Numerical Analysis of Partial Differential Equations

Alfio Quarteroni

MOX, Dipartimento di Matematica
Politecnico di Milano

MODELLISTICA E CALCOLO SCIENTIFICO
MODELING AND SCIENTIFIC COMPUTING

Lecture Notes
A.Y. 2021-2022

Navier-Stokes equations

Cfr ref.[1], Chap. 17

Navier-Stokes equations describe the motion of a fluid with constant density ρ in a domain $\Omega \subset \mathbb{R}^d$ (with $d = 2, 3$). They read as follows

$$\begin{cases} \dfrac{\partial \mathbf{u}}{\partial t} - \mathrm{div}\left[\nu\left(\nabla \mathbf{u} + \nabla \mathbf{u}^T\right)\right] + (\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p = \mathbf{f}, & \mathbf{x} \in \Omega,\ t > 0, \\ \mathrm{div}\mathbf{u} = 0, & \mathbf{x} \in \Omega,\ t > 0, \end{cases} \tag{1}$$

where

- $\mathbf{u}$ is the fluid's velocity
- p is the pressure divided by the density (which will simply be called "pressure")
- ν is the kinematic viscosity
- $\mathbf{f} \in \mathrm{L}^2(\mathbb{R}^+; [\mathrm{L}^2(\Omega)]^d)$ is a forcing term per unit of mass

[1] Alfio Quarteroni. *Numerical models for differential problems*. 3rd. Springer, 2018.

The first equation is that of conservation of linear momentum, the second one that of conservation of mass, which is also called the continuity equation.

- The term $(\mathbf{u} \cdot \nabla)\mathbf{u}$ describes the process of convective transport
- The term $-\mathrm{div}\left[\nu(\nabla\mathbf{u} + \nabla\mathbf{u}^T)\right]$ describes the process of molecular diffusion.

When ν is constant, from the continuity equation we obtain

$$\mathrm{div}\left[\nu(\nabla\mathbf{u}+\nabla\mathbf{u}^T)\right]=\nu\left(\Delta\mathbf{u}+\nabla\mathrm{div}\mathbf{u}\right)=\nu\Delta\mathbf{u}$$

whence system (1) can be written in the equivalent form

$$\begin{cases}\dfrac{\partial\mathbf{u}}{\partial t}-\nu\Delta\mathbf{u}+(\mathbf{u}\cdot\nabla)\mathbf{u}+\nabla p=\mathbf{f}, & \mathbf{x}\in\Omega,\ t>0,\\ \mathrm{div}\mathbf{u}=0, & \mathbf{x}\in\Omega,\ t>0,\end{cases}\tag{2}$$

which is the one that we will consider in the following.

Equations (2) are often called incompressible Navier-Stokes equations. More in general, fluids satisfying the *incompressibility condition* $\mathrm{div}\mathbf{u}=0$ are said to be incompressible.

Constant density fluids necessarily satisfy this condition, however there exist incompressible fluids featuring variable density (e.g., stratified fluids) that are governed by a different system of equations in which the density ρ explicitly shows up.

In order for problem (2) to be well posed it is necessary to assign the initial condition

$$\mathbf{u}(\mathbf{x},0)=\mathbf{u}_0(\mathbf{x}) \quad \forall \mathbf{x}\in\Omega, \tag{3}$$

where $\mathbf{u}_0$ is a given divergence-free vector field, together with suitable boundary conditions, such as, e.g., $\forall t>0$,

$$\begin{cases} \mathbf{u}(\mathbf{x},t)=\boldsymbol{\varphi}(\mathbf{x},t) \quad \forall \mathbf{x}\in\Gamma_D, \\ \\ \left(\nu\dfrac{\partial\mathbf{u}}{\partial\mathbf{n}}-p\mathbf{n}\right)(\mathbf{x},t)=\boldsymbol{\psi}(\mathbf{x},t) \quad \forall \mathbf{x}\in\Gamma_N, \end{cases} \tag{4}$$

where $\boldsymbol{\varphi}$ and $\boldsymbol{\psi}$ are given vector functions, while Γ_D and Γ_N provide a partition of the domain boundary $\partial\Omega$, that is $\Gamma_D\cup\Gamma_N=\partial\Omega$, $\overset{\circ}{\Gamma}_D\cap\overset{\circ}{\Gamma}_N=\emptyset$. Finally, as usual $\mathbf{n}$ is the outward unit normal vector to $\partial\Omega$.

If we use the alternative formulation (1), the second equation in (4) must be replaced by

$$\left[\nu\left(\nabla\mathbf{u}+\nabla\mathbf{u}^T\right)\mathbf{n}-p\mathbf{n}\right](\mathbf{x},t)=\boldsymbol{\psi}(\mathbf{x},t) \quad \forall \mathbf{x}\in\Gamma_N.$$

Denoting with u_i, $i = 1, \dots, d$ the components of the vector $\mathbf{u}$ with respect to a Cartesian frame, and with f_i the components of $\mathbf{f}$, system (2) can be written componentwise as

$$\begin{cases} \dfrac{\partial u_i}{\partial t} - \nu \Delta u_i + \displaystyle\sum_{j=1}^{d} u_j \dfrac{\partial u_i}{\partial x_j} + \dfrac{\partial p}{\partial x_i} = f_i, \quad i = 1, \dots, d, \\ \displaystyle\sum_{j=1}^{d} \dfrac{\partial u_j}{\partial x_j} = 0. \end{cases}$$

Remark: Well-posedness

In the two-dimensional case the Navier-Stokes equations with the boundary conditions previously indicated yield well-posed problems. This means that if all data (initial condition, forcing term, boundary data) are smooth enough, then the solution is continuous together with its derivatives and does not develop singularities in time.

Things may go differently in three dimensions, where existence and uniqueness of classical solutions have been proven only locally in time (that is for a sufficiently small time interval). In the following slides we will introduce the weak formulation of the Navier-Stokes equations, for which existence of a solution has been proven for all times. However, the issue of uniqueness (which is related to that of regularity) is still open, and is actually the central issue of Navier-Stokes theory.

Remark

The Navier-Stokes equations have been written in terms of the *primitive variables* $\mathbf{u}$ and p, but other sets of variables may be used, too. For instance, in the two-dimensional case it is common to see the vorticity ω and the streamfunction ψ, that are related to the velocity as follows

$$\omega = \mathrm{rot}\mathbf{u} = \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2}, \quad \mathbf{u} = \begin{bmatrix} \dfrac{\partial \psi}{\partial x_2} \\ -\dfrac{\partial \psi}{\partial x_1} \end{bmatrix}.$$

The various formulations are in fact equivalent from a mathematical standpoint, although they give rise to different numerical methods. See, e.g., ref.[a].

[a] L. Quartapelle. *Numerical Solution of the Incompressible Navier-Stokes Equations*. Basel: Birkhäuser Verlag, 1993.

Weak formulation of Navier-Stokes equations

A weak formulation of problem (2)–(4) can be obtained by proceeding formally, as follows. Let us multiply the first equation of (2) by a test function $\mathbf{v}$ belonging to a suitable space V that will be specified later on, and integrate in Ω

$$\int_\Omega \frac{\partial \mathbf{u}}{\partial t} \cdot \mathbf{v}\ d\Omega - \int_\Omega \nu \Delta \mathbf{u} \cdot \mathbf{v}\ d\Omega + \int_\Omega [(\mathbf{u} \cdot \nabla)\mathbf{u}] \cdot \mathbf{v}\ d\Omega + \int_\Omega \nabla p \cdot \mathbf{v} d\Omega$$
$$= \int_\Omega \mathbf{f} \cdot \mathbf{v} d\Omega.$$

Using Green's formulae [ref.[2], Chap. 3] we find:

$$\begin{aligned} -\int_\Omega \nu \Delta \mathbf{u} \cdot \mathbf{v}\ d\Omega &= \int_\Omega \nu \nabla \mathbf{u} \cdot \nabla \mathbf{v}\ d\Omega - \int_{\partial\Omega} \nu \frac{\partial \mathbf{u}}{\partial \mathbf{n}} \cdot \mathbf{v}\ d\gamma, \\ \int_\Omega \nabla p \cdot \mathbf{v}\ d\Omega &= -\int_\Omega p \mathrm{div} \mathbf{v}\ d\Omega + \int_{\partial\Omega} p \mathbf{v} \cdot \mathbf{n}\ d\gamma. \end{aligned}$$

[2] Quarteroni, *Numerical models for differential problems*.

Using these relations in the first of (2), we obtain

$$\begin{aligned}&\int_\Omega \frac{\partial \mathbf{u}}{\partial t}\cdot \mathbf{v}\ d\Omega + \int_\Omega \nu \nabla \mathbf{u}\cdot \nabla \mathbf{v}\ d\Omega + \int_\Omega [(\mathbf{u}\cdot\nabla)\mathbf{u}]\cdot \mathbf{v}\ d\Omega \\ &- \int_\Omega p \mathrm{div} \mathbf{v}\ d\Omega = \int_\Omega \mathbf{f}\cdot\mathbf{v}\ d\Omega + \int_{\partial\Omega}\left(\nu\frac{\partial \mathbf{u}}{\partial \mathbf{n}} - p\mathbf{n}\right)\cdot \mathbf{v}\ d\gamma \qquad \forall \mathbf{v}\in V.\end{aligned} \tag{5}$$

(All boundary integrals should indeed be regarded as duality pairings between V' and V.)

Similarly, by multiplying the second equation of (2) by a test function q, belonging to a suitable space Q to be specified, then integrating on Ω it follows

$$\int_\Omega q\, \mathrm{div} \mathbf{u}\ d\Omega = 0 \quad \forall q \in Q. \tag{6}$$

Customarily V is chosen so that the test functions vanish on the boundary portion where a Dirichlet data is prescribed on $\mathbf{u}$, that is

$$V = [\mathrm{H}^1_{\Gamma_D}(\Omega)]^d = \{\mathbf{v} \in [\mathrm{H}^1(\Omega)]^d : \ \mathbf{v}|_{\Gamma_D} = \mathbf{0}\}. \tag{7}$$

It will coincide with $[\mathrm{H}^1_0(\Omega)]^d$ if $\Gamma_D = \partial\Omega$. If Γ_N has positive measure, we can choose $Q = \mathrm{L}^2(\Omega)$. If $\Gamma_D = \partial\Omega$, then the pressure space should be L^2_0 to ensure uniqueness for the pressure p (see later, slide 16).

Moreover, if $t > 0$, then $\mathbf{u}(t) \in [\mathrm{H}^1(\Omega)]^d$, with $\mathbf{u}(t) = \varphi(t)$ on Γ_D, $\mathbf{u}(0) = \mathbf{u}_0$ and $p(t) \in Q$.

Having chosen these functional spaces, we can note first of all that

$$\int_{\partial\Omega} (\nu \frac{\partial \mathbf{u}}{\partial \mathbf{n}} - p\mathbf{n}) \cdot \mathbf{v} d\gamma = \int_{\Gamma_N} \boldsymbol{\psi} \cdot \mathbf{v} d\gamma \quad \forall \mathbf{v} \in V.$$

Notation

For every function $\mathbf{v} \in \mathbf{H}^1(\Omega)$, we denote by

$$\|\mathbf{v}\|_{\mathbf{H}^1(\Omega)} = (\sum_{k=1}^{d} \|v_k\|^2_{\mathrm{H}^1(\Omega)})^{1/2}$$

its norm and by

$$|\mathbf{v}|_{\mathbf{H}^1(\Omega)} = \Big(\sum_{k=1}^{d} |v_k|^2_{\mathrm{H}^1(\Omega)}\Big)^{1/2}$$

its seminorm. The notation $\|\mathbf{v}\|_{\mathbf{L}^p(\Omega)}$, $1 \le p < \infty$, has a similar meaning. The same symbols will be used in case of tensor functions. Thanks to Poincaré's inequality, $|\mathbf{v}|_{\mathbf{H}^1(\Omega)}$ is equivalent to the norm $\|\mathbf{v}\|_{\mathbf{H}^1(\Omega)}$ for all functions belonging to V, provided that the Dirichlet boundary has a positive measure.

All the integrals involving bilinear terms are finite. More precisely, by using the vector notation $\mathbf{H}^k(\Omega) = [\mathrm{H}^k(\Omega)]^d$, $\mathbf{L}^p(\Omega) = [\mathrm{L}^p(\Omega)]^d$, $k \geq 1$, $1 \leq p < \infty$, we find:

$$\left|\nu \int_\Omega \nabla \mathbf{u} \cdot \nabla \mathbf{v} d\Omega\right| \leq \nu |\mathbf{u}|_{\mathbf{H}^1(\Omega)} |\mathbf{v}|_{\mathbf{H}^1(\Omega)},$$

$$\left|\int_\Omega p \mathrm{div} \mathbf{v} d\Omega\right| \leq \|p\|_{\mathrm{L}^2(\Omega)} |\mathbf{v}|_{\mathbf{H}^1(\Omega)},$$

$$\left|\int_\Omega q \nabla \mathbf{u} d\Omega\right| \leq \|q\|_{\mathrm{L}^2(\Omega)} |\mathbf{u}|_{\mathbf{H}^1(\Omega)}.$$

Also the integral involving the trilinear term is finite. To see how, let us start by recalling the following result: if $d \leq 3$,

$$\forall \mathbf{v} \in \mathbf{H}^1(\Omega), \text{ then } \mathbf{v} \in \mathbf{L}^4(\Omega) \text{ and } \exists C > 0 \text{ s.t. } \|\mathbf{v}\|_{\mathbf{L}^4(\Omega)} \leq C \|\mathbf{v}\|_{\mathbf{H}^1(\Omega)}.$$

Using the following three-term Hölder inequality

$$\left| \int_\Omega fgh \; d\Omega \right| \leq \|f\|_{\mathrm{L}^p(\Omega)} \|g\|_{\mathrm{L}^q(\Omega)} \|h\|_{\mathrm{L}^r(\Omega)},$$

valid for all $p, q, r > 1$ such that $p^{-1} + q^{-1} + r^{-1} = 1$, we conclude that

$$\left| \int_\Omega [(\mathbf{u} \cdot \nabla)\mathbf{u}] \cdot \mathbf{v} \; d\Omega \right| \leq \|\nabla \mathbf{u}\|_{\mathbf{L}^2(\Omega)} \|\mathbf{u}\|_{\mathbf{L}^4(\Omega)} \|\mathbf{v}\|_{\mathbf{L}^4(\Omega)} \leq C^2 \|\mathbf{u}\|^2_{\mathbf{H}^1(\Omega)} \|\mathbf{v}\|_{\mathbf{H}^1(\Omega)}.$$

Solution Uniqueness

As for the solution's uniqueness, let us consider again the Navier-Stokes equations in strong form (2) (similar considerations can be made on the weak form (5), (6)).

If $\Gamma_D = \partial\Omega$, when only boundary conditions of Dirichlet type are imposed, the pressure appears merely in terms of its gradient; in such a case, if we call $(\mathbf{u}, p)$ a solution of (2), for any possible constant c the couple $(\mathbf{u}, p + c)$ is a solution too, since $\nabla(p + c) = \nabla p$.

To avoid such indeterminacy one can fix a priori the value of p at one given point $\mathbf{x}_0$ of the domain Ω, that is set $p(\mathbf{x}_0) = p_0$, or, alternatively, require the pressure to have null average, i.e., $\int_\Omega p \; d\Omega = 0$.

The former condition requires to prescribe a pointwise value for the pressure, but this is inconsistent with our ansatz that $p \in \mathrm{L}^2(\Omega)$. (We anticipate, however, that this is admissible at the numerical level when we look for a continuous finite-dimensional pressure).

For this reason we assume from now on that the pressure is average-free. More specifically, we will consider the following pressure space

$$\mathrm{Q} = \mathrm{L}^2_0(\Omega) = \{p \in \mathrm{L}^2(\Omega) : \int_\Omega p \; d\Omega = 0\}.$$

Further, we observe that if $\Gamma_D = \partial\Omega$, the prescribed Dirichlet data $\boldsymbol{\varphi}$ must be compatible with the incompressibility constraint; indeed,

$$\int_{\partial\Omega} \boldsymbol{\varphi} \cdot \mathbf{n} \; d\gamma = \int_{\Omega} \mathrm{div}\mathbf{u} \; \mathrm{d}\Omega = 0.$$

If Γ_N is not empty, i.e. in presence of either Neumann or mixed Dirichlet-Neumann boundary conditions, the problem of pressure indeterminacy (up to an additive constant) no longer exists. In this case we can take $\mathrm{Q} = \mathrm{L}^2(\Omega)$.

In conclusion, from now on we shall implicitly assume

$$Q = \mathrm{L}^2(\Omega) \quad \text{if} \quad \Gamma_N \neq \emptyset, \quad Q = \mathrm{L}^2_0(\Omega) \quad \text{if } \Gamma_N = \emptyset. \tag{8}$$

The weak formulation of the system (2), (3), (4) is therefore:

find $\mathbf{u} \in \mathrm{L}^2(\mathbb{R}^+; [\mathrm{H}^1(\Omega)]^d) \cap C^0(\mathbb{R}^+; [\mathrm{L}^2(\Omega)]^d)$, $p \in \mathrm{L}^2(\mathbb{R}^+; Q)$ such that

$$\begin{cases} \displaystyle\int_\Omega \frac{\partial \mathbf{u}}{\partial t} \cdot \mathbf{v} \; d\Omega + \nu \int_\Omega \nabla \mathbf{u} \cdot \nabla \mathbf{v} \; d\Omega + \int_\Omega [(\mathbf{u} \cdot \nabla)\mathbf{u}] \cdot \mathbf{v} \; d\Omega \\ \displaystyle - \int_\Omega p \; \mathrm{div} \mathbf{v} \; d\Omega = \int_\Omega \mathbf{f} \cdot \mathbf{v} \; d\Omega + \int_{\Gamma_N} \psi \cdot \mathbf{v} d\gamma \quad \forall \mathbf{v} \in V, \\ \displaystyle\int_\Omega q \; \mathrm{div} \mathbf{u} d\Omega = 0 \quad \forall q \in Q, \end{cases} \tag{9}$$

with $\mathbf{u}|_{\Gamma_D} = \varphi_D$ and $\mathbf{u}|_{t=0} = \mathbf{u}_0$. The space V is the one in (7) while Q is the space introduced in (8).

As we have already anticipated, existence of solutions can be proven for this problem for both dimensions $d = 2$ and $d = 3$, whereas uniqueness has been proven only in the case $d = 2$ for sufficiently small data (see, e.g., ref.[3]).

[3] S. Salsa. *Partial Differential Equations in Action - From Modelling to Theory*. Milan: Springer, 2008.

The Reynolds number

Let us define the Reynolds number

$$Re = \frac{|\mathbf{U}|L}{\nu},$$

where L is a representative length of the domain Ω (e.g. the length of a channel where the fluid's flow is studied) and $\mathbf{U}$ a representative fluid velocity.

The Reynolds number measures the extent to which convection dominates over diffusion.

- When $Re \ll 1$ the convective term $(\mathbf{u} \cdot \nabla)\mathbf{u}$ can be omitted, and the Navier-Stokes equations reduce to the so-called Stokes equations, that will be investigated later.
- On the other hand, if Re is large, problems may arise concerning uniqueness of the solution, the existence of stationary and stable solutions, the possible existence of strange attractors, the transition towards turbulent flows.

For more information regarding the topic of turbulence, see e.g. ref.[4] and ref.[5].

When fluctuations of flow velocity occur at very small spatial and temporal scales, their numerical approximation becomes very difficult if not impossible. In those cases one typically resorts to the so-called turbulence models: the latter allow the approximate description of this flow behaviour through either algebraic or differential equations.

[4] R. Temam. *Navier Stokes Equations*. Amsterdam: North-Holland, 2001.

[5] C. Foias et al. *Navier-Stokes Equations and Turbulence*. Cambridge: Cambridge Univ. Press, 2001.

This topic will not be addressed in this course. The interested readers may consult, e.g., ref.[6] for a description of the physical aspects of turbulent flows, ref.[7] for multiscale analysis of incompressible flows, ref.[8] for modelling aspects of multiscale systems, ref.[9] for the analysis of one of the most widely used turbulence models, the so-called $\kappa - \epsilon$ model. Ref.[10] and ref.[11] provide the analysis of the so-called *Large Eddy* model, which is more computationally expensive but in principle better suited to provide a more realistic description of turbulent flow fields.

[6] D. C. Wilcox. *Turbulence Modeling in CFD*. II. La Cañada, CA: DCW Industries, 1998.

[7] T. Y. Hou, D. P. Yang, and H. Ran. "Multiscale Analysis and Computation for the 3D Incompressible Navier-Stokes Equations". In: *SIAM Multiscale Modeling and Simulation* 6 (4) (2008), pp. 1317–1346.

[8] C. Le Bris. *Systemes multiiéchelles: modélisation et simulation*. Vol. 47. Mathématiques et Applications. Paris: Springer, 2005.

[9] B. Mohammadi and O. Pironneau. *Analysis of the K-Epsilon Turbulence Model*. Chichester: John Wiley & Sons, 1994.

[10] P. Sagaut. *Large Eddy Simulation for Incompressible Flows: an Introduction*. III. Berlin Heidelberg: Springer-Verlag, 2006.

[11] L. C. Berselli, T. Iliescu, and W. J. Layton. *Mathematics of Large Eddy Simulation of Turbulent Flows*. Berlin Heidelberg: Springer, 2006.

Divergence free formulation of Navier-Stokes equations

By eliminating the pressure, the Navier-Stokes equations can be rewritten in *reduced form* in the sole variable $\mathbf{u}$. With this aim let us introduce the following subspaces of $[\mathrm{H}^1(\Omega)]^d$:

$$V_{\mathrm{div}} = \{\mathbf{v} \in \left[\mathrm{H}^1(\Omega)\right]^d : \ \mathrm{div}\mathbf{v} = 0\},$$
$$V^0_{\mathrm{div}} = \{\mathbf{v} \in V_{\mathrm{div}} : \ \mathbf{v} = \mathbf{0} \text{ on } \Gamma_D\}.$$

Upon requiring the test function $\mathbf{v}$ in the momentum equation in (9) to belong to the space V_{div}, the term associated to the pressure gradient vanishes, whence we find the following reduced problem for the velocity

find $\mathbf{u} \in \mathrm{L}^2(\mathbb{R}^+; V_{\mathrm{div}}) \cap C^0(\mathbb{R}^+; [\mathrm{L}^2(\Omega)]^d)$ such that

$$\begin{aligned} \int_\Omega \frac{\partial \mathbf{u}}{\partial t} \cdot \mathbf{v} \ d\Omega \quad &+\nu \int_\Omega \nabla \mathbf{u} \cdot \nabla \mathbf{v} \ d\Omega + \int_\Omega [(\mathbf{u} \cdot \nabla)\mathbf{u}] \cdot \mathbf{v} \ d\Omega \\ &= \int_\Omega \mathbf{f} \cdot \mathbf{v} \ d\Omega + \int_{\Gamma_N} \psi \cdot \mathbf{v} \ d\gamma \quad \forall \mathbf{v} \in V^0_{\mathrm{div}}, \end{aligned} \tag{10}$$

with $\mathbf{u}|_{\Gamma_D} = \varphi_D$ and $\mathbf{u}|_{t=0} = \mathbf{u}_0$.

Since this problem is (nonlinear) parabolic, its analysis can be carried out using techniques similar to those applied in for parabolic problems. (See, e.g., ref.[12].)

Obviously, if $\mathbf{u}$ is a solution of (9), then it also solves (10). Conversely, the following theorem holds. For its proof, see, e.g., ref.[13].

Theorem

Let $\Omega \subset \mathbb{R}^d$ be a domain with Lipschitz-continuous boundary $\partial\Omega$. Let $\mathbf{u}$ be a solution to the reduced problem (10). Then there exists a unique function $p \in \mathrm{L}^2(\mathbb{R}^+; Q)$ such that $(\mathbf{u}, p)$ is a solution to (9).

Once the reduced problem is solved, there exists a unique way to recover the pressure p, and hence the complete solution of the original Navier-Stokes problem (9).

[12] Salsa, *Partial Differential Equations in Action - From Modelling to Theory*.

[13] A. Quarteroni and A. Valli. *Numerical Approximation of Partial Differential Equations*. Berlin Heidelberg: Springer, 1994.

In practice, however, this approach can be quite unsuitable from a numerical viewpoint. Indeed, in a Galerkin spatial approximation framework, it would require the construction of finite dimensional subspace, say $V_{\mathrm{div},h}$, of *divergence-free* velocity functions. In this regard, see, e.g., ref.[14] for finite element approximations, and ref.[15] for spectral approximations.

Moreover, the result of the above theorem is not constructive, as it does not provide a way to build the solution pressure p. For these reasons one usually prefers to approximate the complete coupled problem (9) directly.

[14] F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods*. New York: Springer-Verlag, 1991.

[15] C. Canuto et al. *Spectral Methods. Fundamentals in Single Domains*. Berlin Heidelberg: Springer-Verlag, 2006.

Stokes equations and their approximation

In this section we will consider the following *generalized Stokes problem* with homogeneous Dirichlet boundary conditions

$$\begin{cases} \sigma\mathbf{u} - \nu\Delta\mathbf{u} + \nabla p = \mathbf{f} & \text{in } \Omega, \\ \mathrm{div}\mathbf{u} = 0 & \text{in } \Omega, \\ \mathbf{u} = \mathbf{0} & \text{on } \partial\Omega, \end{cases} \tag{11}$$

for a given coefficient $\sigma \geq 0$.

This problem describes the motion of an incompressible viscous flow in which the (quadratic) convective term has been neglected, a simplification that is acceptable when $Re \ll 1$.

Moreover, one can generate a problem like (11) also while using an implicit temporal discretization of the Navier-Stokes equations and by neglecting the convective term (i.e. $(\mathbf{u} \cdot \nabla)\mathbf{u}$).
We have indeed the following scheme, where k denotes the temporal index:

$$\begin{cases} \dfrac{\mathbf{u}^k - \mathbf{u}^{k-1}}{\Delta t} - \nu \Delta \mathbf{u}^k + \nabla p^k = \mathbf{f}(t^k), & \mathbf{x} \in \Omega,\ t > 0, \\ \mathrm{div}\mathbf{u}^k = 0, & \mathbf{x} \in \Omega,\ t > 0 \\ + \text{ B.C.} & \end{cases}$$

Hence, at each time step t^k we need to solve the following Stokes-like system of equations:

$$\begin{cases} \sigma \mathbf{u}^k - \nu \Delta \mathbf{u}^k + \nabla p^k = \tilde{\mathbf{f}}^k & \text{in } \Omega, \\ \mathrm{div}\mathbf{u}^k = 0 & \text{in } \Omega \\ + \text{ B.C.} & \end{cases} \qquad (12)$$

where $\sigma = (\Delta t)^{-1}$ and $\tilde{\mathbf{f}}^k = \tilde{\mathbf{f}}(t^k) + \frac{\mathbf{u}^{k-1}}{\Delta t}$.

The weak formulation of problem (11) reads:
find $\mathbf{u} \in V$ and $p \in Q$ such that

$$\begin{cases} \displaystyle\int_\Omega (\sigma \mathbf{u} \cdot \mathbf{v} + \nu \nabla \mathbf{u} \cdot \nabla \mathbf{v}) \ d\Omega - \int_\Omega p \mathrm{div} \mathbf{v} \ d\Omega = \int_\Omega \mathbf{f} \cdot \mathbf{v} \ d\Omega & \forall \mathbf{v} \in V, \\ \displaystyle\int_\Omega q \, \mathrm{div} \mathbf{u} \ d\Omega = 0 & \forall q \in Q, \end{cases} \tag{13}$$

where $V = [\mathrm{H}_0^1(\Omega)]^d$ and $Q = \mathrm{L}_0^2(\Omega)$.

Now define the bilinear forms $a : V \times V \mapsto \mathbb{R}$ and $b : V \times Q \mapsto \mathbb{R}$ as follows:

$$\begin{aligned} a(\mathbf{u}, \mathbf{v}) &= \int_\Omega (\sigma \mathbf{u} \cdot \mathbf{v} + \nu \nabla \mathbf{u} \cdot \nabla \mathbf{v}) \ d\Omega, \\ b(\mathbf{u}, q) &= - \int_\Omega q \mathrm{div} \mathbf{u} \ d\Omega. \end{aligned} \tag{14}$$

With these notations, problem (13) becomes: find $(\mathbf{u}, p) \in V \times Q$ such that

$$\begin{cases} a(\mathbf{u}, \mathbf{v}) + b(\mathbf{v}, p) = (\mathbf{f}, \mathbf{v}) & \forall \mathbf{v} \in V, \\ b(\mathbf{u}, q) = 0 & \forall q \in Q, \end{cases} \tag{15}$$

where $(\mathbf{f}, \mathbf{v}) = \sum_{i=1}^d \int_\Omega f_i v_i \ d\Omega$.

If we consider non-homogeneous boundary conditions, as indicated in (4), the weak formulation of the Stokes problem becomes: find $(\overset{\circ}{\mathbf{u}}, p) \in V \times Q$ such that

$$\begin{cases} a(\overset{\circ}{\mathbf{u}}, \mathbf{v}) + b(\mathbf{v}, p) = \mathbf{F}(\mathbf{v}) & \forall \mathbf{v} \in V, \\ b(\overset{\circ}{\mathbf{u}}, q) = G(q) & \forall q \in Q, \end{cases} \tag{16}$$

where V and Q are the spaces introduced in (7) and (8), respectively. Having denoted with $\mathbf{R}\varphi \in [\mathrm{H}^1(\Omega)]^d$ a lifting of the boundary datum φ, we have set $\overset{\circ}{\mathbf{u}} = \mathbf{u} - \mathbf{R}\varphi$, while the new terms on the right-hand side have the following expression:

$$\mathbf{F}(\mathbf{v}) = (\mathbf{f}, \mathbf{v}) + \int_{\Gamma_N} \psi \mathbf{v}\, d\gamma - a(\mathbf{R}\varphi, \mathbf{v}), \qquad G(q) = -b(\mathbf{R}\varphi, q). \tag{17}$$

Theorem

The couple $(\mathbf{u}, p)$ solves the Stokes problem (15) if and only if it is a saddle point of the Lagrangian functional

$$\mathcal{L}(\mathbf{v}, q) = \frac{1}{2} a(\mathbf{v}, \mathbf{v}) + b(\mathbf{v}, q) - (\mathbf{f}, \mathbf{v}),$$

or equivalently,

$$\mathcal{L}(\mathbf{u}, p) = \min_{\mathbf{v} \in V} \max_{q \in Q} \mathcal{L}(\mathbf{v}, q).$$

The pressure q hence plays the role of Lagrange multiplier associated to the divergence-free constraint.

Indeed, by (formally) taking the Fréchet derivative of the Lagrangian with respect to the two variables, we get (thanks to the symmetry of $a(\cdot, \cdot)$):

$$\begin{aligned} \langle \frac{\partial \mathcal{L}(\mathbf{u}, p)}{\partial \mathbf{u}}, \mathbf{v} \rangle &= a(\mathbf{u}, \mathbf{v}) + b(\mathbf{v}, p) - (\mathbf{f}, \mathbf{v}) = 0 \qquad \forall \mathbf{v} \in V \\ \langle \frac{\partial \mathcal{L}(\mathbf{u}, p)}{\partial p}, q \rangle &= b(\mathbf{u}, q) = 0 \qquad \forall q \in Q \end{aligned}$$

Definition (Fréchet derivative)

Let $F\colon X \to Y$, with X, Y two normed vector spaces. F is differentiable at $x \in X$ if $\exists L_x\colon X \to Y$ linear and bounded such that

$$\forall\, \epsilon > 0\, \exists\, \delta > 0 \quad \|F(x+h) - F(x) - L_x h\|_Y \leq \epsilon \|h\|_X \quad \forall\, h \in X : \|h\|_X < \delta.$$

$F'(x) := L_x$ is called Fréchet derivative of F at the point x.

Galerkin approximation

The Galerkin approximation of problem (15) has the following form: find $(\mathbf{u}_h, p_h) \in V_h \times Q_h$ such that

$$\begin{cases} a(\mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p_h) = (\mathbf{f}, \mathbf{v}_h) & \forall \mathbf{v}_h \in V_h, \\ b(\mathbf{u}_h, q_h) = 0 & \forall q_h \in Q_h, \end{cases} \tag{18}$$

where $\{V_h \subset V\}$ and $\{Q_h \subset Q\}$ represent two families of finite-dimensional subspaces depending on a real discretization parameter h.

If, instead, we consider problem (16)–(17) corresponding to non-homogeneous boundary data (4), the above formulation needs to be modified by using $\mathbf{F}(\mathbf{v}_h)$ on the right-hand side of the first equation and $G(q_h)$ on that of the second equation. These new functionals can be obtained from (17) by replacing $\mathbf{R}\varphi$ with the interpolant of φ at the nodes of Γ_D (and vanishing at all other nodes), and replacing ψ with its interpolant at the nodes sitting on Γ_N.

Existence and Uniqueness

The following celebrated theorem is due to F. Brezzi, and guarantees uniqueness and existence for problem (18):

Theorem

The Galerkin approximation (18) admits one and only one solution if the following conditions hold:

- The bilinear form $a(\cdot,\cdot)$ is:

  a) coercive, that is $\exists\alpha > 0$ (possibly depending on h) such that

  $$a(\mathbf{v}_h,\mathbf{v}_h) \geq \alpha\|\mathbf{v}_h\|_V^2 \quad \forall\mathbf{v}_h \in V_h^*,$$

  where $V_h^* = \{\mathbf{v}_h \in V_h : \; b(\mathbf{v}_h, q_h) = 0 \; \forall q_h \in Q_h\}$;

  b) continuous, that is $\exists\gamma > 0$ such that

  $$|a(\mathbf{u}_h,\mathbf{v}_h)| \leq \gamma\|\mathbf{u}_h\|_V\|\mathbf{v}_h\|_V \quad \forall\mathbf{u}_h,\mathbf{v}_h \in V_h.$$

- The bilinear form $b(\cdot,\cdot)$ is continuous, that is $\exists\delta > 0$ such that

$$|b(\mathbf{v}_h, q_h)| \leq \delta \|\mathbf{v}_h\|_V \|q_h\|_Q \quad \forall \mathbf{v}_h \in V_h, q_h \in Q_h.$$

- Finally, there exists a positive constant β (possibly depending on h) such that

$$\forall q_h \in Q_h,\ \exists \mathbf{v}_h \in V_h :\ b(\mathbf{v}_h, q_h) \geq \beta \|\mathbf{v}_h\|_{\mathbf{H}^1(\Omega)} \|q_h\|_{\mathrm{L}^2(\Omega)}. \qquad (19)$$

Under the previous assumptions the discrete solution fulfills the following a-priori estimates:

$$\|\mathbf{u}_h\|_V \leq \frac{1}{\alpha}\|\mathbf{f}\|_{V'},$$
$$\|p_h\|_Q \leq \frac{1}{\beta}\left(1 + \frac{\gamma}{\alpha}\right)\|\mathbf{f}\|_{V'},$$

where V' is the dual space of V.

Moreover, the following convergence results hold:

$$\begin{aligned}
\|\mathbf{u}-\mathbf{u}_h\|_V &\leq \left(1+\frac{\delta}{\beta}\right)\left(1+\frac{\gamma}{\alpha}\right)\inf_{\mathbf{v}_h\in V_h}\|\mathbf{u}-\mathbf{v}_h\|_V \\
&\quad +\frac{\delta}{\alpha}\inf_{q_h\in Q_h}\|p-q_h\|_Q, \\
\|p-p_h\|_Q &\leq \frac{\gamma}{\beta}\left(1+\frac{\gamma}{\alpha}\right)\left(1+\frac{\delta}{\beta}\right)\inf_{\mathbf{v}_h\in V_h}\|\mathbf{u}-\mathbf{v}_h\|_V \\
&\quad +\left(1+\frac{\delta}{\beta}+\frac{\delta\gamma}{\alpha\beta}\right)\inf_{q_h\in Q_h}\|p-q_h\|_Q.
\end{aligned}$$

It is worth noticing that condition (19) is equivalent to the existence of a positive constant β such that

$$\inf_{q_h\in Q_h, q_h\neq 0}\ \sup_{\mathbf{v}_h\in V_h, \mathbf{v}_h\neq\mathbf{0}} \frac{b(\mathbf{v}_h,q_h)}{\|\mathbf{v}_h\|_{\mathbf{H}^1(\Omega)}\|q_h\|_{\mathrm{L}^2(\Omega)}} \geq \beta. \qquad (20)$$

For such a reason it is often called the inf-sup condition.

The proof of this theorem requires non-elementary tools of functional analysis (see later).

A general saddle-point problem

Let X and M be two Hilbert spaces endowed with norms $\|\cdot\|_X$ and $\|\cdot\|_M$.

Denoting with X' and M' the corresponding dual spaces (that is the spaces of linear and bounded functionals defined on X and M), we introduce the bilinear forms

- $a(\cdot,\cdot) : X \times X \longrightarrow \mathbb{R}$
- $b(\cdot,\cdot) : X \times M \longrightarrow \mathbb{R}$

that we suppose to be continuous, meaning there exist two constants $\gamma, \delta > 0$ such that for all $w, v \in X$ and $\mu \in M$,

$$|a(w,v)| \leq \gamma \, \|w\|_X \, \|v\|_X, \qquad |b(w,\mu)| \leq \delta \, \|w\|_X \, \|\mu\|_M \tag{21}$$

Consider now the following constrained problem: find $(u, \eta) \in X \times M$ such that

$$\begin{cases} a(u, v) + b(v, \eta) = \langle l, v \rangle & \forall v \in X, \\ b(u, \mu) = \langle \sigma, \mu \rangle & \forall \mu \in M, \end{cases} \tag{22}$$

where $l \in X'$ and $\sigma \in M'$ are two assigned linear functionals, while $\langle \cdot, \cdot \rangle$ denotes the pairing between X and X' or M and M'.

Formulation (22) is general enough to include the formulation (15) of the Stokes problem, that of a generic constrained problem with respect to the bilinear form $a(\cdot, \cdot)$ (with η representing the constraint), or again the formulation which is obtained when mixed finite element approximations are used for various kind of boundary-value problems, for instance those of linear elasticity (see, e.g., ref.[16] and ref.[17]).

[16] Brezzi and Fortin, *Mixed and Hybrid Finite Element Methods*.

[17] Quarteroni and Valli, *Numerical Approximation of Partial Differential Equations*.

In order to analyze problem (22), we introduce the affine manifold

$$X^{\sigma} = \{v \in X \ : \ b(v,\mu) = \langle \sigma, \mu \rangle \ \forall \mu \in M\}. \tag{23}$$

The space X^0 denotes the kernel of B, that is

$$X^0 = \{v \in X \ : \ b(v,\mu) = 0 \ \forall \mu \in M\}$$

This is a closed subspace of X. We can therefore associate (22) with the following reduced problem

$$\text{find} \quad u \in X^{\sigma} \quad \text{such that} \quad a(u,v) = \langle l, v \rangle \quad \forall\, v \in X^0. \tag{24}$$

If (u,η) is a solution of (22), then u is a solution to (24). In the following we will introduce suitable conditions that allow the converse to hold, too. Moreover, we would like to prove uniqueness for the solution of (24). This would allow us to obtain an existence and uniqueness result for the original saddle-point problem (22).

Theorem: Existence, uniqueness and stability (continuous case)

Let the bilinear form $a(\cdot,\cdot)$ satisfy the continuity condition (21) and be coercive on the space X^0, that is

$$\exists\alpha > 0 \;:\; a(v,v) \geq \alpha\|v\|_X^2 \quad \forall v \in X^0. \tag{25}$$

Suppose moreover that the bilinear form $b(\cdot,\cdot)$ satisfies the continuity condition (21) as well as the following compatibility condition: there exists $\beta^* > 0$ such that

$$\forall\ \mu \in M \;\; \exists v \in X, \quad v \neq 0 \;:\; b(v,\mu) \geq \beta^*\|v\|_X\|\mu\|_M. \tag{26}$$

Then for every $l \in X'$ and $\sigma \in M'$, there exists a unique solution $(u,\eta) \in X \times M$ to the saddle-point problem (22).

Moreover, the map $(l, \sigma) \longrightarrow (u, \eta)$ is an isomorphism from $X' \times M'$ onto $X \times M$ and the following a priori estimates hold:

$$\begin{aligned} \|u\|_X &\leq \frac{1}{\alpha}\left[\|l\|_{X'} + \frac{\alpha+\gamma}{\beta^*}\|\sigma\|_{M'}\right], \\ \|\eta\|_M &\leq \frac{1}{\beta^*}\left[\left(1+\frac{\gamma}{\alpha}\right)\|l\|_{X'} + \frac{\gamma(\alpha+\gamma)}{\alpha\beta^*}\|\sigma\|_{M'}\right]. \end{aligned} \qquad (27)$$

The symbols $\|\cdot\|_{X'}$ and $\|\cdot\|_{M'}$ indicate the norms of the dual spaces.

All proofs in Chapter 17 of the textbook.

Galerkin approximation, stability and convergence analysis

To introduce a Galerkin approximation of the abstract saddle-point problem (22), we consider two families of finite-dimensional subspaces X_h and M_h of the spaces X and M, respectively. They can be either finite element piecewise polynomial spaces, or global polynomial (spectral) spaces, or spectral element subspaces.

We look for the solution to the following problem:
given $l \in X'$ and $\sigma \in M'$, find $(u_h, \eta_h) \in X_h \times M_h$ such that:

$$\begin{cases} a(u_h, v_h) + b(v_h, \eta_h) = \langle l, v_h \rangle & \forall v_h \in X_h, \\ b(u_h, \mu_h) = \langle \sigma, \mu_h \rangle & \forall \mu_h \in M_h. \end{cases} \tag{28}$$

We introduce the subspace

$$X_h^\sigma = \{v_h \in X_h \ : \ b(v_h, \mu_h) = \langle \sigma, \mu_h \rangle \ \forall \mu_h \in M_h\} \tag{29}$$

which allows us to introduce the following finite dimensional reduced formulation:

$$\text{find} \quad u_h \in X_h^\sigma \quad \text{such that} \quad a(u_h, v_h) = \langle l, v_h \rangle \quad \forall\, v_h \in X_h^0. \tag{30}$$

Since, in general, M_h is different from M, the space (29) is not necessarily a subspace of X^σ.

Clearly, every solution (u_h, η_h) of (28) yields a solution u_h for the reduced problem (30). In the following we look for conditions that allow us to prove that the converse statement is also true, together with a result of stability and convergence for the solution of problem (28).

Theorem: Existence, uniqueness and stability (discrete case)

Suppose that the bilinear form $a(\cdot,\cdot)$ satisfies the continuity property (21) and that it is coercive on the space X_h^0, that is there exists a constant $\alpha_h > 0$ such that

$$a(v_h, v_h) \geq \alpha_h \|v_h\|_X^2 \quad \forall v_h \in X_h^0. \tag{31}$$

Moreover, suppose that the bilinear form $b(\cdot,\cdot)$ satisfies the continuity condition (21) and that the following discrete compatibility condition holds: there exists a constant $\beta_h > 0$ such that

$$\forall\ \mu_h \in M_h \ \ \exists v_h \in X_h, \quad v_h \neq 0 \ \ : \ \ b(v_h, \mu_h) \geq \beta_h \|v_h\|_X \|\mu_h\|_M. \tag{32}$$

Then, for every $l \in X'$ and $\sigma \in M'$, there exists a unique solution (u_h, η_h) of problem (28).

Moreover, the solution satisfies the following stability conditions:

$$\|u_h\|_X \leq \frac{1}{\alpha_h}\left[\|l\|_{X'} + \frac{\alpha_h+\gamma}{\beta_h}\,\|\sigma\|_{M'}\right], \tag{33}$$

$$\|\eta_h\|_M \leq \frac{1}{\beta_h}\left[\left(1+\frac{\gamma}{\alpha_h}\right)\|l\|_{X'} + \frac{\gamma(\alpha_h+\gamma)}{\alpha_h\,\beta_h}\,\|\sigma\|_{M'}\right]. \tag{34}$$

The coercivity condition (25) does not necessarily guarantee (31), as $X_h^0 \not\subset X^0$, nor does the compatibility condition (26) in general imply the discrete compatibility condition (32), due to the fact that X_h is a proper subspace of X.

Condition (32) represents the well known inf-sup or LBB condition (see ref.[18]). (The condition (19) (or (20)) is just a special case.)

[18] Brezzi and Fortin, *Mixed and Hybrid Finite Element Methods*.

Theorem: Convergence

Let the assumptions of the existence and uniqueness Theorems (continuous and discrete cases) be satisfied. Then the solutions (u, η) and (u_h, η_h) of problems (22) and (28), respectively, satisfy the following error estimates:

$$\|u - u_h\|_X \leq \left(1 + \frac{\gamma}{\alpha_h}\right) \inf_{v_h^* \in X_h^\sigma} \|u - v_h^*\|_X + \frac{\delta}{\alpha_h} \inf_{\mu_h \in M_h} \|\eta - \mu_h\|_M, \quad (35)$$

$$\begin{aligned} \|\eta - \eta_h\|_M \quad &\leq \frac{\gamma}{\beta_h}\left(1 + \frac{\gamma}{\alpha_h}\right) \inf_{v_h^* \in X_h^\sigma} \|u - v_h^*\|_X \\ &+ \left(1 + \frac{\delta}{\beta_h} + \frac{\gamma\delta}{\alpha_h \beta_h}\right) \inf_{\mu_h \in M_h} \|\eta - \mu_h\|_M, \end{aligned} \quad (36)$$

where γ, δ, α_h and β_h are respectively defined by the relations (21), (31) and (32). Moreover, the following error estimate holds

$$\inf_{v_h^* \in X_h^\sigma} \|u - v_h^*\|_X \leq \left(1 + \frac{\delta}{\beta_h}\right) \inf_{v_h \in X_h} \|u - v_h\|_X. \quad (37)$$

Proof. Consider $v_h \in X_h$, $v_h^* \in X_h^\sigma$ and $\mu_h \in M_h$. By (28), we have:

$$a(u_h, v_h) + b(v_h, \eta_h) = \langle l, v_h \rangle. \tag{38}$$

Moreover, being $X_h \subset X$, by (22) we have:

$$a(u, v_h) + b(v_h, \eta) = \langle l, v_h \rangle. \tag{39}$$

By comparing (38) with (39), we get:

$$a(u_h, v_h) + b(v_h, \eta_h) = a(u, v_h) + b(v_h, \eta) \tag{40}$$

Subtracting the quantities $a(v_h^*, v_h)$ and $b(v_h, \mu_h)$ to both sides of the equality, we find

$$a(u_h - v_h^*, v_h) + b(v_h, \eta_h - \mu_h) = a(u - v_h^*, v_h) + b(v_h, \eta - \mu_h).$$

Let us now choose $v_h = u_h - v_h^* \in X_h^0$. By using the continuity of $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ (see (21)):

$$|a(w,v)| \leq \gamma \, \|w\|_X \, \|v\|_X, \quad |b(w,\mu)| \leq \delta \, \|w\|_X \, \|\mu\|_M$$

and the discrete coercivity coercivity of $a(\cdot,\cdot)$ (see (31)):

$$a(v_h, v_h) \geq \alpha_h \|v_h\|_X^2 \quad \forall v_h \in X_h^0,$$

from the definition of the space X_h^0 we find the bound

$$\|u_h - v_h^*\|_X \leq \frac{1}{\alpha_h}\Big(\gamma \, \|u - v_h^*\|_X + \delta \, \|\eta - \mu_h\|_M\Big)$$

from which the estimate (35) immediately follows, as

$$\|u - u_h\|_X \leq \|u - v_h^*\|_X + \|u_h - v_h^*\|_X.$$

Let us prove now the estimate (36). Owing to the compatibility condition (32), for every $\mu_h \in M_h$ we can write

$$\|\eta_h - \mu_h\|_M \leq \frac{1}{\beta_h} \sup_{v_h \in X_h,\ v_h \neq 0} \frac{b(v_h, \eta_h - \mu_h)}{\|v_h\|_X}. \tag{41}$$

Similarly to what done before, by (28), we have:

$$a(u_h, v_h) + b(v_h, \eta_h) = \langle l, v_h \rangle. \tag{42}$$

Moreover, being $X_h \subset X$, by (22) we have:

$$a(u, v_h) + b(v_h, \eta) = \langle l, v_h \rangle. \tag{43}$$

By subtracting side by side (42) and (43), we get:

$$b(v_h, \eta_h) = a(u - u_h, v_h) + b(v_h, \eta). \tag{44}$$

Then, by subtracting the quantity $b(v_h, \mu_h)$ from both sides, we obtain

$$b(v_h, \eta_h - \mu_h) = a(u - u_h, v_h) + b(v_h, \eta - \mu_h).$$

Using this identity in (41) as well as the continuity inequalities (21), that we recall here:

$$|a(w,v)| \leq \gamma \, \|w\|_X \, \|v\|_X, \qquad |b(w,\mu)| \leq \delta \, \|w\|_X \, \|\mu\|_M, \tag{45}$$

it follows that

$$\|\eta_h - \mu_h\|_M \leq \frac{1}{\beta_h} \Big(\gamma \, \|u - u_h\|_X + \delta \|\eta - \mu_h\|_M\Big).$$

This yields the desired result, provided we use the error estimate (35) that was previously derived for the variable u.

Finally, let us prove (37). Using (32), it is possible to prove (proof omitted) that for every $v_h \in X_h$ we can find a unique function $z_h \in (X_h^0)^\perp$ such that

$$b(z_h, \mu_h) = b(u - v_h, \mu_h) \quad \forall \mu_h \in M_h$$

and, moreover,

$$\|z_h\|_X \leq \frac{\delta}{\beta_h} \|u - v_h\|_X.$$

The function $v_h^* = z_h + v_h$ belongs to X_h^σ, as $b(u, \mu_h) = \langle \sigma, \mu_h \rangle$ for all $\mu_h \in M_h$. Moreover,

$$\|u - v_h^*\|_X \leq \|u - v_h\|_X + \|z_h\|_X \leq \left(1 + \frac{\delta}{\beta_h}\right) \|u - v_h\|_X,$$

whence the estimate (37) follows.

□

The inequalities (35) and (36) yield error estimates with optimal convergence rate, provided that the constants α_h and β_h in (31) and (32) are bounded from below by two constants α and β independent of h.

Let us also remark that inequality (35) holds even if the compatibility conditions (26) and (32) are not satisfied.

Remark: spurious pressure modes

The compatibility condition (32) is essential to guarantee the uniqueness of the η_h-component of the solution. Indeed, if (32) does not hold, then

$$\exists\, \mu_h^* \in M_h, \mu_h^* \neq 0 \text{ s.t. } b(v_h, \mu_h^*) = 0 \quad \forall v_h \in X_h.$$

Consequently, if (u_h, η_h) is a solution to problem (28), then $(u_h, \eta_h + \tau\mu_h^*)$, for all $\tau \in \mathbb{R}$, is a solution, too.

Any such function μ_h^* is called spurious mode, or, more specifically, pressure spurious mode when it refers to the Stokes problem (18) in which functions μ_h represent discrete pressures. Numerical instabilities can arise since the discrete problem (28) is unable to detect such spurious modes.

For a given couple of finite dimensional spaces X_h and M_h, proving that the discrete compatibility condition (32) holds with a constant β_h independent of h is not always easy.

Several practical criteria are available, among which we mention those due to Fortin (ref.[19]), Boland and Nicolaides (ref.[20]), and Verfürth (ref.[21]). (See ref.[22].)

[19] M. Fortin. "An Analysis of the Convergence of Mixed Finite Element Methods". In: *R.A.I.R.O. Anal. Numér.* 11 (1977).

[20] J. Boland and R.A. Nicolaides. "Stability of Finite Elements under Divergence Constraints". In: *SIAM J. Numer. Anal.* 20 (1983), pp. 722–731.

[21] R. Verfürth. "Error Estimates for a Mixed Finite Elements Approximation of the Stokes Equations". In: *R.A.I.R.O. Anal. Numér.* 18 (1984), pp. 175–182.

[22] F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods*. Vol. 15. Springer Series in Computational Mathematics. New York: Springer-Verlag, 1991.

Algebraic formulation of the Stokes problem

Let us investigate the structure of the algebraic system associated to the Galerkin approximation (18) to the Stokes problem (or, more generally, to a discrete saddle-point problem like (28)). Denote with

$$\{\varphi_j \in V_h\}, \quad \{\phi_k \in Q_h\},$$

the basis functions of the spaces V_h and Q_h, respectively. Let us expand the discrete solutions $\mathbf{u}_h$ and p_h with respect to such bases,

$$\mathbf{u}_h(\mathbf{x}) = \sum_{j=1}^{N} u_j \varphi_j(\mathbf{x}), \quad p_h(\mathbf{x}) = \sum_{k=1}^{M} p_k \phi_k(\mathbf{x}), \tag{46}$$

having set $N = \dim V_h$ and $M = \dim Q_h$.

By choosing as test functions in (18) the same basis functions we obtain the following block linear system

$$\begin{cases} \mathrm{A}\mathbf{U} + \mathrm{B}^T\mathbf{P} = \mathbf{F}, \\ \mathrm{B}\mathbf{U} = \mathbf{0}, \end{cases} \tag{47}$$

where $\mathrm{A} \in \mathbb{R}^{N\times N}$ and $\mathrm{B} \in \mathbb{R}^{M\times N}$ are the matrices related respectively to the bilinear forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$, whose elements are given by

$$\mathrm{A} = [a_{ij}] = [a(\varphi_j, \varphi_i)], \qquad \mathrm{B} = [b_{km}] = [b(\varphi_m, \phi_k)],$$

while $\mathbf{U}$ and $\mathbf{P}$ are the vectors of the unknowns,

$$\mathbf{U} = [u_j], \quad \mathbf{P} = [p_j].$$

The $(N+M)\times(N+M)$ matrix

$$\mathrm{S} = \begin{bmatrix} \mathrm{A} & \mathrm{B}^{T} \\ \mathrm{B} & 0 \end{bmatrix} \tag{48}$$

is block symmetric (as A is symmetric) and indefinite, featuring real eigenvalues with variable sign (either positive and negative).

S is non-singular iff no eigenvalue is null, a property that follows from the inf-sup condition (20).

To prove the latter statement we proceed as follows.

Since A is non-singular – it is associated to the coercive bilinear form $a(\cdot,\cdot)$ – from the first of (47) we can formally obtain **U** as

$$\mathbf{U} = \mathrm{A}^{-1}(\mathbf{F} - \mathrm{B}^{T}\mathbf{P}). \tag{49}$$

Using (49) in the second equation of (47) yields

$$\mathrm{R}\mathbf{P} = \mathrm{B}\mathrm{A}^{-1}\mathbf{F}, \quad \text{where} \quad \mathrm{R} = \mathrm{B}\mathrm{A}^{-1}\mathrm{B}^{T}. \tag{50}$$

This corresponds to having carried out a block Gaussian elimination on system (48).

In this way we obtain a reduced system for the sole unknown **P** (the pressure), which admits a unique solution in case R is non-singular. Since A is non-singular and positive definite, the latter condition is satisfied iff B^T has a null kernel, that is

$$\mathrm{ker}\mathrm{B}^T = \{\mathbf{0}\}, \tag{51}$$

where $\mathrm{ker}\mathrm{B}^T = \{\mathbf{x} \in \mathbb{R}^M : \ B^T\mathbf{x} = \mathbf{0}\}$.
As a matter of fact:
$R\mathbf{p} = \mathbf{0} \implies \mathbf{p} = \mathbf{0}$ (uniqueness), that is
$\langle \mathrm{B}\mathrm{A}^{-1}\mathrm{B}^T\mathbf{p}, \mathbf{q}\rangle = \mathbf{0} \ \forall \mathbf{q} \implies \mathbf{p} = \mathbf{0}$
Let us take $\mathbf{q} = \mathbf{p}$. We require: $\langle \mathrm{A}^{-1}\mathrm{B}^T\mathbf{p}, \mathrm{B}^T\mathbf{p}\rangle = \mathbf{0} \implies \mathbf{p} = \mathbf{0}$
Since A spd, we have $\langle \mathrm{A}^{-1}\mathbf{w}, \mathbf{w}\rangle = \mathbf{0} \implies \mathbf{w} = \mathbf{0}$
Finally, $(\mathrm{B}^T\mathbf{p} = \mathbf{0} \implies \mathbf{p} = \mathbf{0}) \iff \mathrm{ker}\mathrm{B}^T = \{\mathbf{0}\}$

Remark

Condition (51) is equivalent to the *inf-sup* condition (20).

Proof: note that condition (51) is violated iff $\exists \mathbf{p}^* \neq \mathbf{0}$ with $\mathbf{p}^* \in \mathbb{R}^M$ such that $\mathrm{B}^T \mathbf{p}^* = \mathbf{0}$ or, equivalently, if $\exists p_h^* \in \mathbb{Q}_h$ such that $b(\varphi_n, p_h^*) = 0$ $\forall n = 1, \dots, N$. This is equivalent to $b(\mathbf{v}_h, p_h^*) = 0 \forall \mathbf{v}_h \in V_h$, which in turn is equivalent to violating (20).
Indeed:

$$\exists\, \beta_h > 0 \,\forall\, \mathbf{q}_h \in Q_h \,\exists\, \mathbf{v}_h in X_h \colon \frac{b(\mathbf{v}_h, \mathbf{q}_h)}{\|\mathbf{v}_h\|_V \|\mathbf{q}_h\|_Q} \geq \beta_h$$

is violated if

$$\forall\, \beta_h > 0 \,\exists\, \mathbf{p}_h^* \in Q_h \,\forall\, \mathbf{v}_h in X_h \colon \frac{b(\mathbf{v}_h, \mathbf{p}_h^*)}{\|\mathbf{v}_h\|_V \|\mathbf{p}_h^*\|_Q} < \beta_h$$

Remark (continued)

Take $-\mathbf{v}_h$:

$$\frac{b(-\mathbf{v}_h, \mathbf{p}_h^*)}{\|\mathbf{v}_h\|_V \|\mathbf{p}_h^*\|_Q} = -\frac{b(\mathbf{v}_h, \mathbf{p}_h^*)}{\|\mathbf{v}_h\|_V \|\mathbf{p}_h^*\|_Q} < \beta_h$$

Then

$$-\beta_h < \frac{b(\mathbf{v}_h, \mathbf{p}_h^*)}{\|\mathbf{v}_h\|_V \|\mathbf{p}_h^*\|_Q} < \beta_h$$

Because of the arbitrariness of β_h we conclude that $b(\mathbf{v}_h, \mathbf{p}_h^*) = 0 \, \forall \, \mathbf{v}_h \in X_h$

On the other hand, since A is non-singular, from the existence and uniqueness of **P** we infer that there exists a unique vector **U** which satisfies (49).

In conclusion, system (47) admits a unique solution $(\mathbf{U}, \mathbf{P})$ if and only if condition (51) holds.

Remark

Condition (51) is equivalent to asking that B^T (and consequently B) has full rank, i.e. that $\mathrm{rank}(B^T) = \min(N, M)$, because $\mathrm{rank}(B^T)$ is the maximum number of linearly independent row vectors (or, equivalently, column vectors) of B^T. Indeed, $\mathrm{rank}(B^T) + \dim\ \ker(B^T) = M$.

Let us consider again the Remark about spurious pressure modes concerning the general saddle-point problem and suppose that the *inf-sup* condition (20) does not hold. Then

$$\exists q_h^* \in Q_h : \quad b(\mathbf{v}_h, q_h^*) = 0 \qquad \forall \mathbf{v}_h \in V_h. \tag{52}$$

Consequently, if $(\mathbf{u}_h, p_h)$ is a solution to the Stokes problem (18), then $(\mathbf{u}_h, p_h + q_h^*)$ is a solution too, as

$$\begin{aligned} a(\mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p_h + q_h^*) &= a(\mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p_h) + b(\mathbf{v}_h, q_h^*) \\ &= a(\mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p_h) = (\mathbf{f}, \mathbf{v}_h) \qquad \forall \mathbf{v}_h \in V_h. \end{aligned}$$

Functions q_h^* which fail to satisfy the *inf-sup* condition are invisible to the Galerkin problem(18). For this reason, as already observed, they are called spurious pressure modes, or even parasitic modes. Their presence inhibits the pressure solution from being unique, yielding numerical instabilities. For this reason, those finite-dimensional subspaces that violate the compatibility condition (20) are said to be unstable, or incompatible.

Two strategies are generally adopted in order to guarantee well-posedness of the numerical problem:

- choose spaces V_h and Q_h that satisfy the *inf-sup* condition;
- stabilize (either a priori or a posteriori) the finite dimensional problem by eliminating the spurious modes.

Compatible couples of spaces

Let us analyze the first type of strategy. To start with, we will consider the case of finite element spaces. To characterize Q_h and V_h it suffices to choose on every element of the triangulation their degrees of freedom. Since the weak formulation does not require a continuous pressure, we will consider first the case of discontinuous pressures.

As Stokes equations are of order one in p and order two in $\mathbf{u}$, generally speaking it makes sense to use piecewise polynomials of degree $k \geq 1$ for the velocity space V_h and of degree $k-1$ for the space Q_h.

In particular, we might want to use piecewise linear finite elements $\mathbb{P}_1$ for each velocity component, and piecewise constant finite elements $\mathbb{P}_0$ for the pressure. In fact, this choice, although being quite natural, does not pass the *inf-sup* test (20).

When looking for a compatible couple of spaces, the larger the velocity space V_h, the more likely the *inf-sup* condition is satisfied. Otherwise said, the space V_h should be "rich" enough compared to the space Q_h.

In the following pictures, by means of the symbol □ we indicate the degrees of freedom for the pressure, whereas the symbol • identifies those for each velocity component

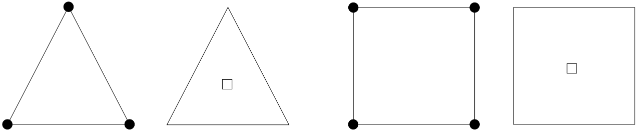

Figure: Case of discontinuous pressure: choices that do not satisfy the *inf-sup* condition, on triangles (left), and on quadrilaterals (right)

In this figure we report three different choices of spaces that fulfill the *inf-sup* condition, still in the case of continuous velocity and discontinuous pressure. Choice (left) is made by $\mathbb{P}_2 - \mathbb{P}_0$ elements, (center) by $\mathbb{Q}_2 - \mathbb{P}_0$ elements, while choice (right) by piecewise linear discontinuous elements for the pressure, while the velocity components are made by piecewise quadratic continuous elements enriched by a cubic bubble function on each triangle – these are the so-called Crouzeix-Raviart elements.

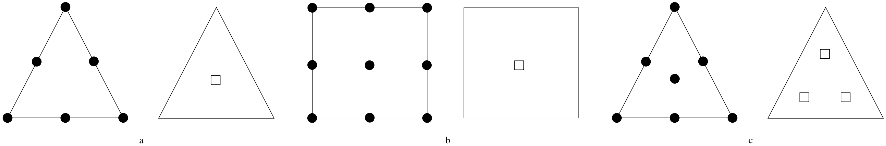

Figure: Case of discontinuous pressure: choices that do satisfy the *inf-sup* condition: on triangles, (left), and on quadrilaterals, (center). Also the couple (right), known as Crouzeix-Raviart elements, satisfies the *inf-sup* condition

In this figure we report two choices of incompatible finite elements in the case of continuous pressure. They consist of piecewise linear elements on triangles (resp. bilinear on quadrilaterals) for both velocity and pressure. More in general, finite elements of the same polynomial degree $k \geq 1$ for both velocity and pressures are unstable (equal order interpolation).

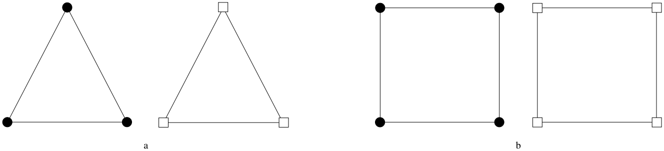

Figure: Case of continuous pressure: the couples displayed in this figure do not satisfy the *inf-sup* condition.

In this figure, the elements displayed are instead stable. In both cases, pressure is a piecewise linear continuous function, whereas velocities are piecewise linear polynomials on each of the four sub-triangles (left), or piecewise linear polynomials enriched by a cubic bubble function (right).

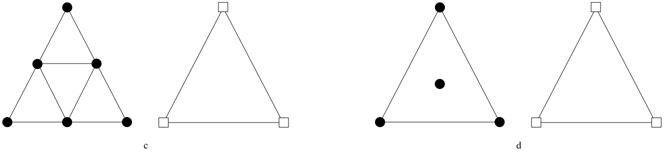

Figure: Case of continuous pressure: the elements used for the velocity components in (left) are known as $\mathbb{P}_1$-*iso*$\mathbb{P}_2$ finite elements, whereas couple (right) is called *mini-element*

The pair $\mathbb{P}_2 - \mathbb{P}_1$ (continuous piecewise quadratic velocities and continuous piecewise linear pressure) is stable. This is the smallest degree representative of the family of the so-called Taylor-Hood elements $\mathbb{P}_k - \mathbb{P}_{k-1}$, $k \geq 2$ (continuous velocities and continuous pressure), that are *inf-sup* stable.

For the proof of the stability results mentioned here, as well for the convergence analysis, we refer to ref.[23].

[23] Brezzi and Fortin, *Mixed and Hybrid Finite Element Methods*.

The case of Spectral Methods

If we use spectral methods, using equal-order polynomial spaces for both velocity and pressure yields subspaces that violate the *inf-sup* condition. Compatible spectral spaces can instead be obtained by using, e.g., polynomials of degree N (≥ 2) for each velocity component, and degree $N-2$ for the pressure, yielding the so-called $\mathbb{Q}_N - \mathbb{Q}_{N-2}$ approximation. The degrees of freedom for each velocity component are represented by the $(N+1)^2$ GLL node (see next figure).

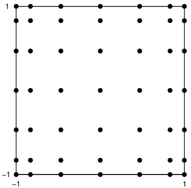

Figure: The $(N+1)^2$ Gauss-Legendre-Lobatto (GLL) nodes (here $N=6$), hosting the degrees of freedom of the velocity components

The case of Spectral Methods

For the pressure, at least two sets of interpolation nodes can be used: either the subset represented by the $(N-1)^2$ internal nodes of the set of $(N+1)^2$ GLL nodes (next figure, left), or the $(N-1)^2$ Gauss nodes (next figure, right). This choice stands at the base of a spectral-type approximation, such as collocation, G-NI (Galerkin with numerical integration), or SEM-NI (spectral element with numerical integration) (see ref.[24]).

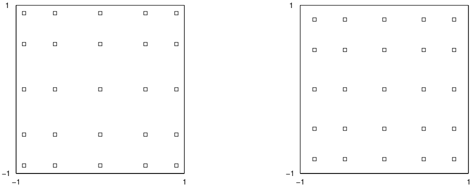

Figure: The $(N-1)^2$ internal Gauss-Legendre-Lobatto (GLL) nodes (left) and the $(N-1)^2$ Gauss-Legendre (GL) nodes (right) (here for $N=6$), hosting the degrees of freedom of the pressure

[24] C. Canuto et al. *Spectral Methods. Evolution to Complex Geometries and*

An example of stabilized problem

We have seen that finite element or spectral methods that make use of equal-degree polynomials for both velocity and pressure do not fulfill the *inf-sup* condition and are therefore unstable.

However, stabilizing them is possible by SUPG or GLS techniques like those encountered in the framework of the numerical approximation of advection-diffusion equations.

For a general discussion on stabilization techniques for Stokes equations, the reader can refer e.g. to ref.[25].

[25] Brezzi and Fortin, *Mixed and Hybrid Finite Element Methods*.

Here we limit ourselves to show how the GLS stabilization can be applied to problem (18) in case piecewise continuous linear finite elements (P1-P1, both continuous) are used for velocity components as well as for the pressure

$$V_h = [\overset{\circ}{X}{}^1_h]^2, \quad Q_h = \{q_h \in X_h^1 : \ \int_\Omega q_h \, d\Omega = 0\}.$$

This choice is urged by the need of keeping the global number of degrees of freedom as low as possible, especially when dealing with three-dimensional problems. However, it will be unstable (as it violates the inf-sup condition) for the pure Galerkin method without stabilization.

We set therefore $W_h = V_h \times Q_h$ and, instead of (18), consider the following problem (we restrict ourselves to the case where $\alpha = 0$):

$$\text{find } (\mathbf{u}_h, p_h) \in W_h : \; A_h(\mathbf{u}_h, p_h; \mathbf{v}_h, q_h) = F_h(\mathbf{v}_h, q_h) \quad \forall (\mathbf{v}_h, q_h) \in W_h. \tag{53}$$

We have set

$$\begin{aligned}
&A_h : W_h \times W_h \to \mathbb{R}, \\
&A_h(\mathbf{u}_h, p_h; \mathbf{v}_h, q_h) = a(\mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p_h) - b(\mathbf{u}_h, q_h) \\
&\qquad + \delta \sum_{K \in \mathcal{T}_h} h_K^2 \int_K (-\nu \Delta \mathbf{u}_h + \nabla p_h)(-\nu \Delta \mathbf{v}_h + \nabla q_h) \, dK,
\end{aligned}$$

$$\begin{aligned}
&F_h : W_h \to \mathbb{R}, \\
&F_h(\mathbf{v}_h, q_h) = (\mathbf{f}, \mathbf{v}_h) \\
&\qquad + \delta \sum_{K \in \mathcal{T}_h} h_K^2 \int_K \mathbf{f}(-\nu \Delta \mathbf{v}_h + \nabla q_h) \, dK
\end{aligned}$$

and we have denoted with δ a positive parameter that must be chosen conveniently.

This is a strongly consistent approximation of problem (11): as a matter of fact, the additional term, which depends on the residual of the discrete momentum equation, is null when calculated on the exact solution as, thanks to (13), $-\nu\Delta\mathbf{u} + \nabla p - \mathbf{f} = \mathbf{0}$.

Note that, since $k = 1$, $\Delta\mathbf{u}_{h|K} = \Delta\mathbf{v}_{h|K} = \mathbf{0}$ $\forall K \in \mathcal{T}_h$ as we are using piecewise linear finite element functions.

From the identity

$$\begin{aligned} A_h(\mathbf{u}_h, p_h; \mathbf{u}_h, p_h) &= \nu \|\nabla \mathbf{u}_h\|^2_{\mathbf{L}^2(\Omega)} + \delta \sum_{k \in \mathcal{T}_h} h_K^2 \|\nabla p_h\|^2_{\mathbf{L}^2(K)} \\ F_h(\mathbf{u}_h, p_h) &= (\mathbf{f}, \mathbf{u}_h) + \delta \sum_{K \in \mathcal{T}_h} h_K^2 \int_K \mathbf{f} \cdot \nabla p_h \; dK, \end{aligned} \tag{54}$$

We have

$$\begin{aligned} |F_h(\mathbf{u}_h, p_h)| &\leq C_p \|\mathbf{f}\|_{L^2(\Omega)} \|\nabla \mathbf{u}_h\|_{L^2(\Omega)} + \delta \sum_{K \in \mathcal{T}_h} h_K^2 \|\mathbf{f}\|_{L^2(K)} \|\nabla p_h\|_{L^2(K)} \\ &\leq \frac{C_p^2}{2\nu} \|\mathbf{f}\|^2_{L^2(\Omega)} + \frac{\nu}{2} \|\nabla \mathbf{u}_h\|^2_{L^2(\Omega)} \\ &+ \frac{\delta}{2} \sum_{K \in \mathcal{T}_h} h_K^2 \|\mathbf{f}\|^2_{L^2(K)} + \frac{\delta}{2} \sum_{K \in \mathcal{T}_h} h_K^2 \|\nabla p_h\|^2_{L^2(K)} \end{aligned}$$

We obtain the following stability inequality

$$\nu\|\nabla \mathbf{u}_h\|^2_{\mathbf{L}^2(\Omega)} + \delta \sum_{K\in\mathcal{T}_h} h_K^2 \|\nabla p_h\|^2_{\mathbf{L}^2(K)} \leq C\|\mathbf{f}\|^2_{\mathbf{L}^2(\Omega)}, \tag{55}$$

C being a constant that depends on ν but not on h.
By applying Strang's Lemma we can now show that the solution to the generalized Galerkin problem (53) satisfies the following error estimate

$$\|\mathbf{u} - \mathbf{u}_h\|_{\mathbf{H}^1(\Omega)} + \left(\delta \sum_{K\in\mathcal{T}_h} h_K^2 \|\nabla p - \nabla p_h\|^2_{\mathbf{L}^2(K)}\right)^{1/2} \leq Ch.$$

We can show that (53) admits the following matrix form

$$\begin{bmatrix} \mathrm{A} & \mathrm{B}^T \\ \mathrm{B} & -\mathrm{C} \end{bmatrix} \begin{bmatrix} \mathbf{U} \\ \mathbf{P} \end{bmatrix} = \begin{bmatrix} \mathbf{F} \\ \mathbf{G} \end{bmatrix}. \qquad (56)$$

This system differs from (47) without stabilization because of the presence of the non-null block occupying the position (2,2), which is associated to the stabilization term. More precisely,

$$C = (c_{km}) \quad , \quad c_{km} = \delta \sum_{K \in \mathcal{T}_h} h_K^2 \int_K \nabla \phi_m \cdot \nabla \phi_k \; dK, \qquad k, m = 1, \dots, M,$$

while the components of the right-hand side **G** are

$$g_k = -\delta \sum_{K \in \mathcal{T}_h} h_K^2 \int_K \mathbf{f} \cdot \nabla \phi_k \; dK, \qquad k = 1, \dots, M.$$

In this case, the reduced system for the pressure unknown reads

$$\mathrm{R}\mathbf{P} = \mathrm{B}\mathrm{A}^{-1}\mathbf{F} - \mathbf{G}.$$

In contrast to (50), this time $\mathrm{R} = \mathrm{B}\mathrm{A}^{-1}\mathrm{B}^{T} + \mathrm{C}$. The matrix R is non-singular as C is a positive definite matrix (even though B^{T} is not full rank).

A numerical example

We want to solve the stationary Navier-Stokes equations in the square domain $\Omega = (0,1) \times (0,1)$ with the following Dirichlet conditions

$$\begin{aligned} \mathbf{u} &= \mathbf{0}, && \mathbf{x} \in \partial\Omega \backslash \Gamma, \\ \mathbf{u} &= (1,0)^T, && \mathbf{x} \in \Gamma, \end{aligned} \tag{57}$$

where $\Gamma = \{\mathbf{x} = (x_1, x_2)^T \in \partial\Omega : x_2 = 1\}$. This problem is known as flow in a lid-driven cavity. We will use continuous piecewise bilinear $\mathbb{Q}_1 - \mathbb{Q}_1$ polynomials on rectangular finite elements. As we know, these spaces do not fulfill the compatibility condition (inf-sup condition).

In this figure, left, we display the spurious pressure modes that are generated by this Galerkin approximation. In the same figure, right, we have drawn the pressure isolines obtained using a GLS stabilization (addressed in the previous section) on the same kind of finite elements. The pressure is now free of numerical oscillations.

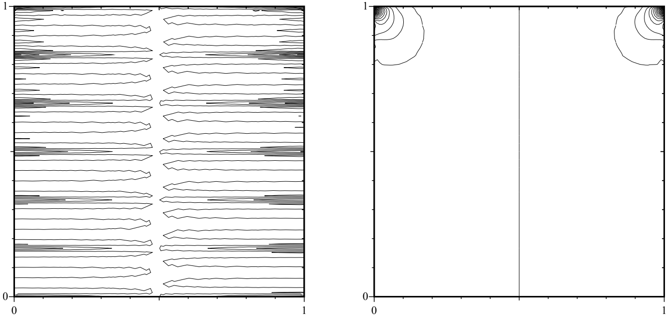

Figure: Pressure isolines for the numerical approximation of the lid-driven cavity problem. Stabilized GLS approximation (on the right); the vertical line corresponds to the null value of the pressure. Non-stabilized approximation (on the left); the presence of a spurious numerical pressure is evident

Still for the stabilized problem, in this figure we display the streamlines for two different values of the Reynolds number, $Re = 1000$ and $Re = 5000$. The stabilization term amends simultaneously pressure instabilities (by getting rid of the spurious modes) and potential instabilities of the pure Galerkin method that develop when diffusion is dominated by convection, an issue that we have extensively addressed in for ADR problems.

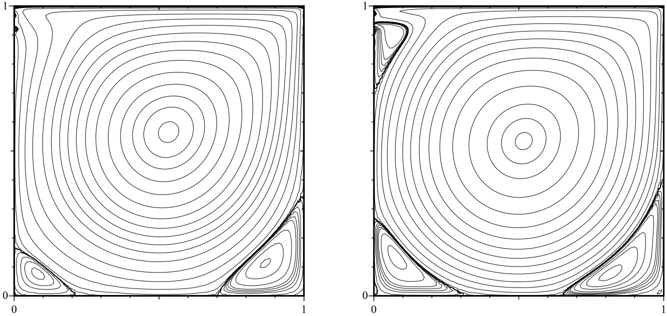

Figure: Streamlines of the numerical solution of the lid-driven cavity problem corresponding to two different values of the Reynolds number: $Re = 1000$, left, and $Re = 5000$, right

For the same problem we consider, as well, a spectral G-NI approximation in which the pressure and each velocity component are polynomials of $\mathbb{Q}_N$ (with $N = 32$). As previously observed, this choice of spaces does not fulfill the *inf-sup* condition, and so it generates spurious pressure modes that are clearly visible in the next figure figure, left.

A GLS stabilization, similar to that previously used for finite elements, can be set up for the G-NI method, too.

The corresponding solution is now stable and free of spurious pressure modes, as the pressure isolines displayed on the right hand of the same figure show.

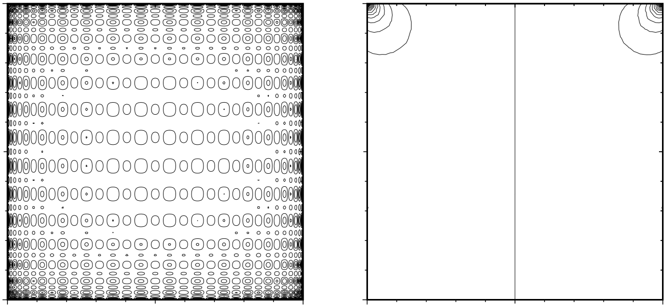

Figure: Pressure isolines obtained by the pure spectral G-NI method (on the left), and by the GLS stabilized spectral G-NI method (on the right). In either case, polynomials of the same degree, $N = 32$, are used for both pressure and velocity. As expected, the pure G-NI method yields spurious pressure solutions. The test case is the same lid-driven cavity problem previously approximated by bilinear finite elements

Time discretization of Navier-Stokes equations

We consider the following semidiscretized formulation:

$$\begin{cases} \mathrm{M}\dfrac{d\mathbf{u}(t)}{dt} + \mathrm{A}\mathbf{u}(t) + \mathrm{C}(\mathbf{u}(t))\mathbf{u}(t) + \mathrm{B}^T\mathbf{p}(t) = \mathbf{f}(t), \\ \mathrm{B}\mathbf{u}(t) = \mathbf{0}, \end{cases} \tag{58}$$

with $\mathbf{u}(0) = \mathbf{u}_0$. $\mathrm{C}(\mathbf{u}(t))$ is in fact a matrix depending on $\mathbf{u}(t)$, whose generic coefficient is $c_{mi}(t) = c(\mathbf{u}(t), \varphi_i, \varphi_m)$.

For the temporal discretization of this system let us use, for instance, the θ-method, that was introduced for parabolic equations. By setting

$$
\begin{aligned}
&\mathbf{u}_\theta^{n+1} = \theta \mathbf{u}^{n+1} + (1-\theta)\mathbf{u}^n,\\
&\mathbf{p}_\theta^{n+1} = \theta \mathbf{p}^{n+1} + (1-\theta)\mathbf{p}^n,\\
&\mathbf{f}_\theta^{n+1} = \theta \mathbf{f}(t^{n+1}) + (1-\theta)\mathbf{f}(t^n),\\
&\mathrm{C}_\theta(\mathbf{u}^{n+1,n})\mathbf{u}^{n+1,n} = \theta \mathrm{C}(\mathbf{u}^{n+1})\mathbf{u}^{n+1} + (1-\theta)\mathrm{C}(\mathbf{u}^n)\mathbf{u}^n,
\end{aligned}
$$

we obtain the following system of algebraic equations

$$
\begin{cases}
\mathrm{M}\dfrac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\Delta t} + \mathrm{A}\mathbf{u}_\theta^{n+1} + \mathrm{C}_\theta(\mathbf{u}^{n+1,n})\mathbf{u}^{n+1,n} + \mathrm{B}^T\mathbf{p}_\theta^{n+1} = \mathbf{f}_\theta^{n+1},\\
\mathrm{B}\mathbf{u}^{n+1} = \mathbf{0}.
\end{cases}
\tag{59}
$$

Except for the special case $\theta = 0$, which corresponds to the forward Euler method, the solution of this system is quite involved.

A possible alternative is to use a semi-implicit scheme, in which the linear part of the equation is advanced implicitly, while nonlinear terms explicitly.

By doing so, if $\theta \geq 1/2$, the resulting scheme is unconditionally stable, whereas it must obey a stability restriction on the time step Δt (depending on h and ν) in all other cases.
For more details, results and bibliographical references, see, e.g., ref.[26], Chap. 13.

[26] Quarteroni and Valli, *Numerical Approximation of Partial Differential Equations*.

Finite difference methods

We consider at first an explicit temporal discretization of the first equation in (58), corresponding to the choice $\theta = 0$ in (59). If we suppose that all quantities are known at the time t^n, we can write the associated problem at time $t^n + 1$ as follows

$$\begin{cases} \mathrm{M}\mathbf{u}^{n+1} = H(\mathbf{u}^n, \mathbf{p}^n, \mathbf{f}^n), \\ \mathrm{B}\mathbf{u}^{n+1} = \mathbf{0}, \end{cases}$$

where M is the mass matrix whose entries are

$$m_{ij} = \int_\Omega \varphi_i \varphi_j \; d\Omega.$$

This system does not allow the determination of the pressure $\mathbf{p}^{n+1}$. In particular, there is no way to enforce the divergence free constraint on $\mathbf{u}^{n+1}$.

However, if we replace $\mathbf{p}^n$ by $\mathbf{p}^{n+1}$ in the momentum equation, we obtain the new linear system

$$\begin{cases} \dfrac{1}{\Delta t}\mathrm{M}\mathbf{u}^{n+1} + \mathrm{B}^T\mathbf{p}^{n+1} = \mathbf{G}, \\ \mathrm{B}\mathbf{u}^{n+1} = \mathbf{0}, \end{cases} \qquad (60)$$

$\mathbf{G}$ being a suitable known vector. This system corresponds to a semi-explicit discretization of (58). Since M is symmetric and positive definite, if condition (51) is satisfied, then the reduced system $\mathrm{BM}^{-1}\mathrm{B}^T\mathbf{p}^{n+1} = \mathrm{BM}^{-1}\mathbf{G}$ is non-singular.

Once solved, the velocity vector $\mathbf{u}^{n+1}$ can be recovered from the first equation of (60). This discretization method is temporally stable provided the time step satisfies the following limitation (of parabolic type)

$$\Delta t \leq C \min\left(\frac{h^2}{\nu}, \frac{h}{\max_{\mathbf{x}\in\Omega}|\mathbf{u}^n(\mathbf{x})|}\right).$$

Let us now consider an implicit discretization of (58), for instance the backward Euler method, which corresponds to choosing $\theta = 1$ in (59). As already observed, this scheme is unconditionally stable. It yields a nonlinear algebraic system which can be regarded as the finite element space approximation to the steady Navier-Stokes problem

$$\begin{cases} -\nu \Delta \mathbf{u}^{n+1} + (\mathbf{u}^{n+1} \cdot \nabla)\mathbf{u}^{n+1} + \nabla p^{n+1} + \dfrac{\mathbf{u}^{n+1}}{\Delta t} = \tilde{\mathbf{f}}, \\ \mathrm{div} \mathbf{u}^{n+1} = 0. \end{cases}$$

The solution of such nonlinear algebraic system can be achieved by Newton-Krylov techniques, that is by using a Krylov method (e.g. GMRES or BiCGStab) for the solution of the linear system that is obtained at each Newton iteration step (see, e.g., ref.[27] or ref.[28]).

[27] Y. Saad. *Iterative Methods for Sparse Linear Systems*. Boston: PWS Publishing Company, 1996.

[28] Quarteroni and Valli, *Numerical Approximation of Partial Differential Equations*, Chap. 2.

We recall that Newton's method is based on the full linearization of the convective term, $\mathbf{u}_k^{n+1} \cdot \nabla \mathbf{u}_{k+1}^{n+1} + \mathbf{u}_{k+1}^{n+1} \cdot \nabla \mathbf{u}_k^{n+1}$.

A popular approach consists in starting Newton iterations after few Piccard iterations in which the convective term is evaluated as follows: $\mathbf{u}_k^{n+1} \cdot \nabla \mathbf{u}_{k+1}^{n+1}$.

This approach entails three nested cycles:

- temporal iteration: $t^n \to t^{n+1}$;
- Newton iteration: $\mathbf{x}_k^{n+1} \to \mathbf{x}_{k+1}^{n+1}$;
- Krylov iteration: $[\mathbf{x}_k^{n+1}]_j \to [\mathbf{x}_k^{n+1}]_{j+1}$;

for simplicity we have called $\mathbf{x}^n$ the couple $(\mathbf{u}^n, \mathbf{p}^n)$. Obviously, the goal is the following convergence result:

$$\lim_{k\to\infty} \lim_{j\to\infty} [\mathbf{x}_k^{n+1}]_j = \left[\begin{array}{c} \mathbf{u}^{n+1} \\ \mathbf{p}^{n+1} \end{array} \right].$$

Finally, let us operate a semi-implicit, temporal discretization, consisting in treating explicitly the nonlinear convective term. The following algebraic linear system, whose form is similar to (47), is obtained in this case

$$\begin{cases} \dfrac{1}{\Delta t}\mathrm{M}\mathbf{u}^{n+1} + \mathrm{A}\mathbf{u}^{n+1} + \mathrm{B}^{T}\mathbf{p}^{n+1} = \mathbf{G}, \\ \mathrm{B}\mathbf{u}^{n+1} = \mathbf{0}, \end{cases} \tag{61}$$

where **G** is a suitable known vector. In this case the stability restriction on the time step takes the following form

$$\Delta t \leq C \frac{h}{\max_{\mathbf{x}\in\Omega} |\mathbf{u}^{n}(\mathbf{x})|}. \tag{62}$$

In all cases, optimal error estimates can be proven.

Fractional step methods (see CFD course)

Let us consider an abstract time dependent problem,

$$\frac{\partial w}{\partial t} + Lw = f,$$

where L is a differential operator that splits into the sum of two operators, L_1 and L_2, that is

$$Lv = L_1 v + L_2 v.$$

Fractional step methods allow the temporal advancement from time t^n to t^{n+1} in two steps (or more).

At first only the operator L_1 is advanced in time implicitly, then the solution so obtained is corrected by performing a second step in which only the other operator, L_2, is in action. This is why these kind of methods are also named operator splitting.

In principle, by separating the two operators L_1 and L_2, a complex problem is split into two simpler problems, each one with its own feature. In this respect, the operators L_1 and L_2 can be chosen on the ground of physical considerations: diffusion can be split from transport, for instance.

In fact, also the solution of Navier-Stokes equations by the characteristic method can be regarded as a fractional step method whose first step operator is expressed by the Lagrangian derivative.

A simple, albeit not optimal fractional step scheme, is the following, known as Yanenko splitting:

1. compute the solution $\tilde{w}$ of the equation

$$\frac{\tilde{w} - w^n}{\Delta t} + L_1 \tilde{w} = 0;$$

2. compute the solution w^{n+1} of the equation

$$\frac{w^{n+1} - \tilde{w}}{\Delta t} + L_2 w^{n+1} = f^n.$$

By eliminating $\tilde{w}$, the following problem is found for w^{n+1}

$$\frac{w^{n+1} - w^n}{\Delta t} + L w^{n+1} = f^n + \Delta t L_1 (f^n - L_2 w^{n+1}).$$

If both L_1 and L_2 are elliptic operators, this scheme is unconditionally stable with respect to Δt.

The term

$$\sigma^{n+1} = \Delta t L_1(f^n - L_2 w^{n+1})$$

is called the splitting error.

Formally speaking, it is of order $O(\Delta t)$. Notice that the error (in the case of the Navier-Stokes equations, the error in the H^1 norm for the velocity and in L^2 norm for the pressure) behave additively w.r.t. the errors in space and time.

More accurate splitting algorithms exist, for instance the Strang splitting, that is of order $O(\Delta t^2)$.

The splitting method of Chorin and Temam (see CFD course)

This strategy can be applied to the Navier-Stokes equations (2), choosing L_1 as $L_1(\mathbf{w}) = -\nu\Delta\mathbf{w} + (\mathbf{w}\cdot\nabla)\mathbf{w}$ whereas L_2 is the operator associated to the remaining terms of the Navier-Stokes problem.

In this way we have split the main difficulties arising when treating Navier-Stokes equations, the nonlinear part from that imposing the incompressibility constraint.

The corresponding fractional step scheme reads:

1. solve the diffusion-transport equation for the velocity $\tilde{\mathbf{u}}^{n+1}$

$$\begin{cases} \dfrac{\tilde{\mathbf{u}}^{n+1} - \mathbf{u}^n}{\Delta t} - \nu \Delta \tilde{\mathbf{u}}^{n+1} + (\mathbf{u}^* \cdot \nabla)\mathbf{u}^{**} = \mathbf{f}^{n+1} & \text{in } \Omega, \\ \tilde{\mathbf{u}}^{n+1} = \mathbf{0} & \text{on } \partial\Omega; \end{cases} \tag{63}$$

2. solve the following coupled problem for the two unknowns $\mathbf{u}^{n+1}$ and p^{n+1}

$$\begin{cases} \dfrac{\mathbf{u}^{n+1} - \tilde{\mathbf{u}}^{n+1}}{\Delta t} + \nabla p^{n+1} = \mathbf{0} & \text{in } \Omega, \\ \mathrm{div}\mathbf{u}^{n+1} = 0 & \text{in } \Omega, \\ \mathbf{u}^{n+1} \cdot \mathbf{n} = 0 & \text{on } \partial\Omega, \end{cases} \tag{64}$$

where $\mathbf{u}^*$ and $\mathbf{u}^{**}$ can be either $\tilde{\mathbf{u}}^{n+1}$ or $\mathbf{u}^n$ depending on whether the nonlinear convective terms are treated explicitly, implicitly or semi-implicitly.

In such a way, in the first step an intermediate velocity $\tilde{\mathbf{u}}^{n+1}$ is calculated, then it is corrected in the second step in order to satisfy the incompressibility constraint.

The diffusion-transport problem of the first step can be successfully addressed by using the approximation techniques investigated in the dedicated chapter.

More involved is the numerical treatment of the problem associated with the second step. By formally applying the divergence operator to the first equation, we obtain

$$\mathrm{div}\frac{\mathbf{u}^{n+1}}{\Delta t} - \mathrm{div}\frac{\tilde{\mathbf{u}}^{n+1}}{\Delta t} + \Delta p^{n+1} = 0,$$

that is an elliptic boundary-value problem with Neumann boundary conditions

$$\begin{cases} -\Delta p^{n+1} = -\mathrm{div}\dfrac{\tilde{\mathbf{u}}^{n+1}}{\Delta t} & \text{in } \Omega, \\ \dfrac{\partial p^{n+1}}{\partial n} = 0 & \text{on } \partial\Omega. \end{cases} \tag{65}$$

The Neumann condition follows from the condition $\mathbf{u}^{n+1} \cdot \mathbf{n} = 0$ on $\partial\Omega$.

From the solution of (65) we obtain p^{n+1}, and thus $\mathbf{u}^{n+1}$ by using the first equation of (64),

$$\mathbf{u}^{n+1} = \tilde{\mathbf{u}}^{n+1} - \Delta t \nabla p^{n+1} \quad \text{in } \Omega. \tag{66}$$

This is precisely the correction to operate on the velocity field in order to fulfill the divergence-free constraint.

In conclusion, the algorithm is structured as follows.

1. At first we solve the elliptic system (63) to obtain the intermediate velocity $\tilde{\mathbf{u}}^{n+1}$

$$\begin{cases} \dfrac{\tilde{\mathbf{u}}^{n+1} - \mathbf{u}^n}{\Delta t} - \nu \Delta \tilde{\mathbf{u}}^{n+1} + (\mathbf{u}^* \cdot \nabla)\mathbf{u}^{**} = \mathbf{f}^{n+1} & \text{in } \Omega, \\ \tilde{\mathbf{u}}^{n+1} = \mathbf{0} & \text{on } \partial\Omega; \end{cases} \tag{67}$$

2. Then the scalar elliptic problem (65) yields the pressure unknown p^{n+1}

$$\begin{cases} -\Delta p^{n+1} = -\mathrm{div} \dfrac{\tilde{\mathbf{u}}^{n+1}}{\Delta t} & \text{in } \Omega, \\ \dfrac{\partial p^{n+1}}{\partial n} = 0 & \text{on } \partial\Omega. \end{cases} \tag{68}$$

3. Finally we obtain the new velocity field $\mathbf{u}^{n+1}$ through the explicit correction equation (66)

$$\mathbf{u}^{n+1} = \tilde{\mathbf{u}}^{n+1} - \Delta t \nabla p^{n+1} \quad \text{in } \Omega. \tag{69}$$

Let us now investigate the main features of this method.
Assume that we take $\mathbf{u}^* = \mathbf{u}^{**} = \mathbf{u}^n$ in the first step; after space discretization, we arrive at a linear system as

$$\left(\frac{1}{\Delta t}\mathrm{M} + \mathrm{A}\right)\tilde{\mathbf{u}}^{n+1} = \tilde{\mathbf{f}}^{n+1}.$$

Because of the explicit treatment of the convective term, the solution undergoes a stability restriction on the time step like (62). On the other hand, this linear system naturally splits into d independent systems of smaller size, one for each spatial component of the velocity field.

If, instead, we use an implicit time advancing scheme, like the one that we would get by setting $\mathbf{u}^* = \mathbf{u}^{**} = \tilde{\mathbf{u}}^{n+1}$, we obtain an unconditionally stable scheme, however with a more involved coupling of all the spatial components due to the nonlinear convective term. This nonlinear algebraic system can be solved by, e.g., a Newton-Krylov method.

In the second step of the method, we enforce a boundary condition only on the normal component of the velocity field. Yet, we lack any control on the behaviour of the tangential component. This generates a so-called splitting error: although the solution is divergence-free, the failure to satisfy the physical boundary condition on the tangential velocity component yields the onset of a pressure boundary layer of width $\sqrt{\nu\,\Delta t}$.

The method just described was originally proposed by Chorin and Temam, and is also called projection method. The reason can be found in the celebrated Helmholtz-Weyl decomposition theorem.

Helmholtz-Weyl decomposition theorem

Let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, be a domain with Lipschitz boundary. Then, for every $\mathbf{v} \in [\mathrm{L}^2(\Omega)]^d$, there exist two (uniquely-defined) functions $\mathbf{w}, \mathbf{z}$,

$$\mathbf{w} \in \mathrm{H}^0_{\mathrm{div}} = \{\mathbf{v} \in \left[\mathrm{L}^2(\Omega)\right]^d : \ \mathrm{div}\mathbf{v} = 0 \text{ in } \Omega, \ \mathbf{v} \cdot \mathbf{n} = 0 \text{ on } \partial\Omega\},$$

$$\mathbf{z} \in [\mathrm{L}^2(\Omega)]^d, \quad \mathrm{rot}\mathbf{z} = \mathbf{0} \quad (\text{so } \ \mathbf{z} = \nabla\psi, \text{ for a suitable } \ \psi \in \mathrm{H}^1(\Omega))$$

such that

$$\mathbf{v} = \mathbf{w} + \mathbf{z}.$$

Owing to this result, any function $\mathbf{v} \in [\mathrm{L}^2(\Omega)]^d$ can be univocally represented as being the sum of a solenoidal (that is, divergence-free) field and of an irrotational field (that is, the gradient of a suitable scalar function).

As a matter of fact, after the first step (63) in which the preliminary velocity $\tilde{\mathbf{u}}^{n+1}$ is obtained from $\mathbf{u}^n$ by solving the momentum equation, in the course of the second step a solenoidal field $\mathbf{u}^{n+1}$ is constructed in (66), with $\mathbf{u}^{n+1} \cdot \mathbf{n} = 0$ on $\partial\Omega$. This solenoidal field is the projection of $\tilde{\mathbf{u}}^{n+1}$, and is obtained by applying the decomposition theorem with the following identifications: $\mathbf{v} = \tilde{\mathbf{w}}^{n+1}$, $\mathbf{v} = \mathbf{u}^{n+1}$, $\psi = +\Delta t p^{n+1}$.

The name projection method is due to the fact that

$$\int_\Omega \mathbf{u}^{n+1} \cdot \boldsymbol{\psi} \; d\Omega = \int_\Omega \tilde{\mathbf{u}}^{n+1} \cdot \boldsymbol{\psi} \; d\Omega \quad \forall \boldsymbol{\psi} \in \mathrm{H}^0_{\mathrm{div}},$$

that is $\mathbf{u}^{n+1}$ is the projection, with respect to the scalar product of $\mathrm{L}^2(\Omega)$, of $\tilde{\mathbf{u}}^{n+1}$ on the space $\mathrm{H}^0_{\mathrm{div}}$.

Remark

Several variants of the projection method have been proposed with the aim of reducing the splitting error on the pressure, not only for the finite element method but also for higher order spectral or spectral element space approximations. The interested reader can refer to, e.g., ref.[a], ref.[b], ref.[c], ref.[d] and ref.[e].

[a] Quarteroni and Valli, *Numerical Approximation of Partial Differential Equations*.

[b] Quartapelle, *Numerical Solution of the Incompressible Navier-Stokes Equations*.

[c] A. Prohl. *Projection and Quasi-Compressibility Methods for Solving the Incompressible Navier-Stokes Equations*. Advances in Numerical Mathematics. Stuttgart: B.G. Teubner, 1997.

[d] G. E. Karniadakis and S. J. Sherwin. *Spectral/hp Element Methods for Computational Fluid Dynamics*. II. New York: Oxford University Press, 2005.

[e] Canuto et al., *Spectral Methods. Evolution to Complex Geometries and Application to Fluid Dynamics*.

Example

In Fig. 10 we display the isolines of the modulus of velocity corresponding to the solution of Navier-Stokes equations in a two-dimensional domain $\Omega = (0, 17) \times (0, 10)$ with five round holes. This can be regarded as the orthogonal section of a three dimensional domain with 5 cylinders. A non-homogeneous Dirichlet condition, $\mathbf{u} = [\arctan(20(5 - |5 - y|)), 0]^T$, is assigned at the inflow, a homogeneous Dirichlet condition is prescribed on the horizontal side as well as on the border of the cylinders, while at the outflow the normal component of the stress tensor is set to zero. For the space discretization the stabilized spectral element method was used, with 114 spectral elements, and polynomials of degree 7 for both the pressure and the velocity components on every element, plus a second-order BDF2 scheme for temporal discretization (see ref.[a]).

[a] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical Mathematics*. II. Berlin and Heidelberg: Springer, 2007.

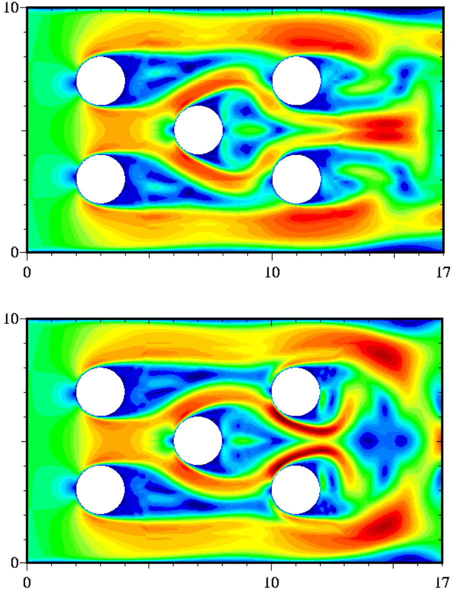

Figure: Isolines of the modulus of the velocity vector for the test case of the Example at the time levels $t = 10.5$ (above) and $t = 11.4$ (below)