabla\mathbf{v}_h\|&\le\frac{1}{32}\alpha\|\mathbf{u}_h\|^2\quad+8\|\mathbf{w}\|_{L^\infty(\Omega)}\frac{1}{\alpha\nu}(\nu\|\nabla\mathbf{v}_h\|^2)\\
\|\mathbf{w}\|_{L^\infty(\Omega)}\|\mathbf{u}_h\|\|\nabla\mathbf{v}_h\|&\le\frac{1}{32}\nu\|\nabla\mathbf{u}_h\|^2+8\|\mathbf{w}\|_{L^\infty(\Omega)}\frac{C_{\mathrm{f}}^2}{\nu^2}(\nu\|\nabla\mathbf{v}_h\|^2)\\
\|\mathbf{w}\|_{L^\infty(\Omega)}\|\mathbf{u}_h\|\|\nabla\mathbf{v}_h\|&\le\frac{1}{32}\alpha\|\mathbf{u}_h\|^2\quad+8\|\mathbf{w}\|_{L^\infty(\Omega)}\frac{C_{\mathrm{in}}^2}{\alpha^2h_{\min}^2}(\alpha\|\mathbf{v}_h\|^2).
\end{aligned}
$$

Combining all three estimates gives

$$
\begin{aligned}
\|\mathbf{w}\|_{L^\infty(\Omega)}\|\nabla\mathbf{v}_h\|\|\mathbf{u}_h\|&\le\frac{1}{32}(\nu\|\nabla\mathbf{u}_h\|^2+\alpha\|\mathbf{u}_h\|^2)\\
&\quad+8\|\mathbf{w}\|_{L^\infty(\Omega)}^2\min\left\{\frac{1}{\alpha\nu},\frac{C_{\mathrm{f}}^2}{\nu^2},\frac{C_{\mathrm{in}}^2}{\alpha^2h_{\min}^2}\right\}(\nu\|\nabla\mathbf{v}_h\|^2+\alpha\|\mathbf{v}_h\|^2).
\end{aligned}
\tag{4.14}
$$

Using same argument to treat the second term on the right-hand side of (4.13), we arrive at

$$
\begin{aligned}
\|\mathbf{w}\|_{L^\infty(\Omega)}\|\nabla\mathbf{u}_h\|\|\mathbf{v}_h\|&\le\frac{1}{32}(\nu\|\nabla\mathbf{u}_h\|^2+\alpha\|\mathbf{u}_h\|^2)\\
&\quad+8\|\mathbf{w}\|_{L^\infty(\Omega)}^2\min\left\{\frac{1}{\alpha\nu},\frac{C_{\mathrm{f}}^2}{\alpha^2},\frac{C_{\mathrm{f}}^2}{\nu^2}\right\}(\nu\|\nabla\mathbf{v}_h\|^2+\alpha\|\mathbf{v}_h\|^2).
\end{aligned}
\tag{4.15}
$$

Hence, we derive using $\min\{a_1,a_2,a_3\}\le 3(a_1^{-1}+a_2^{-1}+a_3^{-1})^{-1}$, the estimate for the first term on the right hand side of (4.12)

$$
\begin{aligned}
&\frac{1}{2}\sum_{\tau\in T_h}(1+\alpha\sigma_\tau)[(\mathbf{w}\cdot\nabla\mathbf{v}_h,\mathbf{u}_h)_\tau-(\mathbf{w}\cdot\nabla\mathbf{u}_h,\mathbf{v}_h)_\tau]\\
&\quad\lesssim\left(1+\frac{\|\mathbf{w}\|_{L^\infty(\Omega)}^2}{\nu\alpha+\nu^2+h_{\min}^2\alpha^2}\right)(\nu\|\nabla\mathbf{v}_h\|^2+\alpha\|\mathbf{v}_h\|^2)+\frac{3}{32}(\nu\|\nabla\mathbf{u}_h\|^2+\alpha\|\mathbf{u}_h\|^2).
\end{aligned}
\tag{4.16}
$$

Now we estimate the second term on the right hand side of (4.12) with the help of the third condition from (4.5):

$$\begin{aligned}
&\sum_{\tau\in T_h}\sigma_\tau\nu[(\Delta\mathbf{v}_h,\mathbf{w}\cdot\nabla\mathbf{u}_h)_\tau-(\Delta\mathbf{u}_h,\mathbf{w}\cdot\nabla\mathbf{v}_h)_\tau]\\
&\le\sum_{\tau\in T_h}[\sigma_\tau\nu\bar{C}_{\rm in}h_\tau^{-1}\|\nabla\mathbf{v}_h\|_\tau\|\mathbf{w}\cdot\nabla\mathbf{u}_h\|_\tau+\sigma_\tau\nu\bar{C}_{\rm in}\|\mathbf{w}\|_{L^\infty(\tau)}h_\tau^{-1}\|\nabla\mathbf{u}_h\|\|\nabla\mathbf{v}_h\|_\tau]\\
&\le\frac{1}{32}(\nu\|\nabla\mathbf{u}_h\|^2+\sum_{\tau\in T_h}\sigma_\tau\|\mathbf{w}\cdot\nabla\mathbf{u}_h\|_\tau^2)+\sum_{\tau\in T_h}8(\sigma_\tau\nu\bar{C}_{\rm in}^2h_\tau^{-2}+\sigma_\tau^2\bar{C}_{\rm in}^2\|\mathbf{w}\|_{L^\infty(\tau)}^2h_\tau^{-2})\nu\|\nabla\mathbf{v}_h\|_\tau^2\\
&\lesssim\frac{1}{32}(\nu\|\nabla\mathbf{u}_h\|^2+\sum_{\tau\in T_h}\sigma_\tau\|\mathbf{w}\cdot\nabla\mathbf{u}_h\|_\tau^2)+(\nu\|\nabla\mathbf{v}_h\|^2+\alpha\|\mathbf{v}_h\|^2).
\end{aligned}\tag{4.17}$$

Summarizing (4.12)– (4.17), we obtain

$$\begin{aligned}
&\frac{7}{8}\left(\alpha\|\mathbf{u}_h\|^2+\nu\|\nabla\mathbf{u}_h\|^2+\sum_{\tau\in T_h}\sigma_\tau\|\mathbf{w}\cdot\nabla\mathbf{u}_h\|_\tau^2\right)+\frac{1}{2}\int_{\Gamma_{\rm N}}(\mathbf{w}\cdot\mathbf{n})|\mathbf{u}_h|^2\,ds\\
&-\sum_{\tau\in T_h}\sigma_\tau(\alpha\mathbf{u}_h-\nu\Delta\mathbf{u}_h,\mathbf{w}\cdot\nabla\mathbf{u}_h)_\tau+\frac{1}{2}\sum_{\tau\in T_h}((\operatorname{div}\mathbf{w})\mathbf{u}_h,\mathbf{u}_h)_\tau\\
&\lesssim\left(1+\frac{\|\mathbf{w}\|_{L^\infty(\Omega)}^2}{\nu\alpha+\nu^2+h_{\min}^2\alpha^2}\right)(\nu\|\nabla\mathbf{v}_h\|^2+\alpha\|\mathbf{v}_h\|^2).
\end{aligned}\tag{4.18}$$

The left-hand side of (4.12) equals

$$\langle A_{\rm S}u,u\rangle-\frac{1}{8}\left(\alpha\|\mathbf{u}_h\|^2+\nu\|\nabla\mathbf{u}_h\|^2+\sum_{\tau\in T_h}\sigma_\tau\|\mathbf{w}\cdot\nabla\mathbf{v}_h\|_\tau^2\right)$$

and due to (4.10) it is estimated from below by $\frac{1}{2}\langle A_{\rm S}u,u\rangle$. Recalling $4\langle A_{\rm S}u,u\rangle\ge\|u\|_*^2=\nu\|\nabla\mathbf{u}_h\|^2+\alpha\|\mathbf{u}_h\|^2$, we obtain with the help of (4.11)

$$C_A\le\|A_{\rm S}^{-1}A_{\rm N}\|_*=\sup_{v\in\mathbb{R}^n}\frac{\|u\|_*}{\|v\|_*}\le 2\sup_{v\in\mathbb{R}^n}\frac{\langle A_{\rm S}u,u\rangle^{\frac{1}{2}}}{\|v\|_*}\lesssim\left(1+\frac{\|\mathbf{w}\|_{L^\infty(\Omega)}}{\sqrt{\nu\alpha}+\nu+h_{\min}\alpha}\right).\tag{4.19}$$

Denote $\tilde{c}_{\mathbf{w}}:=\|\mathbf{w}\|_{L^\infty(\Omega)}$, $\hat{c}_{\mathbf{w}}=\|\operatorname{div}\mathbf{w}\|_{L^\infty(\Omega)}$. To bound from below the ellipticity constant $c_{\rm S}$ for the auxiliary Schur complement matrix S, we first observe the following upper bound

$$\begin{aligned}
\langle A_{\rm S}v,v\rangle=\langle Av,v\rangle&\le 2(\alpha\|\mathbf{v}_h\|^2+\nu\|\nabla\mathbf{v}_h\|^2+\sum_{\tau\in T_h}\sigma_\tau\|\mathbf{w}\cdot\nabla\mathbf{v}_h\|_\tau^2)+C_0\tilde{c}_{\mathbf{w}}\|\nabla\mathbf{v}_h\|^2+\frac{1}{2}\hat{c}_{\mathbf{w}}\|\mathbf{v}_h\|^2\\
&\le 2(\alpha\|\mathbf{v}_h\|^2+\nu\|\nabla\mathbf{v}_h\|^2+\sum_{\tau\in T_h}\sigma_\tau\|\mathbf{w}\|_{L^\infty(\tau)}^2\|\nabla\mathbf{v}_h\|_\tau^2)+C_0\tilde{c}_{\mathbf{w}}\|\nabla\mathbf{v}_h\|^2+\frac{1}{2}\hat{c}_{\mathbf{w}}\|\mathbf{v}_h\|^2\\
&\le 2(\alpha\|\mathbf{v}_h\|^2+\nu\|\nabla\mathbf{v}_h\|^2+\sum_{\tau\in T_h}\frac{h_\tau\|\mathbf{w}\|_{L^\infty(\tau)}}{4C_{\rm in}}\|\nabla\mathbf{v}_h\|_\tau^2)+C_0\tilde{c}_{\mathbf{w}}\|\nabla\mathbf{v}_h\|^2+\frac{1}{2}\hat{c}_{\mathbf{w}}\|\mathbf{v}_h\|^2\\
&\le 2(\alpha\|\mathbf{v}_h\|^2+(\nu+\tilde{c}_{\mathbf{w}})\|\nabla\mathbf{v}_h\|^2)+C_0\tilde{c}_{\mathbf{w}}\|\nabla\mathbf{v}_h\|^2+\frac{1}{2}\hat{c}_{\mathbf{w}}\|\mathbf{v}_h\|^2\\
&\lesssim(\nu+\alpha+\tilde{c}_{\mathbf{w}}+\hat{c}_{\mathbf{w}})\|\nabla\mathbf{v}_h\|^2.
\end{aligned}$$

The above bound and the inf-sup stability of the finite element spaces yield the following relations:

$$\begin{aligned}\langle BA_{\rm S}^{-1}B^Tq,q\rangle &= \sup_{v\in\mathbb{R}^n}\frac{\langle Bv,q\rangle^2}{\langle A_{\rm S}v,v\rangle} \gtrsim \sup_{\mathbf{v}_h\in\mathbb{V}_h}\frac{(\operatorname{div}\mathbf{v}_h,q_h)^2}{(\nu+\alpha+\tilde{c}_{\mathbf{w}}+\hat{c}_{\mathbf{w}})\|\nabla\mathbf{v}_h\|^2}\\ &\gtrsim \frac{\|q_h\|^2}{\nu+\alpha+\tilde{c}_{\mathbf{w}}+\hat{c}_{\mathbf{w}}} = \frac{\langle M_pq,q\rangle}{\nu+\alpha+\tilde{c}_{\mathbf{w}}+\hat{c}_{\mathbf{w}}}.\end{aligned} \tag{4.20}$$

With the help of the first identity from (3.3) and (4.20) we obtain

$$\begin{aligned}\langle Sq,q\rangle &= \langle A^{-1}B^Tq,B^Tq\rangle = \langle (I-(A_{\rm S}^{-\frac12}A_{\rm N}\,A_{\rm S}^{-\frac12})^2)^{-1}A_{\rm S}^{-\frac12}B^Tq,A_{\rm S}^{-\frac12}B^Tq\rangle\\ &\geq \frac{\langle A_{\rm S}^{-\frac12}B^Tq,A_{\rm S}^{-\frac12}B^Tq\rangle}{1+\|(A_{\rm S}^{-\frac12}A_{\rm N}\,A_{\rm S}^{-\frac12})^2\|} = \frac{\langle BA_{\rm S}^{-1}B^Tq,q\rangle}{1+\|(A_{\rm S}^{-\frac12}A_{\rm N}\,A_{\rm S}^{-\frac12})^2\|}\\ &\gtrsim \frac{1}{(\nu+\alpha+\tilde{c}_{\mathbf{w}}+\hat{c}_{\mathbf{w}})(1+\|(A_{\rm S}^{-\frac12}A_{\rm N}\,A_{\rm S}^{-\frac12})\|^2)}\langle M_pq,q\rangle.\end{aligned} \tag{4.21}$$

The desired bound for c_S follows from (4.21).

To estimate ε_E, we use similar technique. For arbitrary given $q\in\mathbb{R}^m$, let $u = A_{\rm S}^{-1}E^Tq$. We have

$$\|A_{\rm S}^{-\frac12}E^Tq\|^2 = \langle A_{\rm S}^{-1}E^Tq,E^Tq\rangle = \langle A_{\rm S}u,u\rangle. \tag{4.22}$$

For arbitrary $v\in\mathbb{R}^n$ it holds $\langle A_{\rm S}u,v\rangle = \langle E^Tq,v\rangle$. For corresponding finite element functions this yields

$$\begin{aligned}&\nu(\nabla\mathbf{u}_h,\nabla\mathbf{v}_h)+\alpha(\mathbf{u}_h,\mathbf{v}_h)+\frac12\int_{\Gamma_{\rm N}}(\mathbf{w}\cdot\mathbf{n})\mathbf{u}_h\cdot\mathbf{v}_h\,ds+\sum_{\tau\in T_h}\sigma_\tau(\mathbf{w}\cdot\nabla\mathbf{u}_h,\mathbf{w}\cdot\nabla\mathbf{v}_h)_\tau\\ &+\frac12\sum_{\tau\in T_h}((\operatorname{div}\mathbf{w})\mathbf{u}_h,\mathbf{v}_h)_\tau+\frac12\sum_{\tau\in T_h}\sigma_\tau[(\alpha\mathbf{u}_h-\nu\Delta\mathbf{u}_h,\mathbf{w}\cdot\nabla\mathbf{v}_h)_\tau+(\alpha\mathbf{v}_h-\nu\Delta\mathbf{v}_h,\mathbf{w}\cdot\nabla\mathbf{u}_h)_\tau]\\ &\qquad=\sum_{\tau\in T_h}\sigma_\tau(\mathbf{w}\cdot\nabla\mathbf{v}_h,\nabla q_h)_\tau \leq \sum_{\tau\in T_h}\sigma_\tau\left(\frac18\|\mathbf{w}\cdot\nabla\mathbf{v}_h\|_\tau^2+2\|\nabla q_h\|_\tau^2\right)\\ &\qquad\leq\sum_{\tau\in T_h}\sigma_\tau\left(\frac18\|\mathbf{w}\cdot\nabla\mathbf{v}_h\|_\tau^2+2C_{\rm in}^2h_\tau^{-2}\|q_h\|_\tau^2\right).\end{aligned}$$

We set $\mathbf{v}_h=\mathbf{u}_h$ and invoke (4.10) to conclude in the vector notation

$$\langle A_{\rm S}u,u\rangle \lesssim \max_\tau(\sigma_\tau h_\tau^{-2})\lambda_{\max}(M_p)\|q\|^2 \leq \frac{\bar\sigma}{2\nu}\lambda_{\max}(M_p)\|q\|^2. \tag{4.23}$$

The last inequality follows from the definition of σ_τ in (2.4) for $\mathrm{Re}_\tau>1$:

$$\sigma_\tau = \bar\sigma\frac{h_{\mathbf{w}}}{2\|\mathbf{w}\|_{L_\infty(\tau)}}\left(1-\frac{1}{\mathrm{Re}_\tau}\right)\leq\bar\sigma\frac{h_{\mathbf{w}}}{2\|\mathbf{w}\|_{L_\infty(\tau)}}\mathrm{Re}_\tau=\bar\sigma\frac{h_{\mathbf{w}}^2}{2\nu}\leq\bar\sigma\frac{h_\tau^2}{2\nu}. \tag{4.24}$$

Recalling the definition of ε_E, the inequality (4.23) together with (4.22) proves the last bound in (4.6). □

The theorem shows that matrices A and $\widetilde{S}$ are positive definite if conditions (4.5) on the parameters of the finite element method are satisfied. In this case, the matrix

in (1.3) admits LU factorization without pivoting. The *first condition* in (4.5) is trivially satisfied with $C_{\mathbf{w}} = 0$ if $\Gamma_{\mathrm{N}} \neq \varnothing$ or the entire Γ_{N} is outflow boundary. The *second condition* may not be restrictive, since $\mathbf{w}$ approximates velocity field of an incompressible fluid and hence $\|\mathrm{div}\,\mathbf{w}\|_{L^\infty(\Omega)}$ decreases for a refined grid. However, the $\mathbf{w}$-divergence norm depends on fluid velocity field and may be large for ν small enough. Fortunately, one can choose such small Δt that the second condition holds due to $\alpha \sim (\Delta t)^{-1}$. The *third condition* in (4.5) puts an upper bound on stabilization parameters. Naturally, the same or a similar condition appears in the literature on the analysis of SUPG stabilized methods for the linearized Navier–Stokes equations, see, e.g., [26]. The reason is that the positive definiteness of A is equivalent to the coercivity of the velocity part of the bilinear form from (2.3), which is crucial for deriving finite element method error estimates. Therefore, stabilization parameter design suggested in the literature typically satisfies $\sigma_\tau \lesssim \frac{h_\tau^2}{\nu}$ and $\sigma_\tau \lesssim \frac{h_\tau}{\|\mathbf{w}\|_{L^\infty(\tau)}}$ *asymptotically*, i.e. up to a scaling factor independent of discretization parameters. As follows from (4.24), the conditions (4.5) on the SUPG stabilization parameters (2.4) are valid if $\bar\sigma \leq \min\{\bar C_{\mathrm{in}}^{-2}, \frac{1}{2} C_{\mathrm{in}}^{-1}\}$. In practice, however, larger values of $\bar\sigma$ are often found optimal for FE solution accuracy. The possible reason of the inconsistency is that smooth harmonics dominate in the solution, and hence the bounds on parameters are less tight. The situation is different when one is concerned with iterative convergence of algebraic solvers, since an algebraic solver has to reduce all possible harmonics in the decomposition of the error vector.

5. A two-parameter threshold ILU factorization. Incomplete LU factorizations of (1.3) can be written in the form $A = LU - E$ with an error matrix E. How small is the matrix E can be ruled by the choice of a threshold parameter $\tau > 0$. The error matrix E is responsible for the quality of preconditioning, see, for example, [19] for estimates on GMRES method convergence written in terms of $\|E\|$ and subject to a proper pre-scaling of A and the diagonalizability assumption. In general, the analysis of ILU factorization is based on the following arguments. For positive definite matrices A one can choose such a small τ that the product LU of its incomplete triangular factors L and U is also positive definite and so estimates from [14] can be applied to assess the numerical stability of the incomplete factorization: for $c_A = \lambda_{\min}(A_{\mathrm{S}})$, the sufficient condition is $\tau < c_A n^{-1}$. In practice, however, larger τ are used, and in the case of non-symmetric matrices non-positive or close to zero pivots may encounter, and breakdown of an algorithm may happen. Although most of remedies were developed for the SPD case [2], some of them are applicable to non-symmetric and non-definite matrices. We use the matrix two-side scaling [20] in our applications.

Stability of ILU factorization for saddle point matrices with positive definite (1,1)-block and $\widetilde B \neq B$ deteriorates in comparison with positive definite matrices and saddle point matrices with $\widetilde B = B$. Theorem 4.1 shows that for certain flow regimes the ellipticity constants c_A, c_S for A and S approach zero. To ameliorate the performance of the preconditioning in such extreme situations, we consider the two-parameter Tismenetsky–Kaporin variant of the threshold ILU factorization. The factorization was introduced and first studied in [18,34,35] for symmetric positive definite matrices and recently for non-symmetric matrices in [20].

Given a matrix $A \in \mathbb{R}^{n\times n}$, the two-parameter factorization can be written as

$$A = LU + LR_u + R_\ell U - E, \tag{5.1}$$

where R_u and R_ℓ are strictly upper and lower triangular matrices, while U and L

are upper and lower triangular matrices, respectively. Given two small parameters $0 < \tau_1 \le \tau_2$ the off-diagonal elements of U and L are either zero or have absolute values greater than τ_1, the absolute values of R_ℓ and R_u entries are either zero or belong to $(\tau_2, \tau_1]$; entries of the error matrix are of order $O(\tau_2)$. We refer to (5.1) as the ILU(τ_1, τ_2) factorization of A. Of course, a generic ILU(τ) factorization can be viewed as (5.1) with $R_u = R_\ell = 0$ and $\tau_1 = \tau_2 = \tau$. The two-parameter ILU factorization goes over a generic ILU(τ) factorization: the fill-in of L and U is ruled by the first threshold parameter τ_1, while the quality of the resulting preconditioner is mainly defined by τ_2, once $\tau_1^2 \lesssim \tau_2$ holds. In other words the choice $\tau_2 = \tau_1^2 := \tau^2$ may provide the fill-in of ILU(τ_1, τ_2) to be similar to that of ILU(τ), while the convergence of preconditioned Krylov subspace method is better and asymptotically (for $\tau \to 0$) can be comparable to the one with ILU(τ^2) preconditioner. For symmetric positive definite matrices this empirical advantages of ILU(τ_1, τ_2) are rigorously explained in [18], where estimates on the eigenvalues and K-condition number of $L^{-1}AU^{-1}$ were derived with $L^T = U$ and $R_\ell^T = R_u$. The price one pays is that computing L, U factors for ILU(τ_1, τ_2) is computationally more costly than for ILU(τ_1), since intermediate calculations involve the entries of R_u. However, this factorization phase of ILU(τ_1, τ_2) is still less expensive than that of ILU(τ_2). We note also that ILU(τ_1,τ_2) with $\tau_1 = \tau_2$ is similar to the ILUT(p,τ) dual parameter incomplete factorization [28] with $p = n$ (all elements passing the threshold criterion are kept in the factors). If no small pivots modification is done, the only differences between the algorithms (for $\tau_1 = \tau_2$ and $p = n$) are a different scaling of pivots and a row dependent scaling of threshold values used in ILUT. A pseudo-code of a row-wise ILU(τ_1, τ_2) can be found in [20].

Analysis of the decomposition (5.1) of a general non-symmetric matrix is limited to simple estimate (2.5) from [15] applied to the matrix $(L + R_\ell)(U + R_u) = A + R_\ell R_u + E$. The low bound for the pivots of the (5.1) factorization is the following

$$|L_{ii}U_{ii}| \ge \min_{v\in\mathbb{R}^n} \frac{\langle (A + R_\ell R_u + E)v, v\rangle}{\|v\|^2} \ge c_A - \|R_\ell R_u\| - \|E\|, \tag{5.2}$$

with the ellipticity constant c_A and the norms $\|R_\ell R_u\|$, $\|E\|$ proportional to τ_1^2 and τ_2, respectively. Hence, we may conclude that the numerical stability of solving for $L^{-1}x$ and $U^{-1}x$ is ruled by the second parameter and the *square* of the first parameter, while the fill-in in both factors is defined by τ_1 rather than τ_1^2. The Oseen problem setup may be such that the estimates from Theorem 4.1 predict that the coercitivity constant c_A and the ellipticity constant c_S are small. This increases the probability of the breakdown of ILU(τ) factorization of the saddle-point matrix $\mathcal{A}$, and demonstrates the benefits of ILU(τ_1, τ_2) factorization.

6. Numerical results. In this section, we demonstrate the performance of the ILU(τ) factorization for different values of discretization, stabilization and threshold parameters. As a testbench, we simulate a blood flow in a right coronary artery within a single cardiac cycle. For numerical test, we use the implementation of ILU(τ_1,τ_2) available in the open source software [21, 22]. The optimal values of ILU thresholds $\tau_1 = 0.03$, $\tau_2 = 7\tau_1^2$ are taken from [20] where detailed analysis of ILU(τ_1,τ_2) and ILU(τ):= ILU(τ,τ) preconditioners for the Oseen systems without stabilization is given. In all experiments we use BiCGstab method with the right preconditioner defined by the ILU(τ_1,τ_2) factorization.

The geometry of the flow domain was recovered from a real patient coronary CT angiography. The diameter of the inlet cross-section is about 0.27 cm and is

TABLE 6.1
The performance of ILU($\tau_1 = 0.03, \tau_2 = 7\tau_1^2$) for right coronary artery. The number and the time of iterations accumulated for 147 time steps

Mesh	$\bar{\sigma}$	#it	T_{it}
63k	0	20908	2267.
63k	1/12	20292	2182.
120k	0	26209	6188.
120k	1/12	26446	6132.

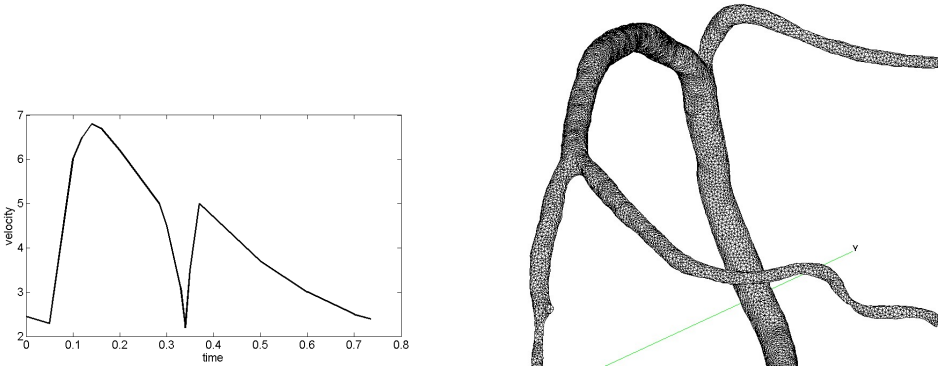

FIG. 6.1. *The velocity waveform on the inflow as a function of time and the coarse grid in the right coronary artery.*

imbedded in the box 6.5 cm × 6.8 cm × 5 cm. The ANI3D package [22] was used to generate two tetrahedral meshes, the coarse mesh is shown in Figure 6.1. The meshes consist of 63k and 120k tetrahedra leading to the discrete (P2-P1 FEM) Navier–Stokes system with about 300k and 600k unknowns, respectively. The Navier–Stokes system (1.1) was integrated in time using a semi-implicit second order method with $\Delta t = 0.005$ and systems (1.3) were solved at every time step. Other model parameters are $\nu = 0.04\,\text{cm}^2/\text{s}$, $\rho = 1\,\text{g/cm}$, one cardiac cycle period was 0.735s. The inlet velocity waveform [17] shown in Figure 6.1 defines the Poiseuille flow rate through the inflow cross-section. The vessel walls were treated as rigid and homogeneous Dirichlet boundary conditions for the velocity are imposed on the vessel walls. On all outflow boundaries we set the normal component of the stress tensor equal zero.

Table 6.1 shows the total number of the preconditioned BiCGstab iterations and the CPU time needed to perform all 147 time steps within a single cardial cycle. For each system, the initial residual due to the solution from the previous time step is reduced by 10 orders of magnitude. We generated sequences of the discrete Oseen problems (1.2) with ($\bar{\sigma} = 1/12$) and without ($\bar{\sigma} = 0$) SUPG-stabilization. The choice of parameters τ_1, τ_2 leads to stable computations over the whole cardiac cycle. The total number of iterations depends on the size of the system and the mesh and appears to be very similar for both examples with and without stabilization. The total number of iterations is 20% larger for the fine grid, which should be expected for the preconditioner based on an incomplete factorization. Over the cardiac cycle, the variations of the iteration counts and CPU times per linear solve are rather modest, see the top and bottom plots in Figures 6.2 and Figures 6.3. The difference in otherwise similar performance of liners solvers for the cases $\bar{\sigma} = 1/12$ and $\bar{\sigma} = 0$ is the following: For $\bar{\sigma} = 1/12$, when the maximum flow rate on the inlet is achieved, the number

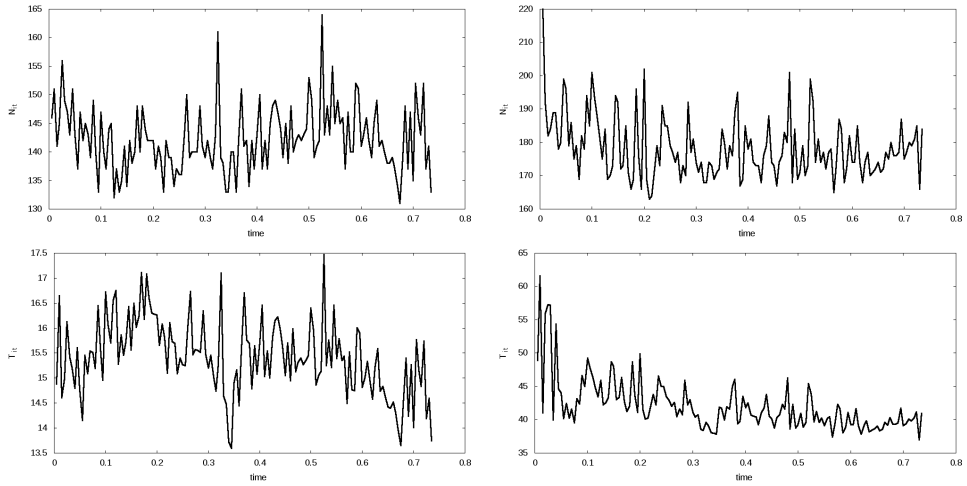

FIG. 6.2. *Right coronary artery, computations on grid 63k (left) and grid 120k (right) without SUPG-stabilization and* $\tau_1 = 0.03$*: The top plots show the number of BiCGStab iterations, the bottom plots show the time of BiCGstab iterations at each time step.*

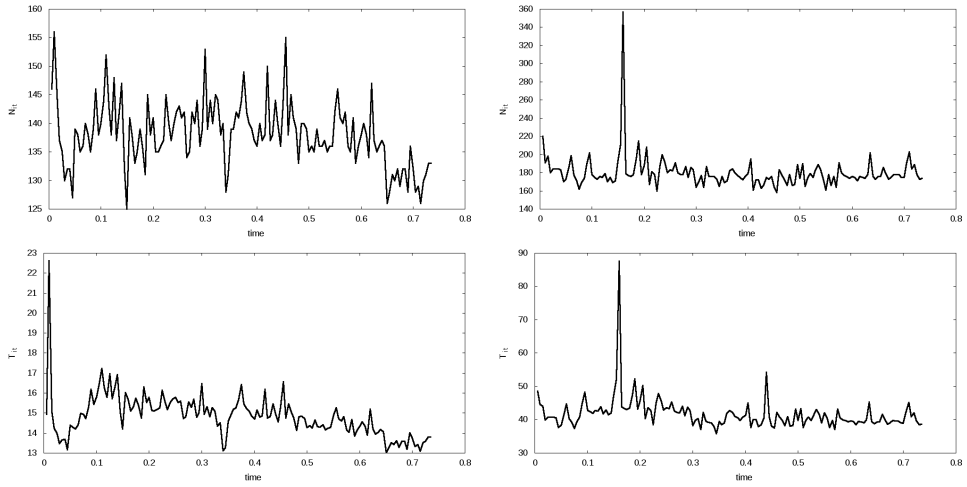

FIG. 6.3. *Right coronary artery, computations on grid 63k (left) and grid 120k (right), SUPG-stabilization with* $\bar{\sigma} = 1/12$ *and* $\tau_1 = 0.03$*: The top plots show the number of BiCGStab iterations, the bottom plots show the time of BiCGstab iterations at each time step.*

of iterations and times needed to build preconditioner increase essentially (approximately twice as much as average). This happens over a few time steps. In these cases when factorization is performed several small pivots occur and their modification is performed during the incomplete factorization.

The next series of experiments shows that restrictions (4.5) on σ_τ are important in practice. According to Theorems 3.2 and 4.1, exact LU factorization of $\mathcal{A}$ without pivoting is stable if σ_τ are small enough. In particular, according to estimate (4.24) sufficient conditions (4.5) are satisfied by parameters from (2.4) if $\bar{\sigma} \le \min\{\bar{C}_{\rm in}^{-2},\ \frac{1}{2}C_{\rm in}^{-1}\}$. In this experiment, we increase $\bar{\sigma}$ two times setting $\bar{\sigma} = 1/6$. It occurs that the matrices associated with the coarse grid are more difficult to solve now. For the first

TABLE 6.2
The performance of ILU($\tau_1, \tau_2 = 7\tau_1^2$) for right coronary artery with less viscous blood $\nu = 0.025\,\mathrm{cm}^2/\mathrm{s}$. Threshold values allowing to run the entire SUPG-stabilized simulation with different stabilization parameters $\bar{\sigma}$. '$\star$' means solution blow-up, '–' means untracktable systems for any applicable τ_1.

$\bar{\sigma}$	0	1/96	1/48	1/24	1/12	1/6	1/3
τ_1	$\star$	0.03	0.03	0.03	0.03	0.003	–

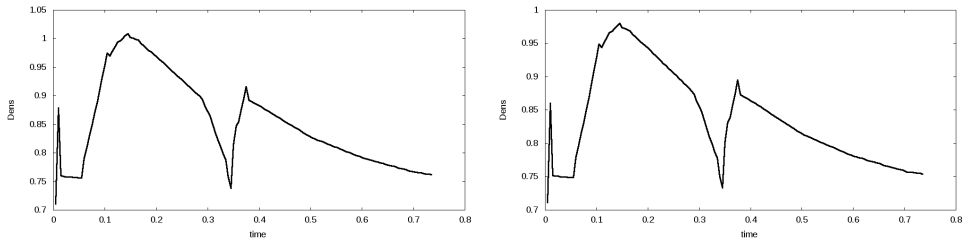

FIG. 6.4. *The fill-in of the LU factors for $\bar{\sigma} = 0$ (left) and $\bar{\sigma} = 1/12$ (right).*

threshold parameter τ_1 as small as 10^{-4}, we observe no pivot modifications and the average number of BiCGstab iterations per linear solve is only 8. This suggests that the exact LU factorization is still stable. Such small τ_1 is non-practical because of enormous memory demands and factorization time. However, already for τ_1 equal $3 \cdot 10^{-4}$ on two time steps the algorithm makes 12 and 4 modifications of nearly zero pivots in order to avoid the breakdown. This caused the convergence slowdown, as many as 135 iterations for one system. Certain Oseen systems with $\bar{\sigma} = 1/6$ on the fine grid can not be solved by the ILU-preconditioned BiCGstab iterations with any values of threshold parameters that we tried. Note that for smaller $\bar{\sigma} = 1/12$ the algorithm performs without pivot modifications even for $\tau_1 = 0.03$.

Further, we decrease the viscosity of the fluid to $\nu = 0.025\,\mathrm{cm}^2/\mathrm{s}$, and try to run the same simulation on the coarse grid. For this value of the viscosity, the simulation without SUPG stabilization fail (solution blow-up is observed on $t = 0.23$ s). Adding SUPG stabilization allows to obtain physiologically meaningful solution, however, for large enough parameter $\bar{\sigma}$ the linear systems are harder to solve: $\bar{\sigma} = 1/6$ requires smaller threshold parameter τ_1, whereas $\bar{\sigma} = 1/3$ generates unsolvable systems, see Table 6.2. This experiment confirms that restrictions on $\bar{\sigma}$ come both from stability of the FE method and algebraic stability of the LU factorization.

We finally note that in experiments with varying inlet velocity, which leads to varying Reynolds number, the two-parameter ILU preconditioner demonstrated a remarkable adaptive property. The fill-in of the L and U blocks decrease or increase depending on the Reynolds number, see Figure 6.4 and compare to the inlet waveform in Figure 6.1. We will study this property of the two-parameter ILU preconditioner in more detail in a subsequent paper.

7. Closing remarks and conclusions. In this paper, we studied the stability of the LU factorization for the stabilized finite element formulations of the incompressible Navier-Stokes equations. Further, the two-parameter threshold ILU factorization was applied to define a preconditioner in the Krylov subspace method. Advantages and shortcomings of incomplete elementwise factorization preconditioners are well known: On the one hand, they are rather insensitive to discretization, boundary con-

ditions for governing PDEs, domain geometry, flow directions; on the other hand, even for discrete elliptic problems, ILU preconditioners do not scale optimally with respect to the number of unknowns. We observed such non-optimality in the numerical experiments for generalized saddle-point problem as well. For 3D problems, when the mesh size is not too small, such dependence can be an acceptable price for other robustness properties of the preconditioner: in our experiments, the two times increase of the number of mesh cells leaded only 20% increase of the iteration counts. Similar to the previous studies in [20] we found that natural **u**-p ordering of unknowns is sufficient for numerical stability of exact LU-factorization, when stabilization parameters satisfy certain bounds. In the algebraic language this translates as the positive definiteness of the A block and the sufficiently small size of perturbation in the (1,2)-block. In this paper, the stability bounds for the factorization are rigorously formulated in terms of algebraic properties of sub-blocks of the original saddle-point matrix.

In general, higher Reynolds numbers lead to efficiency loss for most well-known preconditioners for (1.3). In case of 3D blood flow in coronary arteries, the actual viscosity and velocity are such that P2-P1 stable FE discretization still provides the non-oscillatory solution on tetrahedral meshes with $\sim 10^5$ cells. However, the coronary blood flow parameters are close to the limit of non-oscillatory discretization and SUPG-stabilization may be in-demand. SUPG-stabilization alters the (1,1)-block and (1,2)-block of the Oseen matrix (1.3), and hence changes open new questions about the stability of factorizations. Theorem 4.1 show how the constants in the algebraic stability estimates depend on the flow and discretization parameters. This gives a certain insight into the performance of incomplete factorizations as preconditioners for flow problems. The present numerical analysis of incomplete factorizations for such non-symmetric matrices is still limited to the lower estimate (5.2) of the diagonal entries of the triangular factors.

The two-parameter ILU preconditioner was applied to hemodynamic flow in a right coronary artery reconstructed from a real patient coronary CT angiography. The performance of the preconditioner is good for *a suitable choice* of SUPG-stabilization parameters.

Acknowledgements. The authors thank Tatiana Dobroserdova and Alexander Danilov for the assistance in building tetrahedral meshes and finite element systems, and Sergei Goreinov for sharing his implementation of the row-wise variant of the ILU(τ_1,τ_2) factorization.

REFERENCES

[1] N. Ahmed, T. C. Rebollo, V. John, and S. Rubino, *A review of variational multiscale methods for the simulation of turbulent incompressible flows*, Archives of Computational Methods in Engineering, pp. 1–50.

[2] M. Benzi, *Preconditioning techniques for large linear systems: a survey*, Journal of Computational Physics, 182 (2002), pp. 418–477.

[3] M. Benzi, G. H. Golub, and J. Liesen, *Numerical solution of saddle point problems*, Acta Numerica, 14 (2005), pp. 1–137.

[4] M. Braack, E. Burman, V. John, and G. Lube, *Stabilized finite element methods for the generalized oseen problem*, Computer Methods in Applied Mechanics and Engineering, 196 (2007), pp. 853–866.

[5] M. Braack, P. B. Mucha, and W. M. Zajaczkowski, *Directional do-nothing condition for the Navier–Stokes equations*, J. Comput. Math., 32 (2014), pp. 507–521.

[6] A. N. Brooks and T. J. Hughes, *Streamline upwind/Petrov–Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier–Stokes*

equations, Computer Methods in Applied Mechanics and Engineering, 32 (1982), pp. 199–259.

[7] R. Codina, *Stabilized finite element approximation of transient incompressible flows using orthogonal subscales*, Computer Methods in Applied Mechanics and Engineering, 191 (2002), pp. 4295–4321.

[8] O. Dahl and S. Ø. Wille, *An ILU preconditioner with coupled node fill-in for iterative solution of the mixed finite element formulation of the 2D and 3D Navier-Stokes equations*, International Journal for Numerical Methods in Fluids, 15 (1992), pp. 525–544.

[9] H. C. Elman, D. Silvester, and A. Wathen, *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics*, Oxford University Press, 2014.

[10] H. C. Elman and R. S. Tuminaro, *Boundary conditions in approximate commutator preconditioners for the Navier–Stokes equations*, Electronic Transactions on Numerical Analysis, 35 (2009), pp. 257–280.

[11] L. P. Franca and S. L. Frey, *Stabilized finite element methods: Ii. the incompressible navier-stokes equations*, Computer Methods in Applied Mechanics and Engineering, 99 (1992), pp. 209–233.

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

[13] V. Girault and P.-A. Raviart, *Finite element approximation of the navier-stokes equations*, Lecture Notes in Mathematics, Berlin Springer Verlag, 749 (1979).

[14] G. H. Golub and C. v. Loan, *Matrix computations*, Baltimore, MD: Johns Hopkins University Press, 1996.

[15] G. H. Golub and C. Van Loan, *Unsymmetric positive definite linear systems*, Linear Algebra and its Applications, 28 (1979), pp. 85–97.

[16] T. J. Hughes, G. R. Feijóo, L. Mazzei, and J.-B. Quincy, *The variational multiscale methoda paradigm for computational mechanics*, Computer methods in applied mechanics and engineering, 166 (1998), pp. 3–24.

[17] J. Jung, A. Hassanein, and R. W. Lyczkowski, *Hemodynamic computation using multiphase flow dynamics in a right coronary artery*, Annals of biomedical engineering, 34 (2006), pp. 393–407.

[18] I. E. Kaporin, *High quality preconditioning of a general symmetric positive definite matrix based on its $U^TU + U^TR + R^TU$-decomposition*, Numerical Linear Algebra with Applications, 5 (1998), pp. 483–509.

[19] ———, *Scaling, reordering, and diagonal pivoting in ilu preconditionings*, Russian Journal of Numerical Analysis and Mathematical Modelling rnam, 22 (2007), pp. 341–375.

[20] I. N. Konshin, M. A. Olshanskii, and Y. V. Vassilevski, *ILU preconditioners for nonsymmetric saddle-point matrices with application to the incompressible navier–stokes equations*, SIAM Journal on Scientific Computing, 37 (2015), pp. A2171–A2197.

[21] K. Lipnikov, Y. Vassilevski, A. Danilov, et al., *Advanced Numerical Instruments 2D*, http://sourceforge.net/projects/ani2d.

[22] ———, *Advanced Numerical Instruments 3D*, http://sourceforge.net/projects/ani3d.

[23] M. A. Ol'shanskii and V. M. Staroverov, *On simulation of outflow boundary conditions in finite difference calculations for incompressible fluid*, International Journal for Numerical Methods in Fluids, 33 (2000), pp. 499–534.

[24] M. A. Olshanskii and E. E. Tyrtyshnikov, *Iterative methods for linear systems: theory and applications*, SIAM, 2014.

[25] M. A. Olshanskii and Y. V. Vassilevski, *Pressure Schur complement preconditioners for the discrete Oseen problem*, SIAM Journal on Scientific Computing, 29 (2007), pp. 2686–2704.

[26] H.-G. Roos, M. Stynes, and L. Tobiska, *Numerical methods for singularly perturbed differential equations: convection–diffusion and flow problems*, Springer, Berlin, 1996.

[27] H.-G. Roos, M. Stynes, and L. Tobiska, *Robust numerical methods for singularly perturbed differential equations: convection-diffusion-reaction and flow problems*, vol. 24, Springer Science & Business Media, 2008.

[28] Y. Saad, *Iterative methods for sparse linear systems*, SIAM, 2003.

[29] R. L. Sani and P. M. Gresho, *Résumé and remarks on the open boundary condition minisymposium*, International Journal for Numerical Methods in Fluids, 18 (1994), pp. 983–1008.

[30] J. Scott and M. Tuma, *On signed incomplete Cholesky factorization preconditioners for saddle-point systems*, SIAM Journal on Scientific Computing, 36 (2014), pp. A2984–A3010.

[31] J. Scott and M. Tuma, *Solving symmetric indefinite systems using memory efficient incomplete factorization preconditioners*, tech. rep., STFC Rutherford Appleton Laboratory, 02 2015.

[32] A. Segal, M. ur Rehman, and C. Vuik, *Preconditioners for incompressible Navier–Stokes solvers*, Numerical Mathematics: Theory, Methods and Applications, 3 (2010), pp. 245–275.
[33] J. Stoer and R. Bulirsch, *Introduction to numerical analysis*, Springer, New York, 1993.
[34] M. Suarjana and K. H. Law, *A robust incomplete factorization based on value and space constraints*, Int. Journal for Numerical Methods in Engineering, 38 (1995), pp. 1703–1719.
[35] M. Tismenetsky, *A new preconditioning technique for solving large sparse linear systems*, Linear Algebra and its Applications, 154 (1991), pp. 331–353.
[36] S. Turek, *Efficient Solvers for Incompressible Flow Problems: An Algorithmic and Computational Approache*, vol. 6, Springer Science & Business Media, 1999.
[37] C. Vuik, G. Segal, et al., *A comparison of preconditioners for incompressible Navier–Stokes solvers*, International Journal for Numerical Methods in Fluids, 57 (2008), pp. 1731–1751.
[38] ———, *Simple-type preconditioners for the Oseen problem*, International Journal for Numerical Methods in Fluids, 61 (2009), pp. 432–452.