Computational Fluid Dynamics

Course Notes

September 2023

School of Industrial and Information Engineering -
2023/2024

Contents

Chapter 1
Mathematical models for incompressible flows

In this chapter, the partial differential equations governing incompressible flows, namely the *incompressible Navier–Stokes equations*, will be derived starting from basic physical principles such as the mass conservation principle and the second Newton's law.

1.1 Derivation of the fluid dynamics equations

Let us consider a fluid filling a domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, and moving according to the velocity field $\mathbf{u}(\mathbf{x}, t) : \Omega \times \mathbb{R}^+ \longrightarrow \mathbb{R}^d$. The function $\mathbf{u}(\mathbf{x}, t)$ is called the *Eulerian velocity* of the fluid, since it describes the instantaneous velocity of a particle that, at time t, is located at position $\mathbf{x}$.

The trajectory of each fluid particle $\mathbf{x}(t)$ can be obtained as the solution of the following differential problem:

$$\begin{cases} \dot{\mathbf{x}}(t) = \mathbf{u}(\mathbf{x}, t), \\ \mathbf{x}(0) = \hat{\mathbf{x}}, \end{cases}$$

where $\hat{\mathbf{x}}$ is the location of the particle at the initial time $t = 0$. The function $\mathbf{x} = \mathbf{x}(t; \hat{\mathbf{x}}) = \mathcal{L}_t(\hat{\mathbf{x}})$ is called *Lagrangian map* between the domain $\hat{\Omega}$ occupied by the fluid at the initial time and the domain Ω occupied by the fluid at time t. If the field $\mathbf{u}(\mathbf{x}, t)$ is regular enough, this map is invertible.

Consider a generic function $f : \Omega \times \mathbb{R}^+ \longrightarrow \mathbb{R}$, $f = f(\mathbf{x}, t)$. The quantity

$$\frac{D}{Dt} f(\mathbf{x}, t) = \frac{d}{dt} f(\mathbf{x}(t), t) = \frac{\partial f}{\partial t} + \mathbf{u} \cdot \nabla f$$

is the *material derivative* of f and represents the time derivative of f along the trajectories of the fluid particles. Let V_t be a material volume element

of fluid, namely $V_t = \mathcal{L}_t(\hat{V})$ where $\hat{V}$ is the region occupied by the material volume element at the initial time.

We recall the following important theorem:

Theorem 1.1 (Reynolds transport theorem for a material element). *Given an Eulerian velocity field* $\mathbf{u} : \Omega \times \mathbb{R}^+ \longrightarrow \mathbb{R}^d$*, the correspondent Lagrangian map* $\mathcal{L}_t$*, a material volume element* $V_t = \mathcal{L}_t(\hat{V})$ *transported by the fluid and a differentiable function* $f : \Omega \times \mathbb{R}^+ \longrightarrow \mathbb{R}$*, then*

$$\frac{d}{dt}\int_{V_t} f\, dV = \int_{V_t} \frac{\partial f}{\partial t}\, dV + \int_{\partial V_t} f\mathbf{u}\cdot\mathbf{n}\, dA = \int_{V_t}\left(\frac{\partial f}{\partial t} + \operatorname{div}(f\mathbf{u})\right) dV,$$

where $\mathbf{n}$ *denotes the outward normal of* ∂V_t*.*

1.1.1 Mass conservation

The principle of mass conservation states that the mass $m(V_t)$ of any volume V_t of fluid does not change with time. Denoted with $\rho : \Omega \times \mathbb{R}^+ \longrightarrow \mathbb{R}$, $\rho = \rho(\mathbf{x}, t)$ the *density of mass*, the mass conservation principle can be formulated as

$$\frac{d}{dt}\int_{V_t} \rho\, dV = 0.$$

Using the Reynolds transport theorem, we have

$$\int_{V_t}\left(\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho\mathbf{u})\right) dV = 0$$

which represents the *integral form* of the mass conservation principle. Thanks to the arbitrary of V_t we can get the correspondent *differential form* of the mass conservation principle

$$\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho\mathbf{u}) = 0, \qquad \forall \mathbf{x} \in \Omega, t > 0, \tag{1.1}$$

which is also referred to as the *continuity equation*.

A fluid motion (flow) is called *incompressible* if the volume of an arbitrary region V_t does not change with time. In this case, we have

$$0 = \frac{d}{dt}\int_{V_t} dV = \int_{V_t} \operatorname{div}\mathbf{u}\, dV = 0 \quad \longrightarrow \quad \operatorname{div}\mathbf{u} = 0.$$

Therefore, for incompressible flows, the continuity equation (1.1) is reduced to

$$\frac{\partial \rho}{\partial t} + \mathbf{u}\cdot\nabla\rho = \frac{D\rho}{Dt} = 0$$

which states that density is constant along the flow trajectory. Finally, if the fluid is homogeneous (*i.e.* has a constant density) at the initial time, and its motion is incompressible, the density will remain constant at all time.

1.1.2 Momentum balance

Newton's second law states that the variation of (linear) momentum $\rho\mathbf{u}$ of a system is equal to the resultant of the forces acting on it. The momentum variation of an arbitrary volume of fluid $V_t = \mathcal{L}_t(\hat{V})$ is given by

$$\frac{d}{dt}\int_{V_t} \rho\mathbf{u}\, dV.$$

The forces exerted on V_t can be classified as:

- *volume forces* per unit mass (such as gravity) that will be denoted as $\mathbf{f} : V_t \times \mathbb{R}^+ \longrightarrow \mathbb{R}^d$, whose resultant on V_t will be

$$\mathbf{F} = \int_{V_t} \rho\mathbf{f}\, dV;$$

- *surface forces* per unit area that will be denoted with $\mathbf{t}^s : \partial V_t \times \mathbb{R}^+ \longrightarrow \mathbb{R}^d$; if the volume V_t is completely contained in Ω, then surface forces are only due to the interaction of the volume with the surrounding fluid.

Concerning the surface forces, the Cauchy principle states that there exists a stress field $\mathbf{t} : \Omega \times \mathbb{R}^+ \times S \longrightarrow \mathbb{R}^d$ with $S = \{\mathbf{n} \in \mathbb{R}^d : |\mathbf{n}| = 1\}$, $\mathbf{t} = \mathbf{t}(\mathbf{x}, t, \mathbf{n})$ such that the surface forces $\mathbf{t}^s$ on V_t are given by

$$\mathbf{t}^s = \mathbf{t}(\mathbf{x}, t, \mathbf{n}),$$

that is $\mathbf{t}^s$ does not depend explicitly on the domain V_t but only on its outward normal $\mathbf{n}$. As sketched in Figure 1.1, the surface force per unit area exerted by the surrounding fluid on the surface element dA is the same for both volumes $V_t^{(1)}$ and $V_t^{(2)}$ as long as the outward normals $\mathbf{n}^{(1)}$ and $\mathbf{n}^{(2)}$ coincide.

The stress field is better characterized in the following fundamental theorem:

Theorem 1.2 (Cauchy theorem). *If the stress field* $\mathbf{t}(\mathbf{x}, t, \mathbf{n})$ *is continuous with respect to* $\mathbf{n}$ *and differentiable in* $\mathbf{x}$*, then there exists a tensor field* $\boldsymbol{\sigma} : \Omega \times \mathbb{R}^+ \longrightarrow \mathbb{R}^{d\times d}_{sym}$*, called* Cauchy stress tensor*, such that*

$$\boldsymbol{\sigma}(\mathbf{x}, t) \cdot \mathbf{n} = \mathbf{t}(\mathbf{x}, t, \mathbf{n}).$$

Otherwise stated, the Cauchy theorem affirms that the stress field $\mathbf{t}(\mathbf{x}, t, \mathbf{n})$ is a linear function of $\mathbf{n}$. A proof of the theorem can be found in [41].

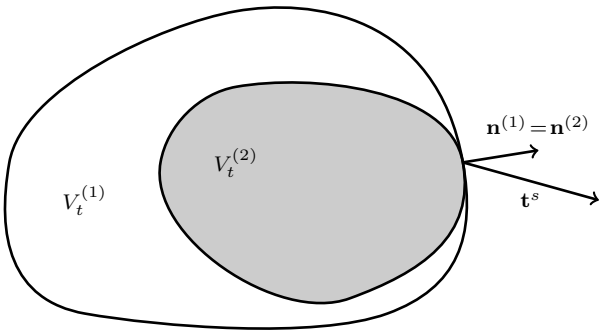


Fig. 1.1: Graphical representation of the Cauchy principle

Once defined the volume and surface forces acting on the fluid volume V_t, we can introduce the *integral form* of the momentum balance, which can be written component-wise as follows:

$$\frac{d}{dt}\int_{V_t} \rho u_i \, dV = \int_{V_t} \rho f_i \, dV + \int_{\partial V_t} \sum_j \sigma_{ij} n_j \, dA, \qquad i = 1, 2, 3. \tag{1.2}$$

The *differential form* of the momentum balance can be easily obtained using the Reynolds transport theorem, the divergence theorem and the arbitrary of volume V_t, and reads

$$\frac{\partial(\rho u_i)}{\partial t} + \mathrm{div}(\rho u_i \mathbf{u}) - \mathrm{div}\boldsymbol{\sigma}_i = \rho f_i, \qquad i = 1, 2, 3. \tag{1.3}$$

We denote with $\mathbf{u} \otimes \mathbf{v}$ the outer product between $\mathbf{u}$ and $\mathbf{v}$ defining the tensor $(\mathbf{u} \otimes \mathbf{v})_{ij} = u_i v_j$. Thus, in vectorial form, equation (1.2) can be rewritten as:

$$\frac{\partial(\rho \mathbf{u})}{\partial t} + \mathrm{div}(\rho \mathbf{u} \otimes \mathbf{u} - \boldsymbol{\sigma}) = \rho \mathbf{f}, \tag{1.4}$$

which is known as the *conservative form of the momentum equation*.

Finally, another useful formulation of the same equation can be obtained using the continuity equation to rewrite the time derivative in (1.4):

$$\frac{\partial(\rho \mathbf{u})}{\partial t} = \rho \frac{\partial \mathbf{u}}{\partial t} + \frac{\partial \rho}{\partial t}\mathbf{u} = \rho \frac{\partial \mathbf{u}}{\partial t} - \mathrm{div}(\rho \mathbf{u})\mathbf{u}.$$

Combining the latter, with the following development of the nonlinear term

$$\mathrm{div}(\rho \mathbf{u} \otimes \mathbf{u})_i = \sum_j \frac{\partial}{\partial x_j}(\rho u_i u_j) = \sum_j \left(\frac{\partial}{\partial x_j}(\rho u_j) u_i + \rho u_j \frac{\partial u_i}{\partial x_j} \right),$$

that yields

$$\mathrm{div}(\rho\mathbf{u}\otimes\mathbf{u}) = \mathrm{div}(\rho\mathbf{u})\mathbf{u} + \rho(\mathbf{u}\cdot\nabla)\mathbf{u},$$

we obtain the *advective (non-conservative) form* of the momentum equation, which reads

$$\rho\left(\frac{\partial\mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u}\right) - \mathrm{div}(\boldsymbol{\sigma}) = \rho\mathbf{f}. \tag{1.5}$$

In equation (1.5), we can recognize the second Newton's law in its form $\mathbf{f} = m\ddot{\mathbf{x}}$, where the acceleration of the fluid particles is given by

$$\frac{\partial\mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u},$$

while

$$\rho\mathbf{f} + \mathrm{div}(\boldsymbol{\sigma})$$

are the forces acting of each particle due to body forces or internal stress related to the interaction with the other particles.

1.1.3 Constitutive laws for Newtonian fluids

In order to complete the definition of the momentum equation, we need to characterize the Cauchy stress tensor $\boldsymbol{\sigma}$. A fluid is called *Stokesian*, if the the following assumptions on the stress tensor $\boldsymbol{\sigma}$ hold:

1. $\boldsymbol{\sigma}$ *is symmetric*. This property is a direct consequence of the balance of angular momentum.
2. *at rest (with no motion), the internal stress are only due to the fluid pressure*. Based on that we consider the following decomposition of the stress tensor:
$$\boldsymbol{\sigma} = -p\mathbf{I} + \boldsymbol{\tau}, \tag{1.6}$$
where $p : \Omega\times\mathbb{R}^+ \longrightarrow \mathbb{R}, p = p(\mathbf{x},t)$ is the pressure and $\boldsymbol{\tau} : \Omega\times\mathbb{R}^+ \longrightarrow \mathbb{R}^{d\times d}_{\mathrm{sym}}$ is the viscous stress tensor.
3. $\boldsymbol{\tau}$ *is continuous with respect to* $\nabla\mathbf{u}$.
4. *frame invariance property*: the same law $\tau(\nabla\mathbf{u})$ should hold in any reference system (even non-inertial), that is, given any orthogonal matrix $Q\in\mathbb{R}^{d\times d}$,
$$Q^T\boldsymbol{\tau}(\nabla\mathbf{u})Q = \boldsymbol{\tau}(Q^T\nabla\mathbf{u}Q).$$
This reflects the fact that no momentum diffusion will be generated by a rigid body motion of the fluid.

From assumptions 1. and 4. it follows that $\boldsymbol{\tau}$ only depends on the symmetric part of $\nabla\mathbf{u}$, that is on the *rate of deformation tensor* defined as

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

.

If the viscous stress tensor $\boldsymbol{\tau}$ depends *linearly* on $\mathbf{D}$, then the fluid is called *Newtonian*. In this case, assumptions $1. - 4.$ imply that

$$\boldsymbol{\tau} = \lambda \mathrm{div}\mathbf{u}\mathbf{I} + 2\mu\mathbf{D}(\mathbf{u}),$$

where $\mu \geq 0$ is the dynamic viscosity of the fluid and λ is the second viscosity coefficient. In order to ensure positive energy dissipation, the two viscosity coefficients should be such that

$$\lambda + \frac{2}{3}\mu \geq 0.$$

Finally, for fluids in thermodynamic equilibrium, the so-called *Stokes relation*

$$\lambda + \frac{2}{3}\mu = 0$$

holds. In this case, we have:

$$\boldsymbol{\sigma} = -(p + \frac{2}{3}\mu \mathrm{div}\mathbf{u})\mathbf{I} + 2\mu\mathbf{D}(\mathbf{u}), \tag{1.7}$$

which reduces to

$$\boldsymbol{\sigma} = -p\mathbf{I} + 2\mu\mathbf{D}(\mathbf{u}) \tag{1.8}$$

for incompressible flows.

1.2 Energy balance

We can now proceed to the derivation of the equation governing the energy conservation. We consider the total specific energy as the sum of internal and kinetic energy, namely

$$e = e_i + \frac{1}{2}|\mathbf{u}|^2.$$

The first law of thermodynamics states that the rate of change of the total energy of an arbitrary volume of fluid $V_t = \mathcal{L}_t(\hat{V})$ is given by is equal to the power of the forces acting on it plus the heat power supplied to it.

The power of the (external) body forces $\mathbf{f}$ acting on V_t is given by

$$\int_{V_t} \rho\mathbf{f} \cdot \mathbf{u}\, dV,$$

while the power of the (internal) surface forces is given by

$$\int_{\partial V_t} (\boldsymbol{\sigma} \cdot \mathbf{n}) \cdot \mathbf{u} \, dA.$$

Denoted with $s(\mathbf{x}, t)$ an internal heat source per unit of mass and unit of time, its contribution as heat power supplied to the system is

$$\int_{V_t} \rho s \, dV.$$

Moreover, denoted with $\mathbf{q}$ the heat flux per unit of time through the boundary $\partial \mathbf{V}$, the heat power entering the system is given by

$$-\int_{\partial V_t} \mathbf{q} \cdot \mathbf{n} \, dA.$$

The heat flux is typically related to the temperature T through the *Fick's law*

$$\mathbf{q} = -k \nabla T,$$

where k is the thermal conductivity.

The energy balance can then be stated as follows

$$\frac{d}{dt} \int_{V_t} \rho e \, dV = \int_{V_t} \rho \mathbf{f} \cdot \mathbf{u} \, dV + \int_{V_t} \rho s \, dV + \int_{\partial V_t} (\boldsymbol{\sigma} \cdot \mathbf{u}) \cdot \mathbf{n} \, dA + \int_{\partial V_t} k \nabla T \cdot \mathbf{n} \, dA,$$

and resorting once more to Reynolds transport theorem and to the divergence theorem we obtain the energy equation in *conservative form*

$$\frac{\partial (\rho e)}{\partial t} + \operatorname{div}(\rho e \mathbf{u} - \boldsymbol{\sigma} \cdot \mathbf{u} - k \nabla T) = \rho \mathbf{f} \cdot \mathbf{u} + \rho s. \tag{1.9}$$

By using the decomposition of the Cauchy stress in pressure and viscous contribution (1.6). the energy equation can be rewritten highlighting the pressure term, as follows:

$$\frac{\partial (\rho e)}{\partial t} + \operatorname{div}((\rho e + p)\mathbf{u} - \boldsymbol{\tau} \cdot \mathbf{u} - k \nabla T) = \rho \mathbf{f} \cdot \mathbf{u} + \rho s. \tag{1.10}$$

1.3 The Navier–Stokes and Euler equations

Collecting equations (1.1), (1.4) and (1.10) which have been derived from the three physical principles (mass conservation, momentum and energy balance), we can finally formulate the Navier-Stokes equation in conservative form

$$\frac{\partial \rho}{\partial t} + \mathrm{div}(\rho \mathbf{u}) = 0, \tag{1.11}$$

$$\frac{\partial (\rho \mathbf{u})}{\partial t} + \mathrm{div}(\rho \mathbf{u} \otimes \mathbf{u} + p\mathbf{I} - \boldsymbol{\tau}) = \rho \mathbf{f}, \tag{1.12}$$

$$\frac{\partial (\rho e)}{\partial t} + \mathrm{div}((\rho e + p)\mathbf{u} - \boldsymbol{\tau} \cdot \mathbf{u} - k\nabla T) = \rho \mathbf{f} \cdot \mathbf{u} + \rho s. \tag{1.13}$$

In $\mathbb{R}^d$, this system is defined by $d+2$ equations and contain $d+4$ independent unknowns, d velocity components, density, pressure, total energy and temperature. In order to close the system suitable state equations relating the thermodynamical variables should be added, such as, for instance

$$p = p(\rho, e),$$
$$T = T(\rho, e).$$

Neglecting all viscous term in equations (1.11)-(1.13), the Euler equations governing inviscid compressible flows are obtained

$$\frac{\partial \rho}{\partial t} + \mathrm{div}(\rho \mathbf{u}) = 0, \tag{1.14}$$

$$\frac{\partial (\rho \mathbf{u})}{\partial t} + \mathrm{div}(\rho \mathbf{u} \otimes \mathbf{u} - p\mathbf{I}) = \rho \mathbf{f}, \tag{1.15}$$

$$\frac{\partial (\rho e)}{dt} + \mathrm{div}((\rho e + p)\mathbf{u}) = \rho \mathbf{f} \cdot \mathbf{u} + \rho s. \tag{1.16}$$

The Euler equations define a system of nonlinear conservation laws for which ad-hoc analytical and numerical tools have been developed (see, e.g., [36] for an introduction to these tools).

1.4 The incompressible Navier–Stokes equations

In the case of flows which are homogeneous (ρ uniform in space) and incompressible ($\mathrm{div}\mathbf{u} = 0$), the momentum equation (1.3) and the continuity equation (1.1) are uncoupled from the energy equation and reduce to the so-called *incompressible Navier-Stokes equations*:

$$\rho \frac{\partial \mathbf{u}}{\partial t} + \rho(\mathbf{u} \cdot \nabla)\mathbf{u} - \mu \Delta \mathbf{u} + \nabla p = \rho \mathbf{f}$$
$$\mathrm{div}\mathbf{u} = 0$$

since

$$\mathrm{div}(2\mu \mathbf{D}(\mathbf{u})) = \mathrm{div}\left(2\mu \frac{\nabla \mathbf{u} + \nabla \mathbf{u}^T}{2}\right) = \mu \Delta \mathbf{u}$$

Dividing by ρ, we get

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u} - \nu\Delta\mathbf{u} + \nabla\left(\frac{p}{\rho}\right) = \mathbf{f} \tag{1.17}$$

$$\mathrm{div}\mathbf{u} = 0 \tag{1.18}$$

where $\nu = \mu/\rho$ is the kinematic viscosity.

1.4.1 Initial conditions

Since equations (1.17)-(1.18) are time-dependent, suitable initial conditions should be defined. However, it should be noted that the pressure does not appear under time derivative. Therefore, initial conditions are only requires for the velocity:

$$\mathbf{u}(\mathbf{x}, t = 0) = \mathbf{u}_0(\mathbf{x}). \tag{1.19}$$

1.4.2 Boundary conditions

If equations (1.17)-(1.18) are solved on a bounded domain $\Omega \subset \mathbb{R}^d$, boundary conditions on $\partial\Omega$ should also be defined. Let us consider a general case in which the boundary is subdivided in a portion Γ_D, where essential (Dirichlet) conditions are prescribed, and a portion Γ_N, where natural (Neumann) conditions are imposed, such that

$$\partial\Omega = \Gamma_D \cup \Gamma_N, \qquad \mathring{\Gamma}_D \cap \mathring{\Gamma}_N = \emptyset.$$

In order to identify which are the correct natural boundary condition associated to equations (1.17)-(1.18), we derive the weak formulation multiplying the momentum and continuity equations by suitable test functions $\mathbf{v} : \Omega \longrightarrow \mathbb{R}^d$ and $q : \Omega \longrightarrow \mathbb{R}$, respectively, and integrating over the domain Ω, as follows:

$$\int_\Omega \left(\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u} - \nu\Delta\mathbf{u} + \nabla\left(\frac{p}{\rho}\right)\right)\cdot\mathbf{v}\, d\Omega = \int_\Omega \mathbf{f}\cdot\mathbf{v}\, d\Omega, \qquad \forall \mathbf{v} \in \mathbf{V}.$$

$$\int_\Omega \mathrm{div}\mathbf{u}\, q\, d\Omega = 0, \qquad \forall q \in Q,$$

where $\mathbf{V}$ and Q are suitable functional spaces (see Section 2.1). To minimize the required regularity of the trial and test functions, the viscous and pressure terms are integrated by parts, namely

$$\begin{aligned}\int_\Omega -\nu\Delta\mathbf{u}\cdot\mathbf{v}\,d\Omega &= \sum_i \int_\Omega -\nu\Delta u_i v_i\,d\Omega = \\ &= \sum_i \int_\Omega \nu\nabla u_i\cdot\nabla v_i\,d\Omega - \sum_i \int_{\partial\Omega} \nu\frac{\partial u_i}{\partial\mathbf{n}}\cdot v_i\,d\gamma = \\ &= \sum_{i,j} \int_\Omega \nu\frac{\partial u_i}{\partial x_j}\frac{\partial v_i}{\partial x_j}\,d\Omega - \sum_{i,j}\int_\Omega \nu\frac{\partial u_i}{\partial x_j}n_j v_i\,d\Omega = \\ &= \int_\Omega \nu\nabla\mathbf{u}:\nabla\mathbf{v}\,d\Omega - \int_{\partial\Omega}\nu\frac{\partial\mathbf{u}}{\partial\mathbf{n}}\cdot\mathbf{v}\,d\gamma,\end{aligned}$$

$$\begin{aligned}\int_\Omega \nabla\frac{p}{\rho}\cdot\mathbf{v}\,d\Omega &= \sum_i \int_\Omega \frac{1}{\rho}\frac{\partial p}{\partial x_i}v_i\,d\Omega = \\ &= -\sum_i\int_\Omega \frac{p}{\rho}\frac{\partial v_i}{\partial x_i}\,d\Omega + \sum_i\int_{\partial\Omega}\frac{p}{\rho}v_i n_i\,d\gamma = \\ &= -\int_\Omega \frac{p}{\rho}\,\mathrm{div}\mathbf{v}\,d\Omega + \int_{\partial\Omega}\frac{p}{\rho}\mathbf{v}\cdot\mathbf{n}\,d\gamma.\end{aligned}$$

The weak form of the original problem (1.17)-(1.18) then reads:

$$\int_\Omega\left(\frac{\partial\mathbf{u}}{\partial t}+(\mathbf{u}\cdot\nabla)\mathbf{u}\right)\cdot\mathbf{v}\,d\Omega + \int_\Omega\nu\nabla\mathbf{u}:\nabla\mathbf{v}\,d\Omega - \int_\Omega\frac{p}{\rho}\,\mathrm{div}\mathbf{v}\,d\Omega$$
$$-\boxed{\int_{\partial\Omega}\left(\nu\frac{\partial\mathbf{u}}{\partial\mathbf{n}}-\frac{p}{\rho}\mathbf{n}\right)\cdot\mathbf{v}}\,d\Omega = \int_\Omega\mathbf{f}\cdot\mathbf{v}\,d\Omega, \quad (1.20)$$

$$\int_\Omega \mathrm{div}\mathbf{u}\;q\,d\Omega = 0. \quad (1.21)$$

In order to make the boundary terms independent of the unknowns $\mathbf{u}$ and p, the essential and natural boundary conditions are defined by:

- *Essential b.c. on* Γ_D: $\quad \mathbf{v}=0, \quad \mathbf{u}=\mathbf{g},$
- *Natural b.c. on* Γ_N: $\quad$ any $\mathbf{v}, \quad \nu\dfrac{\partial\mathbf{u}}{\partial\mathbf{n}}-\dfrac{p}{\rho}\mathbf{n}=\mathbf{d},$

with $\mathbf{g}=\mathbf{g}(\mathbf{x},t)$ and $\mathbf{d}(\mathbf{x},t)$ are two given functions with values in $\mathbb{R}^d$.

With this choice of boundary conditions, the original problem can be written in its complete form as follows

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u} - \nu\Delta\mathbf{u} + \nabla\left(\frac{p}{\rho}\right) = \mathbf{f}, \qquad \text{in } \Omega, \quad t > 0, \tag{1.22}$$

$$\text{div}\mathbf{u} = 0, \qquad \text{in } \Omega, \quad t > 0, \tag{1.23}$$

$$\mathbf{u} = \mathbf{u}_0, \qquad \text{in } \Omega, \quad t = 0, \tag{1.24}$$

$$\mathbf{u} = \mathbf{g}, \qquad \text{on } \Gamma_D, \quad t > 0, \tag{1.25}$$

$$\nu\frac{\partial \mathbf{u}}{\partial \mathbf{n}} - \frac{p}{\rho}\mathbf{n} = \mathbf{d}, \qquad \text{on } \Gamma_N, \quad t > 0, \tag{1.26}$$

This natural boundary conditions (1.26) do not have a real physical meaning. However, they are useful when used as outflow condition on artificial boundaries in internal flows. For instance, in a channel flow, if a Poiseuille velocity profile is imposed at the inflow boundary, imposing condition (1.26) at the outflow permits to recover the exact Poiseuille solution over the entire domain.

Many other different natural conditions, as well as mixed (natural on some velocity components and essential on the others) conditions, can be considered. They are derived from alternative weak formulations of the original problem. In the following sections, we consider some alternative formulation of the problem associated with different natural boundary conditions.

1.4.2.1 Normal stress boundary condition

We recall that the momentum equation (1.17) can be equivalently written as

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u} - \text{div}\frac{\boldsymbol{\tau}}{\rho} + \nabla\left(\frac{p}{\rho}\right) = \mathbf{f}, \tag{1.27}$$

where $\boldsymbol{\tau} = 2\mu\mathbf{D}(\mathbf{u}) = 2\mu\frac{\nabla\mathbf{u}+\nabla^T\mathbf{u}}{2}$ is the deviatoric stress tensor. In this case, the weak form of the momentum equation becomes

$$\int_\Omega \left(\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u}\right)\cdot\mathbf{v}\,d\Omega + \int_\Omega 2\nu\mathbf{D}(\mathbf{u}):\mathbf{D}(\mathbf{v})\,d\Omega - \int_\Omega \frac{p}{\rho}\,\text{div}\mathbf{v}\,d\Omega = \\ = \int_\Omega \mathbf{f}\cdot\mathbf{v}\,d\Omega + \int_{\Gamma_N}\mathbf{d}\cdot\mathbf{v}\,d\gamma,$$

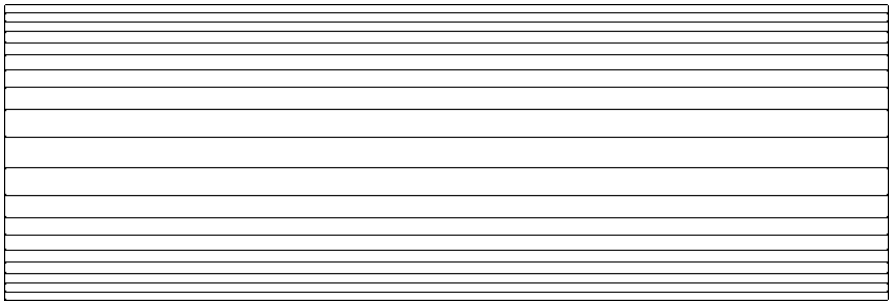
where we have used the identity $\mathbf{D}(\mathbf{u}) : \nabla\mathbf{v} = \mathbf{D}(\mathbf{u}) : \mathbf{D}(\mathbf{v})$. The natural condition imposed at the Neumann boundary is

$$\boldsymbol{\sigma}\cdot\mathbf{n} = -p\mathbf{n} + 2\mu\mathbf{D}(\mathbf{u})\cdot\mathbf{n} = \rho\mathbf{d}, \qquad \text{on } \Gamma_N,$$

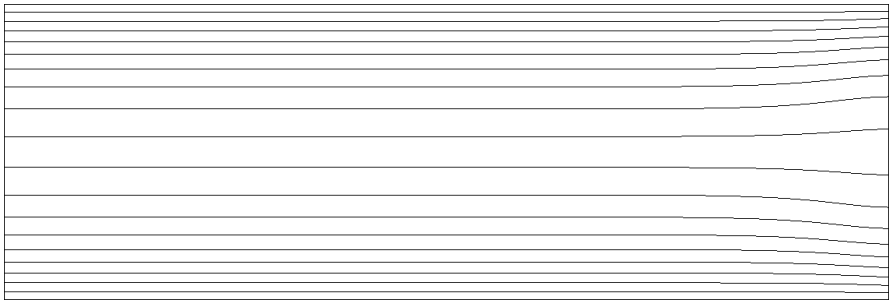
which corresponds to imposing the normal stress tensor on the Neumann boundary.

When this formulation (and the correspondent natural condition at the outflow boundary) is used to simulate a channel flow, the solution that is

obtained presents diverging streamlines at the outflow (Figure 1.2, bottom), while with the standard Laplace formulation of the viscous term and the natural boundary condition (1.26), the Poiseuille solution (with parallel streamlines) is obtained (Figure 1.2, top). This is indeed the correct solution for a finite channel flow ending in the atmosphere.



(a) Laplacian formulation



(b) Stress formulation

Fig. 1.2: Flow in a duct: streamlines obtained with outflow homogeneous boundary condition for different problem formulations of the viscous term.

To recover the Poiseuille flow solution (corresponding to an infinite channel), still using the formulation (1.27), mixed boundary conditions should be imposed at the (artificial) outflow boundary Γ_N, namely

$$\begin{aligned} \mathbf{n}^T \cdot \boldsymbol{\sigma} \cdot \mathbf{n} &= 0, \\ \mathbf{u} \cdot \mathbf{t} &= 0, \quad \forall \mathbf{t} : \mathbf{t} \cdot \mathbf{n} = 0 \end{aligned}$$

Here, the normal component of the normal stress and the tangential component of the velocity are set to zero.

1.4.2.2 Total stress boundary condition

We consider two additional quantities, the total pressure

$$p_T = p + \frac{1}{2}\rho|\mathbf{u}|^2,$$

and the total stress tensor

$$\boldsymbol{\sigma}_T = -p_T\mathbf{I} + 2\mu\mathbf{D}(\mathbf{u}).$$

The boundary condition

$$\boldsymbol{\sigma}_T \cdot \mathbf{n} = -p_T\mathbf{n} + 2\mu\mathbf{D}(\mathbf{u}) \cdot \mathbf{n} = \rho\mathbf{d}, \qquad \text{on } \Gamma_N,$$

can be obtained as natural condition starting from the problem written in rotational form:

$$\frac{\partial \mathbf{u}}{\partial t} + (\boldsymbol{\nabla}\times\mathbf{u})\times\mathbf{u} - \operatorname{div}\left(2\nu\mathbf{D}(\mathbf{u}) - \frac{p_T}{\rho}\mathbf{I}\right) = \mathbf{f}, \tag{1.28}$$

where the following vector identity has been used

$$(\mathbf{u}\cdot\nabla)\mathbf{u} = \frac{1}{2}\nabla(\mathbf{u}\cdot\mathbf{u}) + (\boldsymbol{\nabla}\times\mathbf{u})\times\mathbf{u}.$$

This natural condition can be useful when information on total pressure are available (which is typically the case when geometrical multiscale models are considered [23]). Moreover, the weak problem associated with this formulation displays better stability properties than the analogous problem with boundary condition on the normal stress as it will be discussed in Section 3.1.3.3.

1.4.2.3 Momentum flux boundary condition

An other equivalent formulation can be obtained starting from the incompressible Navier-Stokes equations in conservative form:

$$\rho\frac{\partial \mathbf{u}}{\partial t} + \operatorname{div}\left(\rho\mathbf{u}\otimes\mathbf{u} + p\mathbf{I} - \boldsymbol{\tau}\right) = \rho\mathbf{f}, \tag{1.29}$$

$$\operatorname{div}\mathbf{u} = 0. \tag{1.30}$$

The corresponding weak form of the momentum equation reads

$$\int_\Omega \left[\left(\rho \frac{\partial \mathbf{u}}{\partial t} \cdot \mathbf{v} + \rho(\mathbf{u} \cdot \nabla)\mathbf{u} \right) \cdot \mathbf{v} + 2\mu \mathbf{D}(\mathbf{u}) : \nabla \mathbf{v} - p \mathrm{div} \mathbf{v} \right] d\Omega = \\ = \int_\Omega \mathbf{f} \cdot \mathbf{v} \, d\Omega + \int_{\Gamma_N} \mathbf{d} \cdot \mathbf{v} \, d\gamma,$$

where we have imposed the following natural condition on the Neumann boundary

$$-\rho(\mathbf{u} \cdot \mathbf{n})\mathbf{u} - p\mathbf{n} + 2\mu \mathbf{D}(\mathbf{u}) \cdot \mathbf{n} = \rho \mathbf{d}, \qquad \text{on } \Gamma_N.$$

This condition is useful when working with the conservative form of the equations, which is often the case if the conservation property of the model are considered crucial.

1.4.2.4 Summary of natural boundary conditions

In Table 1.1, we summarize the different options available for the natural boundary conditions and the corresponding formulations of the momentum equations. The latter are written in strong form and it is understood that the "div(·)" terms will be integrated by part in the weak formulation.

Momentum equation	Natural boundary condition
$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla)\mathbf{u} - \mathrm{div}\left(\nu \nabla \mathbf{u} - \frac{p}{\rho}\mathbf{I} \right) = \mathbf{f}$	$-p\mathbf{n} + \mu \nabla \mathbf{u} \cdot \mathbf{n} = \rho \mathbf{d}$
Standard (Laplacian) formulation	
$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla)\mathbf{u} - \mathrm{div}\left(2\nu \mathbf{D}(\mathbf{u}) - \frac{p}{\rho}\mathbf{I} \right) = \mathbf{f}$	$\boldsymbol{\sigma} \cdot \mathbf{n} = -p\mathbf{n} + 2\mu \mathbf{D}(\mathbf{u}) \cdot \mathbf{n} = \rho \mathbf{d}$
Normal stress formulation	
$\frac{\partial \mathbf{u}}{\partial t} + (\boldsymbol{\nabla} \times \mathbf{u}) \times \mathbf{u} - \mathrm{div}\left(2\nu \mathbf{D}(\mathbf{u}) - \frac{p_T}{\rho}\mathbf{I} \right) = \mathbf{f}$	$\boldsymbol{\sigma}_T \cdot \mathbf{n} = -p_T\mathbf{n} + 2\mu \mathbf{D}(\mathbf{u}) \cdot \mathbf{n} = \rho \mathbf{d}$
Total stress formulation	
$\frac{\partial \mathbf{u}}{\partial t} - \mathrm{div}\left(2\nu \mathbf{D}(\mathbf{u}) - \frac{p}{\rho}\mathbf{I} - \mathbf{u} \otimes \mathbf{u} \right) = \mathbf{f}$	$-\rho(\mathbf{u} \cdot \mathbf{n})\mathbf{u} - p\mathbf{n} + 2\mu \mathbf{D}(\mathbf{u}) \cdot \mathbf{n} = \rho \mathbf{d}$
Momentum flux formulation	

Table 1.1: Different formulations of the momentum equations and correspondent natural boundary conditions.

In the following, we will mainly consider the *standard* and *normal stress* cases. As already mentioned, the *standard* natural conditions do not have a physical meaning but they are easy to implement and may be useful for numerical (artificial) boundaries. However, they cannot be used whenever

the physical stress has to be imposed, such as, for instance, in fluid-structure interaction problems.

In the numerical solution of the incompressible Navier-Stokes equations there are 3 main issues that should be faced, namely:

1. how to treat the incompressibility constraint $\mathrm{div}\mathbf{u} = 0$;
2. how to treat the nonlinear term $(\mathbf{u}\cdot\nabla)\mathbf{u}$ (or $(\boldsymbol{\nabla}\times\mathbf{u})\times\mathbf{u}$, $\mathbf{u}\otimes\mathbf{u}$ in the other formulations);
3. how to solve efficiently the time-dependent problem.

All these issues will be discussed in the next two chapters where a general Galerkin framework for the solution of the incompressible Navier-Stokes equations will be introduced, detailing in particular the numerical solution based on the finite-element method.

In Chapter 2, we will start considering the simpler linear Stokes problem, which can be obtained as limit of the incompressible Navier-Stokes equations when viscous effects are dominating.

1.5 Non-dimensional formulation

It is useful to express the incompressible Navier-Stokes equations in non-dimensional form in order to identify the non-dimensional parameters that fully characterize the flow.

For the sake of simplicity, let us consider the incompressible Navier–Stokes equations in Laplacian form with no forcing terms for a given fluid of density ρ and kinematic viscosity ν

$$\frac{\partial\mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u} - \nu\Delta\mathbf{u} + \nabla\left(\frac{p}{\rho}\right) = \mathbf{0}, \tag{1.31}$$

$$\mathrm{div}\mathbf{u} = 0. \tag{1.32}$$

By introducing the following reference quantities

- a reference length L [m],
- a reference velocity magnitude U [m s^{-1}],

we can compute the non-dimensional counterparts of the variables appearing in the equation

$$\mathbf{x}^* = \frac{\mathbf{x}}{L} \qquad \mathbf{u}^* = \frac{\mathbf{u}}{U} \qquad t^* = \frac{t}{\frac{L}{U}}$$

where $\mathbf{x}^*, \mathbf{u}^*$ and t^* denotes the non-dimensional spatial coordinate, velocity and time, respectively. Different scaling can be adopted for the pressure term, depending on the flow regime that is considered.

The derivatives with respect to the non-dimensional time and space coordinates can be computed using the chain rule of derivation:

$$\begin{aligned}
\frac{\partial(\cdot)}{\partial t} &= \frac{\partial(\cdot)}{\partial t^*}\frac{\partial t^*}{\partial t} = \frac{U}{L}\frac{\partial(\cdot)}{\partial t^*},\\
\frac{\partial(\cdot)}{\partial \mathbf{x}} &= \frac{\partial(\cdot)}{\partial \mathbf{x}^*}\frac{\partial \mathbf{x}^*}{\partial \mathbf{x}} = \frac{1}{L}\frac{\partial(\cdot)}{\partial \mathbf{x}^*},\\
\nabla(\cdot) &= \frac{1}{L}\nabla^*(\cdot),\\
\mathrm{div}(\cdot) &= \frac{1}{L}\mathrm{div}^*(\cdot),\\
\Delta(\cdot) &= \frac{1}{L^2}\Delta^*(\cdot).
\end{aligned}$$

1.5.1 Flows with dominant dynamic effects

When dynamic effects are dominating we can consider the following non-dimensional scaling for the pressure

$$p^* = \frac{p}{\rho U^2}.$$

The non-dimensional version of the momentum equation is obtained substituting in (1.31) the scaled variables and derivatives, thus obtaining

$$\frac{U^2}{L}\frac{\partial \mathbf{u}^*}{\partial t^*} + \frac{U^2}{L}(\mathbf{u}^*\cdot\nabla^*)\mathbf{u}^* - \nu\frac{U}{L^2}\Delta^*\mathbf{u}^* + \frac{U^2}{L}\nabla^* p^* = \mathbf{0}.$$

Thus, dividing by U^2/L, we can obtain the non-dimensional incompressible Navier-Stokes equations which depend only on the non-dimensional parameter $\mathrm{Re} = \frac{\rho U L}{\mu} = \frac{U L}{\nu}$

$$\begin{aligned}
\frac{\partial \mathbf{u}^*}{\partial t^*} + (\mathbf{u}^*\cdot\nabla^*)\mathbf{u}^* - \frac{1}{\mathrm{Re}}\Delta^*\mathbf{u}^* + \nabla^* p^* &= \mathbf{0},\\
\mathrm{div}^*\mathbf{u}^* &= 0.
\end{aligned}$$

If we consider the limiting case for $\mathrm{Re}\to\infty$, the viscous term disappears and we obtain the incompressible (inviscid) Euler equations:

$$\begin{aligned}
\frac{\partial \mathbf{u}^*}{\partial t^*} + (\mathbf{u}^*\cdot\nabla^*)\mathbf{u}^* + \nabla^* p^* &= \mathbf{0},\\
\mathrm{div}^*\mathbf{u}^* &= 0.
\end{aligned}$$

1.5.2 Flows with dominant viscous effects

A different scaling for the pressure should be considered when viscous effects are dominating, namely

$$p^* = \frac{p}{\frac{\rho\nu U}{L}}.$$

In this case, the resulting non-dimensional form reads

$$\begin{aligned} \frac{\partial \mathbf{u}^*}{\partial t^*} + (\mathbf{u}^* \cdot \nabla^*)\mathbf{u}^* - \frac{1}{\mathrm{Re}} \Delta^* \mathbf{u}^* + \frac{1}{\mathrm{Re}} \nabla^* p^* &= \mathbf{0}, \\ \mathrm{div}^* \mathbf{u}^* &= 0. \end{aligned}$$

that is

$$\begin{aligned} \mathrm{Re}\left(\frac{\partial \mathbf{u}^*}{\partial t^*} + (\mathbf{u}^* \cdot \nabla^*)\mathbf{u}^* \right) - \Delta^* \mathbf{u}^* + \nabla^* p^* &= \mathbf{0}, \\ \mathrm{div}^* \mathbf{u}^* &= 0. \end{aligned}$$

Where the inertial terms can be neglecting (for the limiting case $\mathrm{Re} \to 0$), the Stokes equations are obtained:

$$\begin{aligned} -\Delta^* \mathbf{u}^* + \nabla^* p^* &= \mathbf{0}, \\ \mathrm{div}^* \mathbf{u}^* &= 0. \end{aligned}$$

Chapter 2
The Stokes problem

The flow equations introduced in Chapter 1 can be solved numerically using the Galerkin method [54]. In this chapter the linear Stokes problem is first considered. The well-posedness of the continuous problem and the convergence properties of the Galerkin approximation will be presented. The discussion will mainly focus on the differences between the discretization of this type of problems and diffusion-advection-reaction problems.

2.1 Weak formulation of the Stokes problem

We start by considering the (linear) steady Stokes problem

$$-\nu\Delta\mathbf{u} + \nabla p = \mathbf{f}, \qquad \text{in } \Omega, \tag{2.1}$$

$$\text{div}\mathbf{u} = 0, \qquad \text{in } \Omega, \tag{2.2}$$

$$\mathbf{u} = \mathbf{g}, \qquad \text{on } \Gamma_D, \tag{2.3}$$

$$-p\mathbf{n} + \nu\frac{\partial\mathbf{u}}{\partial\mathbf{n}} = \mathbf{d}, \qquad \text{on } \Gamma_N, \tag{2.4}$$

where the pressure p has been rescaled by the fluid density.

Let us consider the following functional spaces:

$$Q = L^2(\Omega) = \{q : \Omega \longrightarrow \mathbb{R} : \int_\Omega q^2\, d\Omega < +\infty\}$$

$$\mathbf{V} = [\mathbf{H}^1(\Omega)]^d = \{\mathbf{v} : \Omega \longrightarrow \mathbb{R}^d : \int_\Omega (\mathbf{v}^2 + |\nabla\mathbf{v}|^2)\, d\Omega < +\infty\}$$

$$\mathbf{V}_0 = [\mathbf{H}^1_{\Gamma_D}(\Omega)]^d = \{\mathbf{v} : \Omega \longrightarrow \mathbb{R}^d : \mathbf{v} \in \mathbf{V}, \mathbf{v} = 0 \text{ on } \Gamma_D\}$$

We look for a weak solution $(\mathbf{u}, p) \in \mathbf{V} \times Q$, with $\mathbf{u} = \mathbf{g}$ on Γ_D such that

$$\int_\Omega \nu\nabla\mathbf{u} : \nabla\mathbf{v}\, d\Omega - \int_\Omega p\,\mathrm{div}\mathbf{v}\, d\Omega = \int_\Omega \mathbf{f}\cdot\mathbf{v}\, d\Omega + \int_{\Gamma_N} \mathbf{d}\cdot\mathbf{v}\, d\gamma,\ \forall\mathbf{v}\in\mathbf{V}_0, \tag{2.5}$$

$$\int_\Omega \mathrm{div}\mathbf{u}\, q\, d\Omega = 0, \quad \forall q\in Q. \tag{2.6}$$

$$\tag{2.7}$$

We introduce the bilinear forms :

$$a : \mathbf{V}\times\mathbf{V} \longrightarrow \mathbb{R} : \quad a(\mathbf{u},\mathbf{v}) = \int_\Omega \nu\nabla\mathbf{u} : \nabla\mathbf{v}\, d\Omega,$$
$$b : \mathbf{V}\times Q \longrightarrow \mathbb{R} : \quad b(\mathbf{v},p) = -\int_\Omega \mathrm{div}\mathbf{v}\, p\, d\Omega,$$

and the linear functional:

$$F(\mathbf{v}) = \int_\Omega \mathbf{f}\cdot\mathbf{v}\, d\Omega + \int_{\Gamma_N} \mathbf{d}\cdot\mathbf{v}\, d\gamma, \tag{2.8}$$

The steady Stokes problem in weak form then reads: *find* $(\mathbf{u},p)\in\mathbf{V}\times Q$, $\mathbf{u}=\mathbf{g}$ *on* Γ_D, *such that*

$$a(\mathbf{u},\mathbf{v}) + b(\mathbf{v},p) = F(\mathbf{v}), \qquad \forall\mathbf{v}\in\mathbf{V}_0, \tag{2.9}$$
$$b(\mathbf{u},q) = 0, \qquad \forall q\in Q. \tag{2.10}$$

The Stokes problem (2.9)-(2.10) is a *saddle-point* problem. Saddle-point problems are well-posed if (see [2, 4, 54]):

1. The bilinear form $a(\mathbf{u},\mathbf{v})$ is continuous and coercive on $\mathbf{V}_0$.
2. The bilinear form $b(\mathbf{v},p)$ is continuous on $\mathbf{V}\times Q$.
3. The *inf-sup* condition holds, namely:

$$\exists\beta>0 \text{ s.t.} \qquad \inf_{\substack{q\in Q\\ q\neq 0}} \sup_{\substack{\mathbf{v}\in\mathbf{V}\\ \mathbf{v}\neq 0}} \frac{b(\mathbf{v},q)}{\|\nabla\mathbf{v}\|_{L^2}\|q\|_{L^2}} \geq \beta. \tag{2.11}$$

We show that these conditions hold true for the Stokes problem (2.9)-(2.10).

The bilinear form $a(\mathbf{u},\mathbf{v})$ is continuous and coercive on $\mathbf{V}_0$ since

$$a(\mathbf{v},\mathbf{v}) = \nu\|\nabla\mathbf{v}\|_{L^2}^2 \geq \frac{\nu}{1+C_\Omega^2}\|\mathbf{v}\|_{H^1}^2 = \alpha\|\mathbf{v}\|_{H^1}^2, \qquad \forall\mathbf{v}\in\mathbf{V}_0,$$
$$a(\mathbf{u},\mathbf{v}) \leq \nu\|\nabla\mathbf{u}\|_{L^2}\|\nabla\mathbf{v}\|_{L^2} \leq \nu\|\mathbf{u}\|_{H^1}\|\mathbf{v}\|_{H^1}, \qquad \forall\mathbf{u},\mathbf{v}\in\mathbf{V}_0,$$

where we have used the equivalence of the H^1 semi-norm and the H^1 norm

$$\frac{1}{\sqrt{1+C_\Omega^2}}\|\mathbf{v}\|_{H^1} \leq \|\nabla\mathbf{v}\|_{L^2} \leq \|\mathbf{v}\|_{H^1}, \tag{2.12}$$

thanks to the Poincaré inequality

$$\|\mathbf{v}\|_{L^2} \leq C_\Omega \|\nabla \mathbf{u}\|_{L^2}. \tag{2.13}$$

The bilinear form $b(\mathbf{v}, p)$ is continuous on $\mathbf{V} \times Q$:

$$b(\mathbf{v}, p) \leq \|p\|_{L^2} \|\mathrm{div}\mathbf{v}\|_{L^2} \leq \sqrt{d}\|p\|_{L^2} \|\nabla \mathbf{v}\|_{L^2} \leq \sqrt{d}\|p\|_{L^2} \|\mathbf{v}\|_{H^1}.$$

The *inf-sup* condition (2.11), also known as LBB (Ladyzhenskaya-Babuška-Brezzi) condition [4], is a direct consequence of the following known result (see [25]):

$$\forall q \in L^2(\Omega),\ \exists \mathbf{w} \in \mathbf{V}_0 :\ -\mathrm{div}\mathbf{w} = q,\ \|\mathbf{w}\|_{H^1} \leq C\|q\|_{L^2}.$$

Indeed, using this result, we have that, for any q, there exists $\mathbf{w}$, such that

$$\frac{b(\mathbf{w}, q)}{\|\mathbf{w}\|_{H^1}} = \frac{\|q\|_{L^2}^2}{\|\mathbf{w}\|_{H^1}} \geq \frac{1}{C}\|q\|_{L^2}.$$

Therefore

$$\sup_{\substack{\mathbf{v} \in \mathbf{V} \\ \mathbf{v} \neq 0}} \frac{b(\mathbf{v}, q)}{\|\mathbf{v}\|_{H^1}} \geq \frac{b(\mathbf{w}, q)}{\|\mathbf{w}\|_{H^1}} \geq \frac{1}{C}\|q\|_{L^2}, \qquad \forall q \in Q,$$

which is equivalent to the inf-sup condition

$$\inf_{\substack{q \in Q \\ q \neq 0}} \sup_{\substack{\mathbf{v} \in \mathbf{V} \\ \mathbf{v} \neq 0}} \frac{b(\mathbf{v}, q)}{\|\mathbf{v}\|_{H^1}\|q\|_{L^2}} \geq \frac{1}{C} = \beta.$$

The well-posedness of the continuous Stokes problem is stated in the following theorem.

Theorem 2.1. *The Stokes problem* (2.9)-(2.10) *has a unique solution* $(\mathbf{u}, p) \in \mathbf{V} \times Q$, $\mathbf{u} = \mathbf{g}$ *on* Γ_D. *Moreover, the solution satisfies the following estimate:*

$$\|\nabla \mathbf{u}\|_{L^2} + \|p\|_{L^2} \leq C\left(\|\mathbf{f}\|_{H^{-1}} + \|\mathbf{d}\|_{H^{-1/2}(\Gamma_N)} + \|\mathbf{g}\|_{H^{1/2}(\Gamma_D)}\right).$$

Remark 2.2. In the case of fully Dirichlet problem ($\Gamma_D = \partial\Omega,\ \Gamma_N = \emptyset$), the pressure is defined up to an arbitrary constant, since it appears in the equation only under spatial derivative and it is not fixed by boundary conditions. In this case, the pressure space is replaced by the following quotient space:

$$Q = L_0^2(\Omega) = L^2(\Omega) \setminus \mathbb{R} \equiv \left\{ q \in L^2(\Omega), \int_\Omega q\, d\Omega = 0 \right\}$$

that is the space of square integrable functions with null average. The condition $\int_\Omega q\, d\Omega = 0$ fixes the arbitrary condition. Moreover, the Dirichlet boundary condition $\mathbf{u} = \mathbf{g}$ on $\partial\Omega$ must satisfy a compatibility condition given by

$$\int_\Omega \mathrm{div}\mathbf{u}\, d\Omega = \int_{\partial\Omega} \mathbf{u}\cdot\mathbf{n}\, d\gamma = \boxed{\int_{\partial\Omega} \mathbf{g}\cdot\mathbf{n}\, d\gamma = 0,}$$

which amounts to set to zero the net flux at the boundary.

Remark 2.3. Analogously, in the case of fully Neumann problem ($\Gamma_N = \partial\Omega,\ \Gamma_D = \emptyset$), the velocity is defined up to a constant vector. In this case, we consider the quotient space:

$$\mathbf{V} = \left\{ \mathbf{v} \in [\mathbf{H}^1(\Omega)]^d, \int_\Omega \mathbf{v}\, d\Omega = \mathbf{0} \right\}.$$

The compatibility condition on the Neumann boundary condition is obtained setting $\mathbf{v} = \mathbf{1}$ in the weak form, namely:

$$\int_\Omega \mathbf{f}\, d\Omega = -\int_{\partial\Omega} \mathbf{d}\, d\gamma.$$

Remark 2.4. We have seen that, to impose the physical normal stress as natural boundary condition, the bilinear form $a(\mathbf{u},\mathbf{v})$ should be rewritten as

$$a(\mathbf{u},\mathbf{v}) = \int_\Omega 2\nu\mathbf{D}(\mathbf{u}) : \nabla\mathbf{v}\, d\Omega = \int_\Omega 2\nu\mathbf{D}(\mathbf{u}) : \mathbf{D}(\mathbf{v})\, d\Omega.$$

The latter is coercive thanks to the Korn inequality

$$\|\nabla\mathbf{u}\|_{L^2} \le K\|\mathbf{D}(\mathbf{u})\|_{L^2}.$$

Moreover, if the problem is fully Neumann, the solution is defined up to an arbitrary rigid motion $\mathbf{w} = \mathbf{t} + \boldsymbol{\xi}\times\mathbf{x}, \forall \mathbf{t}, \boldsymbol{\xi} \in \mathbb{R}^d$, (in $\mathbb{R}^2, \boldsymbol{\xi} = (0,0,\xi_3)$. In this case, the quotient space is given by

$$\mathbf{V} \equiv \left\{ \mathbf{v} \in [\mathbf{H}^1(\Omega)]^d, \int_\Omega \mathbf{v}\cdot\mathbf{w}\, d\Omega = 0, \quad \forall \mathbf{w} = \mathbf{t} + \boldsymbol{\xi}\times\mathbf{x} \right\},$$

and the compatibility conditions are

$$\int_\Omega \mathbf{f}\, d\Omega = -\int_{\partial\Omega} \mathbf{d}\, d\gamma,$$
$$\int_\Omega \mathbf{f}\times\mathbf{x}\, d\Omega = -\int_{\partial\Omega} \mathbf{d}\times\mathbf{x}\, d\gamma.$$

2.2 Stokes problem as constrained minimization problem

If we consider the functional space

$$\mathbf{V}_{\text{div}} = \{\mathbf{v} \in [\mathbf{H}^1(\Omega)]^d, \text{div}\mathbf{v} = 0\}$$

and the linear functional

$$\phi(\mathbf{u}) = \frac{1}{2}\int_\Omega \nu\nabla\mathbf{u} : \nabla\mathbf{u}\, d\Omega - \int_\Omega \mathbf{f}\cdot\mathbf{u}\, d\Omega - \int_{\Gamma_N} \mathbf{d}\cdot\mathbf{u}\, d\gamma,$$

the solution of the Stokes problem is simply given by

$$\mathbf{u} = \arg \min_{\substack{\mathbf{v}\in\mathbf{V}_{\text{div}} \\ \mathbf{v}|_{\Gamma_D}=\mathbf{g}}} \phi(\mathbf{v}).$$

Denoted with q the Lagrange multiplier associated to the incompressibility constraint, let us consider the Lagrangian functional

$$\mathcal{L}(\mathbf{v}, q) = \phi(\mathbf{v}) - \int_\Omega q\,\text{div}\mathbf{v}\, d\Omega.$$

Proposition 2.5. *The solution* $(\mathbf{u}, p)$ *of the Stokes problem is a saddle point of* $\mathcal{L}(\mathbf{v}, q)$*, namely*

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

We can first show that the weak solution of the Stokes problem is a stationary point for the Lagrangian $\mathcal{L}(\mathbf{v}, q)$, namely $\forall \mathbf{v} \in \mathbf{V}, q \in Q$:

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\mathbf{v}} &= \lim_{\epsilon\to 0} \frac{\mathcal{L}(\mathbf{u}+\epsilon\mathbf{v}, p) - \mathcal{L}(\mathbf{u}, p)}{\epsilon} \\
&= \int_\Omega \nu\nabla\mathbf{u} : \nabla\mathbf{v}\, d\Omega - \int_\Omega p\,\text{div}\mathbf{v}\, d\Omega - \int_\Omega \mathbf{f}\cdot\mathbf{v}\, d\Omega - \int_{\partial\Omega} \mathbf{d}\cdot\mathbf{v}\, d\gamma = 0 \\
\frac{\partial\mathcal{L}}{\partial q} &= \lim_{\epsilon\to 0} \frac{\mathcal{L}(\mathbf{u}, p+\epsilon q) - \mathcal{L}(\mathbf{u}, p)}{\epsilon} \\
&= -\int_\Omega q\,\text{div}\mathbf{u}\, d\Omega = 0
\end{aligned}$$

We define the variation of the Lagrangian in direction $(\mathbf{v}, q)$ as

$$\begin{aligned}\nabla\mathcal{L} &= \mathcal{L}(\mathbf{u}+\mathbf{v}, p+q) - \mathcal{L}(\mathbf{u}, p) = \\ &= \int_\Omega \nu\nabla\mathbf{u} : \nabla\mathbf{v}\, d\Omega + \frac{1}{2}\int_\Omega \nu\nabla\mathbf{v} : \nabla\mathbf{v}\, d\Omega - \int_\Omega q\,\mathrm{div}\mathbf{u}\, d\Omega \\ &- \int_\Omega p\,\mathrm{div}\mathbf{v}\, d\Omega - \int_\Omega q\,\mathrm{div}\mathbf{v}\, d\Omega - \int_\Omega \mathbf{f}\cdot\mathbf{v}\, d\Omega - \int_{\Gamma_N} \mathbf{d}\cdot\mathbf{v}\, d\gamma \\ &= \frac{1}{2}\int_\Omega \nu\nabla\mathbf{v} : \nabla\mathbf{v}\, d\Omega - \int_\Omega q\mathrm{div}\mathbf{v}\, d\Omega.\end{aligned}$$

The weak solution is a saddle point of $\mathcal{L}(\mathbf{v}, q)$ since moving away form the stationary point $(\mathbf{u}, p)$:

- along directions $(\mathbf{v}, 0)$ (that is $q = 0$) the Lagrangian increases:

$$\nabla\mathcal{L} = \frac{1}{2}\int_\Omega \nu\nabla\mathbf{v} : \nabla\mathbf{v} \geq 0, \quad \forall\mathbf{v} \in \mathbf{V}_0,$$

- along directions $(0, q)$ (that is $\mathbf{v} = \mathbf{0}$) the Lagrangian is stationary:

$$\nabla\mathcal{L} = 0, \quad \forall q \in Q,$$

- along directions $(\mathbf{v}, \frac{1}{\epsilon}\mathrm{div}\mathbf{v})$ the Lagrangian decreases:

$$\nabla\mathcal{L} = \frac{1}{2}\int_\Omega \nu\nabla\mathbf{v} : \nabla\mathbf{v} - \frac{1}{\epsilon}\int_\Omega (\mathrm{div}\mathbf{v})^2 \leq 0, \quad \text{for } \epsilon \to 0.$$

2.3 Galerkin approximation

Given a polygonal domain Ω, let us introduce a computational grid $\mathcal{T}_h$ and two finite dimensional spaces $\mathbf{V}_h \subset \mathbf{V}$ and $Q_h \subset Q$. Moreover, we set $\mathbf{V}_{h,0} = \mathbf{V}_h \cap \mathbf{V}_0$.

The Galerkin approximation of the Stokes problem (2.9)-(2.10) reads: *find* $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$, $\mathbf{u}_h = \mathbf{g}_h$ *on* Γ_D, *such that*

$$a(\mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p_h) = \mathbf{F}(\mathbf{v}_h), \qquad \forall\mathbf{v}_h \in \mathbf{V}_{h,0}, \tag{2.14}$$

$$b(\mathbf{u}_h, q_h) = 0, \qquad \forall q_h \in Q_h. \tag{2.15}$$

where $\mathbf{g}_h$ is an approximation of the Dirichlet data $\mathbf{g}$ in space $\mathbf{V}_h(\Gamma_D)$ (typically obtained by interpolation or projection).

The discrete problem (2.14)-(2.15) is well-posed if the spaces $\mathbf{V}_h$ and Q_h satisfy the discrete inf-sup condition

$$\forall h \exists \beta_h > 0 \text{ s.t.} \qquad \inf_{\substack{q_h \in Q_h \\ q_h \neq 0}} \sup_{\substack{\mathbf{v}_h \in \mathbf{V}_h \\ \mathbf{v}_h \neq 0}} \frac{b(\mathbf{v}_h, q_h)}{\|\nabla\mathbf{v}_h\|_{L^2}\|q_h\|_{L^2}} \geq \beta_h. \tag{2.16}$$

Remark 2.6. The coercivity and continuity of the bilinear form $a(\mathbf{u},\mathbf{v})$, as well as the continuity of $b(\mathbf{v},p)$, which have been proven to hold in $\mathbf{V}$ and Q, still hold in the discrete subspaces $\mathbf{V}_h$ and Q_h. On the other hand, the fact that the continuous inf-sup condition (2.11) is satisfied for the spaces $\mathbf{V}$ and Q does not imply that the discrete condition is also satisfied for the spaces $\mathbf{V}_h$ and Q_h. For any choice of discrete spaces, this condition should be verified.

Remark 2.7. Typically, the constant β_h should be independent of h. If it is not the case and for instance $\beta_h \to 0$ as $h \to 0$, the convergence properties of the method may be affected.

To better understand the meaning of the inf-sup condition, let us see what happens if it is violated. If

$$\inf_{\substack{q_h \in Q_h \\ q_h \neq 0}} \sup_{\substack{\mathbf{v}_h \in \mathbf{V}_h \\ \mathbf{v}_h \neq 0}} \frac{b(\mathbf{v}_h, q_h)}{\|\nabla \mathbf{v}_h\|_{L^2} \|q_h\|_{L^2}} = 0,$$

then there exists at least a function $q^* \in Q_h$ such that

$$\sup_{\substack{\mathbf{v}_h \in \mathbf{V}_h \\ \mathbf{v}_h \neq 0}} \frac{b(\mathbf{v}_h, q^*)}{\|\nabla \mathbf{v}_h\|_{L^2} \|q^*\|_{L^2}} = 0,$$

and, therefore

$$b(\mathbf{v}_h, q^*) = 0, \quad \forall \mathbf{v}_h \in \mathbf{V}_h. \tag{2.17}$$

Given a solution $(\mathbf{u}_h, p_h)$ of the discrete Stokes problem, if (2.17) holds, then also the pair $(\mathbf{u}_h, p_h + q^*)$ is solution. Indeed, we have

$$\begin{aligned} a(\mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p_h + q^*) = \\ = a(\mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p_h) = \mathbf{F}(\mathbf{v}_h), \qquad \forall \mathbf{v}_h \in \mathbf{V}_{h,0}. \end{aligned}$$

Therefore, the solution is not unique anymore. The functions q^* satisfying (2.17) are referred to as *spurious pressure modes*.

2.4 Finite element algebraic formulation

Let us consider the homogeneous Dirichlet case ($\mathbf{g} = \mathbf{0}$) (to which we can always reduce through a lifting procedure). We introduce a (vector) basis $\{\boldsymbol{\phi}_j\}_{j=1}^{\mathcal{N}_\mathbf{u}}$ for the space $\mathbf{V}_h$ and a (scalar) basis $\{\psi_k\}_{k=1}^{\mathcal{N}_p}$ for the space Q_h, where we have denoted with $\mathcal{N}_\mathbf{u} = \dim(\mathbf{V}_h)$ and $\mathcal{N}_p = \dim(Q_h)$ the dimensions of the finite element spaces of velocity and pressure, respectively.

The solution of the discrete problem can then written as linear combination of the basis functions:

$$\mathbf{u}_h(\mathbf{x}) = \sum_{j=1}^{\mathcal{N}_\mathbf{u}} u_j \boldsymbol{\phi}_j(\mathbf{x}), \qquad p_h(\mathbf{x}) = \sum_{k=1}^{\mathcal{N}_p} p_k \psi_k(\mathbf{x}).$$

In the Galerkin formulation (2.14)-(2.15), we take $\mathbf{v}_h = \boldsymbol{\phi}_i$ and $q_h = \psi_l$ as test functions, and we have, $\forall i = 1, \ldots, \mathcal{N}_\mathbf{u}$ and $\forall l = 1, \ldots, \mathcal{N}_p$:

$$\begin{aligned} a(\mathbf{u}_h, \boldsymbol{\phi}_i) + b(\boldsymbol{\phi}_i, p_h) &= \sum_{j=1}^{\mathcal{N}_\mathbf{u}} u_j a(\boldsymbol{\phi}_j, \boldsymbol{\phi}_i) \\ &\quad + \sum_{k=1}^{\mathcal{N}_p} p_k b(\boldsymbol{\phi}_i, \psi_k) = \mathbf{F}(\boldsymbol{\phi}_i), \qquad (2.18) \\ b(\mathbf{u}_h, \psi_l) &= \sum_{j=1}^{\mathcal{N}_\mathbf{u}} u_j b(\boldsymbol{\phi}_j, \psi_l) = 0. \qquad (2.19) \end{aligned}$$

After introducing the matrices

$$\begin{aligned} A \in \mathbb{R}^{\mathcal{N}_\mathbf{u} \times \mathcal{N}_\mathbf{u}}, \quad A_{ij} &= a(\boldsymbol{\phi}_j, \boldsymbol{\phi}_i) = \int_\Omega \nu \nabla \boldsymbol{\phi}_j : \nabla \boldsymbol{\phi}_i \, d\Omega, \\ B \in \mathbb{R}^{\mathcal{N}_p \times \mathcal{N}_\mathbf{u}}, \quad B_{lj} &= b(\boldsymbol{\phi}_j, \psi_l) = -\int_\Omega \psi_l \, \mathrm{div} \boldsymbol{\phi}_j \, d\Omega, \end{aligned}$$

problem (2.18)-(2.19) can be written in algebraic form as

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} U \\ P \end{bmatrix} = \begin{bmatrix} F \\ 0 \end{bmatrix}, \qquad (2.20)$$

where

$$\begin{aligned} U &= [u_1, \ldots, u_{\mathcal{N}_\mathbf{u}}]^T, \\ P &= [u_1, \ldots, p_{\mathcal{N}_p}]^T, \\ F &= [\mathbf{F}(\boldsymbol{\phi}_1), \ldots, \mathbf{F}(\boldsymbol{\phi}_{\mathcal{N}_\mathbf{u}})]^T. \end{aligned}$$

The Dirichlet boundary conditions can be imposed in system (2.20) in different ways:

1. Set to zero the rows corresponding to the boundary nodes; then, on the same rows, set to one the diagonal term and the right-hand-side to the boundary value.
2. Set to zero not only the rows but also the columns corresponding to the boundary nodes, and correct accordingly the right and side. In this way, the symmetry of the matrix is preserved.
3. Eliminate the degrees of freedoms associated to the boundary nodes. The reduced system will read

$$\begin{bmatrix} \tilde{A} & \tilde{B}^T \\ \tilde{B} & 0 \end{bmatrix} \begin{bmatrix} \tilde{U} \\ P \end{bmatrix} = \begin{bmatrix} \tilde{F} + b_1 \\ b_2 \end{bmatrix}, \tag{2.21}$$

where b_1 and b_2 account for the Dirichlet boundary values. Note that, in general, we will have $b_2 \neq 0$.

The easiest and most common way to build the vector space $\mathbf{V}_h$ is to consider, for each velocity component, a scalar finite element space, such as *e.g.*

$$X_h^r = \{v \in C^0(\Omega), v|_K \in \mathbb{P}_r(K), \forall K \in \mathcal{T}_h\}$$

which is the finite element space of piecewise polynomials of degree r. We then set

$$\mathbf{V}_h = [X_h^r]^d,$$

so that each component of the velocity vector $u_{i,h} \in X_h^r$, $i = 1, \ldots, d$.

Denoted by $\mathcal{N} = \dim(X_h^r)$ the dimension of X_h^r, the dimension of the vector space $\mathbf{V}_h$ will be $\mathcal{N}_\mathbf{u} = \dim(\mathbf{V}_h) = 3\mathcal{N}$. Let $\{\phi_j\}_{j=1}^{\mathcal{N}}$ be a basis $\{\boldsymbol{\phi}_j\}_{j=1}^{\mathcal{N}_\mathbf{u}}$ of X_H^r, then a basis of $\mathbf{V}_h$ can be built as

$$\left\{ \begin{bmatrix} \phi_1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \phi_2 \\ 0 \\ 0 \end{bmatrix}, \ldots, \begin{bmatrix} \phi_\mathcal{N} \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ \phi_1 \\ 0 \end{bmatrix}, \ldots, \begin{bmatrix} 0 \\ \phi_\mathcal{N} \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ \phi_1 \end{bmatrix}, \ldots, \begin{bmatrix} 0 \\ 0 \\ \phi_\mathcal{N} \end{bmatrix} \right\}$$

With this choice of basis, the velocity can be expressed as follows

$$\mathbf{u}_h(\mathbf{x}) = \sum_{j=1}^{\mathcal{N}_\mathbf{u}} u_j \boldsymbol{\phi}_j(\mathbf{x}) = \sum_{j=1}^{\mathcal{N}} u_j \begin{bmatrix} \phi_j \\ 0 \\ 0 \end{bmatrix} + \sum_{j=1}^{\mathcal{N}} u_{\mathcal{N}+j} \begin{bmatrix} 0 \\ \phi_j \\ 0 \end{bmatrix} + \sum_{j=1}^{\mathcal{N}} u_{2\mathcal{N}+j} \begin{bmatrix} 0 \\ 0 \\ \phi_j \end{bmatrix}$$

and the vector of degree-of-freedom reads

$$U = [\underbrace{u_1, \ldots, u_\mathcal{N}}_{1^{\text{st}}\text{comp.}}, \underbrace{u_{\mathcal{N}+1}, \ldots, u_{2\mathcal{N}}}_{2^{\text{nd}}\text{comp.}}, \underbrace{u_{2\mathcal{N}+1}, \ldots, u_{3\mathcal{N}}}_{3^{\text{rd}}\text{comp.}}]$$

Moreover, the matrix A features a block structure:

$$A = \begin{bmatrix} K & 0 & 0 \\ 0 & K & 0 \\ 0 & 0 & K \end{bmatrix}$$

where K is the stiffness matrix of the Laplacian operator with $k_{ij} = \int_\Omega \nu \nabla \phi_j \cdot \nabla \phi_i \, d\Omega$, $i, j = 1, \ldots, \mathcal{N}$. Indeed, all the terms which couple two different velocity components, such as

$$\int_\Omega \nu \nabla \begin{bmatrix} \phi_j \\ 0 \\ 0 \end{bmatrix} : \nabla \begin{bmatrix} 0 \\ \phi_i \\ 0 \end{bmatrix} d\Omega$$

are null.

Remark 2.8. When the bilinear form $a(\mathbf{u},\mathbf{v}) = \int_\Omega \nu \mathbf{D}(\mathbf{u}) : \mathbf{D}(\mathbf{v})\, d\Omega$, associated to normal stress natural conditions, is adopted, the block diagonal structure of A is lost. Indeed, if we consider two basis functions

$$\boldsymbol{\phi}_j = \begin{bmatrix} \phi_j \\ 0 \\ 0 \end{bmatrix}, \qquad \boldsymbol{\phi}_{\mathcal{N}+i} = \begin{bmatrix} 0 \\ \phi_i \\ 0 \end{bmatrix},$$

we have

$$\mathbf{D}(\boldsymbol{\phi}_j) = \begin{bmatrix} \dfrac{\partial \phi_j}{\partial x} & \dfrac{1}{2}\dfrac{\partial \phi_j}{\partial y} & \dfrac{1}{2}\dfrac{\partial \phi_j}{\partial z} \\ \dfrac{1}{2}\dfrac{\partial \phi_j}{\partial y} & 0 & 0 \\ \dfrac{1}{2}\dfrac{\partial \phi_j}{\partial z} & 0 & 0 \end{bmatrix},$$

$$\mathbf{D}(\boldsymbol{\phi}_{\mathcal{N}+i}) = \begin{bmatrix} 0 & \dfrac{1}{2}\dfrac{\partial \phi_i}{\partial x} & 0 \\ \dfrac{1}{2}\dfrac{\partial \phi_i}{\partial x} & \dfrac{\partial \phi_i}{\partial y} & \dfrac{1}{2}\dfrac{\partial \phi_i}{\partial z} \\ 0 & \dfrac{1}{2}\dfrac{\partial \phi_i}{\partial z} & 0 \end{bmatrix},$$

so that the terms coupling two different velocity components

$$A_{\mathcal{N}+i,j} = \int_\Omega \frac{\nu}{2} \frac{\partial \phi_j}{\partial y} \frac{\partial \phi_i}{\partial x}\, d\Omega \neq 0$$

are no longer null.

2.4.1 Inf-sup condition at the algebraic level

At the discrete level, the solution $(\mathbf{u}_h, p_h)$ is identified the two algebraic vectors U and P. The bilinear form b can be reformulated as

$$\begin{aligned} b(\mathbf{u}_h, p_h) &= b\left(\sum_j u_j \boldsymbol{\phi}_j, \sum_l p_l \psi_l\right) \\ &= \sum_{j,l} u_j b(\boldsymbol{\phi}_j, \psi_l) p_l = \\ &= U^T B^T P. \end{aligned}$$

If the inf-sup condition is not satisfied, there exists at least a p_h^* such that $b(\mathbf{v}_h, p_h^*) = 0$, $\forall \mathbf{v}_h \in \mathbf{V}_h$. Therefore, in algebraic terms

$$\exists P^* \in \mathbb{R}^{\mathcal{N}_p}, \quad U^T B^T P = 0, \quad \forall U \in \mathbb{R}^{\mathcal{N}_\mathbf{u}},$$

that is, the spurious pressure modes P^* are the members of the kernel of the matrix B^T:

$$P^* \in \ker(B^T).$$

.

Equivalently, if the inf-sup condition is satisfied, then the kernel of B^T is empty:

$$\ker(B^T) = \emptyset \quad \Rightarrow \quad \text{im}(B) = \ker(B^T)^\perp = \mathbb{R}^{\mathcal{N}_p}$$

so that for each $P \in \mathbb{R}^{\mathcal{N}_p}$, there exists $U \in \mathbb{R}^{\mathcal{N}_\mathbf{u}}$ such that $BU = P$. From the *rank-nullity theorem*, we know that

$$\begin{aligned} \dim \mathbb{R}^{\mathcal{N}_u} &= \mathcal{N}_\mathbf{u} = \text{rank}(B) + \dim(\ker(B)), \\ \dim \mathbb{R}^{\mathcal{N}_p} &= \mathcal{N}_p = \text{rank}(B^T) + \dim(\ker(B^T)). \end{aligned}$$

Since $\dim(\ker(B^T)) = 0$, we have $\text{rank}(B^T) = \text{rank}(B) = \mathcal{N}_p$, and

$$\mathcal{N}_\mathbf{u} - \mathcal{N}_p = \dim(\ker(B)). \tag{2.22}$$

Therefore, when the discrete inf-sup condition is satisfied, there exist $\mathcal{N}_\mathbf{u} - \mathcal{N}_p$ functions $\mathbf{u}_h$ which are divergence-free. Moreover, equation (2.22) implies that when the discrete inf-sup condition is satisfied the discrete space of velocities is larger than the space of pressures.

2.4.2 Stability analysis

In the fully Dirichlet case ($\Gamma_D = \partial\Omega$), a stability result for the Galerkin approximation of the Stokes problem can be obtained by setting $\mathbf{v}_h = \mathbf{u}_h$ and $q_h = p_h$ in (2.18)-(2.19):

$$\begin{aligned} \nu \|\nabla \mathbf{u}_h\|^2 = a(\mathbf{u}_h, \mathbf{u}_h) + b(\mathbf{u}_h, p_h) = F(\mathbf{u}_h) = \int_\Omega \mathbf{f} \cdot \mathbf{u}_h \, d\Omega \\ \leq \|\mathbf{f}\|_{H^{-1}} \|\nabla \mathbf{u}_h\| \end{aligned}$$

where we have defined the H^{-1}-norm as

$$\|\mathbf{f}\|_{H^{-1}} = \sup_{\mathbf{v} \in \mathbf{V}_0} \frac{F(\mathbf{v})}{\|\nabla \mathbf{v}\|}.$$

Therefore an energy estimate on the velocity can be obtained, which reads

$$\|\nabla \mathbf{u}_h\| \leq \frac{1}{\nu} \|\mathbf{f}\|_{H^{-1}}, \quad \forall \mathbf{u}_h \in \mathbf{V}_{h,0}. \tag{2.23}$$

In order to obtain a stability result on the pressure we need to resort the discrete inf-sup condition, as follows:

$$\begin{aligned}\|p_h\| &\leq \frac{1}{\beta}\sup_{\mathbf{v}_h\in\mathbf{V}_{h,0}}\frac{b(\mathbf{v}_h,p_h)}{\|\nabla\mathbf{v}_h\|}\leq\frac{1}{\beta}\sup_{\mathbf{v}_h\in\mathbf{V}_{h,0}}\frac{F(\mathbf{v}_h)-a(\mathbf{u}_h,\mathbf{v}_h)}{\|\nabla\mathbf{v}_h\|}\\ &\leq\frac{1}{\beta}\left(\|\mathbf{f}\|_{H^{-1}}+\nu\|\nabla\mathbf{u}_h\|\right)\end{aligned}\tag{2.24}$$

2.4.3 Convergence analysis

We want to prove a convergence estimate for the finite element approximation of the Stokes problem. We first obtain an estimate which does not make use of the inf-sup condition working in the finite dimensional space

$$\mathbf{V}_{h,\mathrm{div}}\equiv\{\mathbf{v}_h\in\mathbf{V}_h, b(\mathbf{v}_h,q_h)=0,\forall q_h\in Q_h\}$$

We have, $\forall\mathbf{w}_h\in\mathbf{V}_{h,\mathrm{div}},\mathbf{w}_h=\mathbf{g}_h$ on Γ_D

$$\begin{aligned}\nu\|\nabla(\mathbf{u}_h-\mathbf{w}_h)\|^2 &= a(\mathbf{u}_h-\mathbf{w}_h,\mathbf{u}_h-\mathbf{w}_h)\\ &= a(\mathbf{u}_h-\mathbf{u},\mathbf{u}_h-\mathbf{w}_h)+a(\mathbf{u}-\mathbf{w}_h,\mathbf{u}_h-\mathbf{w}_h)\\ &= b(\mathbf{u}_h-\mathbf{w}_h,p-p_h)+a(\mathbf{u}-\mathbf{w}_h,\mathbf{u}_h-\mathbf{w}_h)\\ &= b(\mathbf{u}_h-\mathbf{w}_h,p-\pi_h)+b(\mathbf{u}_h-\mathbf{w}_h,\pi_h-p_h)\\ &\quad +a(\mathbf{u}-\mathbf{w}_h,\mathbf{u}_h-\mathbf{w}_h)\\ &\leq\sqrt{d}\|p-\pi_h\|\|\nabla(\mathbf{u}_h-\mathbf{w}_h)\|\\ &\quad +\nu\|\nabla(\mathbf{u}-\mathbf{w}_h)\|\|\nabla(\mathbf{u}_h-\mathbf{w}_h)\|\end{aligned}$$

$$\Rightarrow\qquad\boxed{\|\nabla(\mathbf{u}_h-\mathbf{w}_h)\|\leq\frac{\sqrt{d}}{\nu}\|p-\pi_h\|+\|\nabla(\mathbf{u}-\mathbf{w}_h)\|}$$

By the triangular inequality

$$\|\nabla(\mathbf{u}-\mathbf{u}_h)\|\leq\|\nabla(\mathbf{u}-\mathbf{w}_h)\|+\|\nabla(\mathbf{u}_h-\mathbf{w}_h)\|,$$

we get the following estimate

$$\|\nabla(\mathbf{u}-\mathbf{u}_h)\|\leq 2\inf_{\mathbf{w}_h\in\mathbf{V}_{h,\mathrm{div}}}\|\nabla(\mathbf{u}-\mathbf{w}_h)\|+\frac{\sqrt{d}}{\nu}\inf_{\pi_h\in Q_h}\|p-\pi_h\|,\tag{2.25}$$

where the last term in the right-hand-side is the *best approximation error* of the pressure in the L^2-norm on Q_h (optimality). If we consider the finite element space of piecewise polynomials of degree $(r-1)$, namely $Q_h=X_h^{r-1}$, we have

$$\inf_{\pi_h \in Q_h} \|p - \pi_h\| \leq C h^r \|p\|_{H^r}.$$

The first term in the right-hand-side of (2.25) is the best approximation error of velocity in the H^1-norm in the divergence-free function space $\mathbf{V}_{h,\mathrm{div}}$. It can be shown that if the spaces $(\mathbf{V}_h, Q_h)$ are compatible, that is they satisfy the inf-sup condition, we have the following result:

$$\inf_{\mathbf{w}_h \in \mathbf{V}_{h,\mathrm{div}}} \|\nabla(\mathbf{u} - \mathbf{w}_h)\| \leq \left(1 + \frac{\sqrt{d}}{\beta_h}\right) \inf_{\mathbf{w}_h \in \mathbf{V}_h} \|\nabla(\mathbf{u} - \mathbf{w}_h)\|$$

Moreover, if the inf-sup condition holds, the following convergence estimate for the pressure can be obtained:

$$\begin{aligned}
\|p_h - \pi_h\| &\leq \frac{1}{\beta_h} \sup_{\mathbf{v}_h \in \mathbf{V}_h} \frac{b(\mathbf{v}_h, p_h - \pi_h)}{\|\nabla \mathbf{v}_h\|} \\
&= \frac{1}{\beta_h} \sup_{\mathbf{v}_h \in \mathbf{V}_h} \frac{a(\mathbf{u} - \mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p - \pi_h)}{\|\nabla \mathbf{v}_h\|} \\
&\leq \frac{1}{\beta_h} \sup_{\mathbf{v}_h \in \mathbf{V}_h} \frac{\nu \|\nabla(\mathbf{u} - \mathbf{u}_h)\| \, \|\nabla \mathbf{v}_h\| + \sqrt{d} \|p - \pi_h\| \, \|\nabla \mathbf{v}_h\|}{\|\nabla \mathbf{v}_h\|} \\
&= \frac{1}{\beta_h} (\nu \|\nabla(\mathbf{u} - \mathbf{u}_h)\| + \sqrt{d} \|p - \pi_h\|).
\end{aligned}$$

Therefore

$$\boxed{\|p - p_h\| \leq \left(1 + \frac{\sqrt{d}}{\beta_h}\right) \inf_{\pi_h \in Q_h} \|p - \pi_h\| + \frac{\nu}{\beta_h} \|\nabla(\mathbf{u} - \mathbf{u}_h)\|}$$

In summary, if the spaces $\mathbf{V}_h$ and Q_h satisfy the inf-sup condition, the Galerkin approximation features the following optimal convergence estimates:

$$\begin{aligned}
\|\nabla(\mathbf{u} - \mathbf{u}_h)\| \leq 2 &\left(1 + \frac{\sqrt{d}}{\beta_h}\right) \inf_{\mathbf{w}_h \in \mathbf{V}_h} \|\nabla(\mathbf{u} - \mathbf{w}_h)\| \\
&+ \frac{\sqrt{d}}{\nu} \inf_{\pi_h \in Q_h} \|p - \pi_h\|,
\end{aligned} \tag{2.26}$$

$$\begin{aligned}
\|p - p_h\| \leq 2 \frac{\nu}{\beta_h} &\left(1 + \frac{\sqrt{d}}{\beta_h}\right) \inf_{\mathbf{w}_h \in \mathbf{V}_h} \|\nabla(\mathbf{u} - \mathbf{w}_h)\| \\
&+ \left(1 + 2\frac{\sqrt{d}}{\beta_h}\right) \inf_{\pi_h \in Q_h} \|p - \pi_h\|.
\end{aligned} \tag{2.27}$$

Remark 2.9. If β_h depends on h (*e.g.*, $\beta_h \to 0$, if $h \to 0$), the Galerkin approximation loses its optimality and even the convergence can be compromised. This effect is more evident on the pressure since $\|p - p_h\| \approx 1/\beta_h^2$.

Remark 2.10. If the inf-sup condition is not satisfied, the first estimate still holds, but it is not clear which are the approximation properties of the space $\mathbf{V}_{h,\mathrm{div}}$. In many cases, optimal convergence on the velocity is still observed.

Remark 2.11. When piecewise continuous finite elements of degree $k+1$ for the velocity and k for the pressure are considered then the error estimates (2.26)-(2.27) yield a convergence rate equal h^{r+1} for both velocity and pressure.

2.4.4 Inf-sup compatible finite-elements

Different possible pairs of finite-element spaces that satisfy the discrete inf-sup condition (2.16) are available. A short overview, limited to triangular (in 2D) or tetrahedric (in 3D) finite-elements, is presented in this section. For a complete and detailed discussion on this subject we refer to [26, 57].

The simplest choice when considering continuous finite-element on triangles would be to adopt piecewise linear polynomial functions for both velocity and pressure, that is

$$\begin{aligned}\mathbf{V}_h &= [X_h^1]^d,\\ Q_h &= X_h^1.\end{aligned}$$

In this case, estimates (2.26)-(2.27) would result in a linear convergence:

$$\|\nabla(\mathbf{u}-\mathbf{u}_h)\| + \|p-p_h\| \leq C\,h.$$

However, no convergence can be obtained since it can be proved that this choice does not respect the inf-sup condition and this is also the case for any equal order finite element pairs, such as, $\forall k$

$$\begin{aligned}\mathbf{V}_h &= [X_h^k]^d,\\ Q_h &= X_h^k.\end{aligned}$$

Inf-sup compatible finite-element pairs can be obtained choosing piecewise polynomial with a higher degree for the velocity. For instance $\mathbb{P}_2/\mathbb{P}_1$ finite-element satisfy the discrete inf-sup condition and yield a quadratic convergence:

$$\|\nabla(\mathbf{u}-\mathbf{u}_h)\| + \|p-p_h\| \leq C\,h^2.$$

A higher order extension of the $\mathbb{P}_2/\mathbb{P}_1$ finite-elements are given by the Taylor-Hood $\mathbb{P}_{k+1}/\mathbb{P}_k$ finite-elements defined as:

$$\begin{aligned}\mathbf{V}_h &= [X_h^{k+1}]^d,\\ Q_h &= X_h^k,\end{aligned}$$

for any $k \geq 1$, which guarantees an optimal convergence rate also for higher degrees and yields:

$$\|\nabla(\mathbf{u} - \mathbf{u}_h)\| + \|p - p_h\| \leq C\, h^{k+1}.$$

Nevertheless, the lowest degree case, that is the $\mathbb{P}_1/\mathbb{P}_0$ finite-elements do not fulfil the discrete inf-sup condition.

Additional example of inf-sup compatible finite-element pairs are:

- the $\mathbb{P}_1\text{iso}\mathbb{P}_2/\mathbb{P}_1$ finite-elements where linear elements are used for the pressure and linear elements on a finer nested subgrid are used for the velocity, namely:

$$\mathbf{V}_h = [X^1_{h/2}]^d,$$
$$Q_h = X^1_h,$$

- the $\mathbb{P}_1\text{bubble}/\mathbb{P}_1$ finite elements where the linear velocity space is enriched by additional degrees of freedom (the *bubbles*) which are zero at each element boundary and is either cubic or piecewise linear inside the element.

Both the $\mathbb{P}_1\text{iso}\mathbb{P}_2/\mathbb{P}_1$ finite elements and the $\mathbb{P}_1\text{bubble}/\mathbb{P}_1$ finite elements guarantee a linear convergence rate.

2.4.5 Inf-sup stabilization methods

In certain circumstances, most often for implementation reasons, it would be extremely useful to be able to work with equal order finite-elements, which unfortunately are not stable. Moreover, we have seen the most natural low-order schemes (e.g. $\mathbb{P}_1/\mathbb{P}_0$ or $\mathbb{P}_1/\mathbb{P}_1$) are not stable. Increasing the polynomial degree would result in additional computational costs and may be not the optimal strategy if the solution is not regular enough. Stabilization strategies have been proposed in the literature to make it possible to work with equal order finite-elements fixing the ill-posedness associated to the violation of the discrete inf-sup condition. In this section we introduce the main ideas underlying the most popular stabilizations which are briefly recalled.

We consider a stabilization term (to be added to the standard Galerkin formulation) defined, in general, as a symmetric positive bilinear form on Q_h, denoted as

$$s(p, q) : Q_h \times Q_h \longrightarrow \mathbb{R}.$$

The stabilized discrete formulation then reads: *find* $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$, $\mathbf{u}_h = \mathbf{g}_h$ *on* Γ_D, *such that*

$$a(\mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p_h) = \mathbf{F}(\mathbf{v}_h), \qquad \forall \mathbf{v}_h \in \mathbf{V}_{h,0}, \tag{2.28}$$

$$b(\mathbf{u}_h, q_h) - s(p_h, q_h) = 0, \qquad \forall q_h \in Q_h. \tag{2.29}$$

We consider the case $\mathbf{g}_h = 0$. Taking $\mathbf{v}_h = \mathbf{u}_h$ and $q_h = p_h$, and subtracting (2.29) from (2.28), we get

$$\nu \|\nabla \mathbf{u}_h\|^2 + s(p_h, p_h) = a(\mathbf{u}_h, \mathbf{u}_h) + s(p_h, p_h) \leq F(\mathbf{u}_h),$$

so that we recover a control not only on the velocity but also on the pressure, without resorting to the discrete inf-sup condition. This implies the uniqueness of the solution also in pressure.

At the algebraic level, the stabilization modifies the Stokes matrix adding an additional block, as follows:

$$\begin{bmatrix} A & B^T \\ B & -R \end{bmatrix} \begin{bmatrix} U \\ P \end{bmatrix} = \begin{bmatrix} F \\ 0 \end{bmatrix}, \tag{2.30}$$

where

$$R \in \mathbb{R}^{\mathcal{N}_p \times \mathcal{N}_p}, \quad R_{ij} = s(\psi_j, \psi_i), \qquad i, j = 1, \ldots, \mathcal{N}_p.$$

The choice of the stabilization term $s(\cdot, \cdot)$ is critical since it should be able to guarantee a uniform convergence to the method. Stability and convergence results of stabilized methods can be obtained [16, 6] under the following conditions on the stabilization term:

- *generalized inf-sup condition*, $\exists \beta_h > 0$ s.t.:

$$\sup_{\substack{\mathbf{v}_h \in \mathbf{V}_h \\ \mathbf{v}_h \neq 0}} \frac{b(\mathbf{v}_h, q_h)}{\|\nabla \mathbf{v}_h\|} + s(q_h, q_h)^{1/2} \geq \beta_h \|q_h\|, \quad \forall q_h \in Q_h. \tag{2.31}$$

- *consistency*:

$$s(p_h, q_h) \longrightarrow 0, \quad \text{as } h \longrightarrow 0$$

Optimal convergence rates are guaranteed only if $s(p, q)$ goes to zero as fast as the best approximation error for velocity and pressure, namely

$$\inf_{\mathbf{w}_h \in \mathbf{V}_h} \|\mathbf{u} - \mathbf{w}_h\| + \inf_{\pi_h \in Q_h} \|p - \pi_h\|.$$

This is usually possible only for *strongly consistent stabilization* for which we have $s(p, p) = 0$.

Two classical examples of stabilization are the following:

- *Brezzi-Pitkaranta stabilization* [5]:

$$s(p_h, q_h) = \delta \sum_{K \in \mathcal{T}_h} h_K^2 \int_K \nabla p_h \cdot \nabla q_h \, d\Omega$$

where δ is a stabilization coefficient and h_K is the diameter of element K. This stabilization is uniformly stable for any equal order space pair but, since it guarantees only a linear convergence rate, it's usually worth using with linear finite-element $\mathbb{P}_1/\mathbb{P}_1$, while it's not optimal for higher degree spaces $\mathbb{P}_k/\mathbb{P}_k$ with $k > 1$.

- *Interior-Penalty (IP) stabilization* [7]:

$$s(p_h, q_h) = \delta \sum_{K \in \mathcal{T}_h} h_K^3 \int_{\partial K \setminus \partial \Omega} [\nabla p_h] \cdot [\nabla q_h] \, d\gamma$$

where $[\cdot]$ denotes the jump operator across the interior boundaries. This stabilization is strongly consistent ($s(p, p) = 0$ as long as $p \in H^2(\Omega)$) and yields optimal convergence rates for any equal order finite-element pairs $\mathbb{P}_k/\mathbb{P}_k$.

A common issues in many stabilization procedure is the choice of the stabilization parameter δ. Indeed, optimal values can be obtained a-priori only for very simple geometries. In most cases, some numerical tuning of the parameter has to be carried out to avoid either too small values (which will result in a lack of stability) or to large values (which will give poorly accurate results).

Other strongly consistent stabilizations, such as Streamline Upwind Petrov Galerkin (SUPG) and Galerkin Least Square (GaLS) that will be introduced in Section 3.5 for the stabilization of advection-dominated problems are also effective in the stabilization for the inf-sup condition. Moreover, the equivalence between the $\mathbb{P}_1$bubble$/\mathbb{P}_1$ finite-elements and the $\mathbb{P}_1/\mathbb{P}_1$ with GaLS stabilization has been established.

2.5 Solution of the algebraic Stokes problem

In this section we will describe the different approaches that can be adopted for the solution of the linear system obtained from the discretization of the Stokes problem.

The algebraic formulation of the Stokes problem, in the most general stabilized form, reads:

$$\begin{bmatrix} A & B^T \\ B & -R \end{bmatrix} \begin{bmatrix} U \\ P \end{bmatrix} = \begin{bmatrix} F \\ 0 \end{bmatrix}, \tag{2.32}$$

where R is the stabilization matrix and $R = 0$ if inf-sup compatible spaces are used.

The main objective of this section is to highlight the peculiarity of the linear system(2.32) and to characterize the spectral properties of these block structure matrices in order to devise optimal preconditioners. In this respect, only iterative methods will be considered since they are the most commonly

used for the large linear systems typically arising in the approximation of this kind of differential models.

In the following, we will consider the case of inf-sup compatible spaces ($R = 0$), however most of the results on the preconditioning of the Stokes system can be extended to the stabilized case.

2.5.1 Pressure matrix method

The first approach for the solution of system (2.32) is known as *pressure matrix method* and consists in reducing the system to a problem for the pressure. We Given the Stokes system

$$\begin{aligned} AU + B^T P &= F \\ BU &= 0 \end{aligned}$$

we proceed by formally computing the velocity from the first equation

$$U = A^{-1}(F - B^T P) \tag{2.33}$$

and replacing it in the second equation

$$BA^{-1}(F - B^T P) = 0.$$

We obtain the following algebraic problem for the pressure

$$(BA^{-1}B^T)P = BA^{-1}F. \tag{2.34}$$

The matrix $S = BA^{-1}B^T$, known as the *pressure Schur complement* of the Stokes matrix is a symmetric positive definite matrix.

The solution of the problem requires to compute the pressure solving the linear system (2.34) and then the velocity can be computing through equation (2.33) by solving a system on the (vectorial) stiffness matrix A. While the iterative solution of latter system can be successfully preconditioned by a standard preconditioner (such as, *e.g.*, the Incomplete LU (ILU) or multigrid (MG)), a suitable preconditioner should be identified for the Schur complement.

2.5.1.1 Spectral characterization of the Schur complement

We have already shown that the bilinear form b can be rewritten in algebraic form as:

$$b(\mathbf{u}_h, p_h) = U^T B^T P.$$

Similarly, we can formulate in algebraic form the following quantities

$$\begin{aligned}\nu\|\nabla\mathbf{u}_h\|^2 &= \int_\Omega \nu\nabla\mathbf{u}_h : \nabla\mathbf{u}_h \, d\Omega = a\left(\sum_j u_j\boldsymbol{\phi}_j, \sum_i u_i\boldsymbol{\phi}_i\right) = \\ &= U^T AU = \|A^{1/2}U\|^2, \\ \|p_h\|^2 &= \int_\Omega p_h p_h \, d\Omega = \int_\Omega \left(\sum_k p_k\psi_k\right)\left(\sum_l p_l\psi_l\right) d\Omega = \\ &= P^T M_p P = \|M_p^{1/2}P\|^2\end{aligned}$$

where M_p is the pressure mass matrix:

$$M_p \in \mathbb{R}^{\mathcal{N}_p\times\mathcal{N}_p}, \qquad (M_p)_{lk} = \int_\Omega \psi_k\psi_l \, d\Omega.$$

We note that, $\forall p_h \in P_h$,

$$\begin{aligned}\sup_{\mathbf{u}_h\in V_h} \frac{1}{\sqrt{\nu}}\frac{b(\mathbf{u}_h, p_h)}{\|\nabla\mathbf{u}_h\|} &= \sup_{U\in\mathbb{R}^{\mathcal{N}_\mathbf{u}}} \frac{U^T B^T P}{\|A^{1/2}U\|} = \\ &= \sup_{U\in\mathbb{R}^{\mathcal{N}_\mathbf{u}}} \frac{U^T A^{1/2}A^{-1/2}B^T P}{\|A^{1/2}U\|} \\ &\le \sup_{U\in\mathbb{R}^{\mathcal{N}_\mathbf{u}}} \frac{\|A^{1/2}U\|\,\|A^{-1/2}B^T P\|}{\|A^{1/2}U\|} \\ &= \sqrt{P^T BA^{-1}B^T P}\end{aligned}$$

On the other hand, taking $U = A^{-1}B^T P$, we have

$$\sup_{U\in\mathbb{R}^{\mathcal{N}_\mathbf{u}}} \frac{U^T B^T P}{\|A^{1/2}U\|} \ge \frac{P^T BA^{-1}B^T P}{\sqrt{P^T BA^{-1}B^T P}} = \sqrt{P^T BA^{-1}B^T P}.$$

It follows that

$$\sup_{U\in\mathbb{R}^{\mathcal{N}_\mathbf{u}}} \frac{U^T B^T P}{\|A^{1/2}U\|} = \sqrt{P^T BA^{-1}B^T P}. \tag{2.35}$$

and therefore

$$\beta_h^2 \le \left(\inf_{p_h\in Q_h}\sup_{\mathbf{u}_h\in V_h} \frac{b(\mathbf{u}_h, p_h)}{\|\nabla\mathbf{u}_h\|\,\|p_h\|}\right)^2 = \inf_{P\in\mathbb{R}^{\mathcal{N}_p}} \frac{P^T BA^{-1}B^T P}{P^T\left(\frac{1}{\nu}M_p\right)P}.$$

The inf-sup condition can be associated to the following generalized eigenvalue problem

$$BA^{-1}B^T P = \lambda\left(\frac{1}{\nu}M_p\right)P.$$

Therefore, the generalized eigenvalues of $BA^{-1}B^T$ with respect to $\frac{1}{\nu}M_p$ are bounded from below by

$$\beta_h^2 \leq \lambda_{\min}.$$

Note that the matrix $BA^{-1}B^T$ is symmetric positive definite.

Recalling the continuity of the bilinear form $b(\mathbf{v}_h, p_h)$

$$b(\mathbf{v}_h, p_h) \leq \sqrt{d}\|\nabla \mathbf{v}_h\| \, \|p_h\|, \qquad \forall \mathbf{v}_h \in \mathbf{V}_h, p_h \in Q_h,$$

we have:

$$U^T B^T P \leq \sqrt{\frac{d}{\nu}} \|A^{1/2} U\| \, \|M_P^{1/2} P\|.$$

Therefore

$$\sqrt{P^T B A^{-1} B^T P} = \sup_{U \in \mathbb{R}^{\mathcal{N}_\mathbf{u}}} \frac{U^T B^T P}{\|A^{1/2} U\|} \leq \sqrt{\frac{d}{\nu}} \|M_P^{1/2} P\|$$

and the following upper bound can be established

$$\sup_{P \in \mathbb{R}^{\mathcal{N}_p}} \frac{P^T B A^{-1} B^T P}{P^T \frac{1}{\nu} M_P P} \leq d, \qquad \Rightarrow \boxed{\lambda_{\max} \leq d.}$$

In conclusion, we have

$$\beta_h^2 \leq \frac{P^T B A^{-1} B^T P}{P^T \frac{1}{\nu} M_P P} \leq d, \tag{2.36}$$

which shows that, if the inf-sup constant β_h is independent of h, then the matrix $BA^{-1}B^T$ is spectrally equivalent to $\frac{1}{\nu} M_p$.

Moreover, since

$$C h_{\min}^d \|P\| \leq P^T M_P P \leq C_2 h_{\max}^d \|P\|$$

the spectrum of $BA^{-1}B^T$ can be bounded by

$$\frac{C\beta_h^2}{\nu} h_{\min}^d \leq \lambda(BA^{-1}B^T) \leq \frac{d}{\nu} h_{\max}^d,$$

and, therefore, its condition number is given by

$$\mathcal{K}(BA^{-1}B^T) \leq \frac{d}{\beta_h^2} \frac{C_2}{C_1} \left(\frac{h_{\max}}{h_{\min}} \right)^d$$

.

The same estimate can be proven for the stabilized case ($R \neq 0$) based on the generalized inf-sup condition.

The SPD algebraic pressure problem (2.34) can be solved using the Conjugate Gradient method. If a uniform grid is considered, no preconditioning is needed since the condition number is independent of h and, therefore, the number of iteration required by the Conjugate Gradient will also be inde-

pendent of h. For arbitrary grids, we can get rid of the h-dependence of the eigenvalues $\lambda_{\min}$ and $\lambda_{\max}$ resorting to the (scaled) mass pressure matrix $\frac{1}{\nu}M_P$ as preconditioner.

Indeed, in this case, since the eigenvalues of the preconditioned matrix $\nu M_P^{-1}S$ are equivalent to the generalized eigenvalues of S with respect to $\frac{1}{\nu}M_P$

$$\nu M_P^{-1}BA^{-1}B^TP = \lambda P \quad \longrightarrow \quad BA^{-1}B^TP = \lambda\left(\frac{1}{\nu}M_P\right)P,$$

they are independent of h:

$$\beta_h^2 \leq \lambda \leq d.$$

Therefore, the matrix $\dfrac{1}{\nu}M_P$ is an optimal preconditioner for the Schur complement.

Note that, we can even do better considering as preconditioner the diagonal of the mass pressure matrix, $\hat{M} = \dfrac{1}{\nu}\mathrm{diag}(M_P)$ which is also optimal since it exhibit the same scaling as M_P, namely

$$\hat{M}_{ij} = \frac{1}{\nu}\int_\Omega \psi_j\psi_i \approx \mathcal{O}\left(h^d\right).$$

The solution by the preconditioned Conjugate Gradient method requires matrix-vector products and the computation of the preconditioned residual. Let us analyze the computational effort required by those two operations:

- **Matrix-vector product**: each product $Y = SX = BA^{-1}B^TX$ can be developed as

$$\begin{aligned} W &= B^TX \\ AU &= W \\ Y &= U \end{aligned}$$

  where the solution of the linear system for the velocity is required. When the Laplacian form $-\nu\Delta\mathbf{u}$ of the diffusion term is used, three Laplacian algebraic problems (one for each velocity component) of size $\mathcal{N}\times\mathcal{N}$ are solved. When, instead, the $-\mathrm{div}(\nu\mathbf{D}(\mathbf{u}))$ form is used one should solve a single coupled problem of size $3\mathcal{N}\times 3\mathcal{N} = \mathcal{N}_\mathbf{u}\times\mathcal{N}_\mathbf{u}$.

- **Solution of the preconditioner**: solving $\hat{M}Z = R$ becomes trivial when the diagonal preconditioner $\hat{M} = \dfrac{1}{\nu}\mathrm{diag}(M_P)$ is used.

2.5.2 Uzawa method

A classical iterative method for the solution of saddle-point problems is the so-called *Uzawa method* [17], which requires, at each iteration, the following two steps:

1. solve the unconstrained differential problem;
2. correct the Lagrange multiplier projecting the solution of step 1 on the constrained space.

The Uzawa method applied to the Stokes problem reads: given p_0, for $k = 1, 2, \ldots$

$$\begin{aligned} -\nu\Delta\mathbf{u}^k &= \mathbf{f} - \nabla p^{k-1}, \\ p^k - p^{k-1} &= \alpha \mathrm{div}\mathbf{u}^k. \end{aligned}$$

The algebraic counterpart, in the case of stable velocity/pressure pairs, is the following:

$$AU^k = F - B^T P^{k-1}, \tag{2.37}$$

$$M_P \left(P^k - P^{k-1}\right) = \alpha B U^k. \tag{2.38}$$

Computing U^k from (2.37) and substituting it into (2.38) leads to

$$\begin{aligned} M_P \left(P^k - P^{k-1}\right) &= \alpha B A^{-1} \left(F - B^T P^{k-1}\right) = \\ &= \alpha(BA^{-1}F - BA^{-1}B^T P^{k-1}) = \alpha R^{k-1}, \end{aligned}$$

where R^{k-1} denotes the residual of the pressure equation. At the algebraic level, the Uzawa method can be interpreted as the Richardson method applied to the pressure problem. We recall that the convergence of the Richardson method is guaranteed under the following condition:

$$\alpha \leq \frac{2}{\lambda_{\max}} = \frac{2\nu}{d}$$

.

The Conjugate Gradient method applied to the solution of the pressure problem can then be seen as the natural evolution of the Uzawa method in order to exploit the SPD nature of the Schur complement.

2.5.3 Brief review of Krylov methods

Before analyzing how the monolithic Stokes matrix can be preconditioned, let us briefly recall the main linear algebra results underlying the so-called *Krylov methods* for the iterative solution of linear systems [54].

Definition 2.12. Given a matrix $A \in \mathbb{R}^{N\times N}$, its *characteristic polynomial* is defined as:

$$p_N(\lambda) = \det(A - \lambda I) = \sum_{i=0}^{N} c_i \lambda^i, \qquad c_0 = det(A).$$

Theorem 2.13 (Cayley-Hamilton). *Any square matrix satisfies its own characteristic polynomial, that is*

$$p_N(A) = 0, \quad \forall A \in \mathbb{R}^{N\times N}.$$

Proposition 2.14. *If $A \in \mathbb{R}^{N\times N}$ is invertible, A^{-1} is a polynomial of degree $N-1$ of A.*

Indeed, we have

$$p_N(A) = 0 \quad \longrightarrow \quad A^{-1}p_N(A) = A^{-1}\det(A) + \sum_{i=1}^{N} c_i A^{i-1} = 0,$$

from which it follows

$$A^{-1} = -\frac{1}{\det(A)} \sum_{i=0}^{N-1} c_{i+1} A^i = p_{N-1}(A).$$

Proposition 2.15. *Given a linear system $AU = F$ and an initial vector U_0, let $R_0 = F - AU_0$ denote the initial residual and $\mathbf{E}_0 = U - U_0$ the initial error, then we have*

$$E_0 = A^{-1}R_0 = p_{N-1}(A)R_0 = -\frac{1}{\det(A)} \sum_{i=0}^{N-1} c_{i+1} A^i R_0.$$

We introduce the Krylov space generated by the matrix A starting from the initial vector R_0:

$$K_n(A, R_0) = \text{span}\{R_0, AR_0, \ldots, A^{n-1}R_0\}$$

and we note that $E_0 \in K_N(A, R_0)$.

The Krylov methods for the solution of linear system are based on the construction of a sequence U_k such that, $\forall k = 1, 2, \ldots$:

$$U_k - U_0 \in K_k(A, R_0),$$

or, analogously

$$R_k - R_0 = A(U_k - U_0) \in \text{span}\{AR_0, A^2R_0, \ldots, A^kR_0\}.$$

Depending on the properties of the matrix A and the criteria used to build the sequence U_k, different Krylov method can be defined. Among the most relevant we mention:

- the **Conjugate Gradient method** (for symmetric positive definite matrices), with U_k defined such that

$$\|E_k\|_{A^1} = \|U - U_k\|_A \leq \|V\|_A, \quad \forall V \in K_k(A, R_0)$$

  or equivalently

$$\|R_k\|_{A^{-1}} \leq \|V\|_{A^{-1}}, \quad \forall V \in R_0 + \text{span}\{AR_0, A^2R_0, \ldots, A^kR_0\};$$

- the **MINRES method** (for symmetric matrices) and the **GMRES method** (for arbitrary matrices) in which U_k is defined such that

$$\|R_k\| \leq \|V\|, \quad \forall V \in R_0 + \text{span}\{AR_0, A^2R_0, \ldots, A^kR_0\}.$$

In order to analyses the convergence properties of the different Krylov methods, we first remark that

$$R_k \in R_0 + \text{span}\{AR_0, A^2R_0, \ldots, A^kR_0\}$$

can be expressed as

$$R_k = R_0 + \sum_{i=1}^{k} c_i A^i R_0 = p_k(A)R_0, \quad \text{with } p_k(0) = 1$$

therefore minimizing the residual R_k in a given norm $\|\cdot\|_\star$ implies that

$$\|R_k\|_\star \leq \min_{\substack{p_k \in \mathbb{P}_k \\ p_k(0)=1}} \|p_k(A)R_0\|_\star. \tag{2.39}$$

Since $p_N(A) = 0$, an important direct consequence is that methods which minimize the residual are exact in N steps.

Similarly the error can be expressed as

$$E_k = A^{-1}R_k = E_0 + \sum_{i=1}^{k} c_i A^i E_0 = p_k(A)E_0, \quad \text{with } p_k(0) = 1$$

thus we can minimize E_k in a given norm $\|\cdot\|_\star$ by taking

$$\|E_k\|_\star \leq \min_{\substack{p_k \in \mathbb{P}_k \\ p_k(0)=1}} \|p_k(A)E_0\|_\star. \tag{2.40}$$

When the Conjugate Gradient method is considered, the residual is minimized in the norm $\|\cdot\|_\star = \|\cdot\|_{A^{-1}}$. This is equivalent to minimize the error

in the A-norm. If V_i are the right eigenvectors of A, then

$$A = VDV^{-1},$$

where $D = \text{diag}(\lambda_1, \ldots, \lambda_N)$ and $V = [V_1, \ldots, V_N]$. Developing the initial error as

$$E_0 = \sum_{i=1}^{N} \alpha_i V_i,$$

we have

$$\begin{aligned}\|p_k(A)E_0\|_A &= \|\sum_{i=1}^{N} \alpha_i p_k(A) V_i\|_A = \|\sum_{i=1}^{N} \alpha_i p_k(\lambda_i) V_i\|_A \\ &\leq \max_i(p_k(\lambda_i))\|\sum_{i=1}^{N} \alpha_i V_i\|_A = \max_i(p_k(\lambda_i))\|E_0\|_A\end{aligned}$$

Combining (2.39) and (2.41), we get

$$\|R_k\|_{A^{-1}} = \|E_k\|_A \leq \left(\min_{\substack{p_k \in \mathbb{P}_k \\ p_k(0)=1}} \max_i |p_k(\lambda_i)| \right) \|E_0\|_A \tag{2.41}$$

$$= \left(\min_{\substack{p_k \in \mathbb{P}_k \\ p_k(0)=1}} \max_i |p_k(\lambda_i)| \right) \|R_0\|_{A^{-1}}. \tag{2.42}$$

A similar result can be obtained when considering GMRES (or MINRES) method, minimizing the residual in the Euclidean norm $\|\cdot\|_\star = \|\cdot\|$:

$$\begin{aligned}\|p_k(A)R_0\| &= \|V p_k(D) V^{-1} R_0\| \leq \|V\|\|V^{-1}\|\|p_k(D)\|\|R_0\| \\ &\leq \mathcal{K}(V) \max_i |p_k(\lambda_i)|\|R_0\|,\end{aligned} \tag{2.43}$$

where $\mathcal{K}(V)$ is the condition number of A. Combining (2.39) and (2.43), we get

$$\|R_k\| \leq \mathcal{K}(V) \left(\min_{\substack{p_k \in \mathbb{P}_k \\ p_k(0)=1}} \max_i |p_k(\lambda_i)| \right) \|R_0\|. \tag{2.44}$$

The asymptotic convergence rate ρ is defined as

$$\|R_k\| \approx \rho^k \|R_0\|,$$

and for the GMRES method is given by

$$\rho = \left(\frac{\|R_k\|}{\|R_0\|}\right)^{1/k} = \mathcal{K}(V)^{1/k} \left(\min_{\substack{p_k \in \mathbb{P}_k \\ p_k(0)=1}} \max_i |p_k(\lambda_i)| \right)^{1/k}.$$

The iterative method will converge fast as long as the eigenvalues are clustered so that a low degree polynomial $p_k(\lambda_i)$ will guarantee a small error (see Figure 2.1).

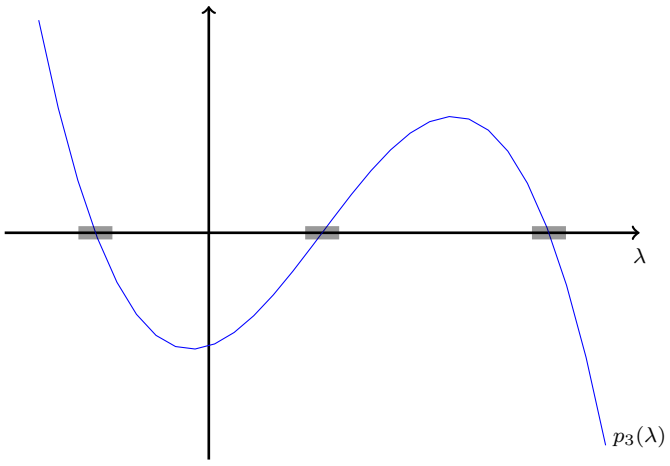


Fig. 2.1: When the eigenvalues are clustered (in the grey intervals) good convergence can be achieved after a few iterations (in this case 3), since $\max_i |p_3(\lambda_i)| << 1$.

2.5.4 Solution of the global Stokes system

We have seen that the Conjugate Gradient method can be successfully applied to the solution of the SPD problem arising in the pressure matrix method. To solve the algebraic Stokes system as a global problem, we may consider the following two formulations:

$$\underbrace{\begin{bmatrix} A & B^T \\ B & -R \end{bmatrix}}_{\Sigma_1} \begin{bmatrix} U \\ P \end{bmatrix} = \begin{bmatrix} F \\ 0 \end{bmatrix}, \qquad \underbrace{\begin{bmatrix} A & B^T \\ -B & R \end{bmatrix}}_{\Sigma_2} \begin{bmatrix} U \\ P \end{bmatrix} = \begin{bmatrix} F \\ 0 \end{bmatrix}, \tag{2.45}$$

where the two systems differ by just the sign of the second equation and give rise to two global matrices Σ_1 and Σ_2 with different properties.

In the first formulation, the matrix Σ_1 is symmetric indefinite, indeed

$$\begin{bmatrix} U^T & P^T \end{bmatrix} \Sigma_1 \begin{bmatrix} U \\ P \end{bmatrix} = \begin{bmatrix} U^T & P^T \end{bmatrix} \begin{bmatrix} AU + B^T P \\ BU - RP \end{bmatrix} = \underbrace{U^T AU}_{\geq 0} + \underbrace{2P^T BU}_{?} - \underbrace{P^T RP}_{\leq 0}.$$

On the other hand, the matrix Σ_2 is positive semi-definite but non-symmetric:

$$\begin{bmatrix} U^T & P^T \end{bmatrix} \Sigma_2 \begin{bmatrix} U \\ P \end{bmatrix} = \begin{bmatrix} U^T & P^T \end{bmatrix} \begin{bmatrix} AU + B^T P \\ -BU + RP \end{bmatrix} = U^T AU + P^T RP \geq 0.$$

In neither case the Conjugate Gradient can be used, however the MINRES and GMRES methods can be used to solve the Σ_1 and Σ_2 problems, respectively. In both case, a suitable preconditioning strategy should be adopted. It turns out that an ideal preconditioner for the global Stokes problem is given by the block diagonal matrix

$$\mathcal{P} = \begin{bmatrix} A & 0 \\ 0 & \frac{1}{\nu} M_P \end{bmatrix}. \tag{2.46}$$

As discussed in Section 2.5.1 for the pressure matrix method, any Krylov-based iterative method requires matrix-vector products and the computation of the preconditioned residual. When systems like (2.45) are solved using the block preconditioner (2.46), the following operations have to be performed:

- **Matrix-vector product**: each product

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} A & B^T \\ B & -R \end{bmatrix} \begin{bmatrix} U \\ P \end{bmatrix}$$

  can be performed without the need of any solution of linear system (while in the pressure matrix method we needed to solve a linear system in A).
- **Solution of the preconditioner**: solving for the preconditioned residual

$$\begin{bmatrix} A & 0 \\ 0 & \frac{1}{\nu} M_P \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} R_1 \\ R_2 \end{bmatrix}$$

  requires the solution of one linear system in A and one linear system in $\frac{1}{\nu} M_P$.

Accounting for the most computational expansive operations (which are the solution of linear system) we may conclude that one iteration of the PCG method applied to Schur complement in the pressure matrix method has a similar computational cost as one iteration of the preconditioned GMRES method applied to the monolithic Stokes systems. However, in this latter case since the linear solution for the matrix A appears in the preconditioning step, one could further improve the computational efficiency by replacing $\mathcal{P}$ with

$$\hat{\mathcal{P}} = \begin{bmatrix} \hat{A} & 0 \\ 0 & \hat{M}_P \end{bmatrix}.$$

where $\hat{A}$ and $\hat{M}_P$ are suitable preconditioners of A and $\frac{1}{\nu}M_P$, respectively. Among the many different possible choices for the preconditioner of the Laplacian $\hat{A}$ we mention those based on multigrid, domain decomposition or incomplete LU factorization.

To prove the optimality of the block diagonal preconditioner, let us consider the case with the symmetric matrix Σ_1. We need to show that the generalized eigenvalues $\mu_i, i = 1, \ldots, N_{\mathbf{u}} + N_p$, of Σ_1 with respect to $\mathcal{P}$, defined by

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} U \\ P \end{bmatrix} = \mu_i \begin{bmatrix} A & 0 \\ 0 & \frac{1}{\nu}M_P \end{bmatrix} \begin{bmatrix} U \\ P \end{bmatrix} \tag{2.47}$$

are bounded independently of h.

The generalized eigenvalues μ_i can be evaluated in two subset. The first subset is obtained considering the case in which $P = 0$ and $U \in \ker(B)$, such that the generalized eigensystem (2.47) reduces to

$$\begin{aligned} AU &= \mu_i AU, \\ 0 &= 0. \end{aligned}$$

Therefore, since for inf-sup compatible spaces $\dim(\ker(B)) = N_{\mathbf{u}} - N_p$, there are $(N_{\mathbf{u}} - N_p)$ coincident generalized eigenvalues $\mu_i = 1$. The correspondent eigenvectors are

$$\begin{bmatrix} U_i \\ 0 \end{bmatrix}$$

where $\{U_i\}, i = 1, \ldots, N_{\mathbf{u}} - N_p$, is a basis of $ker(B)$.

The $2N_p$ missing eigenvalues can be evaluated considering the case $U \notin \ker(B)$, such that

$$AU + B^T P = \mu_i AU, \tag{2.48}$$

$$BU = \mu_i \frac{1}{\nu} M_P \tag{2.49}$$

From (2.48) we get

$$U = \frac{1}{\mu_i - 1} A^{-1} B^T P$$

and substituting it in (2.49), we obtain

$$BA^{-1}B^T P = \mu_i(\mu_i - 1)\frac{1}{\nu} M_P P$$

. Thus, the $2N_p$ missing eigenvalues can be obtained from the generalized eigenvalues λ_i of the Schur complement with respect to $\frac{1}{\nu}M_P$, namely:

$$\mu_i(\mu_i - 1) = \lambda_i, \quad \longrightarrow \quad \mu_i = \frac{1 \pm \sqrt{1 + 4\lambda_i}}{2}.$$

Since, as shown in Section 2.5.1, λ_i are bounded independently of h by

$$\beta_h^2 \leq \lambda_i \leq d$$

then we have a set of N_p negative eigenvalues

$$\mu_i \in \left[\frac{1 - \sqrt{1 + 4\lambda_{\max}}}{2}, \frac{1 - \sqrt{1 + 4\lambda_{\min}}}{2}\right]$$

and a set of N_p positive eigenvalues

$$\mu_i \in \left[\frac{1 + \sqrt{1 + 4\lambda_{\min}}}{2}, \frac{1 + \sqrt{1 + 4\lambda_{\max}}}{2}\right],$$

see Figure 2.2.

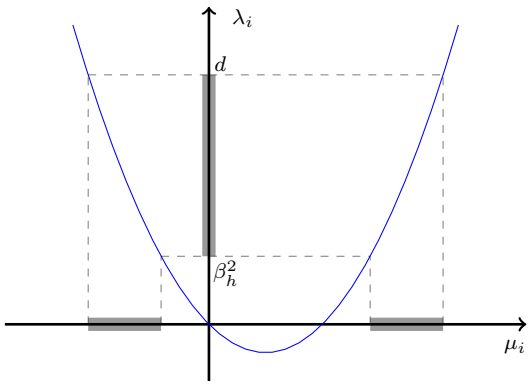


Fig. 2.2: Distribution of the generalized eigenvalues λ_i and μ_i.

Since the generalized eigenvalues are clustered into intervals which do not depend on h, the number of iterations required by the iterative solved will be independent of h, proving that $\mathcal{P}$ is an optimal preconditioner.

A similar analysis can be carried out to prove that the same preconditioner is also optimal for the non-symmetric Stokes matrix Σ_2. In this case, the generalized eigen-problem reads

$$\begin{bmatrix} A & B^T \\ -B & 0 \end{bmatrix} \begin{bmatrix} U \\ P \end{bmatrix} = \mu_i \begin{bmatrix} A & 0 \\ 0 & \frac{1}{\nu} M_P \end{bmatrix} \begin{bmatrix} U \\ P \end{bmatrix}. \tag{2.50}$$

Proceeding as before, a first subset of eigenvalues is obtained considering the case in which $P = 0$ and $U \in \ker(B)$, namely

$$\begin{aligned} AU &= \mu_i AU, \\ 0 &= 0, \end{aligned}$$

that is, also in this case, we have $(N_{\mathbf{u}} - N_p)$ coincident generalized eigenvalues $\mu_i = 1$. From the system

$$AU + B^T P = \mu_i AU, \tag{2.51}$$

$$-BU = \mu_i \frac{1}{\nu} M_P, \tag{2.52}$$

substituting U computed by (2.51) in (2.52), we can compute the missing eigenvalues μ_i such that

$$BA^{-1}B^T P = \mu_i(1 - \mu_i)\frac{1}{\nu} M_P P$$

.

Also in this case, the second subset of $2N_p$ eigenvalues can be obtained from the generalized eigenvalues λ_i of the Schur complement with respect to $\frac{1}{\nu} M_P$, namely:

$$\mu_i(1 - \mu_i) = \lambda_i, \quad \longrightarrow \quad \mu_i = \frac{1 \pm \sqrt{1 - 4\lambda_i}}{2}.$$

For $\beta_h > \frac{1}{4}$ the eigenvalues μ_i are complex conjugate pairs. However, since λ_i are bounded, also μ_i will be clustered, since they will be contained in bounded regions (which do not depend on h) of the complex plane.

Better performances (that is faster convergence) can be usually obtained resorting to more involved preconditioner such as the triangular preconditioner

$$\mathcal{P} = \begin{bmatrix} A & 0 \\ \pm B & \frac{1}{\nu} M_P \end{bmatrix}$$

where the off-diagonal element is positive when solving for Σ_1 and negative for Σ_2.

The optimality of this triangular preconditioner can be proved by computing explicitly its inverse

$$\mathcal{P}^{-1} = \begin{bmatrix} A^{-1} & 0 \\ \mp \nu M_P^{-1} B A^{-1} & \nu M_P^{-1} \end{bmatrix}$$

and the preconditioned Stokes matrix

$$\mathcal{P}^{-1}\Sigma_{1,2} = \begin{bmatrix} A^{-1} & 0 \\ \mp\nu M_P^{-1} B A^{-1} & \nu M_P^{-1} \end{bmatrix} \begin{bmatrix} A & B^T \\ \pm B & 0 \end{bmatrix} = \begin{bmatrix} I & A^{-1}B^T \\ 0 & \mp\nu M_P^{-1} B A^{-1} B^T \end{bmatrix}.$$

Since the preconditioned matrix is block triangular, its eigenvalues are the eigenvalues of its diagonal blocks. In particular, for Σ_1, there are N_u coincident eigenvalues $\mu_i = 1$, while the remaining N_p eigenvalues are the opposite of the eigenvalues λ_i of the Schur complement, such that $-d \leq \mu_i \leq -\beta_h^2$. On the other hand, for Σ_2, there are still N_u coincident eigenvalues $\mu_i = 1$, while the others are exactly the eigenvalues λ_i of the Schur complement, such that $\beta_h^2 \leq \mu_i \leq d$. In both case, the eigenvalues of the preconditioned Stokes matrix are real and clustered independently of h, thus proving the optimality of the triangular preconditioner.

Chapter 3
The stationary Navier-Stokes equations

In this chapter, the nonlinear Navier-Stokes equations are considered. In particular, the numerical solution of the stationary problem (where time evolution is not accounted for) is discussed. The spatial discretization is based on the same finite-element framework introduced in Chapter 2 and will be combined to suitable iterative strategies to treat the nonlinearity of the problem.

3.1 Weak formulation of the steady Navier-Stokes equations

We consider the stationary Navier-Stokes equations

$$-\nu\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p = \mathbf{f}, \qquad \text{in } \Omega \tag{3.1}$$

$$\text{div}\mathbf{u} = 0, \qquad \text{in } \Omega \tag{3.2}$$

$$\mathbf{u} = \mathbf{g}, \qquad \text{on } \Gamma_D \tag{3.3}$$

$$\nu\frac{\partial\mathbf{u}}{\partial\mathbf{n}} - p\mathbf{n} = \mathbf{d}, \qquad \text{on } \Gamma_N \tag{3.4}$$

where the viscosity can be reinterpreted as the inverse of the Reynolds number $\nu = 1/\text{Re}$ when non-dimensional variables are considered.

The weak formulation of problem (3.1)-(3.4) reads: *find* $(\mathbf{u}, p) \in \mathbf{V} \times Q$, $\mathbf{u} = \mathbf{g}$ *on* Γ_D, *such that*

$$a(\mathbf{u},\mathbf{v}) + c(\mathbf{u},\mathbf{u},\mathbf{v}) + b(\mathbf{v},p) = F(\mathbf{v}), \qquad \forall\mathbf{v} \in \mathbf{V}_0, \tag{3.5}$$

$$b(\mathbf{u},q) = 0, \qquad \forall q \in Q. \tag{3.6}$$

which differs from the weak formulation of the Stokes problem only by the additional trilinear form $c : \mathbf{V} \times \mathbf{V} \times \mathbf{V} \longrightarrow \mathbb{R}$ defined by:

$$c(\mathbf{w},\mathbf{u},\mathbf{v}) = \int_\Omega (\mathbf{w}\cdot\nabla)\mathbf{u}\cdot\mathbf{v}\,d\Omega.$$

The bilinear forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ satisfy the same properties as in the Stokes case, namely

$$\begin{aligned}
&a(\mathbf{u},\mathbf{u}) \geq \alpha\|\mathbf{u}\|^2_{H^1(\Omega)},\\
&a(\mathbf{u},\mathbf{v}) \leq \nu\|\nabla\mathbf{u}\|_{L^2(\Omega)}\|\nabla\mathbf{v}\|_{L^2(\Omega)},\\
&b(\mathbf{v},q) \leq \sqrt{d}\|\nabla\mathbf{v}\|_{L^2(\Omega)}\|q\|_{L^2(\Omega)},\\
&\inf_{q\in Q}\sup_{\mathbf{v}\in\mathbf{V}} \frac{b(\mathbf{v},q)}{\|\nabla\mathbf{v}\|_{L^2(\Omega)}\|q\|_{L^2(\Omega)}} \geq \beta > 0.
\end{aligned}$$

We recall that when a fully-Dirichlet problem ($\Gamma_N = \emptyset$) is considered, the pressure is defined up to a constant and the pressure space should be modified accordingly, namely $Q = L^2_0(\Omega)$. Moreover, to impose on the Neumann boundary a physical stress condition such as

$$2\nu D(\mathbf{u})\cdot\mathbf{n} - p\mathbf{n} = \mathbf{d}$$

it is preferable to write the viscous term as

$$a(\mathbf{u},\mathbf{v}) = \int_\Omega 2\nu D(\mathbf{u}) : D(\mathbf{v})\,d\Omega.$$

Let us now analyze the properties of the trilinear form $c(\cdot,\cdot,\cdot)$. First, we can show that $c(\cdot,\cdot,\cdot)$ is continuous. We first recall the three-term Holder inequality which reads: for any $p,r,q>1$ such that $\frac{1}{p}+\frac{1}{r}+\frac{1}{q}=1$, we have

$$\left|\int_\Omega f\,g\,h\,d\Omega\right| \leq \|f\|_{L^p(\Omega)}\|g\|_{L^r(\Omega)}\|h\|_{L^q(\Omega)}. \tag{3.7}$$

The Sobolev embedding theorem guarantees that any $\mathbf{v}\in[H^1(\Omega)]^d$ is also in $[L^4(\Omega)]^d$ and there exists a constant C_S such that

$$\|\mathbf{v}\|_{L^4(\Omega)} \leq C_S\|\mathbf{v}\|_{H^1(\Omega)}.$$

Combining the latter with the three-term Holder inequality, the continuity of the trilinear form can be easily obtained as

$$\begin{aligned}
c(\mathbf{w},\mathbf{u},\mathbf{v}) &= \int_\Omega (\mathbf{w}\cdot\nabla)\mathbf{u}\cdot\mathbf{v}\,d\Omega\\
&\leq \|\mathbf{w}\|_{L^4(\Omega)}\|\nabla\mathbf{u}\|_{L^2(\Omega)}\|\mathbf{v}\|_{L^4(\Omega)}\\
&\leq C_S^2\|\mathbf{w}\|_{H^1(\Omega)}\|\mathbf{u}\|_{H^1(\Omega)}\|\mathbf{v}\|_{H^1(\Omega)},
\end{aligned}$$

where we have identified f,g and h in (3.7) with $\mathbf{w},\nabla\mathbf{u}$ and $\mathbf{v}$, respectively and we have taken $p=q=4$ and $r=2$. Moreover, using Poincarè inequality,

we can also prove that there exists a positive constant $\hat{C}$ such that

$$c(\mathbf{w},\mathbf{u},\mathbf{v}) \leq \hat{C}\|\nabla\mathbf{w}\|_{L^2(\Omega)}\|\nabla\mathbf{u}\|_{L^2(\Omega)}\|\nabla\mathbf{v}\|_{L^2(\Omega)}.$$

3.1.1 Energy estimate for the homogeneous Dirichlet problem

When a homogeneous (fully) Dirichlet problem ($\Gamma_N = \emptyset$) is considered, we have $\forall \mathbf{w} \in V, \mathrm{div}\mathbf{w} = 0$:

$$\begin{aligned} c(\mathbf{w},\mathbf{u},\mathbf{v}) &= \int_\Omega (\mathbf{w}\cdot\nabla)\mathbf{u}\cdot\mathbf{v}\,d\Omega = \int_\Omega \sum_{i,j}^{d} w_j \frac{\partial u_i}{\partial x_j} v_i\,d\Omega \\ &= -\sum_{i,j=1}^{d}\int_\Omega u_i \frac{\partial w_j v_i}{\partial x_j}\,d\Omega + \sum_{i,j=1}^{d}\int_{\partial\Omega} u_i w_j v_i n_j\,d\gamma \\ &= -\int_\Omega \mathrm{div}\mathbf{w}\mathbf{u}\cdot\mathbf{v}\,d\Omega - \int_\Omega (\mathbf{w}\cdot\nabla)\mathbf{v}\cdot\mathbf{u}\,d\Omega + \int_{\partial\Omega}(\mathbf{u}\cdot\mathbf{v})(\mathbf{w}\cdot\mathbf{n})\,d\gamma \\ &= -c(\mathbf{w},\mathbf{v},\mathbf{u}) \end{aligned} \tag{3.8}$$

that is the trilinear form $c(\cdot,\cdot,\cdot)$ is skew-symmetric with respect to the last two arguments. In particular, this implies that

$$c(\mathbf{w},\mathbf{u},\mathbf{u}) = 0, \quad \forall \mathbf{w} \in V, \mathrm{div}\mathbf{w} = 0,$$

which is a key result to prove the well-posedness of the problem.

Thanks to the skew-symmetry of the trilinear form, setting setting $\mathbf{v} = \mathbf{u}$ and $q = p$ in the weak formulation (3.5) and proceeding as in Section 2.4.2 where we obtained an energy estimate for the Stokes problem, we get

$$\|\nabla\mathbf{u}\|_{L^2(\Omega)} \leq \frac{1}{\nu}\|\mathbf{f}\|_{H^{-1}(\Omega)}, \quad \forall \mathbf{u} \in \mathbf{V}_0. \tag{3.9}$$

The inf-sup condition yields immediately an estimate for the pressure:

$$\begin{aligned} \|p\| &\leq \frac{1}{\beta}\sup_{\mathbf{v}\in\mathbf{V}_0}\frac{b(\mathbf{v},p)}{\|\nabla\mathbf{v}\|_{L^2(\Omega)}} \leq \frac{1}{\beta}\sup_{\mathbf{v}\in\mathbf{V}_0}\frac{F(\mathbf{v}) - a(\mathbf{u},\mathbf{v}) - c(\mathbf{u},\mathbf{u},\mathbf{v})}{\|\nabla\mathbf{v}\|_{L^2(\Omega)}} \\ &\leq \frac{1}{\beta}\left[\|\mathbf{f}\|_{H^{-1}(\Omega)} + \nu\|\nabla\mathbf{u}\|_{L^2(\Omega)} + \hat{C}\|\nabla\mathbf{u}\|^2_{L^2(\Omega)}\right] \end{aligned} \tag{3.10}$$

3.1.2 Well-posedness of the homogeneous Dirichlet problem

For a complete analysis of the well-posedness of problem (3.5)-(3.6) we refer to [25, 54, 60]. It can be proven using Lax-Milgram lemma and Brouwer's fixed point theorem that, $\forall \mathbf{f} \in H^{-1}(\Omega)$, the solution $(\mathbf{u}, p)$ of problem (3.5)-(3.6) exists and, as we have seen, satisfies the stability estimates (3.9) and (3.10). However, this solution is not, in general, unique. Uniqueness can only be proven for sufficiently small data (see [25]). Namely, if

$$\frac{\hat{C}\|\mathbf{f}\|_{H^{-1}(\Omega)}}{\nu^2} < 1,$$

then the solution is unique.

3.1.3 More general cases

To prove the stability and the existence and uniqueness of the solution for the homogeneous Dirichlet problem, the skew-symmetry of the trilinear form $c(\cdot,\cdot,\cdot)$ has been exploited. Let us analyze what happens with other kinds of boundary conditions and different formulations of the differential problem.

3.1.3.1 Non-solenoidal advection field

The trilinear form $c(\cdot,\cdot,\cdot)$ in the Dirichlet problem is skew-symmetric for advection fields $\mathbf{w}$ which are solenoidal $\mathrm{div}\mathbf{w} = 0$ (see (3.8)). However, this condition is not always fulfilled. For instance, when the Galerkin formulation of problem (3.5)-(3.6) is considered, the discrete velocity fields is not, in general, pointwise divergence-free. In such cases, it is useful to consider the following modified form of the trilinear form

$$\widetilde{c}(\mathbf{w},\mathbf{u},\mathbf{v}) = c(\mathbf{w},\mathbf{u},\mathbf{v}) + \frac{1}{2}\int_\Omega \mathrm{div}\mathbf{w}\,\mathbf{u}\cdot\mathbf{v}\,d\Omega \tag{3.11}$$

which is always skew-symmetric $\forall \mathbf{w}$ when $\mathbf{u} = 0$ on $\partial\Omega$. Moreover, this modified form is strongly consistent with the continuous problem:

$$\widetilde{c}(\mathbf{u},\mathbf{u},\mathbf{v}) = c(\mathbf{u},\mathbf{u},\mathbf{v}) + \frac{1}{2}\int_\Omega \mathrm{div}\mathbf{u}\,\mathbf{u}\cdot\mathbf{v}\,d\Omega = c(\mathbf{u},\mathbf{u},\mathbf{v})$$

since $\mathrm{div}\mathbf{u} = 0$.

Therefore, to recover the skew-symmetry required to prove the stability and well-posedness results for the Galerkin approximation, the original tri-

linear form $c(\cdot,\cdot,\cdot)$ can be replaced by the modified form $\widetilde{c}(\cdot,\cdot,\cdot)$. Note that, the latter is still continuous

$$\widetilde{c}(\mathbf{w},\mathbf{u},\mathbf{v}) \leq \hat{C}\|\nabla\mathbf{u}\|_{L^2(\Omega)}\|\nabla\mathbf{w}\|_{L^2(\Omega)}\|\nabla\mathbf{v}\|_{L^2(\Omega)}.$$

3.1.3.2 Mixed Dirichlet/Neumann problem

When $\Gamma_N \neq \emptyset$, the skew-symmetry is lost by both $c(\cdot,\cdot,\cdot)$ and $\widetilde{c}(\cdot,\cdot,\cdot)$ since the boundary term in (3.8) is not null. Indeed, we have:

$$\widetilde{c}(\mathbf{u},\mathbf{u},\mathbf{u}) = \frac{1}{2}\int_{\Gamma_N} |\mathbf{u}|^2(\mathbf{u}\cdot\mathbf{n})\,d\gamma.$$

An energy estimate for the mixed problem can be obtained under suitable condition on the advection field at the boundaries. Let us consider the homogeneous mixed Dirichlet/Neumann problem. In this case, we have

$$\begin{aligned}\nu\|\nabla\mathbf{u}\|^2 &= a(\mathbf{u},\mathbf{u}) = F(\mathbf{u}) - c(\mathbf{u},\mathbf{u},\mathbf{u}) - b(\mathbf{u},p)\\ &= \int_\Omega \mathbf{f}\cdot\mathbf{u}\,d\Omega - \frac{1}{2}\int_{\Gamma_N}|\mathbf{u}|^2(\mathbf{u}\cdot\mathbf{n})\,d\gamma\\ &\leq \frac{1}{2\nu}\|\mathbf{f}\|^2_{H^{-1}(\Omega)} + \frac{\nu}{2}\|\nabla\mathbf{u}\|^2_{L^2(\Omega)} - \frac{1}{2}\int_{\Gamma_N}|\mathbf{u}|^2(\mathbf{u}\cdot\mathbf{n})\,d\gamma,\end{aligned}$$

where we have used Young's inequality. Therefore

$$\|\nabla\mathbf{u}\|^2_{L^2(\Omega)} + \underbrace{\frac{1}{\nu}\int_{\Gamma_N}|\mathbf{u}|^2(\mathbf{u}\cdot\mathbf{n})\,d\gamma}_{A} \leq \frac{1}{\nu^2}\|\mathbf{f}\|^2_{H^{-1}(\Omega)}, \tag{3.12}$$

where the sign of A is undefined. The sign of A depends on the direction of the velocity over the Neumann boundary:

- if Γ_N is an *outflow* boundary ($\mathbf{u}\cdot\mathbf{n} > 0$) then A is positive;
- if Γ_N is an *inflow* boundary ($\mathbf{u}\cdot\mathbf{n} < 0$) then A is negative.

Thus, from (3.12) we obtain a meaningful energy estimate whenever we set Neumann boundary conditions on outflow boundaries.

3.1.3.3 Total stress formulation

We have seen in Section 1.4.2.2 that, when the rotational form of the Navier-Stokes equations is considered, the nonlinear term can be rewritten as

$$(\mathbf{u}\cdot\nabla)\mathbf{u} = \frac{1}{2}\nabla|\mathbf{u}|^2 + (\boldsymbol{\nabla}\times\mathbf{u})\times\mathbf{u}.$$

This formulation is best suited when one would like to impose the total stress as boundary condition:

$$\nu \frac{\partial \mathbf{u}}{\partial \mathbf{n}} - p_T \mathbf{n} = \mathbf{d}, \quad \text{on } \Gamma_N,$$

where $p_T = p + \frac{1}{2}|\mathbf{u}|^2$ is the total pressure.

In this case, the weak formulation reads: *find* $(\mathbf{u}, p_T) \in \mathbf{V} \times Q$, *such that*

$$\begin{aligned} a(\mathbf{u}, \mathbf{v}) + c(\mathbf{u}, \mathbf{u}, \mathbf{v}) + b(\mathbf{v}, p) &= F(\mathbf{v}), \qquad \forall \mathbf{v} \in \mathbf{V}_0, \\ b(\mathbf{u}, q) &= 0, \qquad \forall q \in Q. \end{aligned}$$

with

$$c(\mathbf{w}, \mathbf{u}, \mathbf{v}) = \int_\Omega (\boldsymbol{\nabla} \times \mathbf{w}) \times \mathbf{u} \cdot \mathbf{v} \, d\Omega.$$

We note that the trilinear form $c(\mathbf{w}, \mathbf{u}, \mathbf{v})$ is always skew-symmetric since

$$\boldsymbol{\xi} \times \mathbf{u} \cdot \mathbf{v} = -\boldsymbol{\xi} \times \mathbf{v} \cdot \mathbf{u}, \qquad \forall \mathbf{u}, \mathbf{v}, \boldsymbol{\xi}.$$

Therefore, for the rotational formulation of the Navier-Stokes equations with total stress boundary condition, the energy estimate

$$\|\nabla \mathbf{u}\|_{L^2(\Omega)} \leq \frac{1}{\nu} \|\mathbf{f}\|_{H^{-1}(\Omega)}$$

holds.

3.2 Galerkin approximation

Let us consider two finite-element spaces $\mathbf{V}_h \subset \mathbf{V}$ and $Q_h \subset Q$ which satisfy the discrete inf-sup condition:

$$\inf_{q_h \in Q_h} \sup_{\mathbf{v}_h \in \mathbf{V}_h} \frac{b(\mathbf{v}_h, q_h)}{\|\nabla \mathbf{v}_h\|_{L^2(\Omega)} \|q_h\|_{L^2(\Omega)}} \geq \beta_h > 0, \forall h, \tag{3.13}$$

with $\beta_h \to \beta^* > 0$ for $h \to 0$.

The Galerkin approximation of problem (3.5)-(3.6) reads: *find* $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$, $\mathbf{u}_h = \mathbf{g}_h$ *on* Γ_D, *such that*

$$a(\mathbf{u}_h, \mathbf{v}_h) + \widetilde{c}(\mathbf{u}_h, \mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p_h) = F(\mathbf{v}_h), \qquad \forall \mathbf{v}_h \in \mathbf{V}_{h,0}, \tag{3.14}$$

$$b(\mathbf{u}_h, q_h) = 0, \qquad \forall q_h \in Q_h. \tag{3.15}$$

where the skew-symmetric modified trilinear form $\widetilde{c}(\cdot, \cdot, \cdot)$ is adopted. The stability of problem (3.14)-(3.15) can be proved as in the continuous case

and yields the following estimates:

$$\|\nabla \mathbf{u}_h\|_{L^2(\Omega)} \leq \frac{1}{\nu}\|\mathbf{f}\|_{H^{-1}(\Omega)}$$
$$\|p_h\|_{L^2(\Omega)} \leq \frac{1}{\beta_h}\left\{\|\mathbf{f}\|_{H^{-1}(\Omega)} + \nu\|\nabla \mathbf{u}_h\|_{L^2(\Omega)} + \hat{C}\|\nabla \mathbf{u}_h\|^2_{L^2(\Omega)}\right\}$$

3.2.1 Convergence analysis (for small data)

For sufficiently small data, we can show that the solution of the Galerkin approximation (3.14)-(3.15) converges to the solution of the original continuous problem (3.5)-(3.6). To prove this convergence results, let us first introduce the following subspace of $\mathbf{V}_{h,0}$:

$$\mathbf{V}_{h,\text{div}} = \{\mathbf{v}_h \in \mathbf{V}_h, \quad b(\mathbf{v}_h, q_h) = 0, \forall q_h \in Q_h\},$$

whose elements are functions with discrete divergence equal to zero.

Given any $\mathbf{w}_h \in \mathbf{V}_{h,\text{div}}$, we have

$$\begin{aligned}
\nu\|\nabla(\mathbf{u}_h - \mathbf{w}_h)\|^2_{L^2(\Omega)} &= a(\mathbf{u}_h - \mathbf{w}_h, \mathbf{u}_h - \mathbf{w}_h) \\
&= a(\mathbf{u}_h - \mathbf{u}, \mathbf{u}_h - \mathbf{w}_h) + a(\mathbf{u} - \mathbf{w}_h, \mathbf{u}_h - \mathbf{w}_h) \\
&= F(\mathbf{u}_h - \mathbf{w}_h) - b(\mathbf{u}_h - \mathbf{w}_h, p_h) - \widetilde{c}(\mathbf{u}_h, \mathbf{u}_h, \mathbf{u}_h - \mathbf{w}_h) \\
&\quad -F(\mathbf{u}_h - \mathbf{w}_h) + b(\mathbf{u}_h - \mathbf{w}_h, p) + \widetilde{c}(\mathbf{u}, \mathbf{u}, \mathbf{u}_h - \mathbf{w}_h) \\
&\quad +a(\mathbf{u} - \mathbf{w}_h, \mathbf{u}_h - \mathbf{w}_h) \\
&= \widetilde{c}(\mathbf{u}, \mathbf{u}, \mathbf{u}_h - \mathbf{w}_h) - \widetilde{c}(\mathbf{u}_h, \mathbf{u}_h, \mathbf{u}_h - \mathbf{w}_h) \\
&\quad +b(\mathbf{u}_h - \mathbf{w}_h, p) + a(\mathbf{u} - \mathbf{w}_h, \mathbf{u}_h - \mathbf{w}_h)
\end{aligned}$$

Adding and subtracting in the right-hand-side the mixed term $\widetilde{c}(\mathbf{u}_h, \mathbf{u}, \mathbf{u}_h - \mathbf{w}_h)$ and subtracting $b(\mathbf{u}_h - \mathbf{w}_h, \pi_h)$ (which is null $\forall \pi_h \in Q_h$), we get

$$\begin{aligned}
\nu\|\nabla(\mathbf{u}_h - \mathbf{w}_h)\|^2_{L^2(\Omega)} &= \widetilde{c}(\mathbf{u} - \mathbf{u}_h, \mathbf{u}, \mathbf{u}_h - \mathbf{w}_h) + \widetilde{c}(\mathbf{u}_h, \mathbf{u} - \mathbf{u}_h, \mathbf{u}_h - \mathbf{w}_h) \\
&\quad +b(\mathbf{u}_h - \mathbf{w}_h, p - \pi_h) + a(\mathbf{u} - \mathbf{w}_h, \mathbf{u}_h - \mathbf{w}_h) \\
&\leq \hat{C}\|\nabla \mathbf{u}\|_{L^2(\Omega)}\|\nabla(\mathbf{u} - \mathbf{u}_h)\|_{L^2(\Omega)}\|\nabla(\mathbf{u}_h - \mathbf{w}_h)\|_{L^2(\Omega)} \\
&\quad +\hat{C}\|\nabla \mathbf{u}_h\|_{L^2(\Omega)}\|\nabla(\mathbf{u} - \mathbf{u}_h)\|_{L^2(\Omega)}\|\nabla(\mathbf{u}_h - \mathbf{w}_h)\|_{L^2(\Omega)} \\
&\quad +\sqrt{d}\|\nabla(\mathbf{u} - \mathbf{u}_h)\|_{L^2(\Omega)}\|p - \pi_h\|_{L^2(\Omega)} \\
&\quad +\nu\|\nabla(\mathbf{u} - \mathbf{w}_h)\|_{L^2(\Omega)}\|\nabla(\mathbf{u}_h - \mathbf{w}_h)\|_{L^2(\Omega)},
\end{aligned}$$

where we have used the continuity of the bilinear and trilinear forms.

Using the energy estimate for the continuous and discrete problems, we get

$$\|\nabla(\mathbf{u}_h - \mathbf{w}_h)\|_{L^2(\Omega)} \leq \frac{2\hat{C}\|\mathbf{f}\|_{H^{-1}(\Omega)}}{\nu^2}\|\nabla(\mathbf{u} - \mathbf{u}_h)\|_{L^2(\Omega)} + \frac{\sqrt{d}}{\nu}\|p - \pi_h\|_{L^2(\Omega)} + \|\nabla(\mathbf{u} - \mathbf{w}_h)\|_{L^2(\Omega)},$$

which implies, by triangular inequality, that

$$\|\nabla(\mathbf{u} - \mathbf{u}_h)\|_{L^2(\Omega)} \leq \frac{2\hat{C}\|\mathbf{f}\|_{H^{-1}(\Omega)}}{\nu^2}\|\nabla(\mathbf{u} - \mathbf{u}_h)\|_{L^2(\Omega)} + \frac{\sqrt{d}}{\nu}\|p - \pi_h\|_{L^2(\Omega)} + 2\|\nabla(\mathbf{u} - \mathbf{w}_h)\|_{L^2(\Omega)}.$$

From the latter, we can conclude that, if

$$\frac{2\hat{C}\|\mathbf{f}\|_{H^{-1}(\Omega)}}{\nu^2} < 1,$$

then there exist two constants C_1 and C_2 such that

$$\|\nabla(\mathbf{u} - \mathbf{u}_h)\|_{L^2(\Omega)} \leq C_1 \inf_{\mathbf{w}_h \in \mathbf{V}_{h,\mathrm{div}}} \|\nabla(\mathbf{u} - \mathbf{w}_h)\|_{L^2(\Omega)} + C_2 \inf_{\pi_h \in Q_h} \|p - \pi_h\|_{L^2(\Omega)}.$$

Moreover, since the spaces $\mathbf{V}_h$ and Q_h satisfy the inf-sup condition, we have

$$\inf_{\mathbf{w}_h \in \mathbf{V}_{h,\mathrm{div}}} \|\nabla(\mathbf{u} - \mathbf{w}_h)\|_{L^2(\Omega)} \leq \left(1 + \frac{\sqrt{d}}{\beta_h}\right) \inf_{\mathbf{v}_h \in \mathbf{V}_h} \|\nabla(\mathbf{u} - \mathbf{v}_h)\|_{L^2(\Omega)}.$$

With a similar argument, we can obtain a pressure estimate; indeed, $\forall \pi_h \in Q_h$, we have

$$\begin{aligned} b(\mathbf{v}_h, p_h - \pi_h) &= F(\mathbf{v}_h) - a(\mathbf{u}_h, \mathbf{v}_h) - \widetilde{c}(\mathbf{u}_h, \mathbf{u}_h, \mathbf{v}_h) - b(\mathbf{v}_h, \pi_h) \\ &= a(\mathbf{u} - \mathbf{u}_h, \mathbf{v}_h) + \widetilde{c}(\mathbf{u}, \mathbf{u}, \mathbf{v}_h) - \widetilde{c}(\mathbf{u}_h, \mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p - \pi_h) \\ &= a(\mathbf{u} - \mathbf{u}_h, \mathbf{v}_h) + \widetilde{c}(\mathbf{u} - \mathbf{u}_h, \mathbf{u}, \mathbf{v}_h) \\ &\quad + \widetilde{c}(\mathbf{u}_h, \mathbf{u} - \mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p - \pi_h) \\ &\leq \left(\nu + \frac{2\hat{C}\|\mathbf{f}\|_{H^{-1}(\Omega)}}{\nu}\right)\|\nabla(\mathbf{u} - \mathbf{u}_h)\|_{L^2(\Omega)}\|\nabla\mathbf{v}_h\|_{L^2(\Omega)} \\ &\quad + \sqrt{d}\|p - \pi_h\|_{L^2(\Omega)}\|\nabla\mathbf{v}_h\|_{L^2(\Omega)}, \end{aligned}$$

which implies

$$\|p_h - \pi_h\|_{L^2(\Omega)} \leq \frac{1}{\beta_h} \sup_{\mathbf{v}_h \in \mathbf{V}_h} \frac{b(\mathbf{v}_h, p_h - \pi_h)}{\|\nabla \mathbf{v}_h\|_{L^2(\Omega)}}$$
$$\leq \frac{1}{\beta_h} \left[\left(\nu + \frac{2\hat{C}\|\mathbf{f}\|_{H^{-1}(\Omega)}}{\nu} \right) \|\nabla(\mathbf{u} - \mathbf{u}_h)\|_{L^2(\Omega)} + \sqrt{d}\|p - \pi_h\|_{L^2(\Omega)} \right].$$

Using once more the triangular inequality and the small data condition

$$\frac{2\hat{C}\|\mathbf{f}\|_{H^{-1}(\Omega)}}{\nu^2} < 1,$$

the resulting pressure estimate reads

$$\|p - p_h\|_{L^2(\Omega)} \leq \left(1 + \frac{\sqrt{d}}{\beta_h} \right) \inf_{\pi_h \in \mathbb{Q}_h} \|p - \pi_h\|_{L^2(\Omega)} + \frac{2\nu}{\beta_h} \|\nabla(\mathbf{u} - \mathbf{u}_h)\|_{L^2(\Omega)}.$$

In conclusion, under the small data condition, the solution of the Galerkin approximation of the steady Navier-Stokes equations satisfies the following estimate:

$$\|\nabla(\mathbf{u} - \mathbf{u}_h)\|_{L^2(\Omega)} + \|p - p_h\|_{L^2(\Omega)} \leq C_1 \inf_{\mathbf{v}_h \in \mathbf{V}_h} \|\nabla(\mathbf{u} - \mathbf{v}_h)\|_{L^2(\Omega)}$$
$$+ C_2 \inf_{\pi_h \in Q_h} \|p - \pi_h\|_{L^2(\Omega)}.$$

As usual, we can develop the solution as linear combination of a suitable basis function, namely:

$$\mathbf{u}_h(\mathbf{x}) = \sum_{j=1}^{\mathcal{N}_\mathbf{u}} u_j \boldsymbol{\phi}_j, \qquad p_h(\mathbf{x}) = \sum_{k=1}^{\mathcal{N}_p} p_k \psi_k.$$

We take $\mathbf{v}_h = \boldsymbol{\phi}_i$ and $q_h = \psi_l$ as test functions, and we have, $\forall i = 1, \ldots, \mathcal{N}_\mathbf{u}$ and $\forall l = 1, \ldots, \mathcal{N}_p$:

$$a(\mathbf{u}_h, \boldsymbol{\phi}_i) + \widetilde{c}(\mathbf{u}_h, \mathbf{u}_h, \boldsymbol{\phi}_i) + b(\boldsymbol{\phi}_i, p_h) = \tag{3.16}$$
$$= \sum_{j=1}^{\mathcal{N}_\mathbf{u}} u_j a(\boldsymbol{\phi}_j, \boldsymbol{\phi}_i) + \sum_{s,j=1}^{\mathcal{N}_\mathbf{u}} u_s u_j \widetilde{c}(\boldsymbol{\phi}_s, \boldsymbol{\phi}_j, \boldsymbol{\phi}_i) + \sum_{k=1}^{\mathcal{N}_p} p_k b(\boldsymbol{\phi}_i, \psi_k) = \mathbf{F}(\boldsymbol{\phi}_i),$$

$$b(\mathbf{u}_h, \psi_l) = \sum_{j=1}^{\mathcal{N}_\mathbf{u}} u_j b(\boldsymbol{\phi}_j, \psi_l) = 0. \tag{3.17}$$

Using the matrices considered for the Stokes problem and introducing the non-linear term

$$\mathcal{N}(U) \in \mathbb{R}^{\mathcal{N}_\mathbf{u}}, \quad \mathcal{N}(U)_i = \sum_{s,j=1}^{\mathcal{N}_\mathbf{u}} u_s u_j \widetilde{c}(\boldsymbol{\phi}_s, \boldsymbol{\phi}_j, \boldsymbol{\phi}_i),$$

the algebraic formulation of the steady Navier-Stokes equations reads:

$$\begin{aligned} AU + \mathcal{N}(U) + B^T P &= \mathbf{F} \\ BU &= \mathbf{0} \end{aligned}$$

which is a nonlinear system in $\mathcal{N}_{\mathbf{u}} + \mathcal{N}_p$ unknowns.

To face this nonlinear problem different strategies are possible. In the following, we consider two strategies: a fixed-point algorithm and the Newton's method.

3.3 Fixed point method

The Navier–Stokes equations can be seen as an advection-diffusion problem with an additional incompressibility constraint. If we assume that the advection field $\mathbf{w}$ is given, and such that $\mathrm{div}\mathbf{w} = 0$, the following linear problem, known as the *Oseen's problem*, is obtained:

$$-\nu \Delta \mathbf{u} + (\mathbf{w} \cdot \nabla)\mathbf{u} + \nabla p = \mathbf{f}, \qquad \text{in } \Omega \tag{3.18}$$

$$\mathrm{div}\mathbf{u} = 0, \qquad \text{in } \Omega \tag{3.19}$$

$$\mathbf{u} = \mathbf{g}, \qquad \text{on } \Gamma_D \tag{3.20}$$

$$\nu \frac{\partial \mathbf{u}}{\partial \mathbf{n}} - p\mathbf{n} = \mathbf{d}, \qquad \text{on } \Gamma_N \tag{3.21}$$

The corresponding weak formulation reads: *find* $(\mathbf{u}, p) \in \mathbf{V} \times Q$, $\mathbf{u} = \mathbf{g}$ *on* Γ_D, *such that*

$$a(\mathbf{u}, \mathbf{v}) + c(\mathbf{w}, \mathbf{u}, \mathbf{v}) + b(\mathbf{v}, p) = F(\mathbf{v}), \qquad \forall \mathbf{v} \in \mathbf{V}_0, \tag{3.22}$$

$$b(\mathbf{u}, q) = 0, \qquad \forall q \in Q. \tag{3.23}$$

where, as for Navier-Stokes, the trilinear form $c(\cdot, \cdot, \cdot)$ can be replaced by the modified form

$$\widetilde{c}(\mathbf{w}, \mathbf{u}, \mathbf{v}) = c(\mathbf{w}, \mathbf{u}, \mathbf{w}) + \frac{1}{2} \int_\Omega \mathrm{div}\mathbf{w}\mathbf{u} \cdot \mathbf{v} \, d\Omega$$

to recover the skew-symmetry.

Problem (3.22)-(3.23) is well-posed and has a unique solution $\forall \mathbf{f} \in H^{-1}, \forall \mathbf{w} \in \mathbf{V}, \mathrm{div}\mathbf{w} = 0$ (at least for the fully Dirichlet case $\Gamma_N = \emptyset$) thanks to the continuity and coercivity of the bilinear form $a(\mathbf{u}, \mathbf{v}) + c(\mathbf{w}, \mathbf{u}, \mathbf{v})$.

We can therefore define an operator $T : \mathbf{V} \longrightarrow \mathbf{V}$ which associates to any function $\mathbf{w} \in \mathbf{V}$ the solution $\mathbf{u} \in \mathbf{V}$ of the Oseen's problem:

$$T(\mathbf{w}) = \mathbf{u}.$$

It is evident that the solution of the steady Navier-Stokes problem is a fixed-point of T, namely

$$T(\mathbf{u}) = \mathbf{u}.$$

We can show that, for sufficiently small data, the operator is a contraction, that is there exists a constant $\rho < 1$ such that:

$$\|\mathbf{u}_1 - \mathbf{u}_2\|_{H^1(\Omega)} = \|T(\mathbf{w}_1 - \mathbf{w}_2)\|_{H^1(\Omega)} \leq \rho \|\mathbf{w}_1 - \mathbf{w}_2\|_{H^1(\Omega)}, \quad \forall \mathbf{w}_1, \mathbf{w}_2 \in \mathbf{V}.$$

Indeed if we consider $(\mathbf{u}_1, p_1)$ and $(\mathbf{u}_2, p_2)$ solutions of the following Oseen problems

$$-\nu \Delta \mathbf{u}_1 + (\mathbf{w}_1 \cdot \nabla)\mathbf{u}_1 + \nabla p_1 = \mathbf{f}, \qquad \text{in } \Omega \tag{3.24}$$
$$\text{div}\mathbf{u}_1 = 0, \qquad \text{in } \Omega, \tag{3.25}$$

and

$$-\nu \Delta \mathbf{u}_2 + (\mathbf{w}_2 \cdot \nabla)\mathbf{u}_2 + \nabla p_2 = \mathbf{f}, \qquad \text{in } \Omega \tag{3.26}$$
$$\text{div}\mathbf{u}_2 = 0, \qquad \text{in } \Omega. \tag{3.27}$$

The difference between the two problems yields

$$-\nu \Delta(\mathbf{u}_1 - \mathbf{u}_2) + (\mathbf{w}_1 \cdot \nabla)\mathbf{u}_1 - (\mathbf{w}_2 \cdot \nabla)\mathbf{u}_2 + \nabla(p_1 - p_2) = 0, \qquad \text{in } \Omega$$
$$\text{div}(\mathbf{u}_1 - \mathbf{u}_2) = 0, \qquad \text{in } \Omega.$$

Testing against the functions $(\mathbf{u}_1 - \mathbf{u}_2)$ and $(p_1 - p_2)$, respectively, and integrating over Ω, we get

$$\nu \|\nabla(\mathbf{u}_1 - \mathbf{u}_2)\|^2_{L^2(\Omega)} = c(\mathbf{w}_2, \mathbf{u}_2, \mathbf{u}_1 - \mathbf{u}_2) - c(\mathbf{w}_1, \mathbf{u}_1, \mathbf{u}_1 - \mathbf{u}_2).$$

If we add and subtract the mixed term $c(\mathbf{w}_2, \mathbf{u}_1, \mathbf{u}_1 - \mathbf{u}_2)$ from the right hand side, we obtain

$$\begin{aligned}\nu \|\nabla(\mathbf{u}_1 - \mathbf{u}_2)\|^2_{L^2(\Omega)} &= -c(\mathbf{w}_2, \mathbf{u}_1 - \mathbf{u}_2, \mathbf{u}_1 - \mathbf{u}_2) - c(\mathbf{w}_1 - \mathbf{w}_2, \mathbf{u}_1, \mathbf{u}_1 - \mathbf{u}_2) \\ &\leq \hat{C} \|\nabla \mathbf{u}_1\|_{L^2(\Omega)} \|\nabla(\mathbf{w}_1 - \mathbf{w}_2)\|_{L^2(\Omega)} \|\nabla(\mathbf{u}_1 - \mathbf{u}_2)\|_{L^2(\Omega)}\end{aligned}$$

where we have used the continuity and the skew-symmetry of the trilinear functional.

It follows that, under the following small data condition

$$\frac{\hat{C}\|\mathbf{f}\|_{H^{-1}(\Omega)}}{\nu^2} < 1,$$

the operator T is a contraction.

This implies that, the fixed point iteration

$$\mathbf{u}^{k+1} = T(\mathbf{u}^k)$$

converges to the (unique) solution of the steady Navier-Stokes equations, for any initial data $\mathbf{u}^0$.

The fixed-point iteration algorithm is defined as follows: *given* $\mathbf{u}^0 \in \mathbf{V}, \mathbf{u}^0 = \mathbf{g}$ *on* Γ_D, *for* $k = 1, 2, \ldots$, *find* $(\mathbf{u}^k, p^k) \in \mathbf{V} \times Q$, $\mathbf{u}^k = \mathbf{g}$ *on* Γ_D, *such that*

$$a(\mathbf{u}^k, \mathbf{v}) + c(\mathbf{u}^{k-1}, \mathbf{u}^k, \mathbf{v}) + b(\mathbf{v}, p^k) = F(\mathbf{v}), \qquad \forall \mathbf{v} \in \mathbf{V}_0, \tag{3.28}$$

$$b(\mathbf{u}^k, q) = 0, \qquad \forall q \in Q. \tag{3.29}$$

As a convergence criterion for the fixed-point iteration, we can consider the energy norm of the increment, namely the iteration is stopped when

$$\|\nabla(\mathbf{u}^k - \mathbf{u}^{k-1})\|_{L^2(\Omega)} \leq \text{tol},$$

where tol is a suitable tolerance.

The fixed-point method exhibits a linear convergence rate, namely

$$\|\nabla(\mathbf{u} - \mathbf{u}^{k+1})\|_{L^2(\Omega)} \leq \rho \|\nabla(\mathbf{u} - \mathbf{u}^k)\|_{L^2(\Omega)} \leq \rho^k \|\nabla(\mathbf{u} - \mathbf{u}^0)\|_{L^2(\Omega)}.$$

From the algebraic perspective, we notice that each fixed-point iteration requires the solution of an Oseen-type problem:

$$\begin{bmatrix} A + N(U^{k-1}) & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} U^k \\ P^k \end{bmatrix} = \begin{bmatrix} \mathbf{F} \\ \mathbf{0} \end{bmatrix}, \tag{3.30}$$

where the elements of matrix $N(\mathbf{W})$ are defined as follows:

$$N(\mathbf{W})_{ij} = \widetilde{c}(\mathbf{w}, \boldsymbol{\phi}_j, \boldsymbol{\phi}_i) = c(\mathbf{w}, \boldsymbol{\phi}_j, \boldsymbol{\phi}_i) + \frac{1}{2} \int_\Omega \text{div}\mathbf{w} \boldsymbol{\phi}_j \cdot \boldsymbol{\phi}_i \, d\Omega$$

Unfortunately, this matrix has to be recomputed at each iteration. Moreover, $N(\mathbf{W})$ is skew-symmetric (for the fully Dirichlet case):

$$N(\mathbf{W})_{ij} = \widetilde{c}(\mathbf{w}, \boldsymbol{\phi}_j, \boldsymbol{\phi}_i) = -\widetilde{c}(\mathbf{w}, \boldsymbol{\phi}_i, \boldsymbol{\phi}_j) = -N(\mathbf{W})_{ji}.$$

and displays a block diagonal structure; indeed, given

$$\boldsymbol{\phi}_j^{(1)} = \begin{bmatrix} \phi_j \\ 0 \\ 0 \end{bmatrix}, \qquad \boldsymbol{\phi}_i^{(2)} = \begin{bmatrix} 0 \\ \phi_i \\ 0 \end{bmatrix},$$

we have

$$\widetilde{c}(\mathbf{w}, \boldsymbol{\phi}_j^{(1)}, \boldsymbol{\phi}_i^{(2)}) = \int_\Omega (\mathbf{w} \cdot \nabla) \boldsymbol{\phi}_j^{(1)} \cdot \boldsymbol{\phi}_i^{(2)} \, d\Omega = \int_\Omega (\mathbf{w} \cdot \nabla) \begin{bmatrix} \phi_j \\ 0 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ \phi_i \\ 0 \end{bmatrix} \, d\Omega = 0.$$

The resulting Oseen matrix is therefore non-symmetric with a sparsity pattern similar to that of the Stokes matrix

$$\left[\begin{array}{ccc:c} K+N_{\mathbf{W}} & & & B^T \\ & K+N_{\mathbf{W}} & & \\ & & K+N_{\mathbf{W}} & \\ \hdashline B & & & 0 \end{array}\right] \left[\begin{array}{c} U_1 \\ U_2 \\ U_3 \\ \hdashline \mathbf{P} \end{array}\right] = \left[\begin{array}{c} \mathbf{F}_1 \\ \mathbf{F}_2 \\ \mathbf{F}_3 \\ \hdashline \mathbf{0} \end{array}\right].$$

3.4 Newton's method

Before deriving the formulation of the Newton's method for the solution of the nonlinear Navier–Stokes equations (3.5)-(3.6), we recall the main properties of the method on a finite dimensional case.

Let us consider a system of nonlinear equations

$$\mathbf{f}(\mathbf{u}) = \mathbf{0}, \qquad \mathbf{u} \in \mathbb{R}^n, \qquad \mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n. \tag{3.31}$$

Newton's method for the solution of problem (3.31) reads: *given* $\mathbf{u}^0 \in \mathbb{R}^n, \forall k = 1, 2, \ldots$*:*

$$\begin{aligned} J(\mathbf{u}^k)\,\delta\mathbf{u} &= -\mathbf{f}(\mathbf{u}^k) \\ \mathbf{u}^{k+1} &= \mathbf{u}^k + \delta\mathbf{u} \end{aligned} \tag{3.32}$$

where $J(\mathbf{u}^k)$ denotes the Jacobian matrix

$$J_{ij} = (\nabla\mathbf{f})_{ij} = \frac{\partial f_i}{\partial u_j}$$

evaluated in $\mathbf{u}^k$. The left-hand-side of (3.32) represents the directional derivative of $\mathbf{f}$ along the direction $\delta\mathbf{u}$, evaluated in $\mathbf{u}^k$:

$$J(\mathbf{u}^k)\,\delta\mathbf{u} = \nabla\mathbf{f}|_{\mathbf{u}^k} \cdot \delta\mathbf{u}.$$

At each iteration the Newton's method requires the solution of a linear system.

If the Jacobian $J(\mathbf{u})$ is non-singular, then there exists a neighborhood $\mathcal{U}(\mathbf{u})$ around $\mathbf{u}$ such that, $\forall \mathbf{u}^0 \in \mathcal{U}(\mathbf{u})$, the sequence $\mathbf{u}^k$ generated by the Newton's method converges quadratically to the exact solution $\mathbf{u}$, namely:

$$\|\nabla(\mathbf{u} - \mathbf{u}^{k+1})\|_{L^2(\Omega)} \leq C\|\nabla(\mathbf{u} - \mathbf{u}^k)\|^2_{L^2(\Omega)}.$$

Let us now consider the continuous Navier-Stokes problem, which can be reformulated as follows:

$$\boldsymbol{\mathcal{L}}(\mathbf{u}, p) - \boldsymbol{\mathcal{F}} = \mathbf{0}, \tag{3.33}$$

where

$$\boldsymbol{\mathcal{L}}(\mathbf{u},p)=\begin{bmatrix}-\nu\Delta\mathbf{u}+(\mathbf{u}\cdot\nabla)\mathbf{u}+\nabla p\\ \mathrm{div}\mathbf{u}\end{bmatrix},\qquad \boldsymbol{\mathcal{F}}=\begin{bmatrix}\mathbf{f}\\0\end{bmatrix}.$$

The *tangent problem* given by the Newton's method applied to problem (3.33) reads: *given* $(\mathbf{u}^0,p^0)\in\mathbf{V}\times Q$, *for* $k=1,2,\ldots$, *find* $(\delta\mathbf{u},\delta p)\in\mathbf{V}\times Q$, *such that*

$$\mathcal{D}\boldsymbol{\mathcal{L}}_{(\mathbf{u}^k,p^k)}(\delta\mathbf{u},\delta p)=\boldsymbol{\mathcal{F}}-\boldsymbol{\mathcal{L}}(\mathbf{u}^k,p^k), \tag{3.34}$$

$$(\mathbf{u}^{k+1},p^{k+1})=(\mathbf{u}^k,p^k)+(\delta\mathbf{u},\delta p), \tag{3.35}$$

where we have denoted with $\mathcal{D}\boldsymbol{\mathcal{L}}_{(\mathbf{u}^k,p^k)}(\delta\mathbf{u},\delta p)$ the directional derivative of operator $\boldsymbol{\mathcal{L}}$ in direction $(\delta\mathbf{u},\delta p)$, evaluated at $(\mathbf{u}^k,p^k)$. This derivative is known as the *Gâteaux derivative* and is defined as follows:

$$\mathcal{D}\boldsymbol{\mathcal{L}}_{(\mathbf{u}^k,p^k)}(\delta\mathbf{u},\delta p)=\lim_{\varepsilon\to0}\frac{\boldsymbol{\mathcal{L}}(\mathbf{u}^k+\varepsilon\delta\mathbf{u},p^k+\varepsilon\delta p)-\boldsymbol{\mathcal{L}}(\mathbf{u}^k,p^k)}{\varepsilon}=\lim_{\varepsilon\to0}\frac{\Delta\boldsymbol{\mathcal{L}}}{\varepsilon}$$

Evaluating explicitly $\Delta\boldsymbol{\mathcal{L}}$, we get:

$$\begin{aligned}\Delta\mathcal{L}&=\begin{bmatrix}-\nu\Delta(\mathbf{u}^k+\varepsilon\delta\mathbf{u})+(\mathbf{u}^k+\varepsilon\delta\mathbf{u})\cdot\nabla(\mathbf{u}^k+\varepsilon\delta\mathbf{u})+\nabla(p^k+\varepsilon\delta p)\\ \mathrm{div}(\mathbf{u}^k+\varepsilon\delta\mathbf{u})\end{bmatrix}\\&\quad-\begin{bmatrix}-\nu\Delta\mathbf{u}^k+(\mathbf{u}^k\cdot\nabla)\mathbf{u}^k+\nabla p^k\\ \mathrm{div}\mathbf{u}^k\end{bmatrix}\\&=\begin{bmatrix}\varepsilon\{-\nu\Delta\delta\mathbf{u}+(\delta\mathbf{u}\cdot\nabla)\mathbf{u}^k+(\mathbf{u}^k\cdot\nabla)\delta\mathbf{u}+\nabla\delta p\}+\varepsilon^2(\delta\mathbf{u}\cdot\nabla)\delta\delta\mathbf{u}\\ \varepsilon\mathrm{div}\delta\mathbf{u}\end{bmatrix}\end{aligned}$$

and

$$\mathcal{D}\boldsymbol{\mathcal{L}}_{(\mathbf{u}^k,p^k)}(\delta\mathbf{u},\delta p)=\lim_{\varepsilon\to0}\frac{\Delta\boldsymbol{\mathcal{L}}}{\varepsilon}=\begin{bmatrix}-\nu\Delta\delta\mathbf{u}+(\delta\mathbf{u}\cdot\nabla)\mathbf{u}^k+(\mathbf{u}^k\cdot\nabla)\delta\mathbf{u}+\nabla\delta\mathbf{u}\\ \mathrm{div}\delta\mathbf{u}\end{bmatrix}.$$

The tangent problem can be then explicitly written as

$$\begin{aligned}-\nu\Delta\delta\mathbf{u}+(\delta\mathbf{u}\cdot\nabla)\mathbf{u}^k+(\mathbf{u}^k\cdot\nabla)\delta\mathbf{u}+\nabla\delta p&=\mathbf{f}+\nu\Delta\mathbf{u}^k-(\mathbf{u}^k\cdot\nabla)\mathbf{u}^k-\nabla p^k,\\ \mathrm{div}\delta\mathbf{u}&=-\mathrm{div}\mathbf{u}^k,\end{aligned}$$

or, more conveniently, in terms of the unknowns $\mathbf{u}^{k+1}=\mathbf{u}^k+\delta\mathbf{u}$ and $p^{k+1}=p^k+\delta p$, as follows:

$$\begin{aligned}-\nu\Delta\mathbf{u}^{k+1}+(\mathbf{u}^{k+1}\cdot\nabla)\mathbf{u}^k+(\mathbf{u}^k\cdot\nabla)\mathbf{u}^{k+1}+\nabla p^{k+1}&=\mathbf{f}+(\mathbf{u}^k\cdot\nabla)\mathbf{u}^k,\\ \mathrm{div}\mathbf{u}^{k+1}&=0.\end{aligned}$$

Note that $\mathbf{u}^{k+1}$ should satisfy the same boundary conditions as the exact solution. Moreover, comparing one Newton's iteration with the correspondent Oseen-type problem to solve at each fixed-point iteration, we notice that an additional convection term $(\mathbf{u}^{k+1}\cdot\nabla)\mathbf{u}^k$ enters the differential problem.

The weak formulation of the Newton's method then reads: *given* $\mathbf{u}^0 \in \mathbf{V}, \mathbf{u}^0 = \mathbf{g}$ *on* Γ_D*, for* $k = 1, 2, \ldots$*, find* $(\mathbf{u}^k, p^k) \in \mathbf{V} \times Q$*,* $\mathbf{u}^k = \mathbf{g}$ *on* Γ_D*, such that*

$$\begin{aligned} a(\mathbf{u}^k, \mathbf{v}) + c(\mathbf{u}^{k-1}, \mathbf{u}^k, \mathbf{v}) + c(\mathbf{u}^k, \mathbf{u}^{k-1}, \mathbf{v}) + \\ +b(\mathbf{v}, p^k) = F(\mathbf{v}) + c(\mathbf{u}^{k-1}, \mathbf{u}^{k-1}, \mathbf{v}), \quad \forall \mathbf{v} \in \mathbf{V}_0, \quad (3.36) \\ b(\mathbf{u}^k, q) = 0, \qquad \forall q \in Q. \quad (3.37) \end{aligned}$$

Once again, the trilinear form $c(\cdot, \cdot, \cdot)$ can be replaced by its modified version $\widetilde{c}(\cdot, \cdot, \cdot)$ in order to preserve the skew-symmetry also in the Galerkin formulation. As convergence criterion for stopping the Newton's iteration, the same adopted for the fixed-point iteration can be adopted. Namely, the iteration is stopped when

$$\|\nabla(\mathbf{u}^k - \mathbf{u}^{k-1})\|_{L^2(\Omega)} \leq \text{tol}.$$

Other stopping criteria based on either the relative increment or the residual can be also adopted.

Proposition 3.1 (Convergence of the Newton's method). *Given* $(\mathbf{u}, p)$ *solution of the Navier-Stokes equations, if the tangent operator* $\mathcal{D}\boldsymbol{\mathcal{L}}_{(\mathbf{u}^k, p^k)}(\delta\mathbf{u}, \delta p)$ *is non-singular, that is problem*

$$\mathcal{D}\boldsymbol{\mathcal{L}}_{(\mathbf{u}^k, p^k)}(\delta\mathbf{u}, \delta p) = \begin{bmatrix} \mathbf{f} \\ 0 \end{bmatrix}$$

is well-posed $\forall \mathbf{f} \in [H^{-1}(\Omega)]^d$*, for any initial value* $\mathbf{u}^0$ *sufficiently close to* $\mathbf{u}$*, the sequence* $\mathbf{u}^k$ *converges quadratically to* $\mathbf{u}$*, namely:*

$$\|\nabla(\mathbf{u} - \mathbf{u}^k)\|_{L^2(\Omega)} \leq \|\nabla(\mathbf{u} - \mathbf{u}^{k-1})\|^2_{L^2(\Omega)}.$$

For sufficiently small data, that is for

$$\frac{\hat{C}\|\mathbf{f}\|_{H^{-1}(\Omega)}}{\nu^2} < 1,$$

the tangent problem is well-posed. The more general case will be discussed later.

3.4.1 Solution of the linear system

At each iteration of the Newton's method, one needs to solve a linear system of the following form

$$\begin{bmatrix} A + N(U^{k-1}) + M(U^{k-1}) & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} U^k \\ P^k \end{bmatrix} = \begin{bmatrix} \mathbf{F}_N(U^{k-1}) \\ \mathbf{0} \end{bmatrix}, \tag{3.38}$$

where matrix $N(\mathbf{W})$ is the same matrix introduced for the Oseen problem in the fixed-point method, while:

$$M(\mathbf{W})_{ij} = \widetilde{c}(\boldsymbol{\phi}_j, \mathbf{w}, \boldsymbol{\phi}_i) = c(\boldsymbol{\phi}_j, \mathbf{w}, \boldsymbol{\phi}_i) + \frac{1}{2}\int_\Omega \mathrm{div}\boldsymbol{\phi}_j \mathbf{w} \cdot \boldsymbol{\phi}_i \, d\Omega,$$

$$F(\mathbf{W})_i = \int_\Omega \mathbf{f} \cdot \boldsymbol{\phi}_i \, d\Omega + \int_\Omega (\mathbf{w} \cdot \nabla)\mathbf{w} \cdot \boldsymbol{\phi}_i \, d\Omega.$$

The terms depending on U^{k-1} need to be recomputed at each iteration. Moreover, differently from $N(U^{k-1})$, the matrix $M(U^{k-1})$ does not display a block diagonal structure. Indeed, given

$$\boldsymbol{\phi}_j^{(1)} = \begin{bmatrix} \phi_j \\ 0 \\ 0 \end{bmatrix}, \qquad \boldsymbol{\phi}_i^{(2)} = \begin{bmatrix} 0 \\ \phi_i \\ 0 \end{bmatrix},$$

we have

$$\begin{aligned} c(\boldsymbol{\phi}_j^{(1)}, \mathbf{w}, \boldsymbol{\phi}_i^{(2)}) &= \int_\Omega (\boldsymbol{\phi}_j^{(1)} \cdot \nabla)\mathbf{w} \cdot \boldsymbol{\phi}_i^{(2)} \, d\Omega \\ &= \int_\Omega \phi_j \frac{\partial \mathbf{w}}{\partial x} \cdot \boldsymbol{\phi}_i^{(2)} \, d\Omega \\ &= \int_\Omega \phi_j \frac{\partial w_2}{\partial x} \cdot \phi_i \, d\Omega \neq 0. \end{aligned}$$

In the resulting Newton system we loose the sparsity of the matrix:

$$\left[\begin{array}{ccc:c} K + N_{\mathbf{W}} + M_{11} & M_{12} & M_{13} & \\ M_{21} & K + N_{\mathbf{W}} + M_{22} & M_{23} & B^T \\ M_{31} & M_{32} & K + N_{\mathbf{W}} + M_{33} & \\ \hdashline & B & & 0 \end{array}\right] \begin{bmatrix} U_1 \\ U_2 \\ U_3 \\ \hdashline \mathbf{P} \end{bmatrix} = \begin{bmatrix} \mathbf{F}_1 \\ \mathbf{F}_2 \\ \mathbf{F}_3 \\ \hdashline \mathbf{0} \end{bmatrix}$$

making the Newton's method more expansive to be solved than the fixed-point method.

3.5 Stabilization for advection dominated problems

In the case of scalar advection-diffusion problems, it is well-known that the standard Galerkin approximation may present instabilities when $\mathrm{Pe}_h > 1$, having denoted with $\mathrm{Pe}_h = |\beta|h/2D$ the local Peclét number. Suitable stabilization strategies may be introduce to face this problem, ranging from

simple approaches (such as the addition of an *artificial diffusion* or a *streamline diffusion* [54]) to more involved techniques (such as *strongly consistent stabilizations* [57] or *sub-grid stabilizations* [19]).

Similar stabilizations are required when solving by a Galerkin approach the Navier-Stokes in an advection-dominating regime, that is when the Reynolds number is large.

We briefly recall the strongly consistent Streamline-Upwind Petrov Galerkin (SUPG) stabilization and we first apply it to the linearized Oseen problem (3.18)–(3.21). For the sake of simplicity, we will introduce the stabilized formulation for a homogeneous fully Dirichlet case ($\Gamma_D = \partial\Omega, \mathbf{g} = \mathbf{0}$).

3.5.1 SUPG stabilization of the Oseen problem

The strong formulation of the Oseen problem can be written as

$$\boldsymbol{\mathcal{L}}^{(\mathbf{w})}(\mathbf{u}, p) = \boldsymbol{\mathcal{F}} \tag{3.39}$$

where

$$\boldsymbol{\mathcal{L}}^{(\mathbf{w})}(\mathbf{u}, p) = \begin{bmatrix} -\nu\Delta\mathbf{u} + (\mathbf{w}\cdot\nabla)\mathbf{u} + \frac{1}{2}\mathrm{div}\mathbf{w}\mathbf{u} + \nabla p \\ \mathrm{div}\mathbf{u} \end{bmatrix}, \qquad \boldsymbol{\mathcal{F}} = \begin{bmatrix} \mathbf{f} \\ 0 \end{bmatrix}.$$

In the differential operator $\boldsymbol{\mathcal{L}}^{(\mathbf{w})}$ we have also included the consistent term which is used to define the skew-symmetric modified trilinear form (3.11).

The SUPG stabilization of the Galerkin discretization of the Oseen problem then reads: *given* $\mathbf{w}$, *find* $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$, $\mathbf{u}_h = \mathbf{g}_h$ *on* Γ_D, *such that*

$$\begin{aligned} &a(\mathbf{u}_h, \mathbf{v}_h) + \widetilde{c}(\mathbf{w}_h, \mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p_h) - b(\mathbf{u}_h, q_h) + \\ &\quad + \sum_{K\in\mathcal{T}_h} \delta_k \left(\boldsymbol{\mathcal{L}}^{(\mathbf{w}_h)}(\mathbf{u}_h, p_h) - \boldsymbol{\mathcal{F}}, \boldsymbol{\mathcal{L}}_{SS}^{(\mathbf{w}_h)}(\mathbf{v}_h, q_h) \right) d\Omega = F(\mathbf{v}_h), \\ &\qquad\qquad \forall \mathbf{v}_h \in \mathbf{V}_{h,0}, \quad q_h \in Q_h, \end{aligned} \tag{3.40}$$

where $\boldsymbol{\mathcal{L}}_{SS}^{(\mathbf{w})}(\cdot,\cdot)$ denotes the skew-symmetric part of the differential operator $\boldsymbol{\mathcal{L}}^{(\mathbf{w})}(\cdot,\cdot)$.

Let us analyze the differential operator in order to identify its symmetric and skew-symmetric components. Considering the following decomposition

$$\boldsymbol{\mathcal{L}}^{(\mathbf{w})}(\mathbf{u}, p) = \underbrace{\begin{bmatrix} -\nu\Delta\mathbf{u} \\ 0 \end{bmatrix}}_{\boldsymbol{\mathcal{L}}_1} + \underbrace{\begin{bmatrix} (\mathbf{w}\cdot\nabla)\mathbf{u} + \frac{1}{2}\mathrm{div}\mathbf{w}\mathbf{u} \\ 0 \end{bmatrix}}_{\boldsymbol{\mathcal{L}}_2^{(\mathbf{w})}} + \underbrace{\begin{bmatrix} \nabla p \\ \mathrm{div}\mathbf{u} \end{bmatrix}}_{\boldsymbol{\mathcal{L}}_3}.$$

we can easily prove that (in the fully Dirichlet case):

- $\boldsymbol{\mathcal{L}}_1$ is symmetric:

$$< \boldsymbol{\mathcal{L}_1}(\mathbf{u},p),(\mathbf{v},q) > = -\int_\Omega \nu\Delta\mathbf{u}\cdot\mathbf{v}\,d\Omega = \int_\Omega \nu\nabla\mathbf{u} : \nabla\mathbf{v}\,d\Omega$$
$$= -\int_\Omega \nu\mathbf{u}\cdot\Delta\mathbf{v}\,d\Omega =< (\mathbf{u},p), \boldsymbol{\mathcal{L}}_1(\mathbf{v},q) >;$$

- $\boldsymbol{\mathcal{L}}_2^{(\mathbf{w})}$ is skew-symmetric (see (3.11));
- $\boldsymbol{\mathcal{L}}_3$ is skew-symmetric:

$$< \boldsymbol{\mathcal{L}_3}(\mathbf{u},p),(\mathbf{v},q) > = \int_\Omega \nabla p\cdot\mathbf{v}\,d\Omega + \int_\Omega q\,\mathrm{div}\mathbf{u}\,d\Omega$$
$$= -\int_\Omega p\,\mathrm{div}\mathbf{v}\,d\Omega - \int_\Omega \nabla p\cdot\mathbf{u}\,d\Omega$$
$$= - < (\mathbf{u},p), \boldsymbol{\mathcal{L}}_3(\mathbf{v},q) > .$$

Therefore the skew-symmetric part of the operator is

$$\boldsymbol{\mathcal{L}}_{SS} = \boldsymbol{\mathcal{L}}_2^{(\mathbf{w})} + \boldsymbol{\mathcal{L}}_3 = \begin{bmatrix} (\mathbf{w}\cdot\nabla)\mathbf{u} + \frac{1}{2}\mathrm{div}\mathbf{w}\mathbf{u} + \nabla p \\ \mathrm{div}\mathbf{u} \end{bmatrix}$$

and the resulting stabilization term in (3.40) is given by

$$\sum_{K\in\mathcal{T}_h} \delta_k \left(\boldsymbol{\mathcal{L}}^{(\mathbf{w}_h)}(\mathbf{u}_h,p_h) - \boldsymbol{\mathcal{F}}, \boldsymbol{\mathcal{L}}_{SS}^{(\mathbf{w}_h)}(\mathbf{v}_h,q_h) \right) d\Omega =$$
$$= \sum_{K\in\mathcal{T}_h} \delta_k \int_K \left(-\nu\Delta\mathbf{u}_h + (\mathbf{w}_h\cdot\nabla)\mathbf{u}_h + \frac{1}{2}\mathrm{div}\mathbf{w}_h\mathbf{u}_h + \nabla p_h - \mathbf{f} \right) \cdot$$
$$\left((\mathbf{w}_h\cdot\nabla)\mathbf{v}_h + \frac{1}{2}\mathrm{div}\mathbf{w}_h\mathbf{v}_h + \nabla q_h \right) d\Omega + \qquad (3.41)$$
$$+ \sum_{K\in\mathcal{T}_h} \delta_k \int_K \mathrm{div}\mathbf{u}_h\,\mathrm{div}\mathbf{v}_h\,d\Omega$$

Remark 3.2. The stabilization term (3.41) is strongly consistent since it is formulated as a linear function of the residual. The strong consistency is a necessary condition for recovering the optimal convergence rate of the Galerkin approximation.

Remark 3.3. The term

$$\sum_{K\in\mathcal{T}_h} \delta_k \int_K (\mathbf{w}_h\cdot\nabla)\mathbf{u}_h\cdot(\mathbf{w}_h\cdot\nabla)\mathbf{v}_h\,d\Omega$$

in (3.41) corresponds to the streamline diffusion stabilization. Stable results may be obtained adding just this term to the Galerkin formulation; however,

in this case only first order convergence would be obtained regardless the polynomial degree adopted in the finite element discretization.

Remark 3.4. In (3.41), we can recognize another term that we have already encountered in previous section, that is

$$\sum_{K\in\mathcal{T}_h} \delta_k \int_K \nabla p_h \cdot \nabla q_h \, d\Omega$$

which is the Brezzi-Pitkaranta stabilization for the inf-sup stability (introduced in Section 2.4.5). Therefore, when the SUPG stabilization is adopted, it is possible to work with finite-element spaces which are not inf-sup compatible (such as equal order finite-elements for velocity and pressure).

3.5.2 SUPG stabilization of the Navier–Stokes problem

The same stabilization procedure can be adopted for the nonlinear Navier-Stokes problem (3.1)-(3.4). Restricting once more the discussion to the homogeneous fully Dirichlet case, we get the following stabilized discrete formulation: *find* $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$, $\mathbf{u}_h = \mathbf{g}_h$ *on* Γ_D, *such that*

$$\begin{aligned} a(\mathbf{u}_h,\, \mathbf{v}_h\,) + \widetilde{c}(\mathbf{u}_h, \mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p_h) - b(\mathbf{u}_h, q_h) + \\ + \sum_{K\in\mathcal{T}_h} \delta_k \int_K \left(-\nu\Delta\mathbf{u}_h + (\mathbf{u}_h\cdot\nabla)\mathbf{u}_h + \frac{1}{2}\mathrm{div}\mathbf{u}_h\mathbf{u}_h + \nabla p_h - \mathbf{f}\right) \cdot \\ \left((\mathbf{u}_h\cdot\nabla)\mathbf{v}_h + \frac{1}{2}\mathrm{div}\mathbf{u}_h\mathbf{v}_h + \nabla q_h\right) d\Omega + \\ + \sum_{K\in\mathcal{T}_h} \delta_k \int_K \mathrm{div}\mathbf{u}_h \, \mathrm{div}\mathbf{v}_h \, d\Omega = \mathbf{F}(\mathbf{v}_h), \quad \forall \mathbf{v} \in \mathbf{V}_{h,0}, \quad q_h \in Q_h, \end{aligned}$$

Note that the stabilization term contains cubic nonlinear terms. In order to solve this discrete problem, the Newton's method can be adopted deriving a suitable linearized problem to be solved at each iteration. We refer to [57] for a complete discussion of the convergence of strongly coupled stabilization methods.

3.5.3 Choice of the stabilization coefficient

The stabilization coefficient δ_K is used to scale correctly the stabilization with respect to the viscosity, the local velocity and the grid resolution. A local measure of the ratio between advection and diffusion forces is given by

the local Reynolds number

$$\mathrm{Re}_K = \frac{\bar{\mathbf{u}}_K h_K}{\nu},$$

where $\bar{\mathbf{u}}_K = 1/|K| \int_\Omega |\mathbf{u}| \, d\Omega$. When inf-sup compatible finite-element spaces are used, the SUPG stabilization is only required in the regions where advection is dominant and an optimal choice for the stabilization parameter is

$$\delta_K = \begin{cases} \delta \frac{h_K}{\bar{\mathbf{u}}_K}, & \text{for } \mathrm{Re}_K \geq 1, \\ 0, & \text{for } \mathrm{Re}_K < 1, \end{cases}$$

where δ is a scalar that should be chosen case by case.

On the other hand, when equal order finite-elements for are used for velocity and pressure, the inf-sup stabilizing term should work also in the regions where diffusion dominates and an optimal scaling is given by

$$\delta_K = \begin{cases} \delta \frac{h_K}{\bar{\mathbf{u}}_K}, & \text{for } \mathrm{Re}_K \geq 1, \\ \delta \frac{h_K^2}{\nu}, & \text{for } \mathrm{Re}_K < 1, \end{cases}$$

or, in a more compact form, by

$$\delta_K = \delta \frac{h_K}{\bar{\mathbf{u}}_K} \min\{1, \mathrm{Re}_K\}.$$

Chapter 4
The time-dependent Navier-Stokes equations

In this Chapter, we consider the time-dependent Navier-Stokes equations and we discuss their discretization. While the space discretization will be still based on the finite element method, for the time discretization finite difference schemes will be considered following the standard approach used for the approximation of parabolic partial differential equations. Other approaches for the time discretization will also be introduced leading to the so-called *semi-Lagrangian time discretizations* and *projection methods*.

4.1 Weak formulation and Galerkin approximation

The time-dependent Navier-Stokes equations with mixed boundary conditions read, $\forall t > 0$:

$$\frac{\partial \mathbf{u}}{\partial t} - \nu \Delta \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p = \mathbf{f}, \qquad \text{in } \Omega \tag{4.1}$$

$$\text{div}\mathbf{u} = 0, \qquad \text{in } \Omega \tag{4.2}$$

$$\mathbf{u} = \mathbf{g}, \qquad \text{on } \Gamma_D \tag{4.3}$$

$$\nu \frac{\partial \mathbf{u}}{\partial \mathbf{n}} - p\mathbf{n} = \mathbf{d}, \qquad \text{on } \Gamma_N \tag{4.4}$$

$$\mathbf{u}|_{t=0} = \mathbf{u}_0, \qquad \text{in } \Omega \tag{4.5}$$

A discussion of the theoretical results related to problem (4.1)-(4.5) goes far beyond the scope of these notes. A complete and updated review on the well-posedness of this problem can be found in [34].

Let us derive an energy estimate for the continuous problem (4.1)-(4.5) with homogeneous Dirichlet boundary conditions ($\Gamma_N = \emptyset, \mathbf{g} = \mathbf{0}$). In order to do so, the momentum equation (4.1) is multiplied by $\mathbf{u}$ and integrated over the domain Ω. The different terms can be rewritten as follows:

$$\int_\Omega \frac{\partial \mathbf{u}}{\partial t} \cdot \mathbf{u}\, d\Omega = \frac{1}{2}\frac{d}{dt}\|\mathbf{u}(t)\|^2_{L^2(\Omega)},$$
$$\int_\Omega -\nu\Delta\mathbf{u} \cdot \mathbf{u}\, d\Omega = \int_\Omega \nu\nabla\mathbf{u} : \nabla\mathbf{u}\, d\Omega = \nu\|\nabla\mathbf{u}\|^2_{L^2(\Omega)},$$
$$\int_\Omega ((\mathbf{u}\cdot\nabla)\mathbf{u}) \cdot \mathbf{u}\, d\Omega = \mathbf{0},$$
$$\int_\Omega \nabla p \cdot \mathbf{u}\, d\Omega = -\int_\Omega p\,\mathrm{div}\mathbf{u}\, d\Omega = \mathbf{0},$$

where we have used the skew-symmetry of the nonlinear term. Thus, we get

$$\frac{1}{2}\frac{d}{dt}\|\mathbf{u}(t)\|^2_{L^2(\Omega)} + \nu\|\nabla\mathbf{u}\|^2_{L^2(\Omega)} = \int_\Omega \mathbf{f}\cdot\mathbf{u}\, d\Omega.$$

By Young's inequality, we have

$$\int_\Omega \mathbf{f}\cdot\mathbf{u}\, d\Omega \le \|\mathbf{f}\|_{H^{-1}(\Omega)}\|\nabla\mathbf{u}\|_{L^2(\Omega)} \le \frac{1}{2\nu}\|\mathbf{f}\|^2_{H^{-1}(\Omega)} + \frac{\nu}{2}\|\nabla\mathbf{u}\|^2_{L^2(\Omega)},$$

so that

$$\frac{d}{dt}\|\mathbf{u}(t)\|^2_{L^2(\Omega)} + \nu\|\nabla\mathbf{u}\|^2_{L^2(\Omega)} \le \frac{1}{\nu}\|\mathbf{f}\|^2_{H^{-1}(\Omega)}.$$

Integrating the latter in time over the interval $[0,T]$ we get a first energy estimate:

$$\|\mathbf{u}(T)\|^2_{L^2(\Omega)} + \nu\int_0^T \|\nabla\mathbf{u}(s)\|_{L^2(\Omega)}ds \le \|\mathbf{u}_0\|^2_{L^2(\Omega)} + \frac{1}{\nu}\int_0^T \|\mathbf{f}(s)\|^2_{H^{-1}(\Omega)}ds \tag{4.6}$$

which yields a control on the velocity in the norms $L^\infty(0,T;L^2(\Omega))$ and $L^2(0,T;H^1(\Omega))$.

We notice that estimate (4.6) looses significance for vanishing viscosity ($\nu \to 0$), that is for high Reynolds numbers. A weaker energy estimate, independent of the viscosity, can be obtained using Gronwall's lemma, which is recalled hereafter:

Lemma 4.1 (Gronwall's Lemma). *Let $f \in L^1(0,T)$ be a non-negative function, g and ϕ continuous functions on $[0,T]$ and g a non-decreasing function. If ϕ satisfies the inequality*

$$\phi(t) \le g(t) + \int_0^t f(s)\phi(s)ds, \qquad \forall t \in [0,T],$$

then

$$\phi(t) \le g(t)e^{\int_0^t f(s)ds}, \qquad \forall t \in [0,T].$$

Applying Gronwall's Lemma to the following inequality:

$$\underbrace{\frac{1}{2}\|\mathbf{u}(t)\|^2_{L^2(\Omega)} + \nu\int_0^t \|\nabla\mathbf{u}(s)\|^2_{L^2(\Omega)}ds}_{\phi(t)} = \frac{1}{2}\|\mathbf{u}_0\|^2_{L^2(\Omega)} + \int_0^t\int_\Omega \mathbf{f}\cdot\mathbf{u}\,d\Omega\,ds$$
$$\leq \frac{1}{2}\|\mathbf{u}_0\|^2_{L^2(\Omega)} + \frac{1}{2}\int_0^t \|\mathbf{f}(s)\|^2_{L^2(\Omega)}ds + \int_0^t \underbrace{\frac{1}{2}\|\mathbf{u}(s)\|^2_{L^2(\Omega)}}_{\leq\phi(t)}ds,$$

we get, $\forall t\in[0,T]$,

$$\|\mathbf{u}(t)\|^2_{L^2(\Omega)} + 2\nu\int_0^t \|\nabla\mathbf{u}(s)\|^2_{L^2(\Omega)}ds \leq \left(\|\mathbf{u}_0\|^2_{L^2(\Omega)} + \int_o^t \|\mathbf{f}(s)\|^2_{L^2(\Omega)}ds\right)e^t,$$

in which the stability constant grows exponentially in time, but is independent of the viscosity.

Before considering the time discretization of problem (4.1)-(4.5), let us first derive the semi-discrete formulation based on the Galerkin approximation in space.

The weak formulation of problem (4.1)-(4.5) reads: *find* $(\mathbf{u}(t),p(t))\in\mathbf{V}\times Q = [H^1(\Omega)]^d\times L^2(\Omega)$, $\mathbf{u}(0)=\mathbf{u}_0$, $\mathbf{u}=\mathbf{g}$ *on* Γ_D, *such that, for* $t>0$

$$\int_\Omega \frac{\partial\mathbf{u}}{\partial t}\cdot\mathbf{v}\,d\Omega + a(\mathbf{u},\mathbf{v}) + c(\mathbf{u},\mathbf{u},\mathbf{v}) + b(\mathbf{v},p) = F(\mathbf{v}), \tag{4.7}$$
$$b(\mathbf{u},q) = 0, \tag{4.8}$$

$\forall\mathbf{v}\in\mathbf{V}_0 = [H^1_{\Gamma_D}]^d$ and $\forall q\in Q_h$.

Given two subspaces $(\mathbf{V}_h,Q_h)$ such that the discrete inf-sup condition is satisfied, the Galerkin approximation can be formulated as follows: *find* $(\mathbf{u}_h(t),p_h(t))\in\mathbf{V}_h\times Q_h$, $\mathbf{u}_h(0)=\mathbf{u}_{0,h}$ $\mathbf{u}_h|_{\Gamma_D}=\mathbf{g}_h$, *such that, for* $t>0$

$$\int_\Omega \frac{\partial\mathbf{u}_h}{\partial t}\cdot\mathbf{v}_h\,d\Omega + a(\mathbf{u}_h,\mathbf{v}_h) + c(\mathbf{u}_h,\mathbf{u}_h,\mathbf{v}_h) + b(\mathbf{v}_h,p_h) = F(\mathbf{v}_h), \tag{4.9}$$
$$b(\mathbf{u}_h,q_h) = 0, \tag{4.10}$$

$\forall\mathbf{v}_h\in\mathbf{V}_{0,h}$ and $\forall q_h\in Q$.

The latter defines a nonlinear differential algebraic equation (DAE) system given by:

$$M\frac{dU}{dt} + AU + N(U) + B^TP = F,$$
$$BU = \mathbf{0},$$

where M denotes the mass matrix which is defined as

$$M\in\mathbb{R}^{\mathcal{N}_\mathbf{u}\times\mathcal{N}_\mathbf{u}},\quad M_{ij} = \int_\Omega \boldsymbol{\phi}_j\cdot\boldsymbol{\phi}_i\,d\Omega.$$

If we suppose to approximate the problem with $\mathbb{P}_{k+1}/\mathbb{P}_k$, we can expect convergence rates given by

$$\|\nabla \mathbf{u}(t) - \mathbf{u}_h(t)\|_{L^2(\Omega)} \leq C_1(t)h^{k+1}, \quad \|p(t) - p_h(t)\|_{L^2(\Omega)} \leq C_2(t)h^{k+1}, \quad \forall t \in (0, T].$$

However, it can happen that the constants C_1 and C_2 become unbounded as $h \to 0$. This may be the case, for instance, in presence of a loss of regularity of the exact solution $(\mathbf{u}(t), p(t))$ as $t \to 0$, due to a non-compatibility of the initial data (see [30] for a detailed discussion on this topic).

4.2 Time discretization schemes

To complete the discretization one needs to introduce a time approximation of the problem. Different approaches to accomplish this task are possible. Here, we focus our attention on finite difference approximations in time.

Given a uniform time step Δt, we denote with $(\mathbf{u}^n, p^n)$ an approximation of the exact solution $(\mathbf{u}(t^n), p(t^n))$ at time $t^n = n\Delta t$. Note that, for ease of notation, we introduce the time discretization on the continuous problem.

We start considering the time-dependent Stokes problem, defined as:

$$\frac{\partial \mathbf{u}}{\partial t} - \nu \Delta \mathbf{u} + \nabla p = \mathbf{f}, \qquad \text{in } \Omega \tag{4.11}$$

$$\text{div}\mathbf{u} = 0 \tag{4.12}$$

supplemented with suitable boundary and initial conditions.

We first wonder whether or not fully explicit schemes can be adopted for the time-discretization of problem (4.11)-(4.12). Let us suppose, for instance, that the forward Euler discretization is considered for the momentum equation, namely:

$$\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} = \mathbf{f}^n + \nu \Delta \mathbf{u}^n - \nabla p^n.$$

We notice immediately that, in such case, we can get the updated solution $\mathbf{u}^{n+1}$ at the new time step, but this solution is not divergence-free. Indeed, with an explicit treatment of the pressure term (which works as a Lagrange multiplier) there is no way to impose the incompressibility constraint on the updated solution.

We are therefore forced to consider an implicit treatment of the pressure term:

$$\frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} + \nabla p^{n+1} = \mathbf{f}^n + \nu \Delta \mathbf{u}^n, \qquad \text{in } \Omega \tag{4.13}$$

$$\text{div}\mathbf{u}^{n+1} = 0 \tag{4.14}$$

This problem can now be solved and produces a solution which is divergence-free. However, we notice that the explicit treatment of the viscous term implies a very restrictive stability condition on the time step typical of parabolic problems (see, *e.g.*, [57]), namely:

$$\Delta t \leq C\frac{h^2}{\nu}.$$

This condition is very prohibitive in most fluid dynamics applications where very refined grid (with h very small) are required, for instance, to capture the strong gradients in boundary layers. However, this explicit formulation can be convenient whenever a very small time step is required to capture fast physical dynamics, e.g. in turbulent large eddy simulations where small time scales have to be captured correctly.

Let us now consider the (implicit) backward Euler discretization of problem (4.11)-(4.12), which reads:

$$\frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\Delta t} - \nu\Delta\mathbf{u}^{n+1} + \nabla p^{n+1} = \mathbf{f}^{n+1}, \qquad \text{in } \Omega \tag{4.15}$$

$$\text{div}\mathbf{u}^{n+1} = 0. \tag{4.16}$$

A finite element space discretization of problem (4.15)-(4.16) lead to formulate at each time step the following algebraic problem:

$$\begin{bmatrix} C & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} U^{n+1} \\ P^{n+1} \end{bmatrix} = \begin{bmatrix} F^{n+1} + \frac{M}{\Delta t}U^n \\ 0 \end{bmatrix}. \tag{4.17}$$

The matrix C is given by

$$C = \frac{M}{\Delta t} + A = \frac{M}{\Delta t} + \nu K,$$

where K now denotes the block-diagonal vectorial stiffness matrix (highlighting in this way the presence of the viscosity ν).

Problem (4.17) can be solved using the pressure matrix approach applying, for instance, the Conjugate Gradient method to the following pressure system

$$\Sigma P^{n+1} = \widetilde{F}^{n+1},$$

where the Schur complement is defined as

$$\Sigma = BC^{-1}B^T = B\left(\frac{M}{\Delta t} + \nu K\right)^{-1} B^T.$$

When $\nu >> 1/\Delta t$, the contribution of the mass matrix in the Schur complement is small and we expect the optimal preconditioner that we have derived for the steady case, namely

$$P = \frac{1}{\nu} M_P$$

to work properly since $\Sigma \approx B(\nu K)^{-1} B^T$.

On the other extreme, when $\nu << 1/\Delta t$ we should analyse the spectrum of the matrix

$$\Sigma \approx B \Delta t M^{-1} B^T,$$

in order to identify an optimal preconditioner.

For most finite element space pairs which satisfy the discrete inf-sup condition, it can be proven that also the following condition on the "swapped" norms

$$\inf_{\substack{q_h \in Q_h \\ q_h \neq 0}} \sup_{\mathbf{v}_h \in \mathbf{V}_h} \frac{b(\mathbf{v}_h, q_h)}{\|\mathbf{v}_h\|_{L^2(\Omega)} \|\nabla q_h\|_{L^2(\Omega)}} \geq \widetilde{\beta}_h > 0, \quad \forall h, \tag{4.18}$$

holds true.

Following the same argument used in Section 2.4.1 to derive an optimal preconditioner for the Stokes problem based on the discrete inf-sup, the new condition (4.18) can be used to prove that

$$\widetilde{\beta}_h^2 \leq \frac{P^T B M^{-1} B^T P}{P^T K_P P} \leq 1. \tag{4.19}$$

where K_P is the pressure stiffness matrix:

$$K_p \in \mathbb{R}^{\mathcal{N}_p \times \mathcal{N}_p}, \qquad (K_p)_{lk} = \int_\Omega \nabla \psi_k \cdot \nabla \psi_l \, d\Omega.$$

The spectral equivalence between the velocity mass matrix and the pressure stiffness matrix allow us to identify K_P as an optimal preconditioner for $BM^{-1}B^T$.

Therefore, we have the following spectral characterization for the Schur complement of the time dependent Stokes problem:

- for $\nu >> 1/\Delta t$, $\Sigma \approx B(\nu K)^{-1} B^T$ and $P = \frac{1}{\nu} M_P$ is an optimal preconditioner for Σ;
- for $\nu << 1/\Delta t$, $\Sigma \approx B \Delta t (M)^{-1} B^T$ and $P = \Delta t K_P$ is an optimal preconditioner for Σ;

An optimal preconditioner that can be used also in intermediate regimes ($\nu \approx 1/\Delta t$) is the following the one introduced by Cahouet and Chabard [9], defined by:

$$P_{CC} = \left(\nu M_P^{-1} + \frac{1}{\Delta t} K_P^{-1} \right)^{-1}.$$

The solution of this problem by the preconditioned Conjugate Gradient method requires matrix-vector products and the computation of the preconditioned residual. The latter, namely

$$P_{CC} Z = R,$$

can be computed solving two pressure problems for the mass and stiffness matrices, respectively, as follows

$$Z = P_{CC}^{-1} R = \left(\nu M_P^{-1} + \frac{1}{\Delta t} K_P^{-1} \right) R = \nu M_P^{-1} R + \frac{1}{\Delta t} K_P^{-1} R.$$

4.3 Treatment of the nonlinear advection term

When moving from the unsteady Stokes problem (4.11)-(4.12) to the unsteady Navier-Stokes problem, a set of different options are available for the time-discretization of the nonlinear convective term.

Based on the implicit Euler scheme, we consider the following time-discretization:

$$\begin{aligned} \frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} - \nu \Delta \mathbf{u}^{n+1} + (\mathbf{u}^* \cdot \nabla)\mathbf{u}^{**} + \nabla p^{n+1} &= \mathbf{f}^{n+1}, \qquad \text{in } \Omega \\ \text{div} \mathbf{u}^{n+1} &= 0. \end{aligned}$$

where different definitions of the terms $\mathbf{u}^*$ and $\mathbf{u}^{**}$ lead to different schemes. Let us briefly analyse the possible choices.

4.3.1 Implicit nonlinear term

If we consider an implicit treatment for the nonlinear convective term, namely:

$$(\mathbf{u}^* \cdot \nabla)\mathbf{u}^{**} = (\mathbf{u}^{n+1} \cdot \nabla)\mathbf{u}^{n+1},$$

the resulting time-discretization reads:

$$\begin{aligned} \frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} - \nu \Delta \mathbf{u}^{n+1} + (\mathbf{u}^{n+1} \cdot \nabla)\mathbf{u}^{n+1} + \nabla p^{n+1} &= \mathbf{f}^{n+1}, \qquad \text{in } \Omega \\ \text{div} \mathbf{u}^{n+1} &= 0. \end{aligned}$$

In this case, the algebraic system to be solved at each time step is nonlinear, thus requires for its solution a fixed-point or Newton iteration. The computational cost of such approach is further incremented by the need of assembling matrix (and possibly preconditioner) at each time step. On the other hand, the main advantage of an implicit approach is that it is unconditionally stable regardless the time step used.

Let us prove this stability results for the homogeneous Dirichlet case. We multiply the momentum equation by the solution $\mathbf{u}^{n+1}$ and integrate over Ω.

As for the continuous case, we analyse the different resulting terms:

$$\int_\Omega \frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\Delta t}\cdot\mathbf{u}^{n+1}\,d\Omega = \frac{1}{\Delta t}\left(\frac{1}{2}\|\mathbf{u}^{n+1}\|^2-\frac{1}{2}\|\mathbf{u}^n\|^2+\frac{1}{2}\|\mathbf{u}^{n+1}-\mathbf{u}^n\|^2\right),$$
$$\int_\Omega -\nu\Delta\mathbf{u}^{n+1}\cdot\mathbf{u}^{n+1}\,d\Omega = \int_\Omega \nu\nabla\mathbf{u}^{n+1}:\nabla\mathbf{u}^{n+1}\,d\Omega = \nu\|\nabla\mathbf{u}^{n+1}\|^2,$$
$$\int_\Omega((\mathbf{u}^{n+1}\cdot\nabla)\mathbf{u}^{n+1})\cdot\mathbf{u}^{n+1}\,d\Omega = \mathbf{0},$$
$$\int_\Omega \nabla p^{n+1}\cdot\mathbf{u}^{n+1}\,d\Omega = -\int_\Omega p^{n+1}\mathrm{div}\mathbf{u}^{n+1}\,d\Omega = \mathbf{0},$$

where we have used the skew-symmetry of the nonlinear term. Summing up all the terms, we get:

$$\frac{1}{2\Delta t}\|\mathbf{u}^{n+1}\|^2-\frac{1}{2\Delta t}\|\mathbf{u}^n\|^2+\underbrace{\frac{1}{2\Delta t}\|\mathbf{u}^{n+1}-\mathbf{u}^n\|^2}_{\geq 0}+\nu\|\nabla\mathbf{u}^{n+1}\|^2 =$$
$$= \int_\Omega \mathbf{f}^{n+1}\cdot\mathbf{u}^{n+1}\,d\Omega \leq \frac{1}{2\nu}\|\mathbf{f}^{n+1}\|^2_{H^{-1}}+\frac{\nu}{2}\|\nabla\mathbf{u}^{n+1}\|^2$$

which implies, at each time step, the following inequality

$$\|\mathbf{u}^{n+1}\|^2+\nu\Delta t\|\nabla\mathbf{u}^{n+1}\|^2 \leq \frac{\Delta t}{\nu}\|\mathbf{f}^{n+1}\|^2_{H^{-1}}+\|\mathbf{u}^n\|^2$$

The telescoping series obtained summing over n yields the final stability estimate

$$\|\mathbf{u}^N\|^2+\nu\sum_{n=1}^N\Delta t\|\nabla\mathbf{u}^n\|^2 \leq \frac{1}{\nu}\sum_{n=1}^N\Delta t\|\mathbf{f}^n\|^2+\|\mathbf{u}_0\|^2. \tag{4.20}$$

Comparing this result with the energy estimate (4.6) obtained for the continuous problem, we notice that the term $\sum_{n=1}^N\Delta t\|\nabla\mathbf{u}^n\|^2$ can be seen as the discrete counterpart of the norm $L^2(0,T;H^1)$ while the term $\|\mathbf{u}^N\|^2$ corresponds to $\|\mathbf{u}(T)\|^2$, controlling the L^2-norm of the solution at time T.

4.3.2 Semi-implicit nonlinear term

To reduce the computational complexity required by a fully implicit nonlinear term, a semi-implicit treatment can be considered, given by

$$(\mathbf{u}^*\cdot\nabla)\mathbf{u}^{**} = (\mathbf{u}^n\cdot\nabla)\mathbf{u}^{n+1}.$$

The resulting time-discretization of the Navier-Stokes problem reads

$$\frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\Delta t}-\nu\Delta\mathbf{u}^{n+1}+(\mathbf{u}^n\cdot\nabla)\mathbf{u}^{n+1}+\nabla p^{n+1}=\mathbf{f}^{n+1},\qquad \text{in } \Omega$$
$$\text{div}\mathbf{u}^{n+1}=0.$$

The advantage of this approach with respect to the fully implicit case relies on the fact that, in this case, the system to be solved at each time step is linear. However, the matrix depends on the solution $\mathbf{u}^n$ and needs to be recomputed each time and is non-symmetric. The computational cost for one time step is equivalent to one iteration of the fixed-point algorithm for steady problems.

Concerning the stability properties of the semi-implicit scheme, we notice that, when homogeneous Dirichlet boundary conditions are considered, thanks to the skew-symmetry of the convection term, we have

$$\int_\Omega(\mathbf{u}^n\cdot\nabla)\mathbf{u}^{n+1}\cdot\mathbf{u}^{n+1}=0$$

and, therefore, the same energy estimate obtained for the implicit case (4.20) still holds and, therefore, also the semi-implicit scheme is unconditionally stable.

4.3.3 Explicit nonlinear term

To further reduce the computational complexity, one could consider an explicit treatment of the nonlinear term

$$(\mathbf{u}^*\cdot\nabla)\mathbf{u}^{**}=(\mathbf{u}^n\cdot\nabla)\mathbf{u}^n$$

corresponding to the following time discretization

$$\frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\Delta t}-\nu\Delta\mathbf{u}^{n+1}+(\mathbf{u}^n\cdot\nabla)\mathbf{u}^n+\nabla p^{n+1}=\mathbf{f}^{n+1},\qquad \text{in } \Omega$$
$$\text{div}\mathbf{u}^{n+1}=0.$$

The problem to solve at each time step reduces, in this case, to a generalized Stokes problem characterized by a symmetric matrix which does not change at each time step. This strong reduction of the computational complexity required at each time step is, however, balanced by poorer stability properties. In particular, due to the explicit treatment of the convective term, the time step should satisfy a CFL (Courant-Friedrics-Levy) condition given by

$$\Delta t\le\frac{h}{\|\mathbf{u}^n\|_{L^\infty}}$$

which may become too penalizing in presence of high velocities.

4.4 Higher order schemes

So far we have considered implicit, semi-implicit and explicit time discretizations all based on the first order Euler method. Higher order time advancing schemes can also be adopted. Many different options are available for the numerical solution of system of ordinary differential equations, ranging from linear multi-step methods to nonlinear Runge-Kutta schemes. Here we only mention two possible second-order schemes, based on the Backward Difference Formula (BDF2) and the Crank-Nicolson scheme, respectively.

4.4.1 Backward Difference Formula (BDF2)

We first consider a second-order approximation of the time derivative based on the Backward Difference Formula of order 2 (BDF2), namely

$$\frac{\partial \mathbf{u}}{\partial t}\Big|_{t^{n+1}} \approx \frac{1}{\Delta t}\left(\frac{3}{2}\mathbf{u}^{n+1} - 2\mathbf{u}^{n} + \frac{1}{2}\mathbf{u}^{n-1}\right).$$

The resulting time discretization of the Navier-Stokes equations reads:

$$\frac{3\mathbf{u}^{n+1} - 4\mathbf{u}^{n} + \mathbf{u}^{n-1}}{2\Delta t} - \nu\Delta\mathbf{u}^{n+1} + (\mathbf{u}^{*}\cdot\nabla)\mathbf{u}^{**} + \nabla p^{n+1} = \mathbf{f}^{n+1},$$
$$\mathrm{div}\mathbf{u}^{n+1} = 0.$$

where different options are available for the treatment of the nonlinear convective term:

- **implicit:** $(\mathbf{u}^{*}\cdot\nabla)\mathbf{u}^{**} = (\mathbf{u}^{n+1}\cdot\nabla)\mathbf{u}^{n+1}$;
- **semi-implicit:** $(\mathbf{u}^{*}\cdot\nabla)\mathbf{u}^{**} = ((2\mathbf{u}^{n} - \mathbf{u}^{n-1})\cdot\nabla)\mathbf{u}^{n+1}$;
- **explicit:** $(\mathbf{u}^{*}\cdot\nabla)\mathbf{u}^{**} = ((2\mathbf{u}^{n} - \mathbf{u}^{n-1})\cdot\nabla)(2\mathbf{u}^{n} - \mathbf{u}^{n-1})$.

In the explicit and semi-implicit cases a second-order extrapolation of the solution at time t^{n+1}, defined as

$$\mathbf{u}^{n+1} \approx 2\mathbf{u}^{n} - \mathbf{u}^{n-1},$$

is required in order to guarantee the second-order convergence in time. The BDF2 is a two-step method, therefore, in the initialization phase, one step of a one-step method will be required. Notice that a single step of a first order one-step method (such as, e.g., the Euler scheme) will not affect the global second-order time convergence of the scheme.

4.4.2 Crank-Nicolson

A one-step second-order time advancing method is given by the Crank-Nicolson scheme, in which the time derivative at $t^{n+\frac{1}{2}}$ is approximated by centered finite differences, namely

$$\frac{\partial \mathbf{u}}{\partial t}\Big|_{t^{n+\frac{1}{2}}} \approx \frac{1}{\Delta t}\left(\mathbf{u}^{n+1} - \mathbf{u}^{n}\right).$$

leading to the following time discretization of the Navier-Stokes equations:

$$\frac{\mathbf{u}^{n+1} - \mathbf{u}^{n}}{\Delta t} - \nu\Delta\left(\frac{\mathbf{u}^{n+1} + \mathbf{u}^{n}}{2}\right) + (\mathbf{u}^{*}\cdot\nabla)\mathbf{u}^{**} + \nabla\left(\frac{p^{n+1} + p^{n}}{2}\right) = \mathbf{f}^{n+\frac{1}{2}},$$
$$\mathrm{div}\mathbf{u}^{n+1} = 0.$$

where, once again, the nonlinear convective term can be:

- **implicit:** $(\mathbf{u}^{*}\cdot\nabla)\mathbf{u}^{**} = \left(\left(\frac{\mathbf{u}^{n+1}+\mathbf{u}^{n}}{2}\right)\cdot\nabla\right)\left(\frac{\mathbf{u}^{n+1}+\mathbf{u}^{n}}{2}\right)$;
- **semi-implicit:** $(\mathbf{u}^{*}\cdot\nabla)\mathbf{u}^{**} = \left(\left(\frac{3}{2}\mathbf{u}^{n} - \frac{1}{2}\mathbf{u}^{n-1}\right)\cdot\nabla\right)\left(\frac{\mathbf{u}^{n+1}+\mathbf{u}^{n}}{2}\right)$;
- **explicit:** $(\mathbf{u}^{*}\cdot\nabla)\mathbf{u}^{**} = \frac{3}{2}(\mathbf{u}^{n}\cdot\nabla)\mathbf{u}^{n} - \frac{1}{2}(\mathbf{u}^{n-1}\cdot\nabla)\mathbf{u}^{n-1}$.

In this case the explicit and semi-implicit requires the adoption of a second-order extrapolation of the solution at time $t^{n+\frac{1}{2}}$, given by:

$$\mathbf{u}^{n+\frac{1}{2}} \approx \frac{3}{2}\mathbf{u}^{n} - \frac{1}{2}\mathbf{u}^{n-1}.$$

4.5 Semi-Lagrangian methods

An alternative approach in the time discretization of the Navier-Stokes equations relies on the approximation of the *Lagrangian derivative*. Let us first define the trajectory of a fluid particle. We denote with $\mathbf{X}(\mathbf{x}, s; t)$ the trajectory of the particle that at time s is in $\mathbf{x}$. The trajectory is the solution of the following Cauchy problem:

$$\begin{cases} \dfrac{d\mathbf{X}}{dt} = \mathbf{u}(\mathbf{X}, t), \\ \mathbf{X}(\mathbf{x}, s; s) = \mathbf{x}. \end{cases} \tag{4.21}$$

The material (or Lagrangian) derivative denotes the rate of variation of a quantity ϕ along the particle trajectory and is defined as:

$$\frac{D\phi}{Dt} = \frac{\partial\phi}{\partial t} + \mathbf{u}\cdot\nabla\phi = \frac{d}{dt}\phi(\mathbf{X}(t), t).$$

We introduce the following compact notation for the foot of the trajectory

$$\mathbf{X}^n(\mathbf{x}) = \mathbf{X}(\mathbf{x}, t^{n+1}; t^n)$$

that is the position occupied at time t^n by the particle that at time t^{n+1} is in $\mathbf{x}$.

In semi-Lagrangian methods, a discretization of the Lagrangian derivative is considered; for instance using a first-order backward finite difference we have

$$\left.\frac{D\mathbf{u}}{Dt}\right|_{t^{n+1}}(\mathbf{x}) \approx \frac{\mathbf{u}^{n+1}(\mathbf{x}) - \mathbf{u}^n(\mathbf{X}^n(\mathbf{x}))}{\Delta t}.$$

Notice that, in order to compute this approximation, one should integrate backward in time the problem (4.21) and evaluate the root of the trajectory $\mathbf{X}^n(\mathbf{x})$. However, the velocity field $\mathbf{u}^{n+1}$ is unknown.

At the discrete level, this problem is faced by computing an approximation of the trajectory defined, for instance, as

$$\mathbf{X}_h^n \approx \mathbf{x} - \mathbf{u}_h^n(\mathbf{x})\Delta t,$$

which corresponds to the forward Euler discretization of problem (4.21). Since $\mathbf{u}_h^n$ is defined over a triangulation, the trajectory is computed element-wise and one should pay particular care when the trajectory crosses a boundary of the domain.

As far as we are able to compute an approximation of the root of the trajectory, the time discretization of the Navier-Stokes problem requires, at each time step, the solution of the following (linear) problem: *given* $(\mathbf{u}_h^n, p_h^n)$, *find* $(\mathbf{u}_h^{n+1}(t), p_h^{n+1}(t)) \in \mathbf{V}_h \times Q_h$, *such that*

$$\begin{aligned}\frac{1}{\Delta t}(\mathbf{u}_h^{n+1}, \mathbf{v}_h) + a(\mathbf{u}_h^{n+1}, \mathbf{v}_h) + b(\mathbf{v}_h, p_h^{n+1}) &= \frac{1}{\Delta t}(\mathbf{u}_h^n(\mathbf{X}^n), \mathbf{v}_h) + F(\mathbf{v}_h),\\ b(\mathbf{u}_h^{n+1}, q_h) &= 0,\end{aligned}$$

Under suitable conditions on the Lagrangian map (in general not easy to verify a priory), an unconditionally stability result can be proved. For a detailed discussion of semi-Lagrangian methods for linear and nonlinear advection problems, we refer to [20].

4.6 Projection methods

Consider a differential problem of the following type:

$$\frac{\partial w}{\partial t} + L_1 w + L_2 w = f,$$

where L_1 and L_2 represent two differential operators.

A *fractional step method* is based on the idea of splitting the time advancing from t^n to t^{n+1} in two sub-problems each accounting for one differential operator, namely:

$$\frac{\widetilde{w}^{n+1} - w^n}{\Delta t} + L_1 \widetilde{w}^{n+1} = 0$$
$$\frac{w^{n+1} - \widetilde{w}^{n+1}}{\Delta t} + L_2 w^{n+1} = f$$

where $\widetilde{w}^{n+1}$ is an auxiliary intermediate velocity.

The *projection methods* are a particular class of fractional step methods in which the diffusion and advection operators are split from the incompressibility constraint.

4.6.1 The Chorin-Temam method

The original method was introduced by Chorin and Temam [11, 68] and is referred to as the *Chorin-Temam method*. In this case, the splitting of the time-dependent Navier-Stokes equations is given by the following two sub-problems to be solve sequentially at each time iteration:

Step 1

$$\frac{\widetilde{\mathbf{u}} - \mathbf{u}^n}{\Delta t} - \nu \Delta \widetilde{\mathbf{u}} + (\mathbf{u}^* \cdot \nabla)\mathbf{u}^{**} = 0,$$

Step 2

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

The two sub-problems require suitable boundary conditions. Step 1 is an advection-diffusion problem for the intermediate velocity $\widetilde{\mathbf{u}} \in [H^1(\Omega)]^d$ and inherits the same boundary conditions imposed to the original Navier-Stokes problem. On the other hand, step 2 is not diffusive and the solution should be seek in a different functional space, namely $\mathbf{u} \in H_{\mathrm{div}} = \{\mathbf{v} \in [L^2(\Omega)]^d, \mathrm{div}\mathbf{v} = 0\}$. Unfortunately, in this space the trace over the boundary $\mathbf{u}|_{\partial\Omega}$ cannot be defined. We can only define the trace of the normal component of the solution $\mathbf{u} \cdot \mathbf{n}|_{\partial\Omega}$.

This implies that, if in the original problem homogeneous Dirichlet condition are imposed over the boundary ($\mathbf{u}|_{\partial\Omega} = \mathbf{0}$), the problem defined in step 2 is well-posed when only the normal component $\mathbf{u} \cdot \mathbf{n}|_{\partial\Omega} = 0$ is imposed. This mismatch on the boundary conditions of the final (end-of-step) velocity contributes to the *splitting error* associated to the fractional step approach. In this case, due to the different boundary condition imposed on the end-of-

step velocity an artificial boundary layer of thickness $\sqrt{\Delta t}$ often appears in the solution.

The step 2 in the Chorin-Temam projection method is referred to as the *projection step*. Indeed, if we multiply the first equation in step 2 by a test function $\boldsymbol{\phi} \in H^0_{\mathrm{div}} = \{\mathbf{v} \in H_{\mathrm{div}}, \mathbf{v} \cdot \mathbf{n} = 0\}$ (in the case of homogeneous Dirichlet conditions), we get:

$$\int_\Omega \frac{\mathbf{u}^{n+1} - \widetilde{\mathbf{u}}}{\Delta t} \cdot \boldsymbol{\phi}\, d\Omega = -\int_\Omega \nabla p^{n+1} \cdot \boldsymbol{\phi}\, d\Omega = \int_\Omega p^{n+1} \mathrm{div}\mathbf{u}^{n+1}\, d\Omega = 0,$$

which implies that

$$\int_\Omega \mathbf{u}^{n+1} \cdot \boldsymbol{\phi}\, d\Omega = \int_\Omega \widetilde{\mathbf{u}} \cdot \boldsymbol{\phi}\, d\Omega,$$

showing that the end-of-step velocity $\mathbf{u}^{n+1}$ is the orthogonal projection of $\widetilde{\mathbf{u}}$ in the subspace $H^0_{\mathrm{div}} \subset L^2$.

This result is a direct consequence of the following theorem:

Theorem 4.2 (Helmholtz' decomposition). *Given a simply connected domain $\Omega \subset \mathbb{R}^d$, for any function $\mathbf{v} \in [L^2(\Omega)]^d$ there exists a unique decomposition into a solenoidal (divergence-free) and an irrotational (curl-free) part, such that:*

$$\mathbf{v} = \mathbf{w} + \nabla\phi, \qquad \textit{with } \mathbf{w} \in H^0_{div}(\Omega) \quad \textit{and } \phi \in H^1(\Omega).$$

From the projection step of the Chorin-Temam method, the decomposition for the case at hand reads

$$\widetilde{\mathbf{u}} = \underbrace{\mathbf{u}^{n+1}}_{\substack{\text{solenoidal}\\ \text{part}}} + \underbrace{\Delta t \nabla p^{n+1}}_{\substack{\text{irrotational}\\ \text{part}}}.$$

The solution $(\mathbf{u}^{n+1}, p^{n+1})$ of the projection step can be computed more efficiently by reformulating the problem as an elliptic problem on the pressure. Indeed, if we apply the divergence operator to the first equation:

$$\mathrm{div}\left(\frac{\mathbf{u}^{n+1} - \widetilde{\mathbf{u}}}{\Delta t} + \nabla p^{n+1}\right) = \frac{1}{\Delta t}(\underbrace{\mathrm{div}\mathbf{u}^{n+1}}_{=0} - \mathrm{div}\widetilde{\mathbf{u}}) + \mathrm{div}\nabla p^{n+1} = 0,$$

a Poisson problem for the pressure p^{n+1} is obtained, namely

$$\Delta p^{n+1} = \frac{1}{\Delta t}\mathrm{div}\widetilde{\mathbf{u}}.$$

To derive the boundary condition for this equation, we consider the normal component of the first equation evaluated on the boundary:

$$\left(\frac{\mathbf{u}^{n+1}-\widetilde{\mathbf{u}}}{\Delta t}+\nabla p^{n+1}\right)\cdot\mathbf{n}=\frac{1}{\Delta t}(\mathbf{u}^{n+1}\cdot\mathbf{n}-\widetilde{\mathbf{u}}\cdot\mathbf{n})+\nabla p^{n+1}\cdot\mathbf{n}=,0$$

Thus, a Neumann condition on the pressure

$$\frac{\partial p^{n+1}}{\partial\mathbf{n}}=0,$$

is obtained from the velocity boundary condition on the normal component $\mathbf{u}^{n+1}\cdot\mathbf{n}=0$, which, as discussed above, is actually an artificial condition that may lead to the development on numerical boundary layers.

To summarize, the solution of the Chorin-Temam projection method can be computed solving the following steps:

- **1 - Advection-diffusion problem for intermediate velocity**

$$\begin{cases}\dfrac{\widetilde{\mathbf{u}}-\mathbf{u}^n}{\Delta t}+(\mathbf{u}^*\cdot\nabla)\mathbf{u}^{**}-\nu\Delta\widetilde{\mathbf{u}}=0, & \text{in } \Omega, t>0\\ \widetilde{\mathbf{u}}=\mathbf{0} & \text{on } \partial\Omega, t>0\end{cases} \tag{4.22}$$

- **2 - Poisson problem for pressure**

$$\begin{cases}\Delta p^{n+1}=\dfrac{1}{\Delta t}\text{div}\widetilde{\mathbf{u}}, & \text{in } \Omega, t>0\\ \dfrac{\partial p^{n+1}}{\partial\mathbf{n}}=0 & \text{on } \partial\Omega, t>0\end{cases} \tag{4.23}$$

- **3 - Projection step**

$$\mathbf{u}^{n+1}=\widetilde{\mathbf{u}}-\Delta t\nabla p^{n+1}$$

4.6.2 Incremental Chorin-Temam method

One of the main limitations of the original Chorin-Temam method relies on the fact that the method is not *consistent* for steady solutions. Indeed, if $(\mathbf{u},p)$ is a steady solution of the Navier-Stokes equations, it is easy to see that $\mathbf{u}$ does not satisfy the first step of the Chorin-Temam method. To recover consistency for steady solution the *incremental* version of the Chorin-Temam method can be considered. In this case, the pressure is expressed as

$$p^{n+1}=p^n+\delta p,$$

and the problem is reformulated in terms of the pressure increment δp. The incremental Chorin-Temam method is then defined by the following steps:

- **1 - Advection-diffusion problem for intermediate velocity**

$$\begin{cases} \dfrac{\widetilde{\mathbf{u}}-\mathbf{u}^n}{\Delta t}+(\mathbf{u}^*\cdot\nabla)\mathbf{u}^{**}-\nu\Delta\widetilde{\mathbf{u}}=-\nabla p^n, & \text{in } \Omega, t>0 \\ \widetilde{\mathbf{u}}=\mathbf{0} & \text{on } \partial\Omega, t>0 \end{cases} \tag{4.24}$$

- **2 - Poisson problem for pressure**

$$\begin{cases} \dfrac{\mathbf{u}^{n+1}-\widetilde{\mathbf{u}}}{\Delta t}+\nabla\delta p=0 \\ \mathrm{div}\mathbf{u}^{n+1}=0 \end{cases} \rightarrow \begin{cases} \Delta\delta p=\dfrac{1}{\Delta t}\mathrm{div}\widetilde{\mathbf{u}}, & \text{in } \Omega, t>0 \\ \dfrac{\partial\delta p}{\partial\mathbf{n}}=0 & \text{on } \partial\Omega, t>0 \end{cases} \tag{4.25}$$

- **3 - Projection step**

$$\mathbf{u}^{n+1}=\widetilde{\mathbf{u}}-\Delta t\nabla\delta p$$

- **4 - Pressure update**

$$p^{n+1}=p^n+\delta p.$$

It is easy to see that, now, a steady solution $(\mathbf{u}^n, p^n)$ of Navier-Stokes equations is preserved by the scheme. Indeed, in this case, the solution of step 1 is $\widetilde{\mathbf{u}}=\mathbf{u}^n$ the solution of step 2 is $\delta p=0$, and the projection step gives $\mathbf{u}^{n+1}=\widetilde{\mathbf{u}}=\mathbf{u}^n$.

As it was mentioned above, the projection methods introduce a splitting error which is usually reflected in the appearance of artificial boundary layers on pressure. The splitting error is of order $\mathcal{O}(\sqrt{\Delta t})$ for the original Chorin-Temam method and of order $\mathcal{O}(\Delta t)$ for its incremental version. A further improvement can be obtained introducing a pressure correction step (see [70]) which reduces the splitting error to $\mathcal{O}(\Delta t^{3/2})$.

Many different versions of projection methods, and more in general of fractional step methods, for the solution of the time-dependent Navier-Stokes equations have been proposed in the literature over the past three decades. For an interesting review on these methods and for rigorous time convergence analyses, we refer to [28] and references therein.

4.7 Inexact factorization methods

Let us consider again the time-discretization of the Navier-Stokes equations with a semi-implicit (or explicit) treatment of the nonlinear term:

$$\begin{cases} \dfrac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\Delta t}-\nu\Delta\mathbf{u}^{n+1}+(\mathbf{u}^*\cdot\nabla)\mathbf{u}^{**}+\nabla p^{n+1}=\mathbf{f}^{n+1}, \\ \mathrm{div}\mathbf{u}^{n+1}=0. \end{cases} \tag{4.26}$$

with

$$(\mathbf{u}^*\cdot\nabla)\mathbf{u}^{**}=\begin{cases} (\mathbf{u}^n\cdot\nabla)\mathbf{u}^n & \text{(Explicit)} \\ (\mathbf{u}^n\cdot\nabla)\mathbf{u}^{n+1} & \text{(Semi-implicit)} \end{cases}$$

The algebraic counterpart of this problem reads:

$$\begin{bmatrix} C & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} U^{n+1} \\ P^{n+1} \end{bmatrix} = \begin{bmatrix} \widetilde{F}^{n+1} \\ \mathbf{0} \end{bmatrix}, \tag{4.27}$$

with

$$C = \frac{1}{\Delta t} M + \nu K, \qquad \widetilde{F}^{n+1} = F^{n+1} + \frac{1}{\Delta t} M U^n - N(U^n) U^n,$$

for an explicit treatment of the nonlinear term and

$$C = \frac{1}{\Delta t} M + \nu K + N(U^n), \qquad \widetilde{F}^{n+1} = F^{n+1} + \frac{1}{\Delta t} M U^n,$$

for a semi-implicit-treatment of the nonlinear term.

The (exact) LU block factorization of the matrix

$$\Sigma = \begin{bmatrix} C & B^T \\ B & 0 \end{bmatrix}$$

is given by

$$\Sigma = \begin{bmatrix} C & B^T \\ B & 0 \end{bmatrix} = \underbrace{\begin{bmatrix} C & 0 \\ B & -BC^{-1}B^T \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} I & C^{-1}B^T \\ 0 & I \end{bmatrix}}_{U}. \tag{4.28}$$

Most often, this LU factorization can not be used, in practice, for the solution of the problem, since it would require the computation of C^{-1} which is too expensive to afford. The *inexact factorization methods* [56] rely on the idea of replacing C^{-1} with suitable approximations (easier to compute) in the factorization (4.28), namely:

$$\Sigma \approx \hat{\Sigma} = \underbrace{\begin{bmatrix} C & 0 \\ B & -BH_1B^T \end{bmatrix}}_{\hat{L}} \underbrace{\begin{bmatrix} I & H_2B^T \\ 0 & I \end{bmatrix}}_{\hat{U}}, \tag{4.29}$$

where H_1 and H_2 are two approximations of C^{-1}.

To construct suitable approximation of C^{-1}, we can rewrite C (for the semi-implicit case) as follows:

$$C = \frac{1}{\Delta t} M + \nu K + N(U^n) = \frac{1}{\Delta t} M (I + \Delta t M^{-1} \underbrace{(\nu K + N(U^n))}_{W}).$$

Note that, for Δt sufficiently small, we have

$$\rho(\Delta t M^{-1} W) < 1,$$

where $\rho(\cdot)$ denotes the spectral radius. Under this condition, we can use the Neumann expansion to get

$$(I + \Delta t M^{-1}W)^{-1} = \sum_{k=0}^{\infty} (\Delta t M^{-1}W)^k (-1)^k.$$

Taking the first order term in Δt, an approximation of C^{-1} is given by:

$$\begin{aligned} C^{-1} = \Delta t (I + \Delta t M^{-1}W)^{-1} M^{-1} &= \Delta t \left(\sum_{k=0}^{\infty} (\Delta t M^{-1}W)^k (-1)^k \right) M^{-1} \\ &= \Delta t M^{-1} + \mathcal{O}(\Delta t^2). \end{aligned}$$

We could then choose $\Delta t M^{-1}$ in (4.29) as an approximation of H_1 and/or H_2. An even more efficient choice is obtained replacing M with its *lumped* version M_L.

4.7.1 Algebraic Chorin-Temam method

Setting

$$H_1 = H_2 = \Delta t M_L^{-1}$$

the solution of the algebraic problem

$$\hat{A} \begin{bmatrix} U^{n+1} \\ P_{n+1} \end{bmatrix} = \hat{L}\hat{U} \begin{bmatrix} U^{n+1} \\ P_{n+1} \end{bmatrix} = \begin{bmatrix} \widetilde{F}^{n+1} \\ \mathbf{0} \end{bmatrix}$$

can be decoupled in the following two steps:

$$\hat{L} \begin{bmatrix} \widetilde{U} \\ \widetilde{P} \end{bmatrix} = \begin{bmatrix} \widetilde{F}^{n+1} \\ \mathbf{0} \end{bmatrix}, \qquad \hat{U} \begin{bmatrix} U^{n+1} \\ P^{n+1} \end{bmatrix} = \begin{bmatrix} \widetilde{U} \\ \widetilde{P} \end{bmatrix},$$

corresponding to the system:

$$\begin{aligned} C\widetilde{U} &= \widetilde{F}^{n+1}, \\ -\Delta t B M_L^{-1} B^T \widetilde{P} &= -B\widetilde{U}, \\ P^{n+1} &= \widetilde{P}, \\ U^{n+1} &= \widetilde{U} - \Delta t M_L^{-1} B^T P^{n+1}. \end{aligned}$$

Note that the third equation is trivial and can be removed setting $\widetilde{P} = P^{n+1}$ in the whole algorithm. Moreover, we remark that the matrix $BM_L^{-1}B^T$ can be interpreted as a discrete version of the Laplacian operator. The resulting method based on this inexact factorization is referred to as *Algebraic*

Chorin-Temam (ACT) method since it is the algebraic counterpart of the standard Chorin-Temam projection method:

$$
\begin{aligned}
&1) \qquad C\widetilde{U} = \widetilde{F}^{n+1} \longleftrightarrow \begin{cases} \dfrac{\widetilde{\mathbf{u}} - \mathbf{u}^n}{\Delta t} + (\mathbf{u}^* \cdot \nabla)\mathbf{u}^{**} - \nu \Delta \widetilde{\mathbf{u}} = 0, \\ \widetilde{\mathbf{u}} = \mathbf{0} \end{cases} \\
&2) \qquad \Delta t B M_L^{-1} B^T P^{n+1} = B\widetilde{U} \longleftrightarrow \begin{cases} \Delta p^{n+1} = \dfrac{1}{\Delta t} \mathrm{div} \widetilde{\mathbf{u}}, \\ \dfrac{\partial p^{n+1}}{\partial \mathbf{n}} = 0 \end{cases} \\
&3) \; U^{n+1} = \widetilde{U} - \Delta t M_L^{-1} B^T P^{n+1} \longleftrightarrow \mathbf{u}^{n+1} = \widetilde{\mathbf{u}} - \Delta t \nabla p^{n+1}
\end{aligned}
$$

The main difference between the original Chorin-Temam method and it algebraic version is that the latter does not require, in step 2, to impose artificial boundary conditions on the pressure, since the original Dirichlet boundary condition on velocity are embedded into the matrix $BM_L^{-1}B^T$. There is numerical evidence that, thanks to this difference, the algebraic version displays weaker numerical boundary layers in the pressure solution.

4.7.2 Incremental algebraic Chorin-Temam method

As for the differential Chorin-Temam method, an incremental version of the algebraic Chorin-Temam method can be derived, reformulating the problem in terms of the pressure increment $\delta P = P^{n+1} - P^n$, as follows:

$$
\begin{bmatrix} C & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} U^{n+1} \\ P^n + \delta P \end{bmatrix} = \begin{bmatrix} \widetilde{F}^{n+1} \\ \mathbf{0} \end{bmatrix} \quad \longleftrightarrow \quad \begin{bmatrix} C & B^T \\ B & 0 \end{bmatrix} \begin{bmatrix} U^{n+1} \\ \delta P \end{bmatrix} = \begin{bmatrix} \widetilde{F}^{n+1} - B^T P^n \\ \mathbf{0} \end{bmatrix}
$$

The incremental algebraic Chorin-Temam method than reads:

$$
\begin{aligned}
&1) \qquad C\widetilde{U} = \widetilde{F}^{n+1} - B^T P^n \\
&2) \qquad B M_L^{-1} B^T \delta P = \frac{1}{\Delta t} B\widetilde{U} \\
&3) \qquad U^{n+1} = \widetilde{U} - \Delta t M_L^{-1} B^T \delta P \\
&4) \qquad P^{n+1} = P^n + \delta P
\end{aligned}
$$

4.7.3 Yosida method

The inexact factorization approach can be used to recover an algebraic version of the standard and incremental Chorin-Temam methods. Moreover, further methods, with improved convergence properties, have been developed under this framework. In particular, setting

$$
H_1 = \Delta t M_L^{-1}, \qquad H_2 = C^{-1},
$$

we get the so-called *Yosida method* [55]. In this case, the inexact factorization reads

$$\hat{A} = \begin{bmatrix} C & 0 \\ B & -\Delta t B M_L^{-1} B^T \end{bmatrix} \begin{bmatrix} I & C^{-1}B^T \\ 0 & I \end{bmatrix}$$

which leads to the following method:

1) $C\widetilde{U} = \widetilde{F}^{n+1}$

2) $BM_L^{-1}B^T P^{n+1} = \dfrac{1}{\Delta t} B\widetilde{U}$

3) $U^{n+1} = \widetilde{U} - \Delta t C^{-1} B^T P^{n+1} \quad \longrightarrow \quad CU^{n+1} = \widetilde{F}^{n+1} \Delta t B^T P^{n+1}$

The incremental version of the Yosida method can be easily derived following the same procedure used for the Chorin-Temam method.

Comparing Yosida and Chorin Temam methods, we first notice that the former is computational more expansive since it requires, in step 3, the solution of an additional problem on the velocity. However, this additional system makes it possible to impose the exact boundary conditions to the end-of-step velocity U^{n+1}. The splitting error of the different methods depends on the choice of H_1 and H_2. Since

$$\hat{\Sigma} = \begin{bmatrix} C & 0 \\ B & -BH_1B^T \end{bmatrix} \begin{bmatrix} I & H_2B^T \\ 0 & I \end{bmatrix} = \begin{bmatrix} C & CH_2B^T \\ B & B(H_2 - H_1)B^T \end{bmatrix},$$

it is interesting to note that, for the algebraic Chorin-Temam method ($H_1 = H_2$) no error is introduced on the continuity equation (U^{n+1} is discretely divergence free). On the other hand, the Yosida method ($H^2 = C^{-1}$) is exact on the momentum equation and the solution U^{n+1} is not exactly divergence-free.

Higher order versions of the Yosida methods can be derived considering additional terms in the approximation of C^{-1} (see [24]).

4.7.4 Inexact algebraic factorization methods as preconditioners

We have seen how inexact LU factorizations can be used to devise algebraic projection-like methods for the solution of the time-dependent Navier-Stokes equations. The approximate matrix $\hat{\Sigma} = \hat{L}\hat{U}$ can also be considered as a preconditioner for the original complete system

$$\Sigma \begin{bmatrix} U^{n+1} \\ P^{n+1} \end{bmatrix} = \begin{bmatrix} F^{n+1} \\ \mathbf{0} \end{bmatrix}. \tag{4.30}$$

For the sake of simplicity, we consider an iterative solution based on the Richardson method. For a linear system $AX = B$, given an initial guess X_0 and a preconditioner $\mathcal{P}$, the Richardson method generate the sequence

$$X_{k+1} = X_k + \mathcal{P}(B - AX_k) = X_k + \mathcal{P}(R_k), \quad k = 1, 2, \ldots.$$

To apply the Richardson method to system (4.30), we denote by R^u and R^p the residuals of the momentum and continuity equations, respectively, and, given the initial guess $U_0 = U^n$ and $P_0 = P^n$, for any $k = 0, 1, \ldots$ we iterate as follows

$$\hat{\Sigma} \begin{bmatrix} U_{k+1} - U_k \\ P_{k+1} - P_k \end{bmatrix} = \begin{bmatrix} R_k^u \\ R_k^p \end{bmatrix} = \begin{bmatrix} F^{n+1} \\ \mathbf{0} \end{bmatrix} - \Sigma \begin{bmatrix} U_k \\ P_k \end{bmatrix},$$

where the preconditioner $\hat{\Sigma}$ is defined by the inexact factorization

$$\hat{\Sigma} = \hat{L}\hat{U} = \begin{bmatrix} C & 0 \\ B & -BH_1B^T \end{bmatrix} \begin{bmatrix} I & H_2B^T \\ 0 & I \end{bmatrix}.$$

At each iteration, the solution of the following block triangular systems is thus required:

$$\hat{L} \begin{bmatrix} \widetilde{U}_{k+1} - U_k \\ \widetilde{P}_{k+1} - P_k \end{bmatrix} = \begin{bmatrix} R_k^u \\ R_k^p \end{bmatrix}, \qquad \hat{U} \begin{bmatrix} U_{k+1} - U_k \\ P_{k+1} - P_k \end{bmatrix} = \begin{bmatrix} \widetilde{U}_{k+1} - U_k \\ \widetilde{P}_{k+1} - P_k \end{bmatrix},$$

corresponding to the following sequence of linear systems

1) $C\widetilde{U}_{k+1} = F^{n+1} - B^T P_k$,
2) $BH_1B^T(\widetilde{P}_{k+1} - P_k) = B\widetilde{U}_{k+1}$,
3) $P_{k+1} = \widetilde{P}_{k+1}$,
4) $U_{k+1} = \widetilde{U}_{k+1} - H_2B^T(P_{k+1} - P_k)$.

We notice that these linear systems correspond to those required in the algebraic Chorin-Temam and Yosida methods, with the difference that, here, we need to iterate them until convergence. On the other hand, when the inexact factorization is used as a preconditioner, the original problem is solved without introducing any splitting error.

The preconditioner $\hat{\Sigma}$ can be further modified as follows

$$\hat{\Sigma} = \hat{L}\hat{U} = \begin{bmatrix} \hat{C} & 0 \\ B & -BHB^T \end{bmatrix} \begin{bmatrix} I & \hat{C}^{-1}B^T \\ 0 & Q \end{bmatrix} = \begin{bmatrix} C & B^T \\ 0 & (B\hat{C}^{-1}B^T - BHB^TQ) \end{bmatrix},$$

where $\hat{C}$ is a suitable approximation of C and Q is chosen such that the error $\|(B\hat{C}^{-1}B^T - BHB^TQ)\|$ is minimized.

It is important to notice that many classical coupling algorithms (such as SIMPLE, PISO, SIMPLER, SIMPLEC) that have been proposed in the CFD literature in the past 4 decades can be interpreted, in this framework, as inexact factorization preconditioners (see [18]).

Chapter 5
Numerical methods for free-surface flows

In this chapter, we discuss the numerical approximation of fluid dynamic problems with moving boundaries. In particular, we will consider *two-fluid flows* in which two immiscible fluids share a common moving interface which can be referred to as the *free-surface*. Both fluids are considered incompressible and the position of the free-surface interface is unknown and should be determined as part of the problem solution.

5.1 Two-fluids flow equations

We consider a bounded domain $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$. At each time $t > 0$, two different fluids, fluid 1 and fluid 2, fill the sub-domains $\Omega_1(t)$ and $\Omega_2(t)$, respectively, such that $\Omega = \text{int}(\overline{\Omega_1(t) \cup \Omega_2(t)})$.

In each of the two sub-domains $\Omega_i(t), i = 1, 2$, the flow is governed by the incompressible Navier–Stokes equations:

$$\rho_i \partial_t \mathbf{u}_i + \rho_i (\mathbf{u}_i \cdot \nabla)\mathbf{u}_i - \text{div}(\boldsymbol{\sigma}_i(\mathbf{u}_i, p_i)) = \rho_i \mathbf{g}, \tag{5.1}$$

$$\text{div}(\mathbf{u}_i) = 0, \tag{5.2}$$

where $\mathbf{u}_i$ is the velocity field, p_i is the pressure, ρ_i is the density, $\mathbf{g} = (0, 0, g)^T$ is the gravity acceleration, and $\boldsymbol{\sigma}_i(\mathbf{u}_i, p_i) = \mu_i(\nabla \mathbf{u}_i + \nabla \mathbf{u}_i{}^T) - p_i \mathbf{I}$ is the stress tensor with μ_i indicating the dynamic viscosity. Subscript i indicates that all the quantities are restricted to the sub-domain Ω_i and time derivatives are denoted by ∂_t. Equation (5.1) enforces the conservation of linear momentum, while equation (5.2) is the constraint of incompressibility which enforces mass conservation in each sub-domain Ω_i.

The free surface Γ is a sharp interface between Ω_1 and Ω_2 share the sharp interface Γ. We denote with $\mathbf{n} = \mathbf{n}_2 = -\mathbf{n}_1$ the interface normal unit vector pointing from Ω_2 to Ω_1.

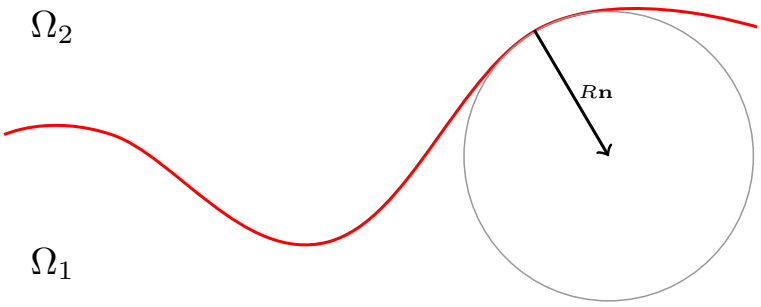


Fig. 5.1: Example of positive local curvature with the chosen convention on the interface normal $\mathbf{n}$

Since there is no flow through the interface, the normal components of the two velocities $\mathbf{u}_1 \cdot \mathbf{n}$ and $\mathbf{u}_2 \cdot \mathbf{n}$ should coincide on Γ. Furthermore, the tangential components must match as well since the two flows are viscous. Thus we have the following kinematic condition

$$\mathbf{u}_1 = \mathbf{u}_2, \qquad \text{on } \Gamma. \tag{5.3}$$

Moreover, the forces acting on the fluid at the free-surface are in equilibrium. This is a dynamic condition and means that the normal forces on either side of Γ are of equal magnitude and opposed direction, while the tangential forces must agree in both magnitude and direction:

$$\boldsymbol{\sigma}_1(\mathbf{u}_1, p_1) \cdot \mathbf{n} - \boldsymbol{\sigma}_2(\mathbf{u}_2, p_2) \cdot \mathbf{n} = -\kappa\gamma\mathbf{n} \quad \text{on } \Gamma, \tag{5.4}$$

where γ is the surface tension coefficient which is here considered constant. The quantity κ in (5.4) is twice the mean curvature of the free-surface, namely:

$$\kappa = \sum_{i=1}^{d-1} \frac{1}{R_i},$$

where R_i denotes the radii of curvature along the principal directions. The sign of R_i is such that the vector $R_i\mathbf{n}$ points from the interface to the center of the circle approximating the interface locally (see Figure 5.1 for a two-dimensional sketch).

Problem (5.1)-(5.2) can be rewritten as a single set of density-dependent Navier-Stokes equations holding on the entire domain Ω:

$$\partial_t \rho(\mathbf{x}) + \mathbf{u} \cdot \nabla\rho(\mathbf{x}) = 0, \tag{5.5}$$

$$\rho(\mathbf{x})\partial_t \mathbf{u} + \rho(\mathbf{x})(\mathbf{u} \cdot \nabla)\mathbf{u} - \mathrm{div}(\boldsymbol{\sigma}(\mathbf{u}, p)) = \rho(\mathbf{x})\mathbf{g} + \mathbf{f}_\Gamma, \tag{5.6}$$

$$\mathrm{div}\mathbf{u} = 0, \tag{5.7}$$

where $\mathbf{u}$, p and $\boldsymbol{\sigma}(\mathbf{u},p) = \mu(\mathbf{x})(\nabla\mathbf{u}+\nabla\mathbf{u}^T) - p\mathbf{I}$ are now defined in the whole Ω and $\rho(\mathbf{x})$ and $\mu(\mathbf{x})$ are variable density and viscosity coefficients. With respect to problem (5.1)-(5.2), the additional equation (5.5) is introduced to express the mass conservation over the whole domain Ω, with the incompressibility constraint (5.7) remaining valid. Equations (5.5)-(5.7) have to be interpreted in the sense of distributions, given the discontinuous nature of $\rho(\mathbf{x})$ and $\mu(\mathbf{x})$.

Note that in (5.6) an additional source term $\mathbf{f}_\Gamma$ appears. It accounts for the jump on the normal stress tensor in the dynamic interface condition (5.4) and it is defined by:

$$\mathbf{f}_\Gamma = \kappa\gamma\delta_\Gamma\mathbf{n}, \tag{5.8}$$

where δ_Γ is the Dirac delta function with support on Γ. For a formal derivation of equation (5.8), we refer to [3].

System (5.5)-(5.7) can be rewritten in conservation form, as follows:

$$\partial_t\rho + \mathrm{div}(\rho\mathbf{u}) = 0, \tag{5.9}$$

$$\partial_t(\rho\mathbf{u}) + \mathrm{div}((\rho\mathbf{u}\otimes\mathbf{u}) - \mathrm{div}(\boldsymbol{\sigma}(\mathbf{u},p)) = \rho\mathbf{g} + \mathbf{f}_\Gamma, \tag{5.10}$$

$$\mathrm{div}\mathbf{u} = 0. \tag{5.11}$$

The above equations have to be complemented with suitable initial and boundary conditions (see 5.1.1). The initial conditions for velocity and density are given as follows

$$\begin{aligned}\mathbf{u}(\mathbf{x},0) &= \mathbf{u}_0, \quad \forall\mathbf{x}\in\Omega,\\ \rho(\mathbf{x},0) &= \rho_0, \quad \forall\mathbf{x}\in\Omega,\end{aligned}$$

where $\mathbf{u}_0$ is a divergence-free velocity field.

A global existence result can be proven for the solution of (5.5)-(5.7) provided Ω is a bounded, connected, open subset of $\mathbb{R}^3$ with smooth boundary (the latter condition is not satisfied in the case at hand, indeed the boundary is only Lipschitz-continuous). In that case, if $\rho_0 \in L^\infty(\Omega)$ and $\mathbf{u}_0 \in (H^1(\Omega))^3$, $\mathbf{m}_0 = \rho\mathbf{u}|_{t=0} \in L^\infty(\Omega)$ a weak solution exists (see [38]) which satisfies

$$\begin{aligned}\rho &\in L^\infty(0,T;L^\infty(\Omega)),\\ \mathbf{u} &\in L^2(0,T;H^1_0(\Omega))^3,\\ \rho|\mathbf{u}|^2 &\in L^\infty(0,T;L^1(\Omega)),\\ \nabla\mathbf{u} &\in L^2(\Omega\times(0,T)),\\ \rho &\in C([0,T];L^p(\Omega)), \qquad \forall p : 1\le p<\infty.\end{aligned} \tag{5.12}$$

Moreover, the following energy inequalities hold

$$\frac{d}{dt}\int_\Omega \rho|\mathbf{u}|^2\ d\Omega + \int_\Omega \mu(\partial_i u_j + \partial_j u_i)^2\ d\Omega \leq 2\int_\Omega \rho\mathbf{g}\cdot\mathbf{u}\ d\Omega \quad \text{in } \mathcal{D}'(0,T), \tag{5.13}$$

$$\int_\Omega \rho|\mathbf{u}|^2\ d\Omega + \int_0^t\int_\Omega \mu(\partial_i u_j + \partial_j u_i)^2\ d\Omega\, ds \leq \int_\Omega |\mathbf{m}_0|^2/\rho_0\ d\Omega + 2\int_0^t\int_\Omega \rho\mathbf{g}\cdot\mathbf{n}\ d\Omega\, ds \quad \text{a.e. } t\in(0,T), \tag{5.14}$$

where ∂_i denotes partial derivative w. r. to x_i, $\mathcal{D}'(0,T)$ is the space of distributions on $(0,T)$ and the summation convention on repeated indexes applies.

5.1.1 Boundary conditions

Suitable boundary conditions have to be imposed to close problem (5.5)-(5.7).

We consider a subdivision of the boundary $\Sigma = \partial\Omega$ in four regions:

- an *inflow* region Σ_{in}, where a Dirichlet boundary condition is imposed
$$\mathbf{u}|_{\Sigma_{\text{in}}} = \mathbf{u}_{\text{in}}(t), \quad \forall t \in]0,T];$$
- an *outflow* region Σ_{out}, where a zero normal stress boundary condition is imposed:
$$\boldsymbol{\sigma}(\mathbf{u},p)\cdot\mathbf{n}|_{\Sigma_{\text{out}}} = \mathbf{0}, \quad \forall t \in]0,T],$$
where $\mathbf{n}$ is the unit outward normal on Σ_{out};
- a *wall* region Σ_{w}, where a no-slip Dirichlet boundary condition is imposed:
$$\mathbf{u}|_{\Sigma_{\text{w}}} = \mathbf{u}_{\text{w}}(t), \quad \forall t \in]0,T];$$
- a *symmetry* region Σ_{sym}, where a symmetry boundary condition is imposed:
$$\begin{aligned}\mathbf{u}\cdot\mathbf{n}|_{\Sigma_{\text{sym}}} &= 0, \quad \forall t \in]0,T],\\ \nabla(\mathbf{u}-(\mathbf{u}\cdot\mathbf{n})\mathbf{n})\cdot\mathbf{n}|_{\Sigma_{\text{sym}}} &= 0, \quad \forall t \in]0,T],\end{aligned}$$
which states that the normal velocity component is zero as well as the normal derivatives of the tangential velocity.

Note that, if Dirichlet boundary conditions on the velocity are imposed on the whole boundary $\Sigma_{\text{in}} = \Sigma$, the initial and boundary conditions should satisfy additional compatibility conditions (see [53]).

A boundary condition on ρ has to be prescribed only at the inflow Σ_{in}:

$$\rho(\mathbf{x},t)|_{\Sigma_{\text{in}}} = \rho_{\text{in}}(t), \qquad \forall t \in]0,T],$$

where we have assumed that no back-flow is present on the outflow region, such that

$$\Sigma_{\text{in}} = \{\mathbf{x} \in \Sigma \mid (\mathbf{u}\cdot\mathbf{n}) < 0\}.$$

5.2 Numerical approaches for free-surface flows

The initial position of the interface is known but its evolution in time has to be computed as part of the solution of problem (5.5)-(5.7).

Several numerical methods for the solution of free-surface problems have been proposed in the literature in the past few decades. These methods can be classified based on their ability to treat different physical situations (*i.e.* flow regimes or types of waves) and on their computational complexity. The choice of the most suitable approach depends therefore on the specific problem at hand.

A popular (although not exhaustive) classification of the numerical methods for free-surface problems divides them in two main categories:

- *Front Tracking methods*: the free-surface interface is explicitly tracked along the trajectory of the fluid particles, making use of the *kinematic* interface condition (5.3). These methods are usually based on a Lagrangian of mixed Eulerian-Lagrangian approach, where the computational grid is adapted to the interface and must be readjusted each time the free-surface is moved (see Fig. 5.2, left).
- *Front Capturing methods*: the free-surface interface is reconstructed from the properties of an appropriate field function (*e.g.* phase volume fraction or density). These methods are based on an Eulerian approach: the computational grid is fixed and both the regions occupied by liquid and gas are modelled (see Fig. 5.2, right).

We refer to [62] for a discussion on the different approaches and for a classification of methods based on the different aspects of the numerical models (flow model, interface model, flow-interface coupling, discretization methods).

5.2.1 Front Tracking methods

Front Tracking methods are often used in two-fluid flow problems involving a liquid phase and a gaseous phase, choosing to discretize only the liquid phase (on a moving domain) and neglecting the contribution of the gaseous phase: the interface is treated as a boundary of the computational domain (see, *e.g*, [46, 33, 73]).

If surface tension effects are neglected, given the position $\Gamma(t)$ of the free-surface at a given time, the flow equations can be defined on the moving (liquid) domain $\Omega(t)$, the dynamic interface condition (5.4) becomes a homogeneous Neumann condition on $\Gamma(t)$ and the kinematic interface condition (5.3) can be used to update the interface condition at the subsequent time step.

Such problem can be solved resorting to the so-called *Arbitrary Lagrangian-Eulerian* approach that will be introduced in Section 5.3. This approach requires to adapt the computational grid to the moving boundary and it is typically adopted when the interface motion has a limited amplitude. Indeed, when large free-surface deformations occur, the grid element can become highly skewed, which is usually a problem for the stability and accuracy of the flow solver. In these cases, local or global remeshing may become necessary.

The intrinsic limitations of front tracking methods when dealing with complex free-surface topologies opened the field to the development of alternative front capturing methods.

5.2.2 Front Capturing methods

Due to their flexibility in dealing with complex free-surface problems, *Front Capturing methods* are receiving an increasing attention in many application domains in which free-surface phenomena occur, ranging from micro-fluidic to naval engineering problems.

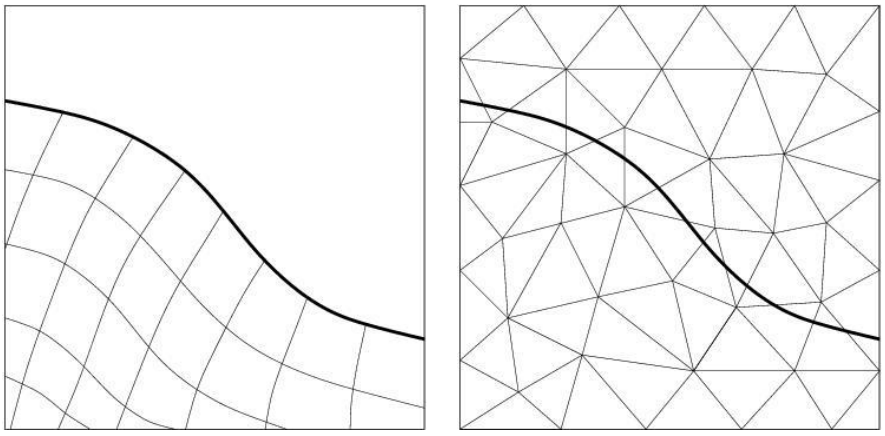


Fig. 5.2: Typical grid topologies in 2D for Lagrangian (left) and Eulerian (right) free-surface methods. The thick line represents the free-surface.

A first approach of this type was considered in the so-called *Marker-and-Cell* method introduced in [29]. In this method, particles with neither mass nor energy are distributed in the whole fluid domain to track the free-surface location. These particles do not play any role in the dynamics of the fluid and are not taken into account in the solution of the flow equations.

Indeed, the Marker-and-Cell method was the first approach able to handle complex and general situations (such as breaking surfaces, splash and fluid detachment) and its original implementation has later been improved by several authors (see, *e.g.*, [43, 71, 45]). On the other hand, its computational cost for the solution of large three-dimensional problems is prohibitive, because of the need of a large number of particles to capture the free-surface shape properly. Moreover, to avoid the generation of false regions of void, the particles need to be periodically redistributed.

The most commonly employed Front Capturing methods for the simulation of free-surface flows with complex interfaces are the so-called *Volume of Fluid* (VOF) method and *Level Set* method. For both, the computational grid is fixed and both the regions occupied by liquid and gas are usually modelled. The interface between the different immiscible fluids is "captured" by solving an additional advection equation for a suitable scalar variable.

Indeed, both methods are based on a splitting of problem (5.5)-(5.7): the momentum equation (5.6) and the incompressibility constraint (5.7) are decoupled from the mass conservation equation (5.5). The latter is replaced by an advection equation for either a discontinuous function (the volume fraction, for the Volume of Fluid method) or a continuous function (the signed distance from the interface, for the Level Set method). In the VOF method the interface is identified as a *discontinuity line* of the volume fraction, while in the Level Set method the interface is implicitly represented by the zero Level Set of the signed distance function.

In the VOF method, originally introduced in [32], the dynamics of the interface is computed by advecting a function $\psi(\mathbf{x}, t)$ which represents the volume fraction of one phase (*e.g.* water) in each cell. The value of ψ is 1 in the cells completely filled by water and 0 in those filled by air. The cells where $0 < \psi < 1$ identify the *interface region* that should be kept as sharp as possible. This method requires the solution of a pure advection equation for the discontinuous function $\psi(\mathbf{x}, t)$, which reads

$$\frac{\partial \psi}{\partial t} + \mathbf{u} \cdot \nabla \psi = 0. \tag{5.15}$$

The Navier–Stokes equations can be either solved in the entire computational domain [59] or only in the region occupied by water [39]. If the first approach is used, the local values of density and viscosity are computed from the volume fraction, as follows

$$\begin{aligned} \rho(\mathbf{x}, t) &= \psi(\mathbf{x}, t)\, \rho_1 + (1 - \psi(\mathbf{x}, t))\, \rho_2, \\ \mu(\mathbf{x}, t) &= \psi(\mathbf{x}, t)\, \mu_1 + (1 - \psi(\mathbf{x}, t))\, \mu_2. \end{aligned} \tag{5.16}$$

Being ψ discontinuous, the numerical solution of equation (5.15) requires special care. If low order schemes are used, the numerical diffusion will lead to a smearing of the interface over several grid elements. On the other hand, standard high-order schemes can lead to the appearance of oscillations in the interface region. To overcome these difficulties, one can resort to high resolution schemes for hyperbolic problems developed for the numerical solution of conservation laws [35] (see, *e.g.*, the high-resolution interface capturing (HRIC) scheme proposed in [44] and the immersed interface method described in [37]). An alternative approach, extensively used in the Volume of Fluid literature, makes use of interface reconstruction techniques based on purely geometrical consideration. Examples are the donor-acceptor algorithm [58, 21], the SLIC algorithm [48, 40] and the PLIC algorithm [1, 59, 52]. However, the implementation of these techniques for three dimensional problems, in particular when unstructured grids are employed, is not always straightforward and the computational cost for large three-dimensional computations can become excessively high. Another aspect that can be critical in VOF methods is the evaluation of the interface curvature, essential in applications where surface tension effects are relevant. Suitable algorithms for the reconstruction of the local curvature from the discontinuous volume fraction field have to be considered (see, *e.g.*, [8]).

The Level Set method, that will be described in details in Section 5.4, is based on the solution of the following advection equation:

$$\frac{\partial \phi}{\partial t} + \mathbf{u} \cdot \nabla \phi = 0, \tag{5.17}$$

where ϕ is a smooth function defined in the whole computational domain as the signed distance function from the interface:

$$\begin{aligned} \phi(\mathbf{x}) &:= \mathrm{dist}(\mathbf{x}, \Gamma), \qquad \forall \mathbf{x} \in \Omega_1, \\ \phi(\mathbf{x}) &:= -\mathrm{dist}(\mathbf{x}, \Gamma), \qquad \forall \mathbf{x} \in \Omega_2. \end{aligned}$$

Negative values of ϕ correspond to fluid 1, while positive values to fluid 2. The zero level set of ϕ implicitly represents the interface.

The idea underlying this type of method was first proposed in [12], where the interface was defined as the zero level set of a continuous *pseudo-density* function. The Level Set method was introduced in [50] for the numerical solution of front-propagating problems with curvature-dependent motion and then extended to a variety of physical applications. We refer to the two books [61] and [49] and to references therein for an overview of the numerical schemes and the description of a large collection of problems treated by the Level Set method.

The property of ϕ being a distance function is not preserved during advection. It has been shown in [66] that, a *reinitialization* procedure is necessary in order to restore this property to the level set function, at least in regions

close to the interface. As a result, this procedure enhances the performance of the numerical algorithm.

In most of the Level Set literature, the finite difference spatial discretization is utilized (see, *e.g.*, [50, 66, 65, 64, 10, 74]. Finite element approximations have also been considered by some authors [72, 62, 27].

The main advantage of the Level Set method, when compared with the VOF method, is that the advection equation (5.17) is solved for a continuous function, rather that a discontinuous one. Moreover, the evaluation of geometrical quantities, such as interface normals and curvature, is much easier. On the other hand, the VOF method guarantees better mass conservation properties [69]. Indeed, in the Level Set method, the mass conservation properties strongly depend on the numerical schemes adopted for the solution of equation (5.17) and on the reinitialization procedure.

5.3 Arbitrary Lagrangian-Eulerian (ALE) method

The Arbitrary Lagrangian-Eulerian (ALE) method [14, 15] is a numerical approach that has been successfully adopted for the numerical approximation of different moving domain problems, including free-surface flows.

The main idea of the method is to combine the advantages of the Eulerian and Lagrangian formulations in a hybrid formulation which allows to treat moving boundary fluid-dynamic problems.

Let us consider the incompressible Navier-Stokes equations defined on a moving domain $\Omega(t)$ such that its boundary $\partial\Omega(t)$ can be subdivided in a moving portion $\Gamma(t)$ (which may represent the moving interface in a free-surface problem) and a fixed portion $\partial\Omega(t) \setminus \Gamma(t)$. Let us also assume that the motion of $\Gamma(t)$ is given (or can be determined by the kinematic interface condition).

The moving domain $\Omega(t)$ is recast at each time t to a reference fixed configuration $\hat{\Omega}$ (typically chosen as the initial configuration $\Omega(t_0)$ through the so-called ALE mapping:

$$\mathcal{A}_t : \hat{\Omega} \longrightarrow \Omega(t).$$

We denote by $\mathbf{x}(\hat{\mathbf{x}}, t) = \mathcal{A}_t(\hat{\mathbf{x}})$ the point $\mathbf{x}$ in the current domain $\Omega(t)$ corresponding to the point $\hat{\mathbf{x}}$ in the reference domain $\hat{\mathbf{x}}$ (see Figure 5.3).

Given the ALE mapping it is possible to introduce the domain (or ALE) velocity defined as

$$\mathbf{w}(\mathbf{x}, t) = \frac{\partial \mathcal{A}_t}{\partial t} \circ \mathcal{A}_t^{-1},$$

that, in turn, can be used to define the so-called ALE derivative:

$$\left.\frac{\partial \mathbf{u}}{\partial t}\right|_{\hat{\mathbf{x}}} = \left.\frac{\partial \mathbf{u}}{\partial t}\right|_{\mathbf{x}} + \mathbf{w}\cdot\nabla\mathbf{u}.$$

The ALE derivative denotes the rate of change of the quantity $\mathbf{u}$ along the trajectory defined by the ALE velocity.

The incompressible Navier-Stokes equations on a moving domain $\Omega(t)$ in ALE form can then be written as follows

$$\rho\left.\frac{\partial \mathbf{u}}{\partial t}\right|_{\hat{\mathbf{x}}} + \rho((\mathbf{u}-\mathbf{w})\cdot\nabla)\mathbf{u} - \mathrm{div}(\boldsymbol{\sigma}(\mathbf{u},p)) = \rho\mathbf{g}, \tag{5.18}$$

$$\mathrm{div}(\mathbf{u}) = 0, \tag{5.19}$$

where the spatial derivatives are computed on the deformed configuration $\Omega(t)$ while the time derivative is computed along the mesh trajectory.

Given the test functions $\hat{\mathbf{v}} \in [H_0^1(\hat{\Omega})]^d$ and $\hat{q} \in L^2(\hat{\Omega})$ defined in the reference domain, the corresponding test functions mapped to the deformed configuration are given by

$$\mathbf{v} = \hat{\mathbf{v}} \circ \mathcal{A}_t^{-1}, \qquad q = \hat{q} \circ \mathcal{A}_t^{-1}.$$

The weak formulation of the ALE problem can be obtained multiplying (5.18)-(5.19) by $\mathbf{v}$ and q, respectively, and integrating over $\Omega(t)$: $\forall \mathbf{v} \in [H_0^1(\Omega(t))]^d$ and $q \in L^2(\Omega(t))$

$$\int_{\Omega(t)} \left(\rho \left(\left.\frac{\partial \mathbf{u}}{\partial t}\right|_{\hat{\mathbf{x}}} + ((\mathbf{u}-\mathbf{w})\cdot\nabla)\mathbf{u} \right) + \boldsymbol{\sigma} : \nabla\mathbf{v} + \mathrm{div}\mathbf{u}q \right) d\Omega = \int_{\Omega(t)} \rho\mathbf{f}\cdot\mathbf{v}\, d\Omega.$$

The ALE finite-element approximation can be obtained starting from a triangulation $\hat{\mathcal{T}}_h$ of the reference domain $\hat{\Omega}$ and chasing a finite-element space pair $(\hat{\mathbf{V}}_h, \hat{Q}_h)$ on $\hat{\mathcal{T}}_h$ for velocity and pressure.

At each time instant, the triangulation $\mathcal{T}_h$ of the current configuration $\Omega(t)$ can be obtained as the image of $\hat{\mathcal{T}}_h$ through the ALE mapping $\mathcal{A}_t$.

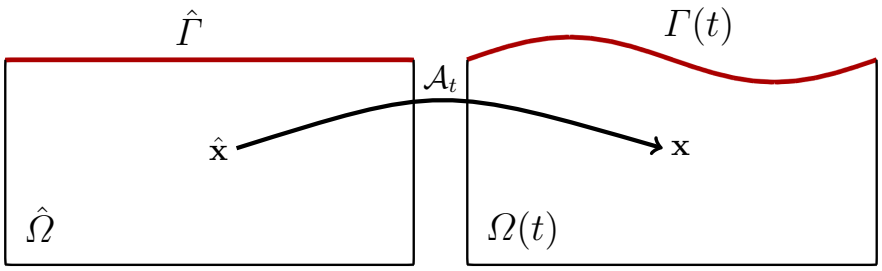


Fig. 5.3: ALE mapping between reference domain $\hat{\Omega}$ and current domain $\Omega(t)$.

Indeed, if the ALE mapping $\mathcal{A}_t$ is a piecewise linear function on $\hat{\mathcal{T}}_h$ than $\mathcal{T}_h$ is a triangulation of $\Omega(t)$. The finite-element spaces $(\mathbf{V}_h, Q_h)$ defined on the triangulation $\mathcal{T}_h$ can also be obtained mapping $(\hat{\mathbf{V}}_h, \hat{Q}_h)$ through $\mathcal{A}_t$ (see Figure 5.4).

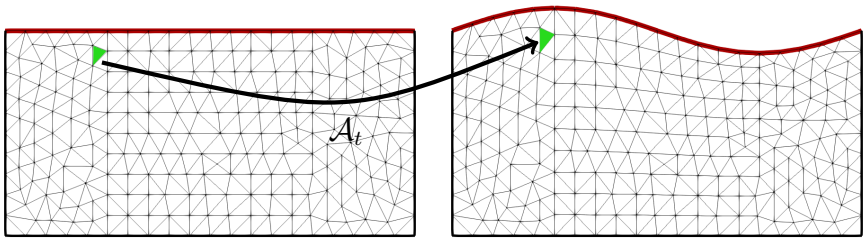


Fig. 5.4: Triangulation of reference domain $\hat{\Omega}$ and current domain $\Omega(t)$.

Note that if $\hat{Q}_h$ (resp. $\hat{\mathbf{V}}$) is the space of piecewise polynomial functions of degree k on $\hat{\mathcal{T}}_h$, then Q_h (resp. $\mathbf{V}$) is the space of piecewise polynomial functions of degree k on $\mathcal{T}_h$.

The ALE finite-element approximation of system (5.18)-(5.19) reads: $\forall \mathbf{v}_h \in V_h$ and $q \in Q_h$

$$\int_{\Omega(t)} \left(\rho \left(\left. \frac{\partial \mathbf{u}_u}{\partial t} \right|_{\hat{\mathbf{x}}} + ((\mathbf{u}_h - \mathbf{w}_h) \cdot \nabla) \mathbf{u}_h \right) + \boldsymbol{\sigma}_h : \nabla \mathbf{v}_h + \mathrm{div} \mathbf{u}_h q_h \right) d\Omega = \int_{\Omega(t)} \rho \mathbf{f} \cdot \mathbf{v}_h \, d\Omega.$$

If we consider an implicit time-discretization with a semi-implicit treatment of the convecting term the resulting algebraic problem reads

$$\begin{bmatrix} \frac{1}{\Delta t} M^n + A^n + N^n(U^{n-1} - W^n) & (B^n)^T \\ B^n & 0 \end{bmatrix} \begin{bmatrix} U^{n+1} \\ P^n \end{bmatrix} = \begin{bmatrix} \frac{1}{\Delta t} M^n U^{n-1} + F^n \\ \mathbf{0} \end{bmatrix}, \tag{5.20}$$

where all matrices have to be recomputed at each time step to account for the domain motion.

At the discrete level, a common strategy to evaluate the ALE mapping relies on the solution of an auxiliary vector elliptic problem which compute the harmonic extension of the displacement η with respect to the reference configuration, defined on the free-boundary $\Gamma(t)$ to the domain $\Omega(t)$. Namely, we compute the new position of any point in $\Omega(t)$ by solving

$$\Delta \mathbf{x}^n = 0, \qquad \text{in } \Omega(t), \tag{5.21}$$

$$\mathbf{x}^n = \hat{\mathbf{x}} + \eta, \quad \text{on } \Gamma(t). \tag{5.22}$$

Solving system (5.21)-(5.22) with P^1 finite elements, the resulting map

$$\mathbf{x}^n(\hat{\mathbf{x}}) = \mathcal{A}_{t^n}(\hat{\mathbf{x}})$$

will be piecewise linear.

Given the ALE map, the ALE velocity can then be evaluated as

$$\mathbf{w}^n = \frac{\mathcal{A}_{t^n} - \mathcal{A}_{t^{n-1}}}{\Delta t}.$$

The harmonic extension approach is rather simple to be implemented and, when the amplitude of the free-surface motion is limited, is typically able to yield an adequate mesh deformation. When the domain motion leads to a poor-quality mesh, a remeshing step may be necessary to allow the simulation to proceed. More robust strategies rely on the solution of the elasticity equations [63] or on Radial Basis Functions [42].

For a detailed analysis of the stability properties of ALE methods, we refer to [47, 22].

5.4 Level-Set method

The density-dependent (inhomogeneous) Navier–Stokes equations (5.5)-(5.7) provide a suitable mathematical model to describe viscous two-fluid flows. Besides, in Section 5.2, we have discussed some possible numerical approaches for the approximation of this problem.

In the *Level Set method*, the interface is defined as the zero-level set of the signed distance function from the interface Γ, designed to be negative in fluid 1 and positive in fluid 2. Given a known velocity field $\mathbf{u}$, the evolution of the interface is determined by solving the following advection equation for ϕ:

$$\phi_t + \mathbf{u} \cdot \nabla\phi = 0. \tag{5.23}$$

The coefficients in the inhomogeneous Navier–Stokes equations (5.5)-(5.7) are functions of ϕ and can be defined as follows:

$$\begin{aligned} \rho(\phi) &:= \rho_1 + (\rho_2 - \rho_1)\, H(\phi), \\ \mu(\phi) &:= \mu_1 + (\mu_2 - \mu_1)\, H(\phi), \end{aligned} \tag{5.24}$$

where H is the Heaviside function.

The two-fluid flow problem in $\Omega \times [0, T]$ can therefore be described, combining system (5.5)-(5.7) with equations (5.23) and (5.24), by the following continuous model:

$$\rho(\phi)\left(\mathbf{u}_t + (\mathbf{u}\cdot\nabla)\mathbf{u}\right) + \mathrm{div}(\boldsymbol{\sigma}) = \rho(\phi)\mathbf{g} + \kappa\gamma\delta(\phi)\mathbf{n}, \tag{5.25}$$

$$\mathrm{div}\mathbf{u} = 0, \tag{5.26}$$

$$\phi_t + \mathbf{u}\cdot\nabla\phi = 0, \tag{5.27}$$

where δ is the Dirac delta distribution. Suitable boundary and initial conditions for $\mathbf{u}$ and ϕ must be imposed to close the problem. In the following, we consider homogeneous Dirichlet boundary condition on the velocity.

To rewrite problem (5.25) in its weak form, we first introduce the following functional spaces

$$\mathbf{V} := (H_0^1(\Omega))^d = \{\mathbf{v} \in (H^1(\Omega))^d \mid \mathbf{v}|_{\partial\Omega} = 0\}, \tag{5.28}$$

$$Q := L_0^2(\Omega) = \left\{ q \in L^2(\Omega) \mid \int_\Omega q = 0 \right\}, \tag{5.29}$$

$$W_\beta(\Omega) := \{\psi \in L^2(\Omega) \mid (\beta\cdot\nabla\psi) \in L^2(\Omega)\}, \tag{5.30}$$

$$W_\beta^0(\Omega) := \{\psi \in V_\beta(\Omega) \mid \psi|_{\partial\Omega^-} = 0\}, \tag{5.31}$$

where β is a vector field such that $\mathrm{div}\beta = 0$ and $\partial\Omega^- = \{\mathbf{x} \in \partial\Omega \mid \beta(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) < 0\}$. $W_\beta(\Omega)$ is a Hilbert space equipped with the norm $\|\psi\|_{1,\beta} = (\|\psi\|_0^2 + \|\beta\cdot\nabla\psi\|_0^2)^{1/2}$.

We introduce the bilinear forms:

$$m_\phi(\mathbf{u},\mathbf{v}) := \int_{\Omega(t)} \rho(\phi)\mathbf{u}\cdot\mathbf{v}\, d\Omega,$$

$$a_\phi(\mathbf{u},\mathbf{v}) := \int_{\Omega(t)} \mu(\phi)(\nabla\mathbf{u} + \nabla\mathbf{u}^T) : \nabla\mathbf{v}\, d\Omega,$$

$$b(\mathbf{u},q) := -\int_{\Omega(t)} \mathrm{div}\mathbf{u} q\, d\Omega,$$

$$l_\mathbf{u}(\phi,\psi) := \int_{\Omega(t)} (\mathbf{u}\cdot\nabla\phi)\psi\, d\Omega,$$

and the trilinear form:

$$c_\phi(\mathbf{u};\mathbf{v},\mathbf{w}) := \int_{\Omega(t)} \rho(\phi)(\mathbf{u}\cdot\nabla)\mathbf{v}\cdot\mathbf{w}\, d\Omega$$

The weak formulation of problem (5.25) reads: find $\mathbf{u}(\mathbf{x},t) \in \mathbf{V}$, $p(\mathbf{x},t) \in Q$ and $\phi(\mathbf{x},t) \in W_\beta^0$ such that:

$$m_\phi(\mathbf{u}_t,\mathbf{v}) + a_\phi(\mathbf{u},\mathbf{v}) + c_\phi(\mathbf{u};\mathbf{u},\mathbf{v}) + b(\mathbf{v},p) = m_\phi(\mathbf{g},\mathbf{v}) + (\mathbf{f}_\Gamma,\mathbf{v}), \quad \forall\mathbf{v} \in \mathbf{V}, \tag{5.32}$$

$$b(\mathbf{u},q) = 0, \quad \forall q \in Q, \tag{5.33}$$

$$(\phi_t,\psi) + l_\mathbf{u}(\phi,\psi) = 0, \quad \forall\psi \in L^2(\Omega), \tag{5.34}$$

for each $t \in (0,T]$.

We introduce now a spatial discretization of problem (5.32)-(5.34) based on a finite element approach. A theoretical analysis concerning the finite element approximation of Navier–Stokes equations with free-surface can be found in [67]. For a general review on numerical methods for the solution of the incompressible Navier–Stokes equations, we refer to [57, 31].

We restrict our attention to the 2-dimensional case. Let us consider a triangulation $\mathcal{T}_h$ of the domain Ω. For the finite element approximation the Navier–Stokes system in problem (5.32)-(5.34), we consider an inf-sup stable space pair $\mathbf{V}_h - Q_h$ for velocity and pressure.

For the approximation of the level-set equation in problem (5.32)-(5.34), we consider the same degrees of freedom as for the velocity.

The semi-discrete formulation of problem (5.32)-(5.34) reads: find $\mathbf{u}_h(t) \in \mathbf{V}_h$, $p_h(t) \in Q_h$ and $\phi_h(t) \in W_h$ such that:

$$\begin{aligned} m_{\phi_h}((\mathbf{u}_h)_t, \mathbf{v}_h) + a_{\phi_h}(\mathbf{u}_h, \mathbf{v}_h) + c_{\phi_h}(\mathbf{u}_h; \mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p_h) \qquad (5.35)\\ = m_{\phi_h}(\mathbf{g}, \mathbf{v}_h) + (\mathbf{f}_{\Gamma_h}, \mathbf{v}_h), \quad \forall \mathbf{v}_h \in \mathbf{V}_h,\\ b(\mathbf{u}_h, q_h) = 0, \quad \forall q_h \in Q_h, \qquad (5.36)\\ ((\phi_h)_t, \psi_h) + l_{\mathbf{u}_h}(\phi_h, \psi_h) + j_h(\phi_h, \psi_h) = 0, \quad \forall \psi_h \in W_h, \qquad (5.37) \end{aligned}$$

for each $t \in (0,T]$. In the level set equation we have added a stabilization term $j_h(\phi, \psi)$. Different choices for the stabilization of the level set equation can be adopted, such as the SUPG stabilization or the sub-grid stabilization proposed in [13]. The purely advection level-set problem (5.37) could also be solved using a semi-Lagrangian approach as the one introduced in 4.5.

The surface tension term in the continuous variational problem (5.32)-(5.34) reads

$$(\mathbf{f}_\Gamma, \mathbf{v}) = \int_\Omega \kappa\gamma\delta(\phi)\mathbf{n}\cdot\mathbf{v}\; d\Omega = \int_\Gamma \kappa\gamma\mathbf{n}\cdot\mathbf{v}\; d\Omega, \qquad \forall \mathbf{v} \in \mathbf{V}, \qquad (5.38)$$

where we have used the fact that the action of δ on a smooth test function w is given by

$$\int_\Omega \delta_\Gamma(x)\, w(x)\, d\Omega = \int_\Gamma v(s)\, ds. \qquad (5.39)$$

One of the major advantages in the level-set method, when compared with other interface capturing methods (*e.g.* VOF) is the possibility of computing the interface normal $\mathbf{n}$ and curvature κ in a very convenient way. Indeed, for any function ϕ normal and curvature of any level-set can be computed as

$$\begin{aligned} \mathbf{n} &= \frac{\nabla\phi}{|\nabla\phi|},\\ \kappa &= -\mathrm{div}\mathbf{n} = -\mathrm{div}\left(\frac{\nabla\phi}{|\nabla\phi|}\right), \end{aligned}$$

and they are defined in the entire domain Ω where the function ϕ is defined. In particular, if ϕ is the signed distance function from the interface Γ the expressions for normal and curvature reduce to

$$\mathbf{n} = \nabla\phi, \tag{5.40}$$
$$\kappa = \mathrm{div}(\nabla\phi), \tag{5.41}$$

At the discrete level, the standard approach in the level set literature relies on the introduction of a smoothed Dirac function δ_ε, defined as:

$$\delta_\varepsilon(\phi(\mathbf{x})) = \left\{ \begin{array}{cc} \frac{1}{2\varepsilon}(1+\cos(\pi\phi(\mathbf{x})/\varepsilon), & \text{if } \phi(\mathbf{x}) \leq \varepsilon \\ 0, & \text{otherwise} \end{array} \right\}, \tag{5.42}$$

where ε is the smoothing parameter that prescribes the artificial thickness of the interface. Typically, the parameter ε is of the order of the grid size h. Given δ_ε, the surface tension term (5.38) can be computed as an integral term over Ω, as follows:

$$(\mathbf{f}_\Gamma, \mathbf{v}) = \int_\Omega \kappa\gamma\delta_\varepsilon(\phi)\mathbf{n}\cdot\mathbf{v}\, d\Omega, \qquad \forall \mathbf{v} \in \mathbf{V}. \tag{5.43}$$

The main drawback of this approach is that it gives rise to an error $\mathcal{O}(h)$ in the interface location. Other approaches based on the evaluation of the surface tension term as a line integral have proved to guarantee second-order accuracy (see, e.g., [62, 51].

Denoting by $N_\mathbf{u}, N_p$ and N_ϕ the number of degrees of freedom of $\mathbf{u}_h, p_h$ and ϕ_h, respectively, problem (5.35)-(5.37) can be rewritten in algebraic form as follows:

Find $\mathbf{U}(t) \in \mathbb{R}^{N_\mathbf{u}}$, $\mathbf{P}(t) \in \mathbb{R}^{N_p}$ and $\Phi(t) \in \mathbb{R}^{N_\phi}$ such that

$$\begin{aligned} M(\Phi(t))\mathbf{U}_t(t) + A(\Phi(t))\mathbf{U}(t) + C(\Phi(t),\mathbf{U}(t))\mathbf{U}(t) + B^T\mathbf{P}(t) \\ = \mathbf{G}(\Phi(t)) + \mathbf{F}_{\Gamma_h}(\Phi(t)), \end{aligned} \tag{5.44}$$
$$B\mathbf{U}(t) = 0, \tag{5.45}$$
$$H\Phi_t(t) + L(\mathbf{U}(t))\Phi(t) + J\Phi(t) = 0. \tag{5.46}$$

The matrices appearing in problem (5.44)-(5.46) are defined as follows:

$$\begin{aligned} M(\Phi(t))_{ij} &= m_{\phi_h}(\boldsymbol{\phi}_j, \boldsymbol{\phi}_i), \\ A(\Phi(t))_{ij} &= a_{\phi_h}(\boldsymbol{\phi}_j, \boldsymbol{\phi}_i), \\ C(\Phi(t),\mathbf{U}(t)))_{ij} &= c_{\phi_h}(\mathbf{u}_h; \boldsymbol{\phi}_j, \boldsymbol{\phi}_i), \\ B_{ij} &= b(\phi_j, \psi_i), \\ H_{ij} &= (\chi_j, \chi_i), \\ L(\mathbf{U})_{ij} &= l_{\mathbf{u}_h}(\chi_j, \chi_i), \\ J_{ij} &= j_h(\chi_j, \chi_i), \end{aligned}$$

where $\{\boldsymbol{\phi}_i\}_{1\le i\le N_{\mathbf{u}}}$, $\{\psi_i\}_{1\le i\le N_p}$ and $\{\chi_i\}_{1\le i\le N_\phi}$ are the nodal basis functions of spaces $\mathbf{V}_h$, Q_h and W_h, respectively.

To reduce the computational complexity of the system, the solution of the Navier–Stokes equations is typically decoupled from that of the level-set equation. Different schemes can be adopted for the time discretization. In this section, we introduce the the first order implicit Euler scheme, although second-order schemes (*e.g.* Crank-Nicholson or Backward Difference Formula (BDF2)) can also be adopted straightforwardly. Nevertheless, the time error introduced by the splitting between Navier–Stokes and level set equations is of first order. The use of higher order time stepping would be justified only if a tighter coupling is adopted.

We consider a uniform decomposition of the time interval $[0,T]$ into N subintervals, namely, if $\Delta t = T/N$ is the time step, we use the index n to denote variables at time $t^n = n\,\Delta t$, with $n = 0,\dots,N$. The time discretization of problem (5.44)-(5.46), combined with the decoupling of the Navier–Stokes equations from the level-set equation, leads to the solution, at each time step, of the algebraic Navier–Stokes system:

$$\begin{aligned}\frac{1}{\Delta t}M(\Phi^n)\mathbf{U}^{n+1} + A(\Phi^n)\mathbf{U}^{n+1} + C(\Phi^n,\mathbf{U}^{n+1})\mathbf{U}^{n+1} + B^T P^{n+1} \qquad (5.47)\\ = \frac{1}{\Delta t}M(\Phi^n)\mathbf{U}^n + \mathbf{G}(\Phi^n) + \mathbf{F}_{\Gamma_h}(\Phi^n),\\ B\mathbf{U}^{n+1} = 0, \qquad (5.48)\end{aligned}$$

followed by the solution of the algebraic level set system:

$$\frac{1}{\Delta t}H\Phi^{n+1} + L(\mathbf{U}^{n+1})\Phi^{n+1} + J\Phi^{n+1} = \frac{1}{\Delta t}H\Phi^n. \qquad (5.49)$$

The nonlinearity in system (5.47) can be treated in different ways. Besides a *fully explicit strategy* (where $C(\Phi^n,\mathbf{U}^{n+1})\mathbf{U}^{n+1}$ is replaced by $C(\Phi^n,\mathbf{U}^n)\mathbf{U}^n$) and the classic *Newton linearization*, a typical approach is the *semi-implicit* one where $C(\Phi^n\mathbf{U}^{n+1})\mathbf{U}^{n+1}$ is replaced by the term $C(\Phi^n,\mathbf{U}^n)\mathbf{U}^{n+1}$).

The inexact-factorization methods introduced in 4.7 can be adopted to reduce the computational complexity associated to the solution of the Navier–Stokes system (5.47).

As mentioned above, when using the Level-Set method accuracy problems may arise as the level-set function lose the property of being a signed distance function from the interface. This phenomenon is clearly depicted in Figure 5.5, where the rising bubble problem is solved without any reinitialization. During the interface evolution, the iso-countours of the level-set function, which are initially equidistant, get closer in the top side of the bubble (where the level-set function becomes steep) and get apart in the bottom side (where the level-set function become flat). This phenomenon can be avoided by resorting to a reinitialization procedure which consists in periodically replacing

the current level-set function the signed-distance function from the current interface, as displayed in Figure 5.6.

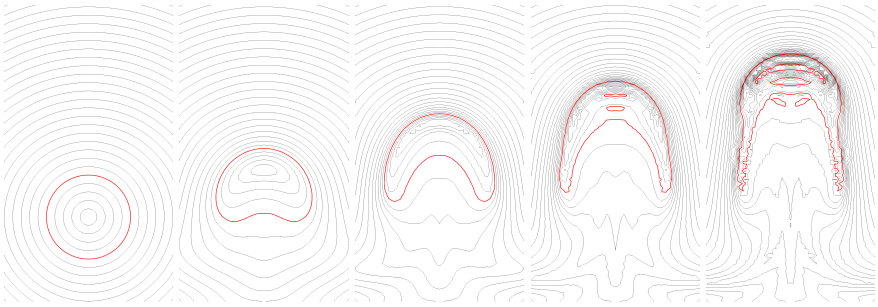


Fig. 5.5: Iso-contours of the level-set function without reinitialization at different time instants. The red contour (zero level-set) represents the interface.

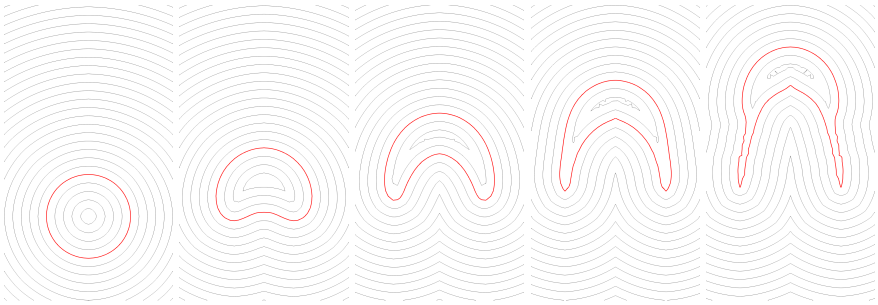


Fig. 5.6: Iso-contours of the level-set function with reinitialization at different time instants. The red contour (zero level-set) represents the interface.

References

1. N. Ashgriz and J. Y. Poo. Flair: fluz line-segment model for advection and interface reconstruction. *J. Comp. Phys.*, 93(2):449–468, 1991.
2. I. Babuska. Error-bounds for finite element method. *Numerische Mathematik*, 16:322–333, 1970/71.
3. J. U. Brackbill, D. B. Kothe, and C. Zemach. A continuum method for modeling surface tension. *J. Comp. Phys.*, 100(2):335–354, 1992.
4. F. Brezzi. On the existence, uniqueness and approximation of saddle-point problems arising from lagrangian multipliers. *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique*, 8(2):129–151, 1974.
5. F. Brezzi and J. Pitkäranta. *On the Stabilization of Finite Element Approximations of the Stokes Equations*, pages 11–19. Vieweg+Teubner Verlag, Wiesbaden, 1984.
6. E. Burman and M. A. Fernández. Galerkin Finite Element Methods with Symmetric Pressure Stabilization for the Transient Stokes Equations: Stability and Convergence Analysis. *SIAM Journal on Numerical Analysis*, 47(1):409–439, 2009.
7. E. Burman and P. Hansbo. Edge stabilization for the generalized stokes problem: a continuous interior penalty method. *Computer methods in applied mechanics and engineering*, 195(19):2393–2410, 2006.
8. A. Caboussat. *Analysis and Numerical Simulation of Free Surface Flows*. Thesis n. 2893, École Polytechnique Fédérale de Lausanne (EPFL), 2003.
9. J. Cahouet and J.-P. Chabard. Some fast 3d finite element solvers for the generalized stokes problem. *International Journal for Numerical Methods in Fluids*, 8(8):869–895, 1988.
10. Y. C. Chang, T. Y. Hou, B. Merriman, and S. Osher. A level set formulation of eulerian interface capturing methods for incompressible fluid flows. *J. Comp. Phys.*, 124(2):449–464, 1996.
11. A. J. Chorin. The numerical solution of the navier-stokes equations for an incompressible fluid. *Bull. Amer. Math. Soc.*, 73(6):928–931, 11 1967.
12. A. Dervieux and F. Thomasset. *Approximation Methods for Navier–Stokes Problems*, volume 771 of *Lecture Notes in Mathematics*, chapter A finite element method for the simulation of Rayleigh-Taylor instability, pages 145–158. Springer-Verlag, Berlin, 1980.
13. D. A. Di Pietro, S. Lo Forte, and N. Parolini. Mass preserving finite element implementations of the level set method. *Applied numerical mathematics*, 56(9):1179–1195, 2006.
14. J. Donea, S. Giuliani, and J. Halleux. An arbitrary lagrangian-eulerian finite element method for transient dynamic fluid-structure interactions. *Computer Methods in Applied Mechanics and Engineering*, 33(1):689–723, 1982.

15. J. Donea, A. Huerta, J.-P. Ponthot, and A. Rodríguez-Ferran. *Arbitrary Lagrangian–Eulerian Methods*, chapter 14. John Wiley & Sons, Ltd, 2004.
16. H. Elman, D. Silvester, and A. Wathen. *Finite Elements and Fast Iterative Solvers: with Applications in Incompressible Fluid Dynamics*. Numerical Mathematics and Scientific Computation. OUP Oxford, 2014.
17. H. C. Elman and G. H. Golub. Inexact and preconditioned uzawa algorithms for saddle point problems. *SIAM Journal on Numerical Analysis*, 31(6):1645–1661, 1994.
18. H. C. Elman, V. E. Howle, J. N. Shadid, R. Shuttleworth, and R. S. Tuminaro. A taxonomy and comparison of parallel block multi-level preconditioners for the incompressible navier-stokes equations. *J. Comput. Phys.*, 227:1790–1808, 2008.
19. A. Ern and J. Guermond. *Theory and Practice of Finite Elements*. Applied Mathematical Sciences. Springer New York, 2004.
20. M. Falcone and R. Ferretti. *Semi-Lagrangian Approximation Schemes for Linear and Hamilton—Jacobi Equations*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2013.
21. G. Fekken, A. E. P. Veldman, and B. Buchner. Simulation of the green water loading using the Navier–Stokes equations. In *Proc. of the 7th International Conference on Numerical Ship Hydrodynamics*, Nantes, 1999.
22. L. Formaggia and F. Nobile. Stability analysis of second-order time accurate schemes for ale–fem. *Computer Methods in Applied Mechanics and Engineering*, 193(39):4097–4116, 2004. The Arbitrary Lagrangian-Eulerian Formulation.
23. L. Formaggia, A. Quarteroni, and A. Veneziani. Cardiovascular mathematics: Modeling and simulation of the circulatory system, 2010.
24. P. Gervasio, F. Saleri, and A. Veneziani. Algebraic fractional-step schemes with spectral methods for the incompressible navier–stokes equations. *Journal of Computational Physics*, 214(1):347–365, 2006.
25. V. Girault and P. Raviart. *Finite Element Methods for Navier-Stokes Equations: Theory and Algorithms*. Springer Series in Computational Mathematics. Springer Berlin Heidelberg, 2012.
26. P. Gresho and R. Sani. *Incompressible Flow and the Finite Element Method, Advection-Diffusion and Isothermal Laminar Flow*. Incompressible Flow and the Finite Element Method. John Wiley & Sons, 2000.
27. S. Gross, V. Reichelt, and A. Reusken. A finite element based level set method for two-phase incompressible flows, 2004. IGPM Report Nr. 243.
28. J. Guermond, P. Minev, and J. Shen. An overview of projection methods for incompressible flows. *Computer Methods in Applied Mechanics and Engineering*, 195(44–47):6011 – 6045, 2006.
29. F. H. Harlow and J. E. Welch. Numerical calculation of time-dependent viscous incompressible flow of fluid with a free interface. *Physics of Fluids*, 8:2182–2189, 1965.
30. J. G. Heywood and R. Rannacher. Finite element approximation of the nonstationary navier–stokes problem. i. regularity of solutions and second-order error estimates for spatial discretization. *SIAM Journal on Numerical Analysis*, 19(2):275–311, 1982.
31. J. G. Heywood and R. Rannacher. Finite element approximation of the nonstationary Navier-Stokes problem, part iii. smoothing property and higher order error estimates for spatial discretization. *SIAM J. Numer. Anal.*, 25(3):489–512, 1988.
32. C. W. Hirt and B. D. Nichols. Volume of Fluid (VOF) method for the dynamics of free boundaries. *J. Comp. Phys.*, 39:201–225, 1981.
33. J. M. Hyman. Numerical methods for tracking interfaces. *Physica*, 12D:396–407, 1984.
34. V. John. *Finite Element Methods for Incompressible Flow Problems*. Springer Series in Computational Mathematics. Springer International Publishing, 2016.
35. R. J. LeVeque. *Numerical Methods for Conservation Laws*. Second edition. Birkhauser Verlag , 1992.
36. R. J. LeVeque. *Numerical Methods for Conservation Laws*. Lectures in mathematics ETH Zürich. Birkhauser Basel, 2013.

37. R. J. LeVeque and Z. Li. Immersed interface methods for stokes flow with elastic boundaries or surface tension. *SIAM J. Sci. Comput.*, 18(3):709–735, 1997.
38. P. L. Lions. *Mathematical Topics in Fluid Mechanics: Incompressible Models*, volume 3 of *Oxford Lecture Series in Mathematics and Its Applications*. Oxford, 1997.
39. V. Maronnier, M. Picasso, and J. Rappaz. Numerical simulation of free surface flows. *J. Comp. Phys.*, 155(2):439–455, 1999.
40. V. Maronnier, M. Picasso, and J. Rappaz. Numerical simulation of three-dimensional free surface flows. *Int. J. Numer. Meth. Fluid*, 42(7):697–716, 2003.
41. J. Marsden and T. Hughes. *Mathematical Foundations of Elasticity*. Dover Civil and Mechanical Engineering Series. Dover, 1994.
42. A. K. Michler. Aircraft control surface deflection using rbf-based mesh deformation. *International Journal for Numerical Methods in Engineering*, 88(10):986–1007, 2011.
43. H. Miyata. Finite-difference simulation of breaking waves. *J. Comp. Phys.*, 65:179–214, 1986.
44. S. Muzaferija and M. Peric. *Nonlinear water wave interaction*, chapter Computation of free surface flows using interface-tracking and interface-capturing methods, pages 59–110. WIT Press, Southampton, 1999.
45. T. Nakayama and M. Mori. An Eulerian finite element method for time-dependent free surface problems in hydrodynamics. *Int. J. Num. Meth. Fluids*, 22:175–194, 1996.
46. B. D. Nichols and C. W. Hirt. Improved free surface boundary conditions for numerical incompressible flow calculation. *J. Comp. Phys.*, 8:434–448, 1973.
47. F. Nobile and L. Formaggia. A stability analysis for the arbitrary lagrangian eulerian formulation with finite elements. *East-West Journal of Numerical Mathematics*, 7(2):105–132, 1999.
48. W. F. Noh and P. Woodward. SLIC (Simple Line Interface Calculation). In A. van de Vooren and P. Zandbergen, editors, *Proc. of the 5th International Conference on Fluid Dynamics*, volume 59 of *Lecture Notes in Physics*, Berlin, 1976. Springer.
49. S. Osher and R. Fedkiw. *The Level Set Method and Dynamic Implicit Surfaces*. Springer-Verlag, New York, 2002.
50. S. Osher and J. A. Sethian. Fronts propagating with curvature-dependent speed: Algorithm based on Hamilton–Jacobi formulations. *J. Comp. Phys.*, 79:12–49, 1988.
51. N. Parolini. Computational fluid dynamics for naval engineering problems, 2004. PhD thesis, EPFL.
52. J. E. Pilliot and E. G. Puckett. Second-order accurate volume-of-fluid algorithms for tracking material interfaces. *J. Copm.Phys.*, 199(2):465–502, 2004.
53. L. Quartapelle. *Numerical solution of the incompressible Navier-Stokes equations*, volume 113 of *International Series of Numerical Mathematics*. Birkhauser Verlag, Basel, 1993.
54. A. Quarteroni. Numerical models for differential problems, 2014.
55. A. Quarteroni, F. Saleri, and A. Veneziani. Analysis of the yosida method for the incompressible navier–stokes equations. *Journal de Mathématiques Pures et Appliquées*, 78(5):473 – 503, 1999.
56. A. Quarteroni, F. Saleri, and A. Veneziani. Factorization methods for the numerical approximation of navier–stokes equations. *Computer Methods in Applied Mechanics and Engineering*, 188(1–3):505 – 526, 2000.
57. A. Quarteroni and A. Valli. *Numerical Approximation of Partial Differential Equations*. Springer Series in Computational Mathematics. Springer Berlin Heidelberg, 2009.
58. J. D. Ramshaw and J. A. Trapp. A numerical technique for low-speed homogeneous two-phase flow with sharp interface. *J. Comp. Phys.*, 21:438–453, 1976.
59. M. Rudman. Volume-tracking methods for interfacial flow calculations. *Int. J. Num. Meth. Fluids*, 24:671–691, 1997.
60. S. Salsa. *Equazioni a Derivate Parziali: Metodi, Modelli E Applicazioni*. Springer-Collana Unitext. Springer, 2004.

61. J. A. Sethian. *Level Set Methods and Fast Marching Methods*. Cambridge University Press, 1999.
62. A. Smolianski. *Numerical modeling of two-fluid interfacial flows*. PhD thesis, University of Jyväskylä, Finland, 2001.
63. K. Stein, T. Tezduyar, and R. Benney. Mesh Moving Techniques for Fluid-Structure Interactions With Large Displacements . *Journal of Applied Mechanics*, 70(1):58–63, 01 2003.
64. M. Sussman and E. Fatemi. An efficient, interface-preserving level set redistancing algorithm and its application to interfacial incompressible fluid flow. *SIAM J. Sci. Comput.*, 20(4):1165–1191, 1999.
65. M. Sussman and P. Smereka. Axisymmetric free boundary problems. *J. Fluid Mech.*, 341:269–294, 1997.
66. M. Sussman, P. Smereka, and S. Osher. A level set approach for computing solutions to incompressible two-phase flow. *J. Comp. Phys.*, 114:146–159, 1994.
67. M. Tabata and D. Tagami. A finite element analysis of a linearized problem of the Navier–Stokes equations with surface tension. *SIAM J. Numer. Anal.*, 38(1):40–57, 2001.
68. R. Temam. Une méthode d'approximation de la solution des équations de navier-stokes. *Bulletin de la Société Mathématique de France*, 96:115–152, 1968.
69. T. G. Thomas, D. C. Leslie, and J. J. R. Williams. Free surface simulations using a conservative 3D code. *J. Comp. Phys.*, 116:52–68, 1995.
70. L. J. P. Timmermans, P. D. Minev, and F. N. van de Vosse. An approximate projection scheme for incompressible flow using spectral elements. *International Journal for Numerical Methods in Fluids*, 22(7):673–688, 1996.
71. M. Tomé and S. McKee. GENSMAC: A computational Marker-And-Cell method for free surface flows in general domains. *J. Comp. Phys.*, 110:171–186, 1994.
72. A.-K. Tornberg. *Interface Tracking Methods with Applications to Multiphase Flows*. PhD thesis, Royal Institute of Technology, Stockholm, Sweden, 2000.
73. G. Tryggvason, B. Bunner, A. Esmaeeli, D. Juric, N. Al-Rawahi, W. Tauber, J. Han, S. Nas, and Y. J. Jan. A front-tracking method for the computations of multiphase flows. *J. Comp. Phys.*, 169:708–759, 2001.
74. H. Zhao, T. Chan, B. Merriman, and S. Osher. A variational level set approach to multiphase motion. *J. Comp. Phys.*, 127:179–195, 1996.