THESE – DISERTAČNÍ PRÁCE

présentée à

L'UNIVERSITÉ MÉDITERRANÉE
Institute de Recherche sur les Phénomènes Hors équilibre
MARSEILLE, FRANCE

et à

UNIVERZITA KARLOVA
Matematicko–fyzikální fakulta
PRAHA, ČESKÁ REPUBLIKA

pour obtenir le titre de

Docteur de l'Université de la Méditerranée, spécialité mécanique
Doktor Univerzity Karlovy, zaměření matematického a numerického modelování

par

Vít Dolejší

Sujet de la thèse:

Sur des méthodes combinant des volumes finis et des éléments finis pour le calcul d'écoulements compressibles sur des maillages non structurés

---

Téma práce:

Kombinovaná metoda konečných objemů a konečných prvků pro řešení stlačitelného proudění na nestrukturovaných sítích

---

Soutenance fixée au 24 November 1998 – Obhajoba práce 24. listopadu 1998

Directeurs de thèse – Školitelé:

Jiří Felcman	**Univerzita Karlova**
Miloslav Feistauer	**Univerzita Karlova**
Philippe Angot	**l'Université de la Mediterranée**

Table des matières

Abstract of the thesis

Combined Finite Volume – Finite Element Methods for Compressible Flow on Unstructured Meshes

We deal with the numerical simulation of compressible flow in two dimensions. Neglecting their microstructure we consider fluids as continuum and therefore the conception of the mechanic of continuum can be applied. Taking into account the mathematical formulation of conservation laws, we obtain a system of differential equations describing the behavior of fluids. To close the system, these equations must be accompanied by constitutive relations. We suppose the perfect gas of Newtonian type, the validity of the Fourier law for a heat transfer and also the validity of the relation between the viscosity coefficients λ and μ for one-atomic gas. External volume forces and heat sources are neglected. Using the mentioned restrictions, we determine the final system of the Navier–Stokes equations.

Numerical simulation of viscous compressible high-speed flow is one of the most difficult areas of Computational Fluid Dynamics because of the fact that convection effects dominate those of diffusion, that shock waves and boundary layers are present and that there is interaction between shocks and boundary layers. Since the viscosity and heat conduction coefficients of gases are small, the viscous dissipative terms are often considered as perturbations of the inviscid Euler system. Very often we are interested in steady flow. In this case, the operator splitting method can be applied. We divide the system of the Navier–Stokes equations into the inviscid system (the Euler equations) and the purely viscous system and discretize these separately. The inviscid system is a hyperbolic system. The nonlinear character of the hyperbolic system causes the possible rise of discontinuities in solutions even in case when the initial condition are smooth, e. g. the Burgers equation. Therefore the finite volume method (FVM) is applied for its discretization. The solution is approximated by a piecewise constant function, which can approximate discontinuities well. The purely viscous system is parabolic and its structure offers us the application of the finite element method (FEM). Furthermore, to solve the complete system of the Navier–Stokes equations, the combined finite volume – finite element method is applied. We have developed two numerical schemes: the former uses the triangular finite volumes and conforming finite elements and the latter the barycentric finite volumes and nonconforming finite elements.

To improve the order of accuracy of our schemes, a higher order reconstruction of piecewise constant functions for FVM is implemented. We have investigated several methods proposed by ourselves and also published in the literature. In order to determine the order of accuracy of our schemes, the computations for Sod's channel were performed. Here the analytical solution is known. The first part of this work is devoted to the numerical schemes developed for the solution of the compressible Navier–Stokes equations. We present a number of numerical examples of inviscid as well as viscous flows. Both the basic test problems and industrial applications are presented. The comparison with experimental measurement is performed.

In the second part we develop the strategies of a mesh adaption. To obtain satisfactory results without high cost of CPU–time and a memory of computers, a suitable mesh should be used for the calculation. We have developed two strategies: multilevel mesh refinement (MMR) and anisotropic mesh adaption (AMA). The former is based on the shock indicator, which seems to be very suitable namely for a precise resolution of shock waves. This indicator searches for regions with physically admissible jumps of state quantities and the corresponding mesh is refined by red-green algorithm. This process can be repeated several times until the

obtained discontinuities are sufficiently precise and sharp. The disadvantage of MMR is that it does not respect the direction of shock waves. To avoid this, the concept of AMA was developed. It can not only refine the triangulation but also remove some elements, swap the diagonal of quadrilateral formed by any pair of adjacent triangles and move the nodes. AMA tries to construct an optimal triangulation of the computational domain in the sense that an "interpolation error" of the solution of the considered problem is uniformly distributed over the whole triangulation. Therefore, we obtain locally refined meshes with no drastically enhanced number of elements. Both of these adaptive strategies are compared.

The last part is devoted to the investigation of the convergence of our numerical method for viscous flow simulations. The attempts to prove the convergence of numerical methods for the solution of complete system describing viscous flow have not yet been successful. Therefore, in order to support our computational results obtained by the combined finite volume–finite element method, we confine our convergence analysis to a model scalar convection–diffusion conservation law equation. Under some restriction of the mesh, with the aid of the discrete maximum principle, a priori estimates and some compactness arguments based on the use of the Fourier transform with respect to time, the convergence of the approximate solution to the exact one is proved, provided the mesh size tends to zero.

Résumé de la Thèse

Sur des méthodes combinant des volumes finis et des éléments finis pour le calcul d'écoulement des fluides compressibles sur des maillages non structurés

On ici travaille de la simulation numérique des fluides compressibles en deux dimensions d'espace. En négligeant leur microstructure on peut considérer des fluides comme continus et la théorie de milieux continus peut être appliquée. En utilisant la formulation mathématique des lois de conservation on déduit un système d'équations décrivant le comportement des fluides. Pour avoir le même nombre d'équations que d'inconnues il faut ajouter les lois de comportement. On suppose un fluide du type newtonien; on suppose que la loi de Fourier pour la transmission de la chaleur s'applique et on suppose que le gaz est polytropique et parfait. On suppose encore que la relation parmi les coefficients visqueux λ et μ pour un gaz uni–atomique est valable. Les forces extérieures et les sources de chaleurs sont négligés. En utilisant ces restrictions on obtient le système des équations de Navier–Stokes.

On s'intéresse à l'écoulement à petite viscosité, où l'influence des termes diffusifs est petite en comparaison avec les termes convectifs. On considère donc les termes visqueux comme une perturbation du système non visqueux. Si nous nous intéressons aux écoulements stationnaires, la méthode de la division de l'opérateur peut être appliquée, c'est-à-dire on divise les équations de Navier – Stokes en une partie convective (les équations d'Euler) et une partie diffusive et chaque partie est discrétisée séparément. La première partie est un système hyperbolique. La propriété caractéristique de ces systèmes est l'existence de discontinuités malgré des conditions initiales régulières. C'est la raison pour laquelle on utilise la méthode des volumes finis (MVF), où la solution cherchée est approchée par une fonction constante par parties, qui saisit mieux les discontinuités. Le deuxième système est du type parabolique et il est discrétisé par la méthode des éléments finis (MEF). Pour ces raisons nous avons appliqué la méthode combinant des volumes finis et des éléments finis pour la résolution du système des équations de Navier – Stokes. Nous avons développé deux méthodes. La première utilise des volumes finis triangulaires et des éléments finis conformes et la deuxième utilise des volumes finis barycentriques et des éléments finis non conformes. Pour augmenter la précision d'une solution, l'utilisation de la reconstruction d'ordre plus élevé pour MVF est appliquée. On a examiné plusieurs méthodes proposées dans la littérature et toutes ont été testées sur le tube à chocs de Sod, où la solution analytique est connue et où l'ordre de la précision peut être déterminé. La première partie de cette thèse est consacrée aux méthodes numériques pour la résolution des équations de Navier–Stokes. Ici on présente plusieurs résultats numériques de l'écoulement visqueux et aussi non visqueux. Quelques résultats concernant des problèmes académiques et d'autres des problèmes appliqués. Pour ces problèmes une comparaison avec les résultats expérimentaux est présentée.

Dans la partie deuxième on développe une stratégie de l'adaptation du maillage. Pour obtenir des résultats satisfaisants sans l'utilisation de mémoire et temps de CPU trop grand, un maillage convenable doit être utilisé. Nous avons développé deux stratégies, la technique multiniveau de raffinement (TMR) et l'adaptation des maillages anisotropiques (AMA). La première est basée sur l'indicateur de choc et elle est très convenable pour la résolution des problèmes avec plusieurs onde de choc. Cet indicateur indique les régions où l'on observe de fortes variations des quantités physiques (ondes de choc) et ces régions sont raffinées par la technique "red–green". Après plusieurs raffinements la netteté des ondes du chocs est suffisante. Le désavantage de TMR est l'impossibilité de grossière du maillage. C'est la raison

pour laquelle la méthode d'AMA a été développée. Cet algorithme permet non seulement de raffiner le maillage, mais aussi d'éliminer un côté du maillage, basculer d'une diagonale d'un quadrangle créé par des triangles voisins et bouger des noeuds. AMA construit un maillage optimal dans le sens que "l'erreur" de la solution est uniformément distribuée sur tout le maillage. De cette façon on obtient un maillage avec un numéro des éléments pas trop grands, mais sur lequel la bonne approximation numérique peut être obtenue. Les deux méthodes d'adaptation sont comparées.

La dernière partie est consacrée à la démonstration de convergence de notre méthode pour la simulation des écoulements visqueux. Comme l'existence de la solution des équations de Navier – Stokes est encore une question ouverte, l'analyse numérique des méthodes est effectuée pour le problème convectif – diffusif scalaire. En supposant quelques restrictions sur un maillage, utilisant le principe de maximum discrète, des estimations a priori et des proprétés de compacité basées sur l'utilisation de la transformation de Fourier, la convergence de la solution approximative vers la solution précise est démontrée en supposant que le pas du maillage tend vers zéro.

Abstract disertační práce

Kombinovaná metoda konečných prvků a konečných objemů pro řešení stlačitelného proudění na nestrukturovaných sítích.

V této práci se věnujeme numerické simulaci stlačitelného rovinného proudění. Zanedbáme-li mikrostructuru tekutin, tak je lze považovat za kontinuum a tedy můžeme použít přístup mechaniky kontinua. Pomocí matematické formulace základních zákonů zachování dostaneme systém diferenciálních rovnic popisujících proudění tekutin. Abychom měli stejný počet rovnic a neznámých, je potřeba k této soustavě rovnic přidat tzv. konstituční vztahy. V této práci se omezíme na Newtonovské tekutiny, pro které platí navíc Fourierův zákon vedení tepla. Dále předpokládáme, že náš plyn splňuje stavovou rovnici pro ideální plyn a uvažujeme platnost vztahu mezi koeficienty vazkosti λ a μ, který platí přesně pouze pro jednoatomové plyny. Rovněž neuvažujeme působení vnějších sil a zdrojů tepla. Za uvedených předpokladů získáme konečný systém Navier–Stokesových rovnic.

Zabýváme se převážně prouděním tekutin s malou vazkostí v transonickém režimu, kde vliv difuzních členů je malý v porovnání s konvektivními. Proto považujeme vazké členy za jakousi perturbaci nevazkého systému rovnic. Pokud uvažujeme stacionární proudění, tak můžeme použít metodu operátorového rozštěpení. Systém Navier–Stokesových rovnic rozdělíme na dva podsystémy: nevazký a čistě vazký a každý z těchto systémů diskretizujeme samostatně. První z nich je hyperbolický a charakteristickou vlastností takovýchto systémů je existence nespojitých řešení i pro hladká data. Proto jsme k diskretizaci použili metodu konečných objemů (MKO), kde hledané řešení je aproximováno funkcí po částech konstantní, což umožnuje dobře zachytit nespojitosti. Druhý podsystém je parabolický, a proto jsme ho diskretizovali pomocí metody konečných prvků (MKP). Pak úplný systém Navier–Stokesových rovnic je řešen pomocí kombinované metody konečných prvků a konečných objemů. Vypracovali jsme dvě numerická schémata: první užívá trojúhelníkové konečné objemy a konformní konečné prvky a druhé z nich barycentrické konečné objemy a nekonformní konečné prvky.

Pro zvýšení přesnosti našich schémat jsme použili pro MKO rekonstrukci vyššího řádu. Vyšetřovali jsme několik metod jak již publikovaných v literatuře, tak také námi odvozených. Pro určení řádu přesnosti jsme tyto schémata testovali na Sodově trubici, kde je známo analytické řešení nevazkého jednorozměrného proudění.

První část této práce je věnována numerickým schématům odvozených pro řešení Navier–Stokesovych rovnic popisujících stlačitelné proudění. Zde publikujeme několik příkladů nevazkého i vazkého prudění. Některé z těchto příkladů jsou základní testovací příklady a některé jsou dány přímo průmyslovou praxí. Pro tyto příklady máme i porovnání s experimentálním měřením.

Ve druhé části se věnujeme strategiím adaptace sítí. Pro získání uspokojivých výsledků bez enormních požadavků na CPU čas a pamět počítače je potřeba k výpočtům použít vhodnou sít. Zabýváme se zde dvěma přístupy: násobné zjemňování sítí (NZS) a anisotropní adaptace sítí (AAS). První z těchto metod je založena na indikátoru rázové vlny. Tato metoda je vhodná zvláště pro případy s několika rázovými vlnami. Tento inikátor vyhledává oblasti, kde dochází k fyzikálně přípustným skokům stavových veličin a odpovídající část sítě je pak zjemněna pomocí "red–green" algoritmu. Po několika aplikacích zjemnění získají rázové vlny požadovanou přesnost a ostrost. Nevýhoda NZS je, že nerespektuje směr rázových vln, což vede k prudkému nárůstu počtu elementů po několikanásobném zjemnění. Toto ovšem umožňuje AAS, která kromě klasického zjemňování je schopna sít také zhrubit, přehodit

úhlopříčku čtyřúhelníka tvořeného dvojicí sousedních trojúhelníků a taktéž posouvat uzly sítě.

AAS generuje optimální triangulaci dané oblasti tím způsobem, že tzv. interpolační chyba řešení uvažovaného problému je rovnoměrně rozprostřena přes celou oblast. Tímto způsobem získáme triangulaci s nepříliš velkým počtem elementů, ale takovou, že na ní lze získat dobrou numerickou aproximaci. Oba dva adaptivní přístupy jsou zde porovnány.

Poslední část této práce je věnována vyšetřování konvergence jedné z našich numerických metod. Vzhledem k tomu, že doposud nebyl vyřešen problém ohledně existence a jednoznačnosti řešení Navier–Stokesových rovnic, tak numerická analýza je prováděna pouze pro skalární konvektivně–difusní problém. Za jistých omezeních na triangulace výpočetní oblasti, za užití diskrétního principu maxima, některých apriorních odhadů a jistých tvrzeních o kompaktnosti založených na Fourierově transformaci vzhledem k času, jsme dokázali pro druhou z našich metod, že přibližné řešení konverguje k přesnému za předpokladu, že krok sítě jde k nule.

Première partie

Les méthodes numériques

Dans cette partie nous nous intéressons à une simulation numérique des fluides compressibles. Les fluides sont composés de particules discrètes en très grand nombre, mais dans les conditions habituelles ils apparaissent comme continus. Dans la plupart des problèmes pratiques, on s'intéresse au mouvement d'ensembles composés d'un très grand nombre de molécules et la matière peut alors être considérée comme un **milieu continu**. Ce concept n'est utile que si les dimensions caractéristiques du problème considéré sont grandes par rapport à la distance moyenne qui sépare les molécules composant le milieu.

En utilisant le concept de la mécanique d'un milieu continu on déduit la formulation mathématique des lois physiques fondamentales décrivant un comportement des fluides. La formulation mathématique est présentée par des équations différentielles. Ces équations sont du type hyperbolique et il n'y a pas beaucoup de résultats théoriques, mais néanmoins nous essayons de les résoudre par une méthode numérique. Les résultats obtenus peuvent être comparés avec des mesures expérimentales.

Le premier chapitre est consacré à une détermination brève des équations décrivant des fluides et à la formulation d'un problème d'écoulement des fluides. Le Chap. 2 décrit les schémas numériques pour la résolution des équations décrivant l'écoulement. Le Chap. 3 est consacré aux méthodes d'ordre plus élevé qui permettent l'augmentation d'une précision de la solution. Les exemples numériques et leurs comparaisons des mesurages expérimentales sont présentés dans le Chap. 4.

Chapitre 1

Les équations fondamentales décrivant d'écoulement

Nous nous occupons de la description du comportement des fluides. Dans ce chapitre on présente la formulation mathématique des équations de bilan – les lois fondamentales de la physique. En y ajoutant les lois de comportement on obtient le système des équations de Navier–Stokes et d'Euler, qui décrivent l'écoulement compressible. Après la présentation des lois physiques, la formulation mathématique du problème va être donnée.

Dans cette partie les notations suivantes sont utilisées:

$\Omega \subset I\!R^2$	domaine occupé par un fluide
$(0, T)$	intervalle du temps
$Q_T = \Omega \times (0, T)$	cylindre d'espace et du temps
$\boldsymbol{x} = (x_1, x_2) \in \Omega$	coordonnés spatiales à Ω
$t \in (0, T)$	instant du temps
$\rho = \rho(\boldsymbol{x}, t)$	densité
$\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x}, t) = (u_1, u_2)$	vecteur vitesse
$p = p(\boldsymbol{x}, t)$	pression
$\theta = \theta(\boldsymbol{x}, t)$	température absolue
$e = e(\boldsymbol{x}, t)$	énergie totale
u	énergie interne
c	vitesse du son
$\boldsymbol{g} = \boldsymbol{g}(\boldsymbol{x}, t) = (g_1, g_2)$	force extérieure
$q = q(\boldsymbol{x}, t)$	source de chaleur
$\boldsymbol{q} = \boldsymbol{q}(\boldsymbol{x}, t) = (q_1, q_2)$	flux de chaleur
γ	constante de Poisson adiabatique
$c_p,\ c_V$	chaleurs spécifiques à pression et volumes constantes
$\mu,\ \lambda$	coefficients visqueux
k	conductivité thermique
$\boldsymbol{e} = \{e_{ij}\}_{i,j=1}^2$	tenseur de déformation de vitesse
$\tau = \{\tau_{ij}\}_{i,j=1}^2$	tenseur des contraintes
δ_{ij}	symbole de Kronecker, $\delta_{ij} = 1$ si $i = j$, $\delta_{ij} = 0$ sinon
R	constante molaire des gaz parfaits

Re	nombre de Reynolds
Pr	nombre de Prandtl
M	nombre de Mach

1.1 Modèle physique

Soit $\Omega \subset I\!R^2$ une domaine borné occupé par un fluide et $T > 0$. On pose $Q_T = \Omega \times (0, T)$. L'écoulement d'un fluide compressible en 2 dimensions est décrit par les équations de bilan (pour exemple voir [17]): conservation de la masse (1.1), conservation de la quantité de mouvement (1.2) et conservation de l'énergie (1.3), que l'on peut présenter sous la forme suivante:

$$\frac{\partial \rho}{\partial t} + \sum_{i=1}^{2} \frac{\partial (\rho u_i)}{\partial x_i} = 0 \qquad \text{dans } \Omega \times (0,T)\,, \tag{1.1}$$

$$\frac{\partial \rho u_i}{\partial t} + \sum_{j=1}^{2} \frac{\partial (\rho u_i u_j)}{\partial x_j} = \rho g_i + \sum_{j=1}^{2} \frac{\partial \tau_{ij}}{\partial x_j}, \qquad \text{dans } \Omega \times (0,T)\,, \qquad i = 1,2 \tag{1.2}$$

$$\frac{\partial e}{\partial t} + \sum_{i=1}^{2} \frac{\partial (e u_i)}{\partial x_i} = \sum_{i=1}^{2} \rho g_i u_i + \sum_{i,j=1}^{2} \frac{\partial (\tau_{ij} u_i)}{\partial x_j} + \rho q - \sum_{i=1}^{2} \frac{\partial q_i}{\partial x_i} \qquad \text{dans } \Omega \times (0,T)\,. \tag{1.3}$$

Dans nos problèmes, nous négligerons les forces extérieures et des sources de chaleur, c'est-à-dire $\boldsymbol{g} = 0$, $q = 0$.

Nous avons quatre équations (le relations (1.1) – (1.3)) pour plusieurs quantités inconnues. Il est nécessaire de les accompagner par d'autres relations qui complètent le système. Nous appelons ces relations *les lois de comportement.*

Nous supposerons un fluide de type newtonien, ainsi le tenseur des contraintes a la forme

$$\tau_{ij} = -p\delta_{ij} + \tau_{ij}^V, \quad \tau_{ij}^V = \lambda \mathrm{div}\boldsymbol{u}\ \delta_{ij} + 2\mu e_{ij}, \quad i,j = 1,2, \tag{1.4}$$

où $\boldsymbol{e} = \{e_{ij}\}_{i,j=1}^2$ est le tenseur de déformation de vitesse

$$e_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right). \tag{1.5}$$

Ici, μ, λ sont les coefficients de viscosité. Nous supposerons les relations

$$3\lambda + 2\mu = 0, \quad \mu \geq 0, \tag{1.6}$$

qui sont valables pour un gaz uni-atomique. La relation suivante exprime la loi de Fourier

$$\boldsymbol{q} = -k\nabla\theta, \tag{1.7}$$

où $k \geq 0$ est *la conductivité thermique.* Nous considérerons toujours un fluide parfait, c'est-à-dire

$$p = R\rho\theta. \tag{1.8}$$

Cette équation est appelée *l'équation d'état pour le gaz parfait.* $R > 0$ est *la constante molaire des gaz parfaits*, que on peut exprimer sous la forme

$$R = c_p - c_V, \tag{1.9}$$

où c_p et c_V sont les chaleurs spécifiques à pression et volumes constants. Nous supposerons que c_p et c_V sont constantes. Il est facile de montrer que $c_p > c_V$. La quantité

$$\gamma = \frac{c_p}{c_V} > 1 \tag{1.10}$$

est appelée *la constante de Poisson.*

L'énergie totale consiste en l'énergie interne u et l'énergie cinétique. Nous supposons le gaz polytropique, c'est-à-dire $u = \rho c_V \theta$. Nous obtenons

$$e = \rho c_V \theta + \frac{1}{2}\rho|\boldsymbol{u}|^2. \tag{1.11}$$

Les équations (1.1) – (1.3) et les relations (1.4), (1.5), (1.7), (1.8) et (1.11) donnent

$$\frac{\partial \rho}{\partial t} + \sum_{i=1}^{2} \frac{\partial (\rho u_i)}{\partial x_i} = 0, \tag{1.12}$$

$$\frac{\partial \rho u_i}{\partial t} + \sum_{j=1}^{2} \frac{\partial (\rho u_i u_j)}{\partial x_j} + \frac{\partial p}{\partial x_i} = \frac{\partial}{\partial x_i}(\lambda \operatorname{div} \boldsymbol{u}) + 2\sum_{j=1}^{2} \frac{\partial}{\partial x_j}(\mu e_{ij}), \tag{1.13}$$

$$\frac{\partial e}{\partial t} + \sum_{j=1}^{2} \frac{\partial (e+p) u_j}{\partial x_j} = \sum_{i,j=1}^{2} \frac{\partial (\tau_{ij}^V \ u_i)}{\partial x_j} + k\Delta\theta. \tag{1.14}$$

Le système (1.12) – (1.14) peut être transcrit sous la forme conservative

$$\frac{\partial}{\partial t}\begin{pmatrix} \rho \\ \rho u_1 \\ \rho u_2 \\ e \end{pmatrix} + \sum_{i=1}^{2} \frac{\partial}{\partial x_i}\begin{pmatrix} \rho u_i \\ \rho u_i u_1 + p\delta_{i1} \\ \rho u_i u_2 + p\delta_{i2} \\ (e+p)\, u_i \end{pmatrix} = \sum_{i=1}^{2} \frac{\partial}{\partial x_i}\begin{pmatrix} 0 \\ \tau_{1i}^V \\ \tau_{2i}^V \\ \sum_{j=1}^{2} \tau_{ij}^V u_j + k\frac{\partial \theta}{\partial x_i} \end{pmatrix} \quad \text{dans } Q_T. \tag{1.15}$$

Pour les calculs numériques, il est convenable d'utiliser les quantités adimensionalles. Soit $\bar{\rho} > 0$ densité caractéristique, $\bar{l} > 0$ longueur caractéristique et $\bar{u} > 0$ vitesse caractéristique (on rappelle que $\bar{u}$ est une grandeur scalaire). On multiplie la première l'équation du système (1.15) par $\bar{l}/(\bar{\rho}\bar{u})$, la deuxième et la troisième par $\bar{l}/(\bar{\rho}\bar{u}^2)$ et la quatrième par $\bar{l}/(\bar{\rho}\bar{u}^3)$. On obtient

$$\frac{\partial \boldsymbol{w}}{\partial t^*} + \sum_{i=1}^{2} \frac{\partial}{\partial x_i^*}\boldsymbol{f}_i(\boldsymbol{w}) = \sum_{i=1}^{2} \frac{\partial}{\partial x_i^*}\mathbf{R}_i(\boldsymbol{w}, \nabla \boldsymbol{w}) \quad \text{dans } Q_T, \tag{1.16}$$

où

$$\boldsymbol{w} \;:=\; \begin{pmatrix} \rho^* \\ \rho^* u_1^* \\ \rho^* u_2^* \\ e^* \end{pmatrix}, \qquad \boldsymbol{f}_i(\boldsymbol{w}) := \begin{pmatrix} \rho^* u_i^* \\ \rho^* u_i^* u_1^* + p^*\delta_{i1} \\ \rho^* u_i^* u_2^* + p^*\delta_{i2} \\ (e^* + p^*)\, u_i^* \end{pmatrix},$$

$$\mathbf{R}_i(\boldsymbol{w}, \nabla \boldsymbol{w}) := \begin{pmatrix} 0 \\ \tau_{1i}^* \\ \tau_{2i}^* \\ \sum_{j=1}^{2} \tau_{ij}^* u_j^* + \frac{\gamma}{Re\ Pr}\frac{\partial \theta^*}{\partial x_i^*} \end{pmatrix}, \quad i = 1, 2, \tag{1.17}$$

$$\tau_{ij}^* = \frac{1}{Re}\left[\left(\frac{\partial u_i^*}{\partial x_j^*} + \frac{\partial u_j^*}{\partial x_i^*}\right) - \frac{2}{3}\operatorname{div} \boldsymbol{u}^* \delta_{ij}\right].$$

et où

$$\begin{aligned} \rho^* &= \frac{\rho}{\bar{\rho}}, \qquad p^* = \frac{p}{\bar{\rho}\bar{u}^2}, \qquad \boldsymbol{u}^* = \frac{\boldsymbol{u}}{\bar{u}}, \qquad e^* = \frac{e}{\bar{\rho}\bar{u}^2}, \\ \boldsymbol{x}^* &= \frac{\boldsymbol{x}}{\bar{l}}, \qquad \theta^* = \frac{\theta}{\bar{u}^2}, \qquad t^* = t\frac{\bar{u}}{\bar{l}}, \qquad Re = \frac{\bar{\rho}\bar{u}\bar{l}}{\mu}, \qquad Pr = \frac{\gamma\mu}{k}, \end{aligned} \tag{1.18}$$

On appelle Re le nombre de Reynolds et Pr le nombre de Prandtl. Il'est facile de voir que les équations pour les quantités non-dimensionales (1.16) – (1.18) ont le même caractère que les équations pour les quantités dimensionales (1.15). Dans les chapitres suivants on n'utilise pas d'indice * et $\rho, \boldsymbol{u}, p, \ldots$ vont signifier toujours les quantités adimensionalles. Les fonctions $\boldsymbol{f}_i,\ i = 1, 2$ sont appelées *les termes convectifs* et $\mathbf{R}_i,\ i = 1, 2$ sont appelées *les termes diffusifs*.

Les équations (1.8), (1.11) et l'utilisation des quantités adimensionalles donnent les relations entre e, p et θ par

$$p = (\gamma - 1)\left(e - \frac{1}{2}\rho|\boldsymbol{u}|^2\right) \qquad \text{et} \qquad \theta = \frac{p}{\rho(\gamma - 1)}. \tag{1.19}$$

Les relations (1.16) et (1.17) sont appelées le système des *équations de Navier–Stokes*.

1.2 Formulation mathématique du problème

Ici on présente la formulation du problème des l'écoulements visqueux et non visqueux. Le système des équations (1.16) – (1.17) doit être accompagné par des conditions initiales et aux limites. Le caractère des équations décrivant l'écoulement visqueux et non visqueux est très différent et il est nécessaire de les discuter séparément.

1.2.1 Ecoulement non visqueux

Si nous nous intéressons à l'écoulement non visqueux, c'est-à-dire l'influence des termes visqueux peut être négligée, on pose $\mu = 0 \Rightarrow Re = \infty$, qui donne $\mathbf{R}_i = 0,\ i = 1, 2$ et le système (1.16) a la forme

$$\frac{\partial \boldsymbol{w}}{\partial t} + \sum_{i=1}^{2} \frac{\partial}{\partial x_i}\boldsymbol{f}_i(\boldsymbol{w}) = 0 \quad \text{dans } Q_T, \tag{1.20}$$

où $\boldsymbol{f}_i$ sont données par (1.17). Le système (1.20), qui est appelé *les équations d'Euler*, est hyperbolique non linéaire. Il est connu de la théorie des équations hyperboliques linéaires, combien de conditions frontières doivent être prescrites pour que le problème soit bien posé. Ces résultats sont utilisés pour la déduction heuristique des conditions frontières de notre problème. Le nombre des conditions nécessaires est différent pour l'entrée et la sortie et pour l'écoulement subsonique ($|\boldsymbol{u}| < c$) et supersonique ($|\boldsymbol{u}| > c$), où $\boldsymbol{u}$ est le vecteur vitesse d'écoulement et c la *vitesse du son* donnée par $c = (\gamma p/\rho)^{1/2}$. L'analyse des équations d'Euler sera présentée dans le Chap. 2.

Au contraire, pour les parois imperméables la situation est plus facile. Grâce à l'imperméabilité des parois on pose la composante normale de la vitesse égale à zéro, c'est-à-dire

$$u_n = \boldsymbol{u} \cdot \boldsymbol{n} = 0 \qquad \text{sur parois imperméables}, \tag{1.21}$$

où $\boldsymbol{n}$ est le vecteur normal à $\partial\Omega$.

Les conditions frontières, que l'on va prescrire sont les suivantes:

1. Ecoulement entrant ($\boldsymbol{u} \cdot \boldsymbol{n} < 0$) supersonique ($|\boldsymbol{u} \cdot \boldsymbol{n}| > c$): 4 conditions par exemple ρ, u_1, u_2, p.
2. Ecoulement entrant ($\boldsymbol{u} \cdot \boldsymbol{n} < 0$) subsonique ($|\boldsymbol{u} \cdot \boldsymbol{n}| < c$): 3 conditions par exemple ρ, u_1, u_2.
3. Ecoulement sortant ($\boldsymbol{u} \cdot \boldsymbol{n} > 0$) supersonique ($|\boldsymbol{u} \cdot \boldsymbol{n}| > c$): 0 condition.
4. Ecoulement sortant ($\boldsymbol{u} \cdot \boldsymbol{n} > 0$) subsonique ($|\boldsymbol{u} \cdot \boldsymbol{n}| < c$): 1 condition par exemple p.
5. Ecoulement glissant sur une paroi ($\boldsymbol{u} \cdot \boldsymbol{n} = 0$): 1 condition, $\boldsymbol{u} \cdot \boldsymbol{n} = 0$.

On prend des conditions initiales sous la forme

$$\boldsymbol{w}(\boldsymbol{x}, 0) = \boldsymbol{w}_0(\boldsymbol{x}) \qquad \forall \boldsymbol{x} \in \Omega. \tag{1.22}$$

1.2.2 Ecoulement visqueux

L'écoulement visqueux est décrit par

$$\frac{\partial \boldsymbol{w}}{\partial t} + \sum_{i=1}^{2} \frac{\partial}{\partial x_i} \boldsymbol{f}_i(\boldsymbol{w}) = \sum_{i=1}^{2} \frac{\partial}{\partial x_i} \mathbf{R}_i(\boldsymbol{w}, \nabla \boldsymbol{w}) \quad \text{dans } Q_T, \tag{1.23}$$

où $\boldsymbol{f}_i$ et $\mathbf{R}_i$ sont donnés par (1.17). Le problème des conditions aux limites est plus compliqué que pour le cas non visqueux, parce que (1.23) est du type parabolique–hyperbolique. Nous divisons la frontière $\partial\Omega$ en plusieurs parties, qui sont disjointes. On désigne par Γ_I, Γ_O, Γ_W, Γ^+,Γ^- entrée, sortie, parois imperméables et les frontières périodiques, respectivement (voir Fig. 1.1). Si $\Gamma^+ \neq \emptyset$ on a

$$\Gamma^+ = \{(x_1 + d_1, x_2 + d_2);\ (x_1, x_2) \in \Gamma^-\}, \tag{1.24}$$

où $\boldsymbol{d} = (d_1, d_2)$, $|\boldsymbol{d}| > 0$ est le vecteur de la période.

Grâce à la viscosité un liquide adhère aux parois et il faut prescrire la condition "no-slip", c'est-à-dire

$$\boldsymbol{u} = 0 \qquad \text{sur } \Gamma_W \times (0, T). \tag{1.25}$$

Comme la première équation de (1.23) est hyperbolique, la densité ρ est prescrite seulement sur Γ_I tandis que $\boldsymbol{u}$, θ sont prescrites sur tout $\partial\Omega$ grâce à la parabolicité de la deuxième, troisième et quatrième équation de (1.23). Utilisant cette déduction heuristique et aussi les expériences numériques, les conditions frontières suivantes sont considérées:

$$\begin{aligned} \rho = \tilde{\rho}, \quad \boldsymbol{u} = \tilde{\boldsymbol{u}}, \quad & \frac{\partial \theta}{\partial n} = 0 && \text{sur } \Gamma_I, \\ \boldsymbol{u} = 0, \quad & \frac{\partial \theta}{\partial n} = 0 && \text{sur } \Gamma_W, \\ \sum_{j=1}^{2} \tau_{ij} u_j = 0, \quad i = 1, 2, \quad & \frac{\partial \theta}{\partial n} = 0 && \text{sur } \Gamma_O, \\ \boldsymbol{w}|_{\Gamma^+} = \boldsymbol{w}|_{\Gamma^-} & && \text{sur les frontière périodiques} \end{aligned} \tag{1.26}$$

Ici, $\tilde{\rho}$, $\tilde{\boldsymbol{u}}$ sont les fonctions données et $\partial/\partial n$ signifie une dérivée à une direction normale.

On prend bien sûr des conditions initiales comme pour le cas non visqueux (1.22).

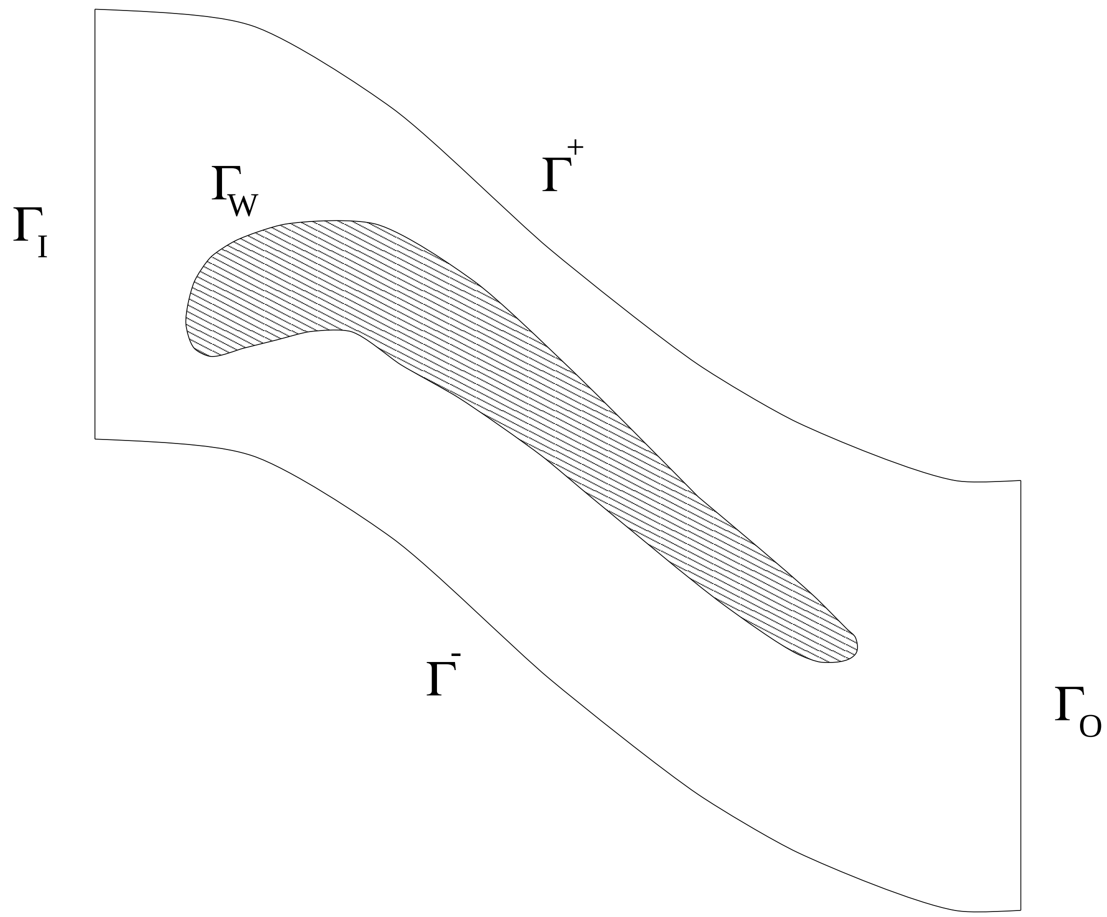


FIG. 1.1: La division de la frontière sur $\Gamma_I,\ \Gamma_O,\ \Gamma_W, \Gamma^+$ et Γ^-

Chapitre 2

Schémas numériques fondamentaux

Dans ce chapitre on présente les méthodes numériques pour la résolution des problèmes donnés dans le Chap. 1. En régime transonique et supersonique, l'influence des termes visqueux reste très petite en comparaison à celle des termes convectifs. Si nous nous intéressons à l'écoulement stationnaire, la méthode de la *division de l'opérateur* peut être appliquée. Par cet effet, on divise les équations de Navier–Stokes (1.16) en une partie convective (les équations d'Euler)

$$\frac{\partial \boldsymbol{w}}{\partial t} + \sum_{i=1}^{2} \frac{\partial}{\partial x_i} \boldsymbol{f}_i\left(\boldsymbol{w}\right) = 0 \quad \text{dans } Q_T. \tag{2.1}$$

et une partie diffusive

$$\frac{\partial \boldsymbol{w}}{\partial t} = \sum_{i=1}^{2} \frac{\partial}{\partial x_i} \mathbf{R}_i\left(\boldsymbol{w}, \nabla \boldsymbol{w}\right) \quad \text{dans } Q_T \tag{2.2}$$

et les systèmes (2.1) et (2.2) sont discrétisés séparément. Le système des équations d'Euler (2.1) est du type hyperbolique. La propriété caractéristique de ces systèmes est l'existence de discontinuités malgré des conditions initiales régulières. C'est la raison pour laquelle on utilise *la méthode des volumes finis* (MFV), où la solution cherchée est approximée par une fonction constante par parties, qui saisit mieux les discontinuités. Le système diffusif (2.2) est du type parabolique. Pour cette classe des équations les résultats théoriques aussi numériques avaient été développés en utilisant la méthode des éléments finis (MEF).

Pour ces raisons nous avons appliqué la méthode combinant des volumes finis et des éléments finis pour la résolution du système des équations de Navier–Stokes. Dans la Sec. 2.1 on décrit la simulation numérique d'écoulements non visqueux par la méthode des volumes finis. Dans la Sec. 2.2 l'utilisation de MEF et deux méthodes numériques pour le calcul d'écoulement visqueux sont présentées.

2.1 Discrétisation du système non visqueux

2.1.1 Résultats théoriques

Ici on répète brièvement les propriétés fondamentaux du système (2.1). Soit

$$\frac{\partial \boldsymbol{w}}{\partial t} + \sum_{i=1}^{2} \frac{\partial \boldsymbol{f}_i\left(\boldsymbol{w}\right)}{\partial x_i} = 0 \qquad \text{dans } Q_T, \tag{2.3}$$

$$\boldsymbol{w} \;=\; (\rho, \rho u_1, \rho u_2, e)^T, \quad \boldsymbol{f}_i(\boldsymbol{w}) = (\rho u_i, \rho u_1 u_i + \delta_{i1} p, \rho u_2 u_i + \delta_{i2} p, (e+p)u_i)^T \;, \; i = 1, 2$$

le système des équations d'Euler qui est accompagné par la condition initiale

$$\boldsymbol{w}(\cdot, 0) = \boldsymbol{w}_0 \qquad \text{dans } \Omega \tag{2.4}$$

et par la condition frontière qui peut être représentée schématiquement par l'opérateur de la frontière B

$$B(\boldsymbol{w}) = \mathbf{b} \qquad \text{sur } \partial\Omega \times (0, T). \tag{2.5}$$

Le système (2.3) sera fermé en additionnant la relation

$$p = (\gamma - 1)\left(e - \frac{1}{2}\rho|\boldsymbol{u}|^2\right), \tag{2.6}$$

qui représente l'équation d'état.

La fonction vectorielle $\boldsymbol{w}$ est l'opérateur $\boldsymbol{w} : Q_T \to F$, où

$$F = \left\{ \boldsymbol{w} \in I\!R^4;\; w_1 > 0,\; w_4 > \frac{w_2^2 + w_3^2}{2w_1} \right\}, \tag{2.7}$$

parce que seulement la situation physique est considérée, c'est-à-dire $\rho > 0$ et $p > 0$. On peut facilement voir que $\boldsymbol{f}_j \in C^1(F, I\!R^4)$ pour $j = 1, 2$. Utilisant cette propriété on obtient

$$\frac{\partial \boldsymbol{w}}{\partial t} + \sum_{i=1}^{2} \mathbf{A}_i(\boldsymbol{w}) \frac{\partial \boldsymbol{w}}{\partial x_i} = 0, \tag{2.8}$$

où

$$\mathbf{A}_i(\boldsymbol{w}) = \frac{D\boldsymbol{f}_i(\boldsymbol{w})}{D\boldsymbol{w}}, \qquad i = 1, 2 \tag{2.9}$$

sont les 4×4 matrices jacobiennes de $\boldsymbol{f}_i(\boldsymbol{w}),\ i = 1, 2$.

Définition 1 *On dit, que le système* (2.8) *est du type* hyperbolique, *si pour quelconques vecteurs $\boldsymbol{w} \in F$ et $\nu = (\nu_1, \nu_2) \in I\!R^2$ la matrice*

$$\mathbf{P}(\mathbf{w}, \nu) = \sum_{j=1}^{2} \nu_j \mathbf{A}_j(\boldsymbol{w}) \tag{2.10}$$

a les valeurs propres $\lambda_i = \lambda_i(\mathbf{w}, \nu),\ i = 1, \ldots, 4$, et elle peut être diagonalisée, c'est-à-dire il y a une matrice régulière $\mathbf{T} = \mathbf{T}(\mathbf{w}, \nu)$, telle que

$$\mathbf{T^{-1}} \cdot \mathbf{P} \cdot \mathbf{T} = \mathbf{D}(\mathbf{w}, \nu) = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 \\ 0 & 0 & 0 & \lambda_4 \end{pmatrix}. \tag{2.11}$$

Lemme 1 *Le système des équations d'Euler* (2.3) *est du type hyperbolique.*

Démonstration. Voir l'exemple [17].

Remarque 1 *Pour les équations d'Euler, les valeurs propres de* $\mathbf{P}(\mathbf{w}, \nu)$ *sont*

$$\begin{aligned}
\lambda_2 = \lambda_3 &= u_1\nu_1 + u_2\nu_2 \\
\lambda_1 &= \lambda_2 + c|\nu| \\
\lambda_4 &= \lambda_2 - c|\nu|
\end{aligned} \tag{2.12}$$

où

$$c = \sqrt{\frac{\gamma p}{\rho}} \tag{2.13}$$

est la vitesse locale du son *et* $\boldsymbol{u} = (u_1, u_2)$ *est le vecteur vitesse.*

Dans la littérature on peut trouver beaucoup de propriétés des équations du type hyperbolique. Ici on en présente quelques. La solution classique existe seulement pour des cas particuliers. La solution faible peut être discontinue malgré les conditions initiales régulières. La solution n'est pas toujours unique et une condition additionnelle doit être prescrite. Elle est appelée *la condition éntropique.* On ne présente pas ici la définition précise de la solution faible–éntropique, mais on fait seulement mention que l'éntropie mathématique est en corrélation de l'éntropie physique qui est bien connue de thermodynamique et physique statistique. Pour plus de détails voir [17].

L'existence de solution du système (2.3) est encore une question ouverte. Il y a les résultats particuliers, par exemple l'existence et l'unicité locale en temps de la solution classique (voir [10], [64], [56]). Les résultats plus progressifs sont obtenus par le problème plus simple – problème de Cauchy, c'est-à-dire le système (2.3) est considéré sur $I\!R^2 \times I\!R$ et la condition initiale (2.4) sur $I\!R^2$. La condition frontière manque grâce à l'infinité du domaine Ω. Malgré cette simplification l'existence et l'univocité de la solution a été prouvé seulement pour le cas scalaire (voir [42]). Pour trouver plus de propriétés des équations hyperboliques voir [17].

Nous présentons encore une propriété de (2.3) qui est importante pour la détermination du schéma numérique.

Lemme 2 *On pose*

$$G(\boldsymbol{w}, \boldsymbol{n}) = n_1 \boldsymbol{f}_1(\boldsymbol{w}) + n_2 \boldsymbol{f}_2(\boldsymbol{w}), \tag{2.14}$$

où $\boldsymbol{f}_1$, $\boldsymbol{f}_2$ *sont donnés par (2.3) et* $\boldsymbol{n} = (n_1, n_2)$, $n_1^2 + n_2^2 = 1$. *Le système (2.8) est* invariant par rotater*, c'est-à-dire si on pose*

$$Q = Q(\boldsymbol{n}) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & n_1 & n_2 & 0 \\ 0 & -n_2 & n_1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \tag{2.15}$$

on obtient

$$\begin{aligned}
G(\mathbf{w}, \mathbf{n}) &= Q^{-1} \boldsymbol{f}_1(Q\boldsymbol{w}) \\
\mathbf{P}(\mathbf{w}, \mathbf{n}) &= Q^{-1} \mathbf{A}_1(Q\boldsymbol{w}) Q,
\end{aligned} \tag{2.16}$$

où $\mathbf{A}_1$ *est donnée par (2.9) et* $\mathbf{P}$ *par (2.10).*

D é m o n s t r a t i o n . Voir pour exemple [17].

2.1.2 Méthode de volumes finis

Dans cette section on décrit le schéma numérique pour la résolution du problème (2.3) – (2.5) utilisant MVF. Cette méthode paraît convenable parce qu'elle capture mieux des discontinuités. L'idée essentielle de MVF est suivante. Le domaine Ω est replacé par un domaine polygonal Ω_h, qui est divisé sur des volumes finis D_i, $i \in J$, qui créent le maillage $\mathcal{D}_h = \{D_i\}_{i \in J}$. ($J$ est un ensemble d'indices.) La division a les propriétés suivantes:

1. D_i est un polygone fermé $\forall i \in J$
2. $\cup_{i \in J} D_i = \overline{\Omega}_h$, où $h = \sup_{i \in J}$ diam D_i, diam signifie le diamètre
3. $D_i, D_j \in \mathcal{D}_h, \quad D_i \neq D_j$
   $D_i \cap D_j = \emptyset$ ou
   $D_i \cap D_j$ est un sommet commun ou
   $D_i \cap D_j$ est un côté commun, en ce cas D_i est D_j sont appelés voisins.

On introduit la notation suivante

$$
\begin{array}{rcl}
|D_i| & & \text{la superficie de } D_i \\
\Gamma_{ij} & & \text{un côté commun parmi les volumes } D_i \text{ et } D_j \text{ voisins} \\
|\Gamma_{ij}| & & \text{la grandeur de } \Gamma_{ij} \\
s(i) & = & \{j \in J; D_i \cap D_j \text{ est un côté commun}\} \\
S(i) & = & s(i) \cup \{-1\} \quad \text{si } D_i \text{ est un volume frontière,} \\
 & & \qquad \text{c'est-à-dire } D_i \cap \partial\Omega_h \text{ est le côté } \Gamma_{i,-1} \text{ du } D_i \\
 & = & s(i) \qquad \text{sinon}
\end{array}
\tag{2.17}
$$

Il est évident que

$$\partial D_i = \cup_{j \in S(i)} \Gamma_{ij}, \quad i \in J. \tag{2.18}$$

L'intervalle du temps $(0,T)$ est divisé sur des intervalles (t_k, t_{k+1}), $k = 0, 1, \ldots, N-1$, $0 = t_0 < t_1 < \ldots < t_N = T$. Intégrant (2.3) dans $D_i \times (t_k, t_{k+1})$, on obtient

$$\int_{t_k}^{t_{k+1}} \int_{D_i} \frac{\partial \boldsymbol{w}}{\partial t}\, d\boldsymbol{x}\, dt + \int_{t_k}^{t_{k+1}} \int_{D_i} \sum_{l=1}^{2} \frac{\partial \boldsymbol{f}_l(\boldsymbol{w})}{\partial x_l}\, d\boldsymbol{x}\, dt = 0. \tag{2.19}$$

L'utilisation du théorème de Green donne

$$\int_{D_i} [\boldsymbol{w}(\boldsymbol{x}, t_{k+1}) - \boldsymbol{w}(\boldsymbol{x}, t_k)]\, d\boldsymbol{x} = -\int_{t_k}^{t_{k+1}} \int_{\partial D_i} \left[\sum_{l=1}^{2} \boldsymbol{f}_l(\boldsymbol{w})\, n_l \right] dS\, dt, \tag{2.20}$$

où $\boldsymbol{n} = (n_1, n_2)$ est le vecteur normal extérieur de D_i, $|\boldsymbol{n}| = 1$. Nous approchons la fonction $\boldsymbol{w}(\boldsymbol{x}, t)$ par une fonction constante par parties sur chaque volume fini D_i et nous notons

$$\boldsymbol{w}_i^k \approx \boldsymbol{w}(\boldsymbol{x}, t_k), \qquad \forall\, \boldsymbol{x} \in D_i, \quad i \in J \qquad \forall\, k = 0, \ldots, N. \tag{2.21}$$

Un schéma explicite sera utilisé ici, donc

$$\int_{t_k}^{t_{k+1}} \int_{\partial D_i} [\boldsymbol{f}_1(\boldsymbol{w})\, n_1 + \boldsymbol{f}_2(\boldsymbol{w})\, n_2]\, dS\; dt \approx \tau_k \int_{\partial D_i} [\boldsymbol{f}_1(\boldsymbol{w}(\boldsymbol{x}, t_k))\, n_1 + \boldsymbol{f}_2(\boldsymbol{w}(\boldsymbol{x}, t_k))\, n_2]\, dS, \tag{2.22}$$

où $\tau_k = t_{k+1} - t_k$ est le pas du temps. Les relations (2.20) – (2.22) et (2.18) impliquent un schéma explicite

$$\boldsymbol{w}_i^{k+1} = \boldsymbol{w}_i^k - \frac{\tau_k}{|D_i|} \sum_{j \in S(i)} |\Gamma_{ij}| \boldsymbol{H}(\boldsymbol{w}_i^k, \boldsymbol{w}_j^k, \boldsymbol{n}_{ij}), \quad i \in J, \tag{2.23}$$

où $\boldsymbol{n}_{ij}$ est le vecteur normal à Γ_{ij} extérieur de D_i, $|\boldsymbol{n}_{ij}| = 1$ et $\boldsymbol{H}(\boldsymbol{w}_i^k, \boldsymbol{w}_j^k, \boldsymbol{n}_{ij})$ approche l'intégrale frontière

$$\int_{\Gamma_{ij}} \sum_{l=1}^{2} \boldsymbol{f}_l(\boldsymbol{w}(\boldsymbol{x}, t_k)) n_l \; dS, \quad i \in J, \; j \in S(i). \tag{2.24}$$

La fonction $\boldsymbol{H}$ s'appelle *le flux numérique.* Le schéma (2.23) a besoin d'une condition initiale qui est posée comme

$$\boldsymbol{w}_i^0 := \frac{1}{|D_i|} \int_{D_i} \boldsymbol{w}_0(\boldsymbol{x}) \; dx, \quad i \in J. \tag{2.25}$$

Maintenant, la question est de choisir le flux numérique. Nous exigerons que $\boldsymbol{H}$ ait les propriétés convenables suivantes

1. $\boldsymbol{H}(\mathbf{v_1}, \mathbf{v_2}, \mathbf{n}) : F \times F \times S_1 \to I\!R^4$ est localement lipschitzien

$$\begin{aligned} &\forall\, R > 0, \quad \exists\, c(R) > 0, \quad \forall\, \mathbf{v_1}, \mathbf{v_2}, \mathbf{v_1^*}, \mathbf{v_2^*} \in B_R \\ &|\boldsymbol{H}(\mathbf{v_1}, \mathbf{v_2}, \mathbf{n}) - \boldsymbol{H}(\mathbf{v_1^*}, \mathbf{v_2^*}, \mathbf{n})| \le c(R)\,(|\mathbf{v_1} - \mathbf{v_1^*}| + |\mathbf{v_2} - \mathbf{v_2^*}|)\,, \end{aligned} \tag{2.26}$$

   où $S_1 = \{(n_1, n_2), n_1^2 + n_2^2 = 1\}$, $B_R = \{\boldsymbol{x} \in F, \; |\boldsymbol{x}| < R\}$ et F est un domaine auquel le vecteur $\boldsymbol{w}$ est défini

$$F = \{\boldsymbol{w} \in I\!R^4 : \; w_1 > 0, \; w_4 > \frac{w_2^2 + w_3^2}{2w_1}\}. \tag{2.27}$$

   (C'est équivalent aux relations $p > 0, \; \rho > 0$.)

2. $\boldsymbol{H}$ est *consistant* avec l'équation (2.3)

$$\boldsymbol{H}(\mathbf{v_1}, \mathbf{v_1}, \mathbf{n}) = \boldsymbol{f}_1(\mathbf{v_1}) n_1 + \boldsymbol{f}_2(\mathbf{v_1}) n_2, \qquad \mathbf{v_1} \in F, \; \boldsymbol{n} \in S_1. \tag{2.28}$$

3. $\boldsymbol{H}$ est *conservatif*

$$\boldsymbol{H}(\mathbf{v_1}, \mathbf{v_2}, \mathbf{n}) = -\boldsymbol{H}(\mathbf{v_2}, \mathbf{v_1}, -\mathbf{n}) \qquad \mathbf{v_1}, \mathbf{v_2} \in F, \; \boldsymbol{n} \in S_1. \tag{2.29}$$

On a observé que pour obtenir le flux numérique avec bonnes propriétés il faut incorporer à $\boldsymbol{H}$ l'information qui élargit le long des caractéristiques. Cette approche mène au schéma d' "upwinding". La détermination précise de ce schéma se trouve par exemple dans [17] et ici on trouve seulement la brève description.

Utilisant les relations (2.15) – (2.16) du Lemme 2, on peut écrire le flux numérique de la forme

$$\boldsymbol{H}(\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{n}) = Q^{-1} f_R(Q\boldsymbol{w}_1, Q\boldsymbol{w}_2), \tag{2.30}$$

où f_R est un *solveur de Riemann approché* correspondant au système avec une variable d'espace $\tilde{x}_1$, que l'on obtient par la transformation des coordonnées données par la matrice Q:

$$\tilde{x} = (\tilde{x}_1, \tilde{x}_2) = Q(\boldsymbol{n}_{ij})x + \tilde{0}, \qquad \tilde{0} = \text{le centre de } \Gamma_{ij}, \tag{2.31}$$

et les dérivations à la direction $\tilde{x}_2$ sont négligées. La transformation du vecteur d'état $\boldsymbol{w} = (\rho, \rho u_1, \rho u_2, e)$ conduit

$$\begin{aligned} \boldsymbol{q} &= Q\boldsymbol{w} = (\rho, \rho\tilde{u}, \rho\tilde{v}, e) \\ \tilde{u} &= u_1 n_1 + u_2 n_2 \\ \tilde{v} &= -u_1 n_2 + u_2 n_1, \end{aligned} \tag{2.32}$$

qù $\boldsymbol{n}_{ij} = (n_1, n_2)$. On écrit symboliquement

$$f_R(\boldsymbol{q}_1, \boldsymbol{q}_2) = \boldsymbol{f}_1(\boldsymbol{q}_2) - \int_{\boldsymbol{q}_1}^{\boldsymbol{q}_2} \mathbf{A}^+(\boldsymbol{q})\ d\boldsymbol{q}. \tag{2.33}$$

où

$$\mathbf{A}^+ = \mathbf{T}\mathbf{D}^+\mathbf{T}^{-1}, \qquad \mathbf{D}^+ = \mathrm{diag}(\lambda_1^+, \lambda_2^+, \lambda_3^+, \lambda_4^+), \qquad \lambda^+ = \max(\lambda, 0) \tag{2.34}$$

et $\lambda_1, \ldots, \lambda_4$ sont les valeurs propres de la matrice $\mathbf{P}(\mathbf{w}, \mathbf{n})$ de Déf. 1 où on pose $\boldsymbol{n} = (1, 0)$ et $\mathbf{T}$ est la matrice de Déf. 1. Pour plus de détails voir [17].

L'integrale dans (2.33) peut être approximée soit par une quadrature numérique convenable soit peut être calculée sur une chemin intégrale convenable. La première approche est utilisée par le flux numérique de Vijayasundaram [66] et le deuxième par le flux numérique de Osher–Solomon [51].

(i) *Le flux numérique de Vijayasundaram*

Le solveur du Riemann approximatif du schéma de Vijayasundaram a la forme suivante (voir [66]):

$$f_{VS}(\boldsymbol{q}_1, \boldsymbol{q}_2) = \mathbf{A}^+\left(\frac{\boldsymbol{q}_1 + \boldsymbol{q}_2}{2}\right)\boldsymbol{q}_1 + \mathbf{A}^-\left(\frac{\boldsymbol{q}_1 + \boldsymbol{q}_2}{2}\right)\boldsymbol{q}_2, \qquad \mathbf{A}^- = \mathbf{A} - \mathbf{A}^+. \tag{2.35}$$

Si $\Gamma_{ij} \in \partial\Omega_h$ (c'est-à-dire $j = -1$) il faut déterminer l'état frontière $\boldsymbol{w}_B$ pour que l'on puisse calculer le flux numérique $H(\boldsymbol{w}_i^k, \boldsymbol{w}_B, \boldsymbol{n}_{ij})$ sur la frontière $\partial\Omega$. On utilise l'approche heuristique proposée par [17]. Les conditions frontières qui mènent au bien–posé du problème de Riemann linéaire sont appliquées.

On prescrit m_n composantes du vecteur d'état frontière $\boldsymbol{w}_B$ et on extrapole m_p composantes du $\boldsymbol{w}_B$ de D_i, où m_n, m_p est le nombre des valeurs propres négatives, positives de la matrice $\mathbf{P}(\boldsymbol{w}_i, \boldsymbol{n}_{ij})$, respectivement. Il est convenable de distinguer plusieurs cas:

1. *Entrée subsonique* ($-c \leq \mathbf{u} \cdot \mathbf{n} < 0$ sur D_i): On prescrit ρ, u_1, u_2 sur Γ_{ij} et extrapole p de D_i on Γ_{ij}.
2. *Entrée supersonique* ($\mathbf{u} \cdot \mathbf{n} \leq -c$ sur D_i): On prescrit ρ, u_1, u_2, p sur Γ_{ij}.
3. *Sortie subsonique* ($0 < \mathbf{u} \cdot \mathbf{n} < c$ sur D_i): On prescrit p sur Γ_{ij} et extrapole ρ, u_1, u_2 de D_i sur Γ_{ij}.
4. *Sortie supersonique* ($\mathbf{u} \cdot \mathbf{n} \geq c$ sur D_i): On extrapole ρ, u_1, u_2, p de D_i sur Γ_{ij}.
5. *Parois imperméables*: On extrapole p de D_i sur Γ_{ij} par p_i^k et (en considérant la condition $\mathbf{u} \cdot \mathbf{n} = 0$) on pose
$$H(\boldsymbol{w}_i^k, \boldsymbol{w}_B, \boldsymbol{n}_{ij}) = p_i^k(0, n_1, n_2, 0)^T. \tag{2.36}$$

Pour plus de détails voir [66] ou [47].

(ii) *Le flux numérique de Osher–Solomon*

Le schéma de Osher–Solomon a été établi en utilisant les invariants de Riemann des vecteurs propres de la matrice $A_1(\boldsymbol{q})$. Le solveur de Riemann approximatif peut être exprimé comme une combinaison linéaire des valeurs de la fonction vectoriel le $f := f_1$ dans les points déterminés uniquement et exprimés analytiquement, comme on peut voir dans le tableau suivant:

	$\tilde{u}_2 \geq -c_2$	$\tilde{u}_2 \geq -c_2$	$\tilde{u}_2 < -c_2$	$\tilde{u}_2 < -c_2$
	$\tilde{u}_1 \leq c_1$	$\tilde{u}_1 > c_1$	$\tilde{u}_1 \leq c_1$	$\tilde{u}_1 > c_1$
$c_A \leq \tilde{u}_A$	$f(q_1^S)$	$\mathrm{f}(q_1)$	$f(q_2) - f(q_2^S) + f(q_1^S)$	$f(q_1) - f(q_1^S) + f(q_2)$
$0 < \tilde{u}_A$ $\tilde{u}_A < c_A$	$f(q_A)$	$f(q_1) - f(q_1^S) + f(q_A)$	$f(q_2) - f(q_2^S) + f(q_A)$	$f(q_1) - f(q_2^S) + f(q_2) -$ $- f(q_1^S) + f(q_A)$
$-c_B \leq \tilde{u}_A$ $\tilde{u}_A \leq 0$	$f(q_B)$	$f(q_1) - f(q_1^S) + f(q_B)$	$f(q_2) - f(q_2^S) + f(q_B)$	$f(q_1) - f(q_2^S) + f(q_2) -$ $- f(q_1^S) + f(q_B)$
$\tilde{u}_A < -c_B$	$f(q_2^S)$	$f(q_1) - f(q_1^S) + f(q_2^S)$	$f(q_2)$	$f(q_1) + f(q_2) - f(q_1^S)$

Les états q_A, q_B, q_1^S, q_2^S sont définis de la manière suivante. On pose

$$\begin{aligned} c_i &= (\gamma p_i/\rho_i)^{1/2}, \quad s_i = p_i/\rho_i^{\gamma}, \quad i = 1, 2, \\ z_1 &= (\gamma - 1)\tilde{u}_1/2 + c_1, \qquad z_2 = (\gamma - 1)\tilde{u}_2/2 + c_2, \qquad \alpha = (s_2/s_1)^{(1/2\gamma)}. \end{aligned} \tag{2.37}$$

Après

$$\begin{array}{llll} c_A = (z_1 - z_2)/(1 + \alpha) & \rho_A = (c_A/c_1)^{2/(\gamma-1)}\rho_1 & \tilde{u}_A = 2(z_1 - c_A)/(\gamma - 1) & \tilde{v}_A = \tilde{v}_1 \\ c_B = \alpha c_A & \rho_B = \rho_A/\alpha^2 & \tilde{u}_B = \tilde{u}_A & \tilde{v}_B = \tilde{v}_2 \\ c_1^S = 2z_1/(\gamma + 1) & \rho_1^S = (c_1^S/c_1)^{2/(\gamma-1)}\rho_1 & \tilde{u}_1^S = c_1^S & \tilde{v}_1^S = \tilde{v}_1 \\ c_2^S = -2z_2/(\gamma + 1) & \rho_2^S = (c_2^S/c_2)^{2/(\gamma-1)}\rho_2 & \tilde{u}_2^S = -c_2^S & \tilde{v}_2^S = \tilde{v}_2 \end{array} \tag{2.38}$$

Pour le cas où $\Gamma_{ij} \in \partial\Omega_h$ on définit le solveur de Riemann approximatif $f_{OS}(\boldsymbol{q}_i, \boldsymbol{q}_B) := f_1(\boldsymbol{q}_B)$ à la manière

1. *Entrée subsonique* : On prescrit $\rho_B, \tilde{u}_B, \tilde{v}_B$ et détermine p_B des relations:

$$p_B = \rho_I c_I^2/\gamma, \qquad \rho_I = (c_I^2 \rho_i/\gamma p_i)^{1/(\gamma-1)}\rho_i, \qquad c_I = c_i + \frac{\gamma - 1}{2}(\tilde{u}_i - \tilde{u}_B). \tag{2.39}$$

2. *Entrée supersonique*: On prescrit $\rho_B, \tilde{u}_B, \tilde{v}_B, p_B$.
3. *Sortie subsonique*: On prescrit p_B et détermine $\rho_B, \tilde{u}_B, \tilde{v}_B$ des relations

$$\rho_B = \rho_i(p_B/p_i)^{1/\gamma}, \qquad \tilde{u}_B = \tilde{u}_i + \frac{2}{\gamma - 1}(c_i - \sqrt{\gamma p_B/\rho_B}), \qquad \tilde{v}_B = \tilde{v}_i. \tag{2.40}$$

4. *Sortie supersonique*: On pose $\boldsymbol{q}_B = \boldsymbol{q}_i$.
5. *Parois imperméables*: On pose

$$\tilde{u}_B = 0, \quad c_B = c_i + (\gamma - 1)\tilde{u}_i/2, \quad \rho_B = (c_B^2 \rho_i/\gamma p_i)^{1/(\gamma-1)}\rho_i, \quad p_B = \rho_B c_B^2/\gamma \tag{2.41}$$

et on a $f_{OS}(\boldsymbol{q}_i, \boldsymbol{q}_B) = (0, p_B, 0, 0)^T$.

Pour plus de détails voir [51] ou [39].

Il est possible de démontrer que les flux numériques

$$\begin{aligned} \boldsymbol{H}_{VS}(\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{n}) &= Q^{-1} f_{VS}(Q\boldsymbol{w}_1, Q\boldsymbol{w}_2), \\ \boldsymbol{H}_{OS}(\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{n}) &= Q^{-1} f_{OS}(Q\boldsymbol{w}_1, Q\boldsymbol{w}_2) \end{aligned} \tag{2.42}$$

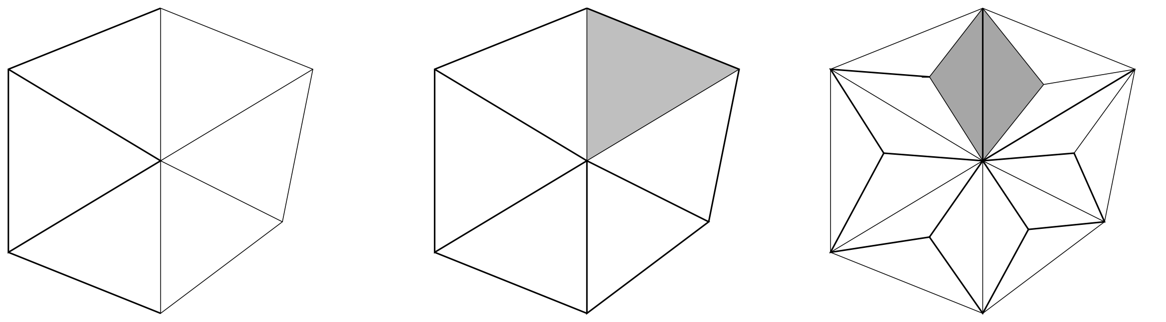


FIG. 2.1: Construction des volumes finis

satisfont les conditions (2.26), (2.28) et (2.29). Sur la base des calculs numériques le schéma d'Osher–Solomon paraît meilleur et plus robuste.

Comme la méthode (2.23) et (2.25) est explicite, le pas du temps doit être limité par une condition additionnelle. La détermination heuristique basée sur le cas linéaire et aussi les expériences numériques donnent que, la stabilité de la schéma est garantie par

$$\frac{\tau_k}{|D_i|} \max_{j \in S(i)} \left\{ |\partial D_i| \max_{m=1,\dots,4} \left\{ \lambda_m \left(\mathbf{P}(\boldsymbol{w}_i^k, \boldsymbol{n}_{ij}) \right) \right\} \right\} \leq CFL \approx 0.85, \qquad i \in J, \tag{2.43}$$

où $\lambda_m,\ m = 1, \dots, 4$ sont les valeurs propres de la matrice $\mathbf{P}$ donnée par (2.9) et (2.10). Nous avons utilisé deux types de volumes finis, des volumes finis triangulaires et barycentriques. Pour la première méthode, les volumes de contrôle utilisés pour MVF sont les triangles de la grille fondamentale. Les valeurs de la solution sont associées aux centres de gravité des triangles. La deuxième méthode utilise des volumes de contrôle barycentriques, qui sont construit au fond de la grille triangulaire fondamentale, où les centres de gravité sont joints avec les sommets de la grille fondamentale triangulaire. Les valeurs de la solution sont associées aux centres des arêtes de la grille triangulaire. (Voir Fig. 2.1.) Plus de détails sont dans la section 2.2.2.

Convergence de MVF On fait déjà mention du manque d'existence de solution des équations d'Euler. C'est la raison pour laquelle l'analyse numérique s'occupe aux problèmes plus simples, concrètement du problème scalaire de Cauchy. Pour ce problème existe la solution u et on s'intéresse si la solution numérique u_h obtenues par MVF converge vers u quand le pas du maillage va vers zéro, c'est-à-dire $u_h \to u$ si $h \to 0$. Les résultats fondamentaux ont été fait dans [41], où la convergence est démontrée dans quelques restrictions du maillage et en utilisant les propriétés du flux numérique similaire à (2.26), (2.28) et (2.29).

2.2 Discrétisation du système visqueux

Dans cette section nous nous intéressons à la solution du système des équations de Navier–Stokes (voir le Chap. 1)

$$\frac{\partial \boldsymbol{w}}{\partial t} + \sum_{i=1}^{2} \frac{\partial \boldsymbol{f}_i(\boldsymbol{w})}{\partial x_i} = \sum_{i=1}^{2} \frac{\partial \mathbf{R}_i(\boldsymbol{w}, \nabla \boldsymbol{w})}{\partial x_i} \qquad \text{dans } Q_T, \tag{2.44}$$

$$\begin{aligned}
\boldsymbol{w} &= (\rho, \rho u_1, \rho u_2, e)^T, \\
\boldsymbol{f}_i(\boldsymbol{w}) &= (\rho u_i, \rho u_1 u_i + \delta_{i1} p, \rho u_2 u_i + \delta_{i2} p, (e+p) u_i)^T, \ i = 1,2 \\
\mathbf{R}_i(\boldsymbol{w}, \nabla \boldsymbol{w}) &= \left(0, \tau_{1i}, \tau_{2i}, \sum_{j=1}^{2} \tau_{ij} u_j + \frac{\gamma}{Re\ Pr} \frac{\partial \theta}{\partial x_i}\right)^T, \quad i = 1,2 \\
\tau_{ij} &= \frac{1}{Re} \left[\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) - \frac{2}{3} \mathrm{div} \boldsymbol{u}\, \delta_{ij} \right], \quad i,j = 1,2
\end{aligned}$$

avec les conditions initiales

$$\boldsymbol{w}(\cdot, 0) = \boldsymbol{w}_0 \qquad \text{dans } \Omega \tag{2.45}$$

et les conditions frontières

$$\begin{aligned}
\rho = \tilde{\rho}, \quad \boldsymbol{u} = \tilde{\boldsymbol{u}}, \quad & \frac{\partial \theta}{\partial n} = 0 \qquad \text{sur } \Gamma_I, \\
\boldsymbol{u} = 0, \quad & \frac{\partial \theta}{\partial n} = 0 \qquad \text{sur } \Gamma_W, \\
\sum_{j=1}^{2} \tau_{ij} u_j = 0, \quad i = 1,2, \quad & \frac{\partial \theta}{\partial n} = 0 \qquad \text{sur } \Gamma_O, \\
\boldsymbol{w}|_{\Gamma+} = \boldsymbol{w}|_{\Gamma-} \quad & \qquad \text{sur les frontière périodiques,}
\end{aligned} \tag{2.46}$$

où $\tilde{\rho}$ et $\tilde{\boldsymbol{u}}$ sont les fonctions données. Le système (2.3) sera fermé en additionnant les relations

$$p = (\gamma - 1)\left(e - \frac{1}{2}\rho |\boldsymbol{u}|^2\right) \qquad \text{et} \qquad \theta = \frac{p}{\rho(\gamma - 1)}, \tag{2.47}$$

qui représentent l'équation d'état et la définition d'énergie.

Au contraire de la théorie mathématique pour l'écoulement visqueux incompressible (voir [17], [44], [62]), la théorie pour le cas compressible est pauvre. Le problème d'existence et d'unicité global (en temps) est encore une question ouverte. Néanmoins, il y a des résultats particuliers d'existence et d'unicité, qui peuvent être divisés en deux groupes:

1. l'existence et d'unicité locale (en temps)
2. l'existence et d'unicité globale dans la supposition petites données

Les résultats du premier groupe se trouvent dans [61] et du deuxième dans [46]. Ces résultats avaient été élargis dans [64], [65], [10], [52], [45]. Néanmoins le cas du système complet est encore ouvert.

En premier lieu nous nous occupons de la discrétisation du système purement visqueux (2.2). Ces équations, de type parabolique, sont résolues par *la méthode des éléments finis* (MEF), où la solution est approchée par une fonction régulière par morceaux (ici linéaire). Deux types de MEF sont utilisés: les éléments finis conformes et les éléments finis non conformes. Après on introduit deux méthodes combinantes les éléments finis et les volumes finis qui ont été déterminés pour la résolution numérique de (2.44) – (2.46).

2.2.1 La méthode des éléments finis

La MEF est une méthode très standard et sa description précise se trouve par exemple dans [8] ou [53]. C'est parce qu'ici on décrit seulement brièvement l'utilisation de MVF. L'idée est la suivante: on introduit un espace $\boldsymbol{X}_h$ des fonctions dans lequel la solution est cherchée. On utilise la formulation faible des équations (2.44), où les fonctions testées appartiennent dans un espace $\boldsymbol{V}_h \subset \boldsymbol{X}_h$. Si l'espace $\boldsymbol{V}_h$ est choisi convenablement on obtient un système de relations algébriques facilement à résoudre.

Soit $\mathcal{T}_h = \{T_k,\ k \in I\}$ une triangulation avec des propriétés standards:

1.
$$T_i \text{ est une triangle fermé } \forall i \in I \tag{2.48}$$

2.
$$\cup_{i\in T} T_i = \overline{\Omega}_h,\ \text{ où } h = \sup_{i\in I}\ \text{diam } T_i,\ \text{ diam signifie diamètre} \tag{2.49}$$

3.
$$T_i, T_j \in \mathcal{T}_h,\ T_i \neq T_j \Rightarrow \begin{cases} T_i \cap T_j = \emptyset \\ \quad \text{or} \\ T_i \cap T_j \text{ est un sommet commun} \\ \quad \text{or} \\ T_i \cap T_j \text{ est un arête commun} \end{cases} \tag{2.50}$$

On va définir l'espace des fonctions $\boldsymbol{X}_h$ dans lequel on va chercher la solution. Cet espace est différent pour les éléments finis conformes et non conformes.

Les éléments finis conformes Nous définissons les espaces $X_h, \boldsymbol{X}_h, \boldsymbol{V}_h$ comme

$$\begin{aligned} X_h &= \{v \in C(\overline{\Omega}),\ v|_{T_i} \text{ est linéaire } \forall i \in I\} \\ \boldsymbol{X}_h &= [X_h]^4 \\ \boldsymbol{V}_h &= \{\phi_h \in \boldsymbol{X}_h,\ i\text{-ème élément de } \phi_h \text{ est égal à 0 sur la partie de frontière } \partial\Omega_h, \\ &\qquad \text{où la condition Dirichlet est donnée}\}. \end{aligned} \tag{2.51}$$

Nous cherchons la solution numérique $\boldsymbol{w}_h$ de (2.2) dans l'espace $\boldsymbol{X}_h$. En multipliant (2.2) considéré sur niveau du temps t_k par $\varphi_h = (\varphi_{h,1}, \varphi_{h,2}, \varphi_{h,3}, \varphi_{h,4}) \in \boldsymbol{V}_h$, en intégrant dans Ω_h, en utilisant le théorème de Green et en approximant la dérivation du temps on obtient

$$\int_{\Omega_h} \boldsymbol{w}_h^{k+1} \varphi_h d\boldsymbol{x} = \int_{\Omega_h} \boldsymbol{w}_h^k \varphi_h d\boldsymbol{x} - \tau_k \int_{\Omega_h} \sum_{s=1}^{2} \mathbf{R}_s(\boldsymbol{w}_h^k, \nabla \boldsymbol{w}_h^k) \frac{\partial \varphi_h}{\partial x_s} d\boldsymbol{x}, \quad \forall\ \varphi_h \in \mathbf{V}_h, \tag{2.52}$$

où $\boldsymbol{w}_h^k = \boldsymbol{w}_h(\cdot, t_k),\ k = 0, \ldots, r$. Les intégrales sur $\partial\Omega_h$ néant par l'utilisation du théorème de Green ont disparu grâce aux conditions aux limites (2.46) et aux propriétés de $\varphi \in \boldsymbol{V}_h$. Nous utilisons la quadrature numérique (mass lumping)

$$\int_T f\ dx \approx \frac{1}{3}|T| \sum_{i=1}^{3} f(P_T^i), \tag{2.53}$$

où f est une fonction définite sur T, $P_T^i,\ i = 1, 2, 3$ sont les sommets du triangle T. Cette quadrature est précisée pour toutes les fonctions linéaires (voir [8]).

Nous avons

$$(\boldsymbol{w}_h^{k+1}, \varphi_h)_h = (\boldsymbol{w}_h^k, \varphi_h)_h - \tau_k a_h(\boldsymbol{w}_h^k, \varphi_h), \quad \forall\ \varphi_h \in \mathbf{V}_h, \tag{2.54}$$

où

$$(\boldsymbol{w}_h, \varphi_h)_h = \frac{1}{3} \sum_{T \in \mathcal{T}_h} |T| \sum_{i=1}^{3} \boldsymbol{w}_h(P_T^i)\varphi_h(P_T^i), \qquad \boldsymbol{w}_h, \varphi_h \in \mathbf{X}_h \tag{2.55}$$

et

$$\begin{aligned}
a_h(\boldsymbol{w}_h, \varphi_h) &= \left(a_h^1(\boldsymbol{w}_h, \varphi_h), \ldots, a_h^4(\boldsymbol{w}_h, \varphi_h)\right)^T, \\
a_h^1(\boldsymbol{w}, \varphi_h) &\equiv 0, \\
a_h^2(\boldsymbol{w}, \varphi_h) &= \sum_{T \in \mathcal{T}_h} |T| \frac{1}{Re} \Bigg\{ 2 \frac{\partial u_{h,1}}{\partial x_1}\Big|_T \frac{\partial \varphi_{h,2}}{\partial x_1}\Big|_T - \\
&- \frac{2}{3}(\mathrm{div}\ \mathbf{u}_h)|_T \frac{\partial \varphi_{h,2}}{\partial x_1}\Big|_T + \left(\frac{\partial u_{h,2}}{\partial x_1}\Big|_T + \frac{\partial u_{h,1}}{\partial x_2}\Big|_T\right) \frac{\partial \varphi_{h,2}}{\partial x_2}\Big|_T \Bigg\}, \\
a_h^3(\boldsymbol{w}, \varphi_h) &= \sum_{T \in \mathcal{T}_h} |T| \frac{1}{Re} \Bigg\{ 2 \frac{\partial u_{h,2}}{\partial x_2}\Big|_T \frac{\partial \varphi_{h,3}}{\partial x_2}\Big|_T - \\
&- \frac{2}{3}(\mathrm{div}\ \mathbf{u}_h)|_T \frac{\partial \varphi_{h,3}}{\partial x_2}\Big|_T + \left(\frac{\partial u_{h,2}}{\partial x_1}\Big|_T + \frac{\partial u_{h,1}}{\partial x_2}\Big|_T\right) \frac{\partial \varphi_{h,3}}{\partial x_1}\Big|_T \Bigg\}, \\
a_h^4(\boldsymbol{w}, \varphi_h) &= \sum_{T \in \mathcal{T}_h} \Bigg\{ \frac{1}{3}|T| \left(\tau_{h,11}|_T \sum_{i=1}^{3} u_{h,1}(P_T^i) + \tau_{h,12}|_T \sum_{i=1}^{3} u_{h,2}(P_T^i)\right) \frac{\partial \varphi_{h,4}}{\partial x_1}\Big|_T + \\
&+ \frac{1}{3}|T| \left(\tau_{h,21}|_T \sum_{i=1}^{3} u_{h,1}(P_T^i) + \tau_{h,22}|_T \sum_{i=1}^{3} u_{h,2}(P_T^i)\right) \frac{\partial \varphi_{h,4}}{\partial x_2}\Big|_T + \\
&+ \frac{\gamma}{Re\ Pr}|T| \sum_{j=1}^{2} \frac{\partial \theta_h}{\partial x_j}\Big|_T \frac{\partial \varphi_{h,4}}{\partial x_j}\Big|_T \Bigg\}, \\
\tau_{h,rs}|_T &= \frac{1}{Re}\left[-\frac{2}{3}\ \mathrm{div}\ \boldsymbol{u}_h\ \delta_{rs} + \left(\frac{\partial u_{h,r}}{\partial x_s} + \frac{\partial u_{h,s}}{\partial x_r}\right)\right]\Bigg|_T = \ \mathrm{const.}\ ,
\end{aligned} \tag{2.56}$$

où $u_{h,s}$, $s = 1, 2$ et θ_h sont les fonctions de l'espace X_h approchant des composantes de vitesse et température. Ce système a une unique solution. Nous venons d'obtenir un schéma explicite et pour cette raison le pas du temps doit être limité par une condition de stabilité. La détermination heuristique basée sur le cas scalaire et aussi les expériences numériques donnent que la stabilité du schéma est garantie par

$$3 \frac{\tau_k}{|T|} \frac{1}{Re} \leq CFL \leq 1 \quad \forall T \in \mathcal{T}_h. \tag{2.57}$$

La méthode des éléments finis non conformes Soit $\mathcal{T}_h = \{T_k,\ k \in I\}$ une triangulation avec les propriétés (2.48) – (2.50). Soit $\mathcal{S}_h$ un ensemble des arêtes S_k des triangles $T_i \in \mathcal{T}_h$ et L son ensemble d'indices, c'est-à-dire

$$\mathcal{S}_h = \{S_k,\ k \in L\}. \tag{2.58}$$

Le centre de S_k est désigné par Q_k. Nous définissons les espaces $X_h, \boldsymbol{X}_h, \boldsymbol{V}_h$ comme

$$\begin{aligned} X_h &= \{v \in L^2(\Omega),\ v|_{T_i} \text{ est linéaire } \forall i \in I,\ v \text{ est continues en } Q_k\ \forall k \in L\} \\ \boldsymbol{X}_h &= [X_h]^4 \\ \boldsymbol{V}_h &= \{\phi_h \in \boldsymbol{X}_h,\ i\text{-ème élément de } \phi_h \text{ est égal à 0 dans } Q_k \in \partial\Omega_h, \\ &\qquad \text{où la condition Dirichlet est donnée}\}. \end{aligned} \tag{2.59}$$

Maintenant, nous pouvons suivre la procédure pour les éléments conformes. A la place (2.53) nous utilisons

$$\int_T f\ dx \approx \frac{1}{3}|T| \sum_{i=1}^{3} f(Q_T^i), \tag{2.60}$$

où $Q_T^i,\ i = 1, 2, 3$ sont les centres des arêtes de T. L'approximation (2.60) est précise pour tout polynômes de degré au plus égal à 2. (Voir [8]). Les relations (2.54), (2.55) et (2.56) restent les mêmes.

Les relations (2.54) – (2.56) représentent l'utilisation de MEF pour (2.2)

2.2.2 Discrétisation du système complet

Maintenant nous pouvons procéder à la discrétisation du problème (2.44) – (2.46). Nous avons divisé la système (2.44) en systèmes (2.1) et (2.2), que nous avons discrétisé séparément et *la méthode de la division de l'opérateur* a été appliquée.

Nous avons développé deux méthodes. La première utilise des volumes finis (VF) triangulaires et des éléments finis (EF) conformes et la deuxième utilise des volumes finis barycentriques et des éléments finis non conformes.

Méthode VF triangulaire et EF conformes Soit $\mathcal{T}_h$ une triangulation du domaine Ω_h avec les propriétés (2.48) – (2.50). Les volumes finis sont les triangles et on pose $\mathcal{D}_h := \mathcal{T}_h$. Maintenant, nous construisons une nouvelle triangulation $\mathcal{T}_h'$. Les sommets de $\mathcal{T}_h'$ sont les centres de gravité de la grille fondamentale $\mathcal{D}_h$ et les sommets de $\mathcal{D}_h$ qui appartiennent à la frontière $\partial\Omega_h$. (Voir Fig. 2.2)

La triangulation $\mathcal{T}_h'$ s'appelle *la triangulation associée*. On désigne par $\mathcal{P}_h$ l'ensemble de toutes sommets de $\mathcal{T}_h'$ et $\mathcal{P}_h = \{P_k,\ k \in K\}$. Il est évident que

$$K = J \cup B,\ J \cap B = \emptyset, \tag{2.61}$$

où J est l'ensemble de indices de $\mathcal{D}_h$ et B est l'ensemble de indices du sommet de $\mathcal{D}_h$ qui appartient à la frontière $\partial\Omega_h$. Nous avons $\mathcal{D}_h = \{D_j,\ j \in J\}$, $\mathcal{T}_h' = \{T_i,\ i \in I\}$. Nous définissons les espaces de fonctions suivants

$$\begin{aligned} X_h &= \{u_h \in C(\overline{\Omega}_h), \quad u_h|_{T_i} \text{ linéaire } \quad \forall i \in I\}, \\ \boldsymbol{X}_h &= [X_h]^4, \\ Z_h &= \{u_h \in L^2(\Omega_h), \quad u_h|_{D_j} \text{ constant } \quad \forall j \in J\}, \\ \boldsymbol{Z}_h &= [Z_h]^4. \end{aligned} \tag{2.62}$$

La base de X_h est créée par les fonctions $\varphi_j \in X_h$

$$\varphi_i(P_j) = \delta_{ij}, \quad i, j \in K, \tag{2.63}$$

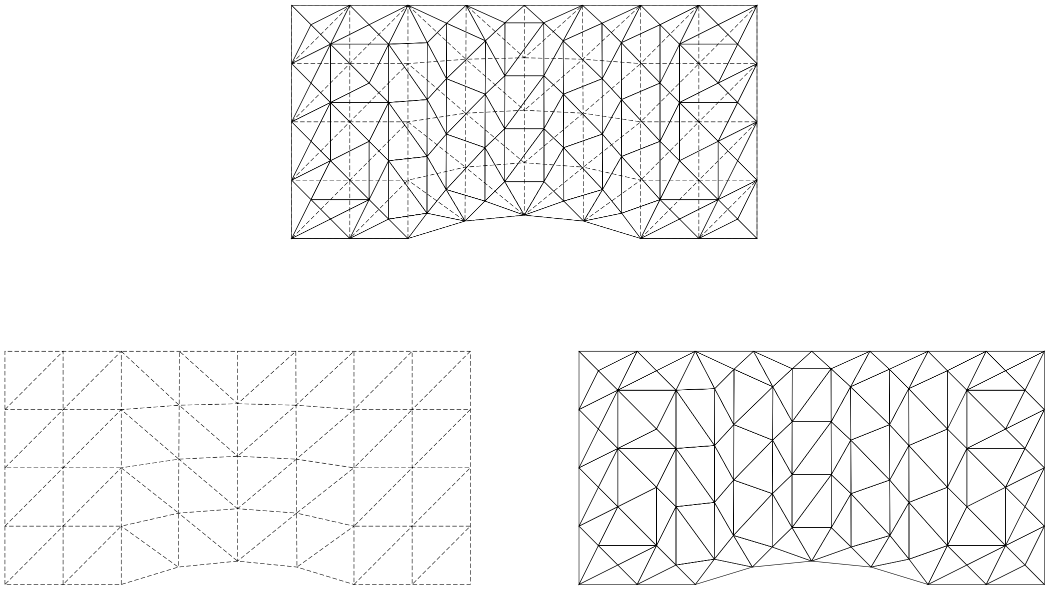


FIG. 2.2: Un maillage triangulaire et un maillage triangulaire associé: ensemble et séparément

où $P_i \in \mathcal{P}_h$. La base de Z_h est créée par $\psi_i \in Z_h$

$$\psi_i = \chi_{D_i}, \quad i \in J, \tag{2.64}$$

où χ_{D_i} est la fonction caractéristique de D_i

$$\chi_{D_i}(x) = \begin{cases} 1 & x \in D_i, \\ 0 & x \notin D_i. \end{cases} \tag{2.65}$$

Après, il est facile de voir que

$$w_h = \sum_{i \in J} w_i \psi_i \qquad w_h \in Z_h \tag{2.66}$$

$$\tilde{w}_h = \sum_{j \in K} \tilde{w}_h(P_j)\varphi_j \qquad \tilde{w}_h \in X_h. \tag{2.67}$$

Nous pouvons définir l'opérateur $\Pi_h : Z_h \to X_h$ par

$$\begin{aligned} &\Pi_h w_h = \tilde{w}_h, \quad w_h \in Z_h, \ \tilde{w}_h \in X_h, \\ &\tilde{w}_h(P_j) = \begin{cases} w_h|_{D_j}, & j \in J \quad (P_j \text{ est centre de gravité de } D_j), \\ Lw_h, & j \in B \quad (P_j \in \partial\Omega_h), \end{cases} \\ &\tilde{w}_h = \sum_{j \in K} \tilde{w}_h(P_j)\varphi_j, \end{aligned} \tag{2.68}$$

qui suit de (2.61). Lw_h signifie une reconstruction de la fonction w_h sur la frontière $\partial\Omega_h$. L'opérateur $\Pi_h^{-1} : X_h \to Z_h$ est prescrit par

$$\Pi_h^{-1}\tilde{w}_h = w_h, \quad \tilde{w}_h \in X_h, \ w_h \in Z_h, \tag{2.69}$$

$$w_h|_{D_j} = \tilde{w}_h(P_j),$$
$$w_h = \sum_{j\in J} w_h|_{D_j}\psi_j.$$

(Les opérateurs Π_h, Π_h^{-1} ne sont pas inverses.)

Pour les vecteurs $\boldsymbol{w}_h = ((w_h)_1, (w_h)_2, (w_h)_3, (w_h)_4)$, $\tilde{\boldsymbol{w}}_h = ((\tilde{w}_h)_1, (\tilde{w}_h)_2, (\tilde{w}_h)_3, (\tilde{w}_h)_4)$ les opérateurs Π_h, Π_h^{-1} sont définis séparément pour chacun composant.

Nous déduisons comment nous obtenons les valeurs $\boldsymbol{w}_i^{k+1}$ si nous connaissons les valeurs $\boldsymbol{w}_i^k$. Chacune itération a quatre pas suivants

1. $\boldsymbol{w}_i^k \to \boldsymbol{w}_i^{k+\frac{1}{2}}$, $i \in J$ utilisant le schéma MVF (2.23), (2.25), $\qquad \boldsymbol{w}_h^{k+\frac{1}{2}} = \sum_{j\in J} \boldsymbol{w}_j^{k+\frac{1}{2}}\psi_j$.
2. $\boldsymbol{w}_h^{k+\frac{1}{2}} \to \tilde{\boldsymbol{w}}_h^{k+\frac{1}{2}}$, $\quad \tilde{w}_h = \Pi_h w_h$ définit par (2.68).
3. $\tilde{\boldsymbol{w}}_h^{k+\frac{1}{2}} \to \tilde{\boldsymbol{w}}_h^{k+1}$ utilisant la schéma MEF (2.54) – (2.56)
4. $\tilde{\boldsymbol{w}}_h^{k+1} \to \boldsymbol{w}_h^k$, $\quad w_h^{k+1} = \Pi_h^{-1}\tilde{w}_h^{k+1}$ définit par (2.69), $\qquad \boldsymbol{w}_h^{k+1} = \sum_{j\in J} \boldsymbol{w}_j^{k+1}\psi_j$.

Méthode VF barycentriques et EF non conformes Soit $\mathcal{T}_h = \{T_i\}_{i\in I}$ une triangulation du domaine Ω_h satisfaisant (2.48) – (2.50). Soit $\mathcal{S}_h$ un ensemble de toutes arêtes S_j des triangles $T_i \in \mathcal{T}_h$ et J son ensemble d'indices, c'est-à-dire

$$\mathcal{S}_h = \{S_j,\ j \in J\}. \tag{2.70}$$

Le centre de S_j est désigné par Q_j. Maintenant, nous construisons une grille barycentrique qui naît par connexion des centres de gravité de triangles de $\mathcal{T}_h$ avec leurs sommets. (Voir Fig. 2.3.)

Il est simple à voir que pour chaque Q_j, $j \in J$ existe un volume fini barycentrique D_j qui est un quadrangle ou un triangle. On pose $\mathcal{D}_h = \{D_j, j \in J\}$. Le procédé ultérieur est analogue avec la première méthode.

Nous définissons les espaces de fonctions suivantes

$$\begin{aligned}
X_h &= \{u_h \in L^2(\Omega_h),\quad u_h|_{T_i} \text{ linéaire} \quad \forall i \in I,\ u_h \text{ est continué à } Q_j\ \forall j \in J\},\\
\boldsymbol{X}_h &= [X_h]^4,\\
Z_h &= \{u_h \in L^2(\Omega_h),\quad u_h|_{D_j} \text{ constant} \quad \forall j \in J\},\\
\boldsymbol{Z}_h &= [Z_h]^4.
\end{aligned} \tag{2.71}$$

La base de X_h est créée par les fonctions $\varphi_j \in X_h$

$$\varphi_i(Q_j) = \delta_{ij}, \quad i, j \in J. \tag{2.72}$$

La base de Z_h est créée par $\psi_i \in Z_h$

$$\psi_i = \chi_{D_i}, \quad i \in J. \tag{2.73}$$

Les relations (2.66) et (2.67) sont valables aussi pour les bases définies par (2.72) et (2.73). Nous définissons l'opérateur $\Pi_h : Z_h \to X_h$ par

$$\Pi_h w_h = \tilde{w}_h, \quad w_h \in Z_h,\ \tilde{w}_h \in X_h \tag{2.74}$$
$$\tilde{w}_h(Q_j) = w_h|_{D_j},\ j \in J$$
$$\tilde{w}_h = \sum_{j\in J} \tilde{w}_h(Q_j)\varphi_j.$$

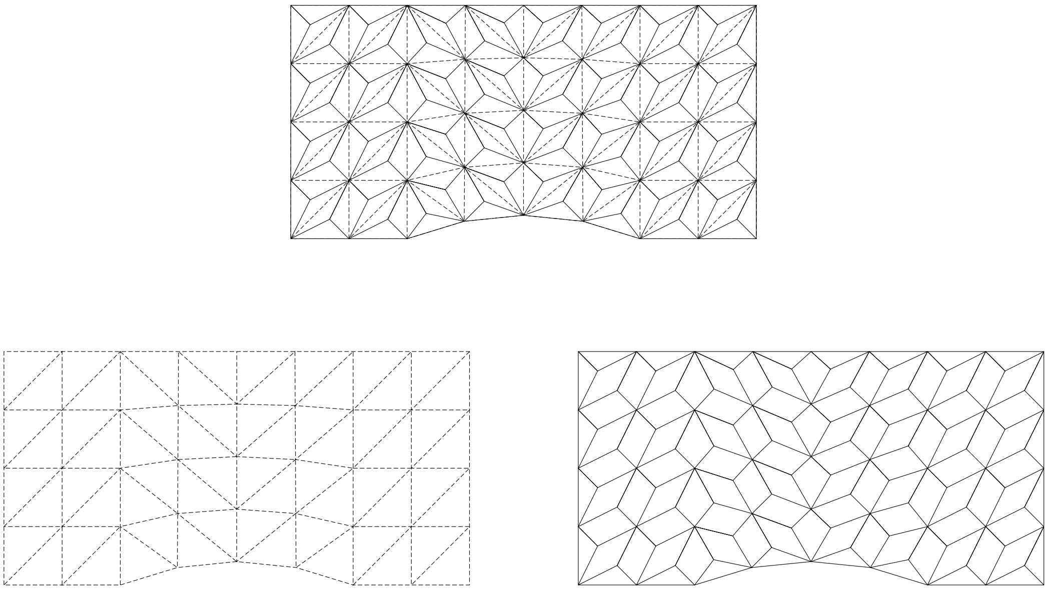


FIG. 2.3: Un maillage triangulaire et un maillage barycentrique: ensemble et séparément

L'opérateur inverse vers Π_h est désignée par Π_h^{-1}:

$$\begin{aligned} &\Pi_h^{-1}\tilde{w}_h = w_h, \quad \tilde{w}_h \in X_h,\ w_h \in Z_h, \\ &w_h|_{D_j} = \tilde{w}_h(Q_j), \\ &w_h = \sum_{j\in J} w_h|_{D_j}\psi_j. \end{aligned} \tag{2.75}$$

On rappelle que pour les vecteurs $\boldsymbol{w}_h = ((w_h)_1, \ldots, (w_h)_4)$, $\tilde{\boldsymbol{w}}_h = ((\tilde{w}_h)_1, \ldots, (\tilde{w}_h)_4)$ les opérateurs Π_h, Π_h^{-1} sont définis séparément pour chacune composante.

Maintenant, il est possible utiliser une procédure semblable, comme pour la première méthode.

1. $\boldsymbol{w}_i^k \to \boldsymbol{w}_i^{k+\frac{1}{2}}$, $i \in J$ utilisant le schéma MVF (2.23), (2.25), $\quad \boldsymbol{w}_h^{k+\frac{1}{2}} = \sum_{j\in J} \boldsymbol{w}_j^{k+\frac{1}{2}}\psi_j$.
2. $\boldsymbol{w}_h^{k+\frac{1}{2}} \to \tilde{\boldsymbol{w}}_h^{k+\frac{1}{2}}$, $\quad \tilde{w}_h = \Pi_h w_h$ définit par (2.74).
3. $\tilde{\boldsymbol{w}}_h^{k+\frac{1}{2}} \to \tilde{\boldsymbol{w}}_h^{k+1}$ utilisant la schéma MEF (2.54) – (2.56)
4. $\tilde{\boldsymbol{w}}_h^{k+1} \to \boldsymbol{w}_h^k$, $\quad w_h^{k+1} = \Pi_h^{-1}\tilde{w}_h^{k+1}$ définit par (2.75), $\quad \boldsymbol{w}_h^{k+1} = \sum_{j\in J} \boldsymbol{w}_j^{k+1}\psi_j$.

Les méthodes combinantes VF et EF peuvent être utilisées pour le calcul d'écoulement stationnaire par la méthode de stabilisation du temps, c'est-à-dire le calcul est effectué pour $t \to \infty$. En commençant de la condition initiale on utilise le processus itératif. La solution approximative est considérée stationnaire si le résidu est petit:

$$\|\boldsymbol{w}_h^{k+1} - \boldsymbol{w}_h^k\|/\tau_h < \varepsilon, \tag{2.76}$$

où ε est une tolérance donnée.

L'application des méthodes pour les problèmes de mécanique des fluides est dans le Chap. 4. La question de la convergence de ces méthodes est discutée dans la Partie: Analyses théoriques.

Chapitre 3

Les méthodes d'ordre plus élevé

L'augmentation de la précision d'une solution est possible par l'utilisation de méthodes d'ordre plus élevé. Dans nos problèmes, l'influence des termes de la diffusion est petite en comparaison à celle des termes convectifs. C'est la raison pour laquelle nous essayons d'augmenter l'ordre de précision de la méthode volumes finis qui est utilisée pour l'approximation des termes convectifs.

Dans la littérature on peut trouver plusieurs méthodes d'ordre plus élevé. Ces méthodes sont basées sur la généralisation de la reconstruction linéaire (schémas ENO et TVD) pour des maillages non structurés. Nous avons modifié quelques de ces méthodes pour des maillages triangulaires et tous étaient testées sur la tube de Sod, où la solution analytique est connue et où l'ordre de la précision peut être déterminé.

Ici on présente l'article qui a été accepté pour la présentation dans le *"Journal of Engineering Mechanics"*, num. 5/1998. Cet article est consacré aux méthodes d'ordre plus élevé sur des maillages triangulaires.

L'idée fondamentale d'utilisation des méthodes d'ordre plus élevé pour des maillages barycentriques se trouve dans la Sec. 3.5 (voir aussi [13]).

INVESTIGATION OF ORDER OF ACCURACY FOR HIGHER ORDER FINITE VOLUME SCHEMES

V. Dolejší, J. Felcman*

The paper is concerned with the higher order extension of the cell centred finite volume scheme for the numerical modelling of inviscid gas flow. In the smooth part of the flow field, first order accuracy is not accurate enough. Furthermore, the shocks and contact discontinuities are smeared out because of the numerical viscosity hidden in the first order scheme. Therefore, we wish to improve the order of accuracy on one hand and to resolve discontinuities without introducing over or undershoots on the other hand. In the last several years there has been considerable effort aimed at constructing and analyzing high order accurate, non-oscillatory approximations to hyperbolic conservation laws. Here we investigate several methods and determine the order of accuracy for Sod's channel where the analytical solution is known. The computation is performed for three different triangulations with sizes of mesh h, $h/2$ and $h/4$ and the determination of the order of accuracy is done using the method of the least squares.

Keywords: compressible flow, Euler equations, finite volume method, upwind flux vector splitting schemes of Godunov type, higher order method

3.1 Formulation of problem

We consider the first order hyperbolic system in a space-time cylinder $Q_T = \Omega \times (0,T)$

$$\frac{\partial w}{\partial t} + \sum_{s=1}^{2} \frac{\partial f_s(w)}{\partial x_s} = 0 \tag{3.1}$$

describing the inviscid compressible flow. Here

$$\begin{aligned} &w = (\rho,\, \rho v_1,\, \rho v_2,\, e)^T, \quad w = w(x,t), \quad x \in \Omega, \quad t \in (0,T), \\ &f_s(w) = (\rho v_s,\, \rho v_s v_1 + \delta_{s1} p,\, \rho v_s v_2 + \delta_{s2} p,\, (e+p)\, v_s)^T \end{aligned} \tag{3.2}$$

and p is given by the state equation

$$p = (\gamma - 1)\,(e - \rho|\boldsymbol{v}|^2/2).$$

We use the standard notation: t – time, x_s $(s = 1, 2)$ – Cartesian coordinates, ρ – density, p – pressure, e – total energy, $\boldsymbol{v} = (v_1, v_2)$ – velocity, δ_{ij} – Kronecker delta, $\gamma > 1$ is the Poisson adiabatic constant.

The initial condition and a set of boundary conditions complete the above system. For their description see e.g. [17].

* Mgr. V. Dolejší, Dr. J. Felcman, Faculty of Mathematics and Physics, Charles University Prague, Malostranské nám. 25, 118 00 Praha 1, Czech Republic

Let the domain Ω be approximated by a polygonal domain Ω_h. We construct a triangulation $\mathcal{D}_h$ of Ω_h with the usual properties (in the sense of the finite element method), $\mathcal{D}_h = \{D_i\}_{i\in I}$, I is a suitable index set. We denote by h the mesh size parameter, i.e.

$$h = \max_{D_i \in \mathcal{D}_h} \operatorname{diam} D_i. \tag{3.3}$$

The boundary ∂D_i of the triangle D_i can be expressed in the form

$$\partial D_i = \bigcup_{j\in S(i)} \Gamma_{ij}, \tag{3.4}$$

where Γ_{ij} is either the common side of two neighbouring triangles D_i and D_j or $\Gamma_{ij} \subset \partial\Omega_h$. $S(i)$ is an index set. Furthermore, we introduce the following notation: $|D_i|$ = area of D_i, $n_{ij} = (n_1, n_2)$ = unit outer normal to ∂D_i on Γ_{ij}, ℓ_{ij} = length of Γ_{ij}, $0 = t_0 < t_1 < \ldots$ is a partition of the time interval $(0, T)$ and $\tau_k = t_{t+1} - t_k$ is a time step, $k = 0, 1, \ldots$.

The *finite volume (cell centred) discretization* is based on the integration of (3.1) over every set $D_i \times (t_k, t_{k+1})$, the use of Green's theorem, the approximation of the exact solution w averaged at time t_k over the triangle D_i by constant values w_i^k and the approximation of the flux

$$\int_{\Gamma_{ij}} \sum_{s=1}^{2} f_s(w)\, n_s \; dS$$

of the quantity w through the segment Γ_{ij} in the direction n_{ij} per second on the time level t_k with the aid of the so-called *numerical flux* $H(w_i^k, w_j^k, n_{ij})$ calculated from w_i^k, w_j^k and n_{ij}. In this way we obtain the following explicit numerical scheme:

$$\begin{aligned} w_i^{k+1} \quad &= \quad w_i^k - \frac{\tau_k}{|D_i|} \sum_{j\in S(i)} H(w_i^k, w_j^k, n_{ij})\, \ell_{ij}, \\ &\qquad\qquad D_i \in \mathcal{D}_h \text{ (i. e., } i \in I), \; k = 0, 1, \ldots. \end{aligned} \tag{3.5}$$

In case that $\Gamma_{ij} \subset \partial\Omega_h$, it is necessary to determine the boundary state on Γ_{ij}. For the evaluation of the numerical flux in (3.5) the *Osher–Solomon scheme* is used (see [51]).

3.2 Higher order reconstruction

The higher order extension of the scheme (3.5) lies in the nature of the preprocessing of the cell averaged data, which is accomplished via a nearest neighbour linear interpolation followed by a slope limiting procedure. The numerical flux $H(w_i^k, w_j^k, n_{ij})$ in the scheme (3.5) (this scheme is denoted as BASIC) is replaced by the flux $H(\hat{w}_{ij}^k, \hat{w}_{ji}^k, n_{ij})$, where values $\hat{w}_{ij}^k$, $\hat{w}_{ji}^k$ are reconstructed component-wise from the cell averaged data in the following form

$$\hat{w}_{ij}^k = w_i^k + \nabla L_i \cdot (x_i - x_{ij}, y_i - y_{ij})^T, \quad i \in I, j \in S(i). \tag{3.6}$$

Here (x_i, y_i) is the center of gravity of D_i, (x_{ij}, y_{ij}) is the center of Γ_{ij} and L_i is a linear function defined on $D_i \in \mathcal{D}_h$ (see Figure 1). We call L_i a linear reconstruction of w on D_i. We have proposed and tested several different methods and completed them by the comparison with the methods from [15, 30]. For all of this methods the order of accuracy is numerically determined and methods are mutually compared. Now we describe the computation of L_i for each of tested methods (we omit the "time" index k).

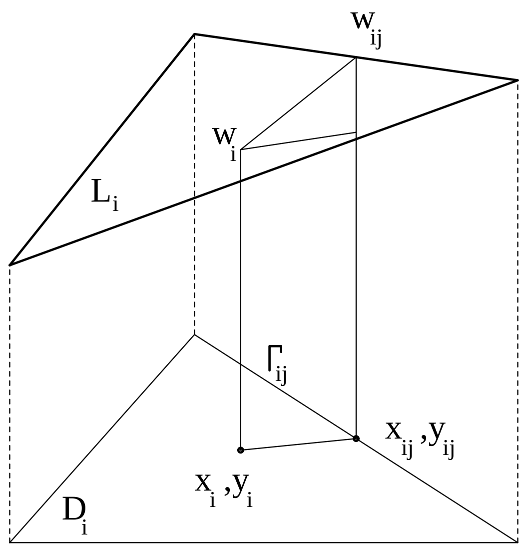


Fig. 1: Linear reconstruction on D_i

3.2.1 Description of higher order schemes

1. **DEO–M**. This method was proposed in [15]. We associate the value of w_i with the center of gravity of D_i. For each finite volume $D_i \in \mathcal{D}_h$, $D_i \cap \partial\Omega_h = \emptyset$, there exist three neighbouring triangles: D_{i1}, D_{i2}, D_{i3}. Let D_{i1}, D_{i2}, D_{i3} be the neighbours of D_i and $w_i, w_{i1}, w_{i2}, w_{i3}$ the corresponding values of the approximate solution. By $P_i, P_{i1}, P_{i2}, P_{i3}$ we denote the centers of gravity of $D_i, D_{i1}, D_{i2}, D_{i3}$, respectively. The centers of edges of D_i are denoted by z_{i1}, z_{i2}, z_{i3} (see Figure 2). There exist linear functions L_{i1}, L_{i2}, L_{i3} such that

$$\begin{array}{lll} L_{i1}(P_i) = w_i, & L_{i1}(P_{i1}) = w_{i1}, & L_{i1}(P_{i2}) = w_{i2}, \\ L_{i2}(P_i) = w_i, & L_{i2}(P_{i2}) = w_{i2}, & L_{i2}(P_{i3}) = w_{i3}, \\ L_{i3}(P_i) = w_i, & L_{i3}(P_{i3}) = w_{i3}, & L_{i3}(P_{i1}) = w_{i1}. \end{array} \tag{3.7}$$

We put

$$|L_{ij}| := \left[\left(\frac{\partial L_{ij}}{\partial x_1} \right)^2 + \left(\frac{\partial L_{ij}}{\partial x_2} \right)^2 \right]^{\frac{1}{2}}. \tag{3.8}$$

The following procedure for construction of L_i is used:

(a) We reorder (using a permutation) $L_{ij},\ j = 1,2,3$ onto $L^{ij},\ j = 1,2,3$ such that $|L^{i1}| \geq |L^{i2}| \geq |L^{i3}|$ and set $L_i := L^{i1}$.

(b) We verify the following conditions

$$\begin{aligned} &L_i(z_{i1}) \text{ is between } w_i \text{ and } w_{i1}, \\ &L_i(z_{i2}) \text{ is between } w_i \text{ and } w_{i2}, \\ &L_i(z_{i3}) \text{ is between } w_i \text{ and } w_{i3}. \end{aligned} \tag{3.9}$$

If all conditions (3.9) are satisfied then the procedure is finished. If this is not the case we continue to the next step.

(c) We put $L_i := L_{i2}$ and we test the condition (3.9) again. If the conditions (3.9) are satisfied for the second candidate we take it, if not we set $L_i := L^{i3}$ and finish the procedure , i.e. for the L^{i3} conditions (3.9) are not tested.

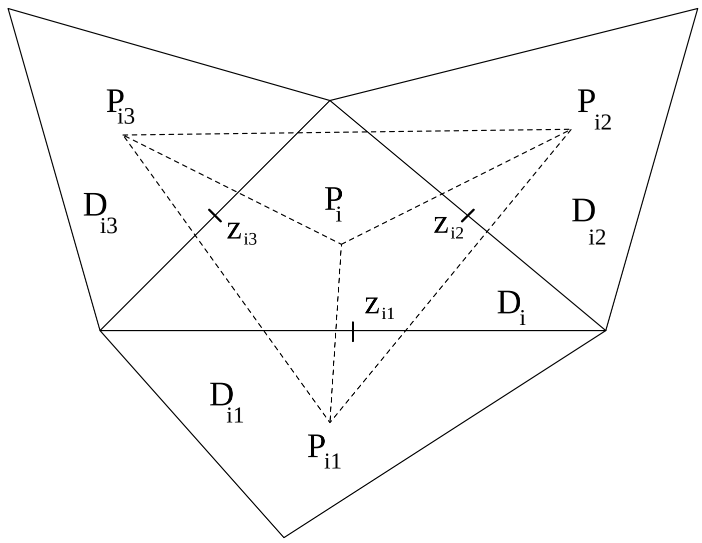


Fig. 2: Triangle D_i and its neighbours D_{i1}, D_{i2}, D_{i3}

2. **DEO–B**. This method is very similar to DEO-M. The condition (3.9) are tested also for the case when we chose $L_i := L^{i3}$. If L^{i3} does not satisfy the conditions (3.9) then we put $L_i := 0$, i.e. we use $\hat{w}_{ij} = w_i$.
3. **MIN**. This method is also a modification of **DEO–M**. We put directly $L_i := L^{i3}$, i.e. from three candidates we choose this one whose absolute value of the gradient is minimal. This construction is analogous to ENO-schemes.
4. **AVER**. The weighted average of three linear functions is chosen. Using the previous notation we set

$$L_i := \frac{L_{i1}|P_iP_{i1}P_{i2}| + L_{i2}|P_iP_{i2}P_{i3}| + L_{i3}|P_iP_{i3}P_{i1}|}{|P_iP_{i1}P_{i2}| + |P_iP_{i2}P_{i3}| + |P_iP_{i3}P_{i1}|}, \tag{3.10}$$

where $|P_iP_{i1}P_{i2}|$ is the area of triangle $P_iP_{i1}P_{i2}$ and similarly for triangles $P_iP_{i2}P_{i3}$ and $P_iP_{i3}P_{i1}$.
5. **EXTRAP**. In order to involve more information into the scheme, we extrapolate cell average values into vertices and construct their linear interpolant. Let $D_i \in \mathcal{D}_h$ and $X_{i_1}, X_{i_2}, X_{i_3}$ be the vertices of D_i. For each $X_{i_j},\ j = 1,2,3,\ i \in J$ (= an index set of all nodes from $\mathcal{D}_h$) we define a set $Y(X_{i_j}) = \{k \in I;\ X_{i_j}$ is a vertex of $D_k\}$. Then we extrapolate the cell data into the vertices $X_{i_j},\ j = 1,2,3,\ i \in J$ setting

$$\bar{w}(X_{i_j}) = \frac{\sum_{k\in Y(X_{i_j})} w_k|D_k|}{\sum_{k\in Y(X_{i_j})} |D_k|}, \tag{3.11}$$

where $|D_k|$ denotes the area of D_k. Now we construct a linear function $L_i,\ i \in I$ satisfying

$$L_i(X_{i_j}) = \bar{w}(X_{i_j}),\ j = 1,2,3. \tag{3.12}$$

6. **GREEN**. This method uses the Green's theorem in the evaluation of the average value of the gradient of a function to be recovered and was proposed in [30]. The idea of this

method is based on the following approximations:

$$\nabla w|_{D_i} \approx \frac{1}{|D_i|}\int_{D_i} \nabla w \; dx = \frac{1}{|D_i|}\int_{\partial D_i} w \; n \; dx = \frac{1}{|D_i|}\sum_{j\in S(i)}\int_{\Gamma_{ij}} w \; n_{ij} \; dx.$$

For the function w on Γ_{ij} the weighted average is used

$$w|_{\Gamma_{ij}} \approx \frac{w_i|D_j| + w_j|D_i|}{|D_j| + |D_i|}. \tag{3.13}$$

Using the above approximations we obtain

$$\nabla w|_{D_i} \approx \frac{1}{|D_i|}\sum_{j\in S(i)} |\Gamma_{ij}| \frac{w_i|D_j| + w_j|D_i|}{|D_j| + |D_i|} n_{ij}, \tag{3.14}$$

This approximation of the gradient of w is then used in (3.6) as ∇L_i.

7. **MLS.** Using notation from the method DEO–M and Figure 2 we look for a linear function L_i which minimizes the functional

$$S(L_i) = (L_i(P_i) - w_i)^2 + \sum_{j=1}^{3} (L_i(P_{ij}) - w_{ij})^2 . \tag{3.15}$$

This minimization is performed by *the method of the least squares* (see Section 3.3.1).

3.2.2 Slope limiter

To avoid spurious oscillations in the solution we use the Barth–Jesperson slope limiter from [6]. We set

$$\hat{w}_{ij}^{old} = w_i + \nabla L_i \cdot (x_i - x_{ij}, y_i - y_{ij})^T, \quad j \in S(i), \; i \in I, \tag{3.16}$$

and compute

$$\hat{w}_{ij} = w_i + \Phi_i \nabla L_i \cdot (x_i - x_{ij}, y_i - y_{ij})^T, \quad j \in S(i), \; i \in I, \tag{3.17}$$

where

$$\Phi_i = \min_{j\in S(i)} \Phi_{ij},$$
$$\Phi_{ij} = \begin{cases} 1, & \hat{w}_{ij}^{old} - w_i = 0, \\ \min(1, \frac{w_i^{max} - w_i}{\hat{w}_{ij}^{old} - w_i}), & \hat{w}_{ij} - w_i > 0, \\ \min(1, \frac{w_i^{min} - w_i}{\hat{w}_{ij}^{old} - w_i}), & \hat{w}_{ij} - w_i < 0. \end{cases} \tag{3.18}$$

Here $\hat{w}_{ij}^{old}$ is obtained from (3.16) and

$$w_i^{min} = \min(w_i, \min_{j\in S(i)} w_j), \qquad w_i^{max} = \max(w_i, \max_{j\in S(i)} w_j). \tag{3.19}$$

The application of the slope limiter guarantees that the following condition is satisfied

$$w_i^{min} \le \hat{w}_{ij} \le w_i^{max}, \quad j \in S(i), \; i \in I. \tag{3.20}$$

3.3 Determination of the order of accuracy

Let w be the exact solution of a physical problem and w_h its numerical approximation on $\mathcal{D}_h$. We suppose that

$$\|w - w_h\|_{L^p} \approx C \cdot h^\alpha, \tag{3.21}$$

where h denotes the mesh size parameter, C is a positive constant and α is called the *order of accuracy* of a scheme in L^p – norm, $(1 \leq p \leq \infty)$. If we know the analytical solution of the problem, we can compute the approximate solution w_h for different mesh size parameters h and using the method of the least squares to determine α and C.

3.3.1 Method of the least squares

Now we will suppose that the relation (3.21) holds precisely, i.e.

$$\|w - w_h\|_{L^p} = C \cdot h^\alpha. \tag{3.22}$$

We define the *error of the method* as

$$e_h \equiv \|w - w_h\|_{L^p}. \tag{3.23}$$

Our aim is to determine the constant C and the order of accuracy α provided the analytical solution is known. We compute the approximate solution w_h for N different meshes $(N \geq 2)$, i.e. we have $h_i, e_{h_i}, i = 1, \ldots, N$. Now it is possible to use the *method of the least squares.*

From (3.22) and (3.23) we have the relation

$$e_h = C \cdot h^\alpha. \tag{3.24}$$

On (3.24) we apply the operator of logarithm and it leads to the relation

$$\log e_h = \log C + \alpha \cdot \log h. \tag{3.25}$$

If we denote

$$\begin{aligned} y &= \log e_h, \\ x &= \log h, \\ \beta &= \log C. \end{aligned} \tag{3.26}$$

we get in (3.25) the linear dependence of y on x given by the parameters α and β. Due to the relation (3.21) this dependence is linear only in the sense of least squares method which leads us to the determination of α and β in such a way, that the sum

$$S(\alpha, \beta) = \sum_{i=1}^{N} \left(y_i - \beta - \alpha \cdot x_i\right)^2, \tag{3.27}$$

where $y_i = \log e_{h_i}$, $x_i = \log h_i, i = 1, \ldots, N$, is minimal.

The conditions of the minimum of S lead to the relation for α, β

$$\alpha = \frac{N \sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2}, \tag{3.28}$$

$$\beta = \frac{\sum_{i=1}^{N} y_i - \alpha \sum_{i=1}^{N} x_i}{N}. \tag{3.29}$$

Further we define the *correlation coefficient* r

$$r = \frac{N \sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{\sqrt{\left(N \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)\right)\left(N \sum_{i=1}^{N} y_i^2 - \left(\sum_{i=1}^{N} y_i\right)\right)}}. \tag{3.30}$$

We have $r \in [-1, 1]$ and $r \approx 1$ means that the relation

$$y_i \approx \beta + \alpha \cdot x_i, \quad i = 1, \ldots, N \tag{3.31}$$

is valid almost precisely (for $r = 1$ it is valid exactly) and if $r \ll 1$ than this relation is valid inaccurately.

3.3.2 Sod's channel

To determine the order of accuracy of proposed methods we have solved the Sod's channel problem (see [57])

$$\frac{\partial}{\partial t}\begin{pmatrix} \rho \\ \rho v \\ e \end{pmatrix} + \frac{\partial}{\partial x}\begin{pmatrix} \rho v \\ \rho v^2 + p \\ (e+p)v \end{pmatrix} = 0, \qquad x \in \mathbf{R}, \tag{3.32}$$

where

$$p = (\kappa - 1)\left(e - \frac{1}{2}\rho v^2\right), \tag{3.33}$$

with the initial conditions

$$\begin{aligned} v(x,0) &= 0, \\ \rho(x,0) &= \begin{cases} \rho_L, & x < x_0 \\ \rho_R, & x > x_0 \end{cases}, \\ p(x,0) &= \begin{cases} p_L, & x < x_0 \\ p_R, & x > x_0 \end{cases}, \end{aligned} \tag{3.34}$$

where we suppose $p_L > p_R$.

The exact analytical solution can be expressed in the following way:

$$p(x,t) = \begin{cases} p_L \quad , & x < x_0 - c_L t \\ p_L\left[1 - \frac{\kappa-1}{\kappa+1}\left(1 + \frac{x-x_0}{c_L t}\right)\right]^{\frac{2\kappa}{\kappa-1}}, & x_0 - c_L t < x < x_1 \\ p_2, & x_1 < x < x_o + v_c t \\ p_2, & x_0 + v_c t < x < x_0 + v_s t \\ p_R, & x > x_0 + v_s t \end{cases}, \tag{3.35}$$

$$v(x,t) = \begin{cases} 0, & x < x_0 - c_L t \\ \frac{2}{\kappa+1}\left(\frac{x-x_0}{t} + c_l\right), & x_0 - c_L t < x < x_1 \\ v_c, & x_1 < x < x_o + v_c t \\ v_c, & x_0 + v_c t < x < x_0 + v_s t \\ 0, & x > x_0 + v_s t \end{cases}, \tag{3.36}$$

$$\rho(x,t) = \begin{cases} \rho_L, & x < x_0 - c_L t \\ \rho_L \left[1 - \frac{\kappa-1}{\kappa+1}\left(1 + \frac{x-x_0}{c_L t}\right)\right]^{\frac{2}{\kappa-1}}, & x_0 - c_L t < x < x_1 \\ \rho_L \left(\frac{p_2}{p_L}\right)^{\frac{1}{\kappa}}, & x_1 < x < x_o + v_c t \\ \rho_R \frac{1+\frac{\kappa+1}{\kappa-1}\frac{p_2}{p_R}}{\frac{\kappa+1}{\kappa-1}+\frac{p_2}{p_R}}, & x_0 + v_c t < x < x_0 + v_s t \\ \rho_R, & x > x_0 + v_s t \end{cases} . \tag{3.37}$$

Here

$$x_1 = x_0 + \left(\frac{\kappa+1}{2} v_c - c_L\right) t, \tag{3.38}$$

for the speed of the *shock wave* v_s we have

$$v_s = c_R^2 \frac{\frac{p_2}{p_R} - 1}{\kappa v_c} \tag{3.39}$$

and for the speed of *contact discontinuity* it holds

$$v_c = c_R \sqrt{\frac{2}{\kappa(\kappa-1)}} \frac{\frac{p_2}{p_R} - 1}{\sqrt{1 + \frac{\kappa+1}{\kappa-1}\frac{p_2}{p_R}}}. \tag{3.40}$$

Moreover the pressure p_2 has to be determined. For p_2 one can find the following equation

$$\sqrt{\frac{2}{\kappa(\kappa-1)}} \frac{\frac{p_2}{p_R} - 1}{\sqrt{1 + \frac{\kappa+1}{\kappa-1}\frac{p_2}{p_R}}} = \frac{2}{\kappa-1}\frac{c_L}{c_R}\left[1 - \left(\frac{p_2}{p_L}\right)^{\frac{\kappa-1}{2\kappa}}\right], \tag{3.41}$$

which can be solved numerically. The local speed of sound c_L and c_R are obtained from the initial conditions.

The analytical solution of the problem (3.32) – (3.34) is demonstrated for the density in Figure 3. The contact discontinuity (between regions 3 and 2) and the shock wave (between regions 2 and R) propagate in the positive direction of the axis x from the point $(x_0, 0)$ with speed v_c and v_s, respectively. The expansion region (region 4) increases in the negative direction of the axis x, its edges propagate with the speed c_L and $(\kappa+1)v_c/2 - c_L$, respectively. In regions L and R, there are constant values given by initial conditions. The Figure 4 shows 3-D plot of the density in the (x, t) plane.

3.4 Numerical results

To determine the order of accuracy of our schemes we solved the Sod's channel problem in the domain $\Omega = [-3 : 3] \times [0 : 0.1]$ using the 2D code described in [20] for the numerical solution of the Euler equations (3.1) equipped by the initial conditions (3.34), where

$$\begin{aligned} \rho_L &= 1.00, & p_L &= 2.00, \\ \rho_R &= 1.101463, & p_R &= 1.00. \end{aligned} \tag{3.42}$$

The analytical and approximate solution are computed at time $t = 1$. We have used three triangulations containing 75×4, 150×8 and 300×16 nodes, respectively, computed the error $e_h = \|\rho - \rho_h\|_{L^1(\Sigma)}$ and using the method of the least squares we determined the order of accuracy α and the constant C. For the domain Σ where we investigate the order of accuracy, two cases were distinguished

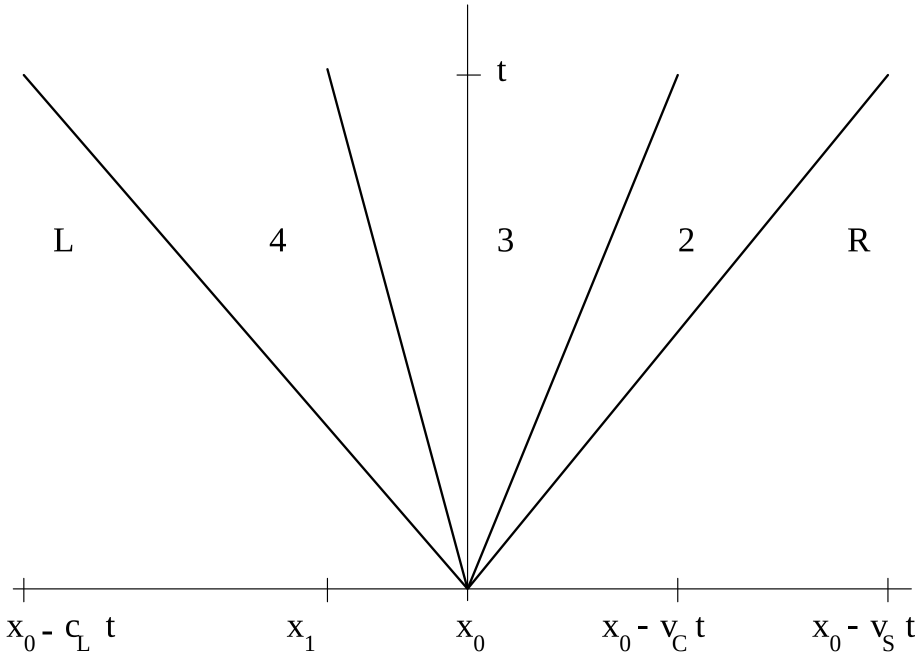


Fig. 3: Propagation of discontinuities for Sod's channel

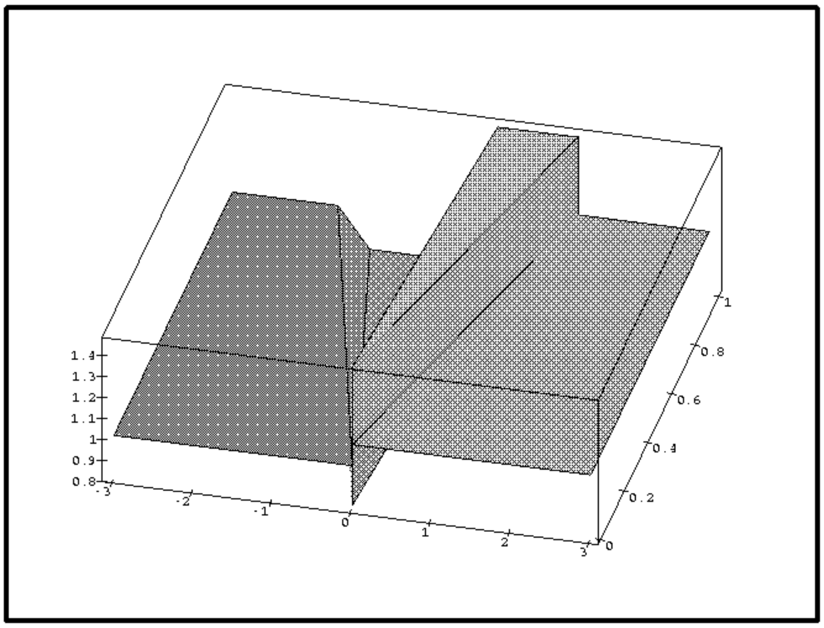


Fig. 4: Density distribution in (x, t) plane

1. **whole** – we put $\Sigma = \Omega$
2. **smooth** – we put $\Sigma = \Omega - \sigma$, where $\sigma \subset \Omega$ is a "narrow corridor" along discontinuities, i.e. Σ is a subdomain of Ω where the solution u is smooth.

The results are in the following table (α is the order of accuracy, C the constant from (3.22) and r the correlation coefficient defined by (3.30).

			whole			smooth	
num.	method	α	C	r	α	C	r
0.	BASIC	0.627	0.065	0.999	1.062	0.106	0.999
1.	DEO–M	0.210	0.021	0.981	0.074	0.009	0.546
2.	DEO–B	0.203	0.016	0.630	0.010	0.004	0.019
3.	MIN	0.848	0.063	0.991	1.399	0.094	0.987
4.	AVER	- 0.222	0.008	-0.757	- 0.298	0.005	-0.794
5.	EXTRAP	-0.032	0.009	-0.520	-0.375	0.002	-0.915
6.	GREEN	0.599	0.061	0.999	1.022	0.097	0.999
7.	MLS	0.594	0.044	0.998	1.270	0.113	0.999

Tab. 1: Order of accuracy of investigated methods

The dependency of the error e_h on the mesh size parameter h for the latter case is depicted on the Figure 5, where log-log scale is used.

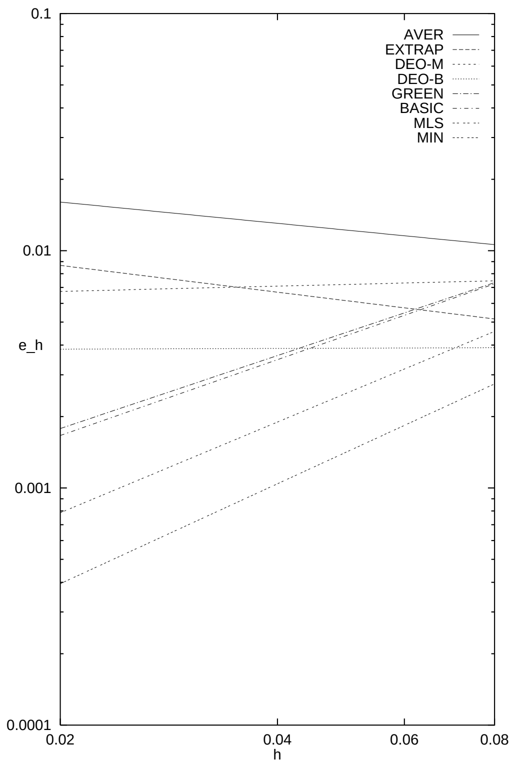


Fig. 5: Behaviour of the error e_h in dependence on h

We see from the Figure 5 that only 3 methods give reasonable results: MIN, GREEN and MLS. The method GREEN gives results corresponding to the first order and only methods MIN and MLS bring some increasing of order of accuracy. The best one seems to be MIN.

On the other hand the scheme is not a second order one. Figures 6 shows the comparison of analytical solution (dashed line), numerical solution obtained by the basic scheme (pointed line) and numerical solution obtained by MIN (full line) for three different meshes.

Acknowledgement

This work was supported under the grant No. 201/96/0313 of the Czech Grant Agency.

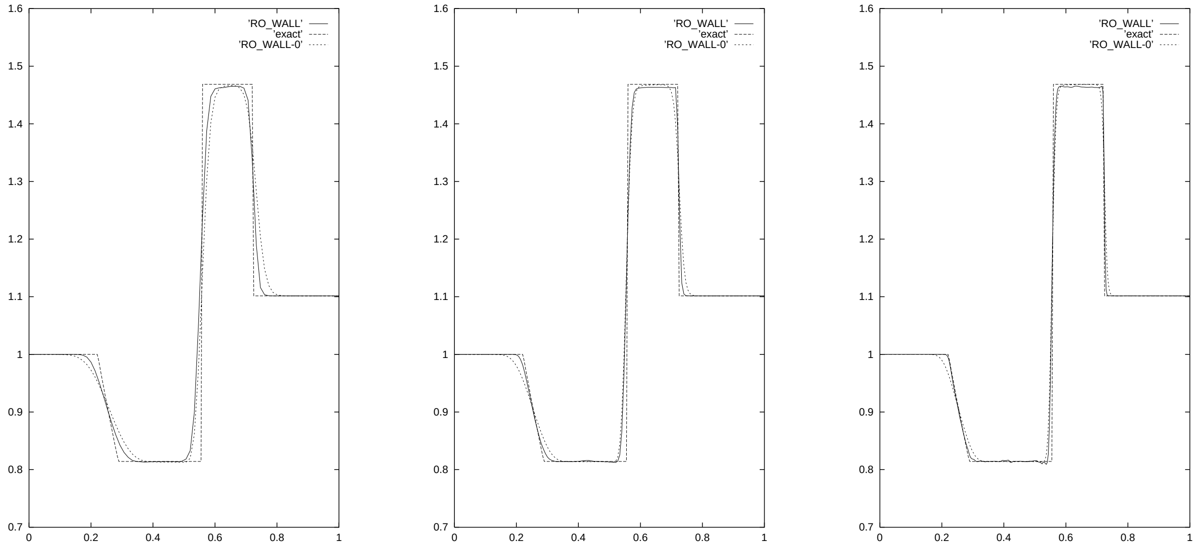


Fig. 6: Distribution of density on mesh with 75×4, 150×8 and 300×16 nodes, respectively

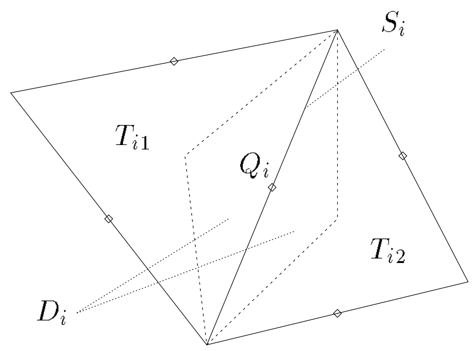


FIG. 3.1: Le reconstruction d'ordre plus élevé pour les volumes finis barycentriques.

3.5 Schéma d'ordre plus élevé pour des volumes finis barycentriques

On peut utiliser pour les volumes finis barycentric une approche très similaire à celle utilisés pour les volumes finis triangulaires. Le schéma donné par les relations (3.5) et (3.6) est utilisé. La reconstruction linéaire est effectuée sous la forme suivante. Soit $D_i \in \mathcal{D}_h$ un volume fini barycentrique. Alors il existe deux triangles T_{i1}, T_{i2} de la triangulation fondamentale, tels que $D_i \subset T_{i1} \cup T_{i2}$, $T_{i1} \cap T_{i2} = S_i$, $Q_i \in S_i$. Voir Fig. 3.1

On identifie les valeurs de la fonction w sur D_i à des valeurs de w dans $Q_i \in S_i$ avec les notation de Chap. 2. Puis, pour chacun triangle T_k, $k \in I$ existe une fonction linéaire L_k pour laquelle $L_k(Q_{kj}) = w(Q_{kj})$, $j = 1, 2, 3$, où Q_{k1}, Q_{k2}, Q_{k3} sont les centres des côtés du T_k. Ensuite pour D_i, $i \in \mathcal{D}_h$, on a deux candidats pour la reconstruction linéaire L_i, les fonctions $\tilde{L}_{i1}, \tilde{L}_{i2}$ correspondantes aux triangles T_{i1}, T_{i2}, respectivement. Pour les volumes finis frontières il y a seulement un candidat $\tilde{L}_{i1}$ et on pose $\tilde{L}_{i2} = \tilde{L}_{i1}$. Nous avons testé les quatre possibilités suivantes

1. *Le gradient minimal (MIN)*: On choisit la fonction de $\tilde{L}_{i1}, \tilde{L}_{i2}$, qui a la plus petite valeur du gradient, c. à. d. **if** $|\nabla \tilde{L}_{i1}| < |\nabla \tilde{L}_{i2}|$ **then** $L_i := \tilde{L}_{i1}$ **else** $L_i := \tilde{L}_{i2}$.

2. *Le gradient maximal (MAX)*: On choisit la fonction de $\tilde{L}_{i1}, \tilde{L}_{i2}$, qui a la plus grand valeur du gradient, c. à. d. **if** $|\nabla \tilde{L}_{i1}| < |\nabla \tilde{L}_{i2}|$ **then** $L_i := \tilde{L}_{i2}$ **else** $L_i := \tilde{L}_{i1}$.

3. *Le gradient moyen (AVER)*: Pour la reconstruction linéaire on utilise le gradient moyen des fonctions $\tilde{L}_{i1}, \tilde{L}_{i2}$, c. à. d. $L_i := (\tilde{L}_{i1}|T_{i1}| + \tilde{L}_{i2}|T_{i2}|)/(|T_{i1}| + |T_{i2}|)$, où $|T_{i1}|, |T_{i2}|$ sont les surfaces des triangles T_{i1}, T_{i2}.

4. *Le théorème de Green (GREEN)*: Ici procéde de la même manière que pour le maillage triangulaire.

Pour obtenir des meilleur résultats il faut appliquer le limiter de Barth–Jesperson (voir la Sec. 3.2.2). On a aussi testé ces differentes méthodes sur le tube de Sod et on a déterminé les ordres de précision suivants:

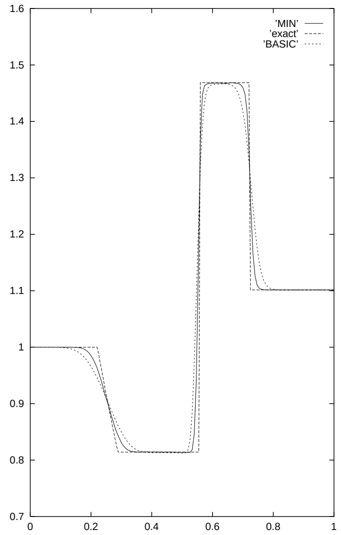


FIG. 3.2: La distribution de la densité pour le schéma fondamental (BASIC) et le schéma d'ordre plus élevé (MIN) en comparaison avec la solution analytique (exact).

			whole			smooth	
num.	méthode	α	C	r	α	C	r
0.	de base.	0.588	0.081	1.000	0.920	0.121	0.999
1.	MIN	0.771	0.066	1.000	1.096	0.069	0.996
2.	MAX	0.803	0.081	0.999	0.951	0.063	0.999
3.	AVER	0.795	0.070	1.000	0.963	0.052	0.999
4.	GREEN	0.587	0.080	1.000	0.919	0.120	0.999

Les meilleurs résultats ont été obtenus par la méthode MIN, mais en comparaison avec les schémas sur des maillages triangulaires les ordres de précision sont plus petits. Fig. 3.2 montre la distribution de la densité pour le schéma fondamental et le schéma d'ordres plus élevé.

Chapitre 4

Exemples numériques

Dans ce chapitre nous présentons plusieurs résultats numériques pour l'écoulement visqueux et non visqueux. Des calculs ont été exécutés par le flux numérique Osher – Solomon. Des maillages utilisés pour les calculs ont été construits par le générateur du maillage anisotropique (Chap. 6).

On obtient une solution stationnaire par la méthode de la stabilisation en temps pour $t \to \infty$. En commençant par la condition initiale on utilise le processus itératif en temps. La solution approchée est considérée stationnaire si le résidu est petit:

$$\|\boldsymbol{w}_h^{k+1} - \boldsymbol{w}_h^k\| / \tau_h < \varepsilon, \tag{4.1}$$

où ε est une tolérance donnée. Dans nos expériences on utilise une condition plus faible que (4.1). L'erreur relative à la densité, donnée par

$$e^k := \left\| \frac{\rho_h^{k+1} - \rho_h^k}{\rho_h^k} \right\|_{L^1(\Omega_h)} = \sum_{j \in J} |D_j| \left| \frac{\rho_h^{k+1} - \rho_h^k}{\rho_h^k} \right|, \tag{4.2}$$

a été utilisée. Dans tout le chapitre le nombre de Poisson $\gamma = 1.4$ et le nombre de Prandtl $Pr = 0.72$ sont considérés.

Ici on présente les résultats obtenus par les méthodes présentées dans le Chap. 2. Pour le calcul d'écoulement non visqueux, la méthode des volumes finis a été utilisée. On a testé des volumes finis triangulaires et des volumes finis barycentriques. Les méthodes combinant volumes finis – éléments finis ont été utilisées pour la calculation d'écoulement visqueux. Aussi les schémas d'ordre plus élevés ont été testés. Le schéma MIN a été utilisé pour les volumes finis triangulaires et aussi barycentriques.

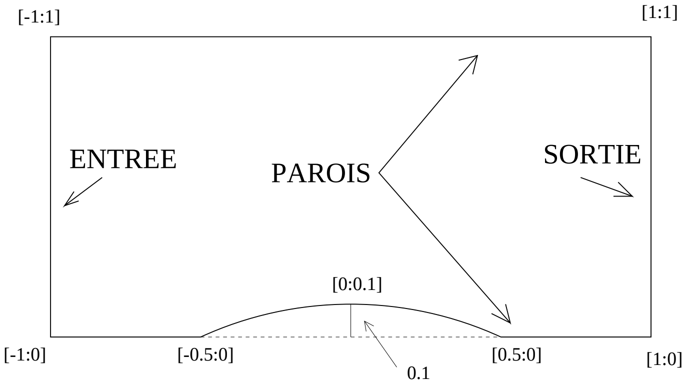


FIG. 4.1: GAMM – tube

4.1 GAMM – tube

Cet exemple est très utilisé pour le calcul non visqueux transonique. Le GAMM – tube est une tube de dimensions 2×1 m avec 10% d'arc circulaire (voir Fig. 4.1). Les conditions initiales sont les suivantes:

$$\rho_0 = 1, \qquad \boldsymbol{u}_0 = (1, 0), \qquad p_0 = 1.591191166. \tag{4.3}$$

Les mêmes données sont utilisées pour les conditions frontières. Les valeurs (4.3) donnent le nombre de Mach entrée

$$M_{ent} := \frac{|\boldsymbol{u}|}{c_0} = 0.67, \tag{4.4}$$

où c_0 est la vitesse du son local sur l'entrée.

Pour le calcul la méthode de volumes finis triangulaires et barycentriques a été utilisée (voir la Sec. 2.1). Ici on présente les résultats obtenus pour le schéma de base et pour le schéma d'ordre plus élevé (voir Chap. 3). Pour chacun de ces cas on présente le maillage du calcul, la distribution de nombre de Mach M sur les parois imperméables et les lignes du niveau du nombre de Mach. Les figures sont classées de la manière suivante:

volumes finis	schéma	figures	numéro d'éléments
triangulaires	de base	Fig. 4.2	2 600
triangulaires	plus élevé	Fig. 4.3	2 780
barycentriques	de base	Fig. 4.4	3 600
barycentriques	plus élevé	Fig. 4.5	3 924

Fig. 4.6 montre la distribution de nombre de Mach sur les parois imperméables pour tous les quatre cas. Ici l'onde de choc est très bien capturée et aussi le minimum local (la singularité de Zierep) pour $x \approx 0.6$ est évident. Le capture de cette singularité est bon indicateur de la qualité de la méthode. On peut facilement voir que la qualité des résultats est bonne pour toutes les méthodes. Les schémas d'ordre plus élevé donnent des résultats un peu meilleurs que les schémas fondamentaux.

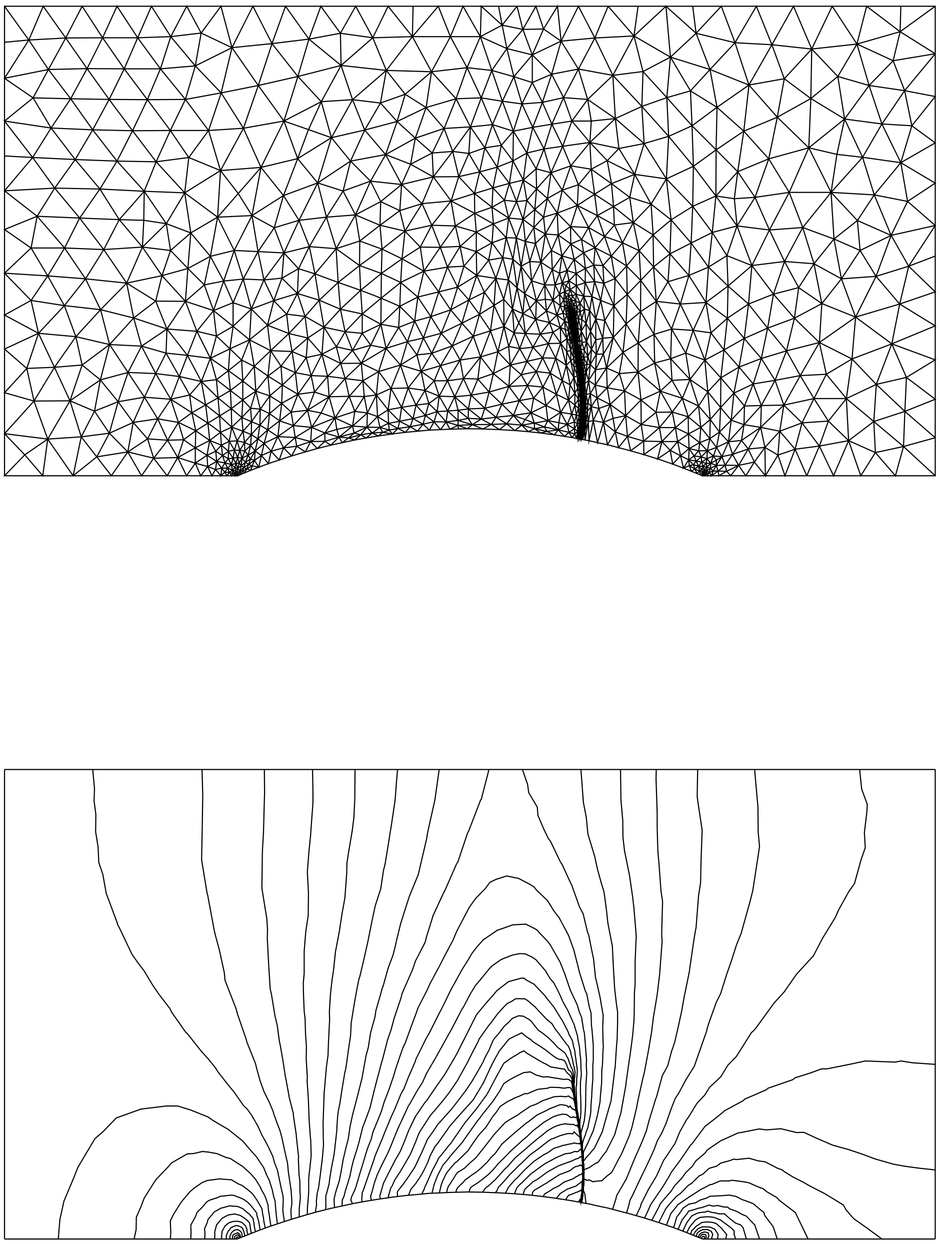
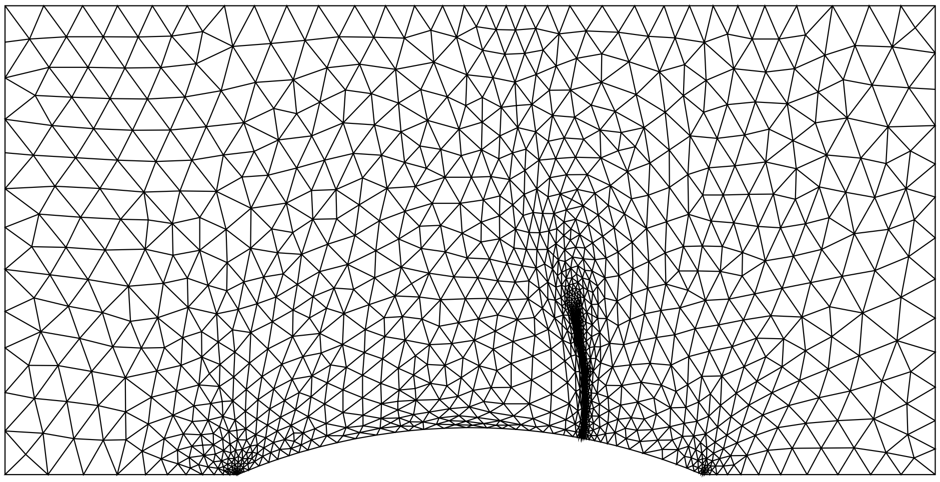


Fig. 4.2: Volumes finis triangulaires, schéma de base: Maillage (en haut) et lignes du niveau du nombre de Mach (en bas).



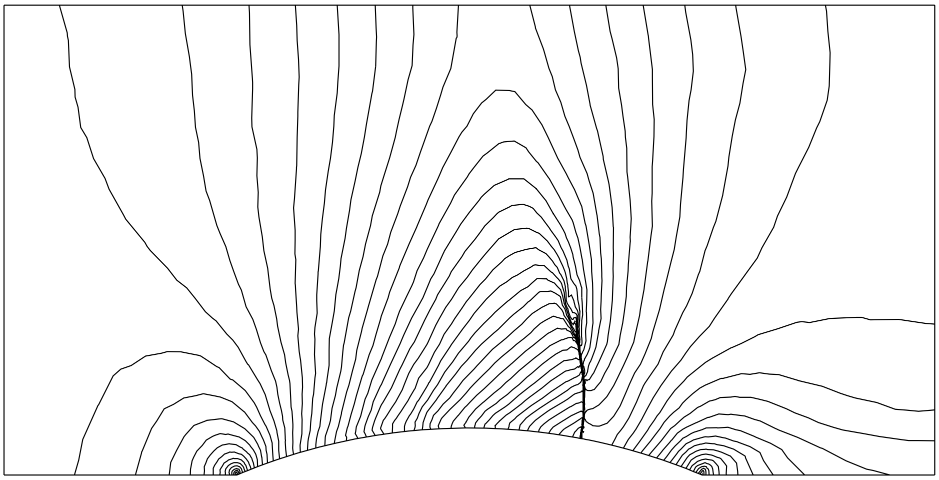


FIG. 4.3: Volumes finis triangulaires, schéma d'ordre plus élevé: Maillage (en haut) et lignes du niveau du nombre de Mach (en bas).

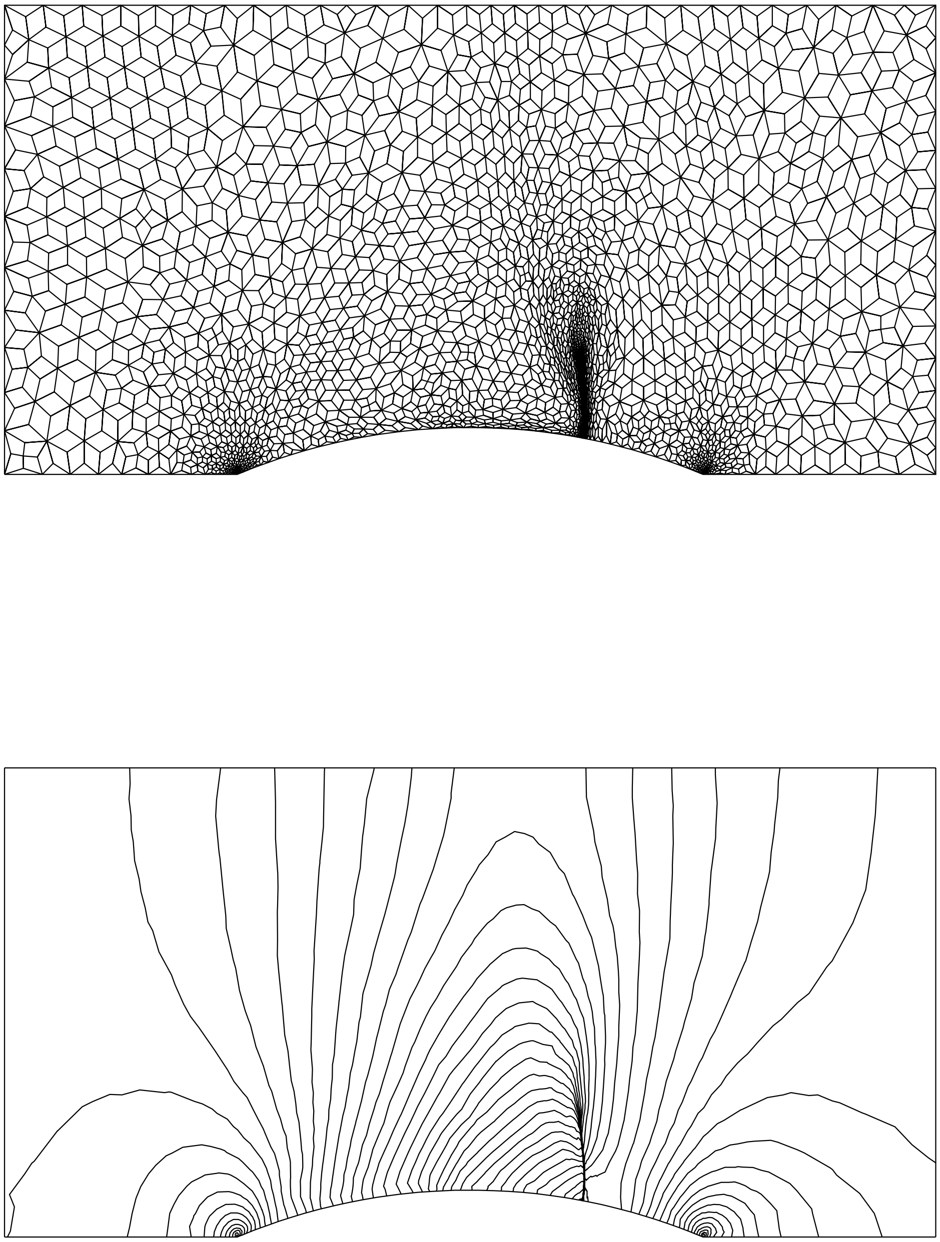


FIG. 4.4: Volumes finis barycentriques, schéma de base: Maillage (en haut) et lignes du niveau du nombre de Mach (en bas).

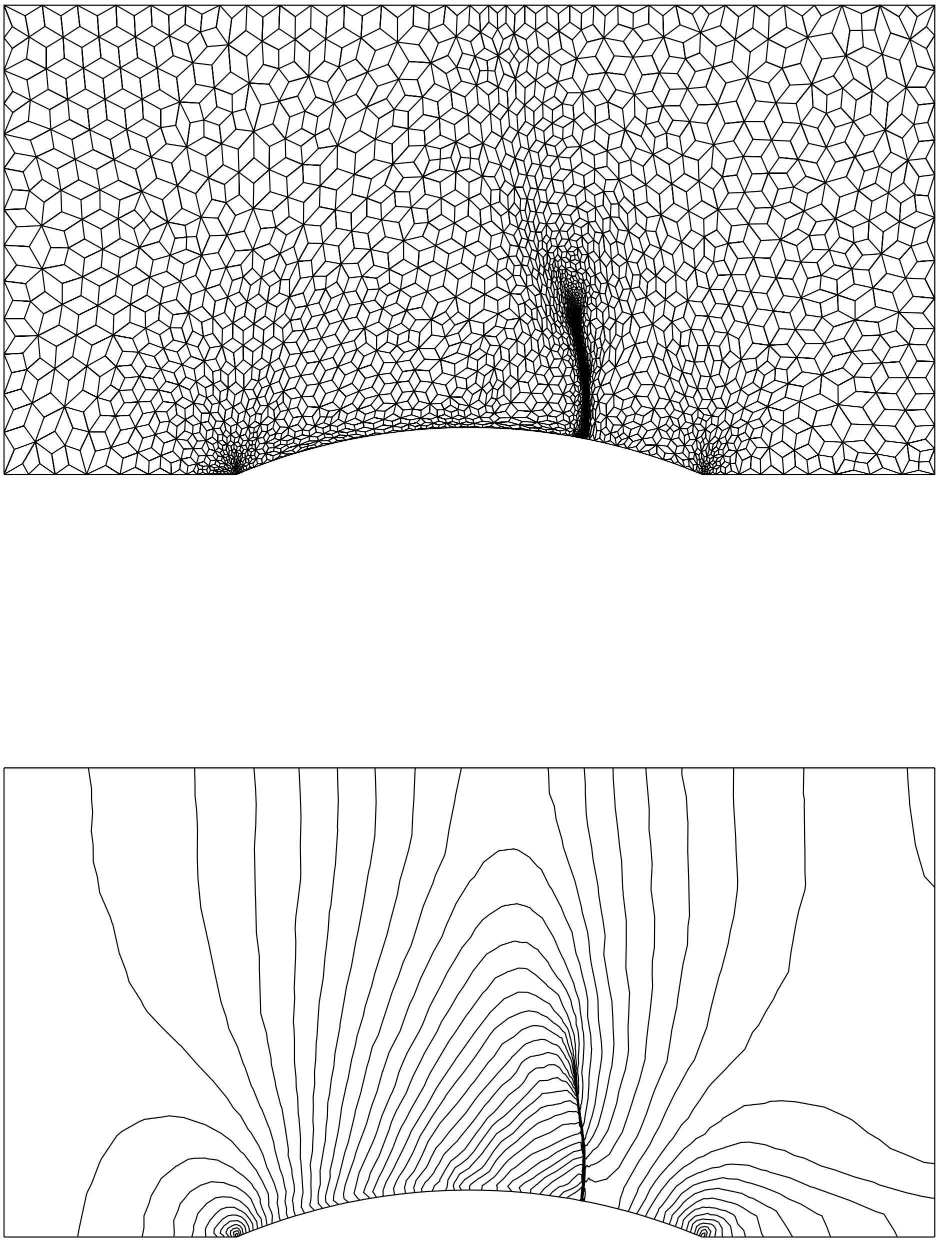


FIG. 4.5: Volumes finis barycentriques, schéma d'ordre plus élevé: Maillage (en haut) et lignes du niveau du nombre de Mach (en bas).

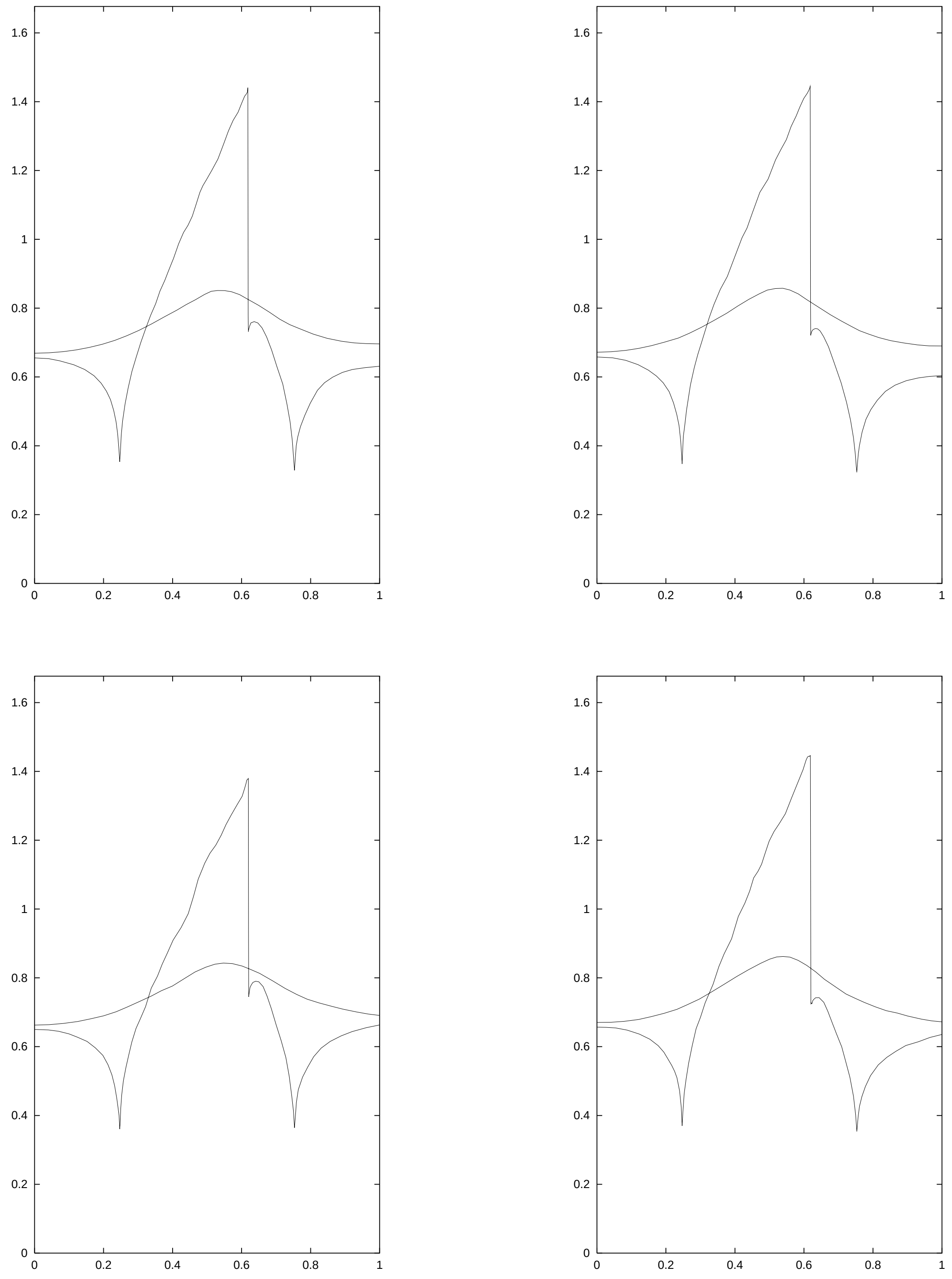


FIG. 4.6: Distribution du nombre de Mach sur parois: volumes finis triangulaires (en haut), volumes finis barycentriques (en bas), schémas de base (à gauche) et schémas d'ordre plus élevé (à droite).

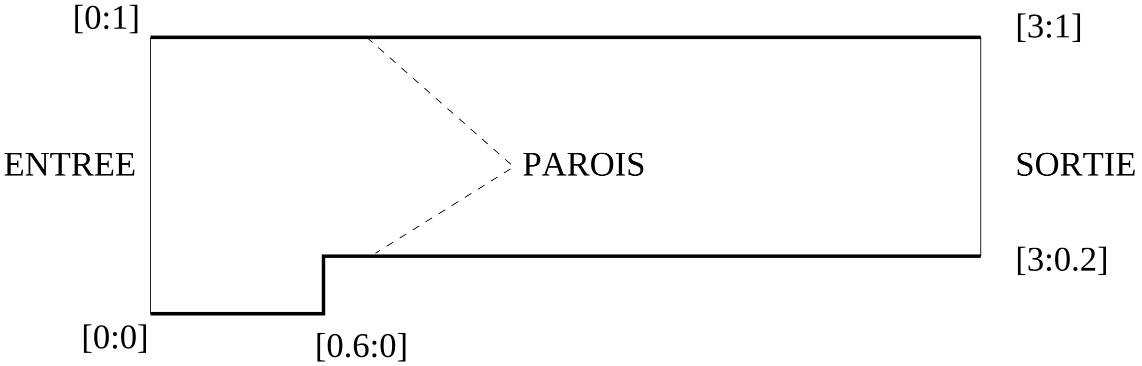


FIG. 4.7: La marche en avant

4.2 Le problème de la marche en avant

Ce problème est souvent utilisé pour un contrôle des méthodes pour l'écoulement non visqueux non stationnaire. Le domaine du calcul est très simple (voir Fig. 4.7). Les conditions initiales sont

$$\rho_0 = 1.4, \qquad \boldsymbol{u}_0 = (3, 0), \qquad p_0 = 1, \tag{4.5}$$

qui sont utilisées aussi comme les conditions frontières. Nous nous intéressons à la solution au temps $T = 2.5$. Fig. 4.8 montre un maillage et les lignes du niveau du nombre de Mach. Le maillage a été obtenu par l'algorithme AMA (Chap. 6) et les schémas de base ont été utilisés. On observe plusieurs ondes de chocs et aussi la discontinuité du contact. Les résultats sont en bon accord avec [67].

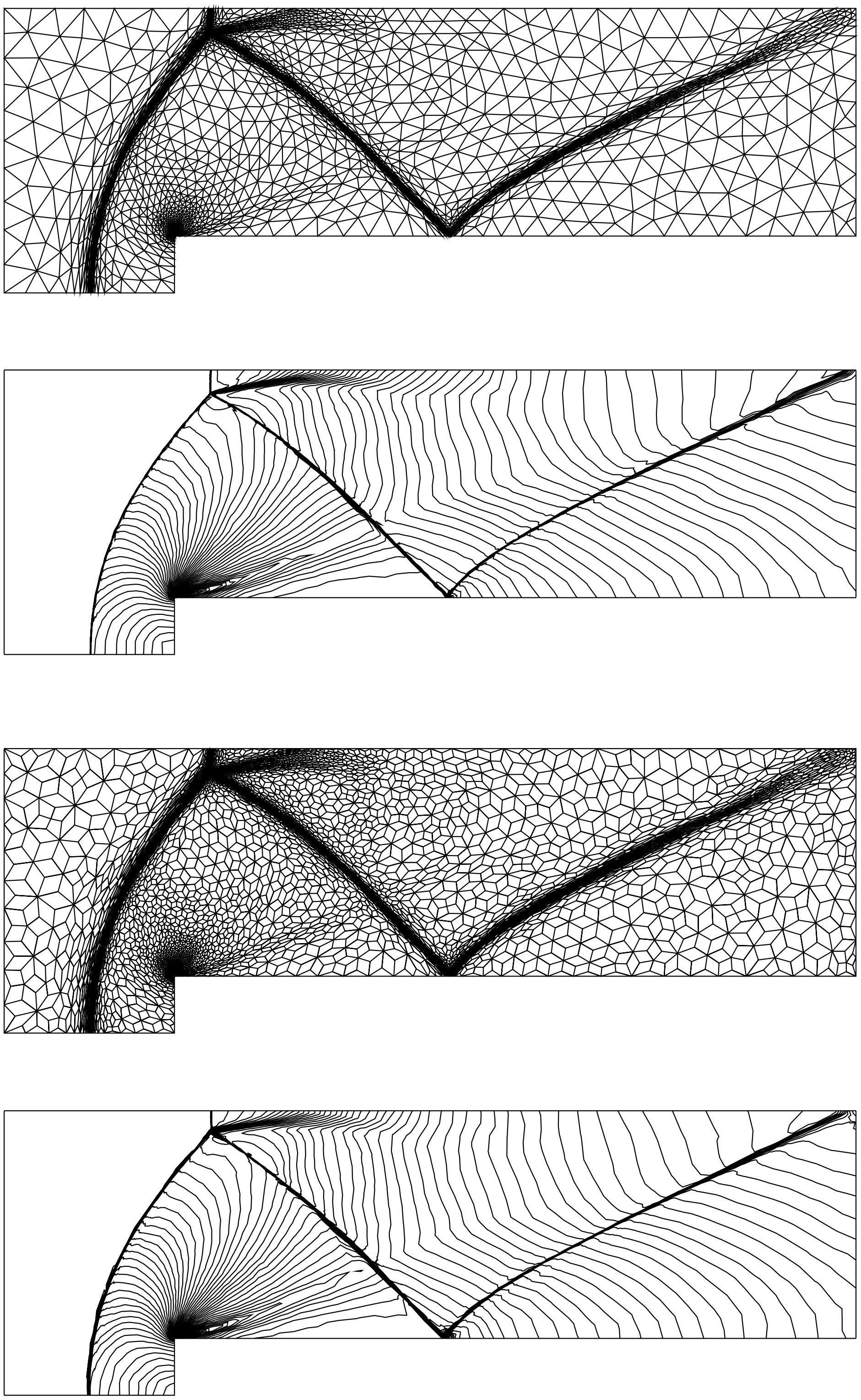


FIG. 4.8: L'escalier en avant: le maillage et les lignes du niveau du nombre de Mach. Volumes finis triangulaires (en haut) et volumes finis barycentriques (en bas).

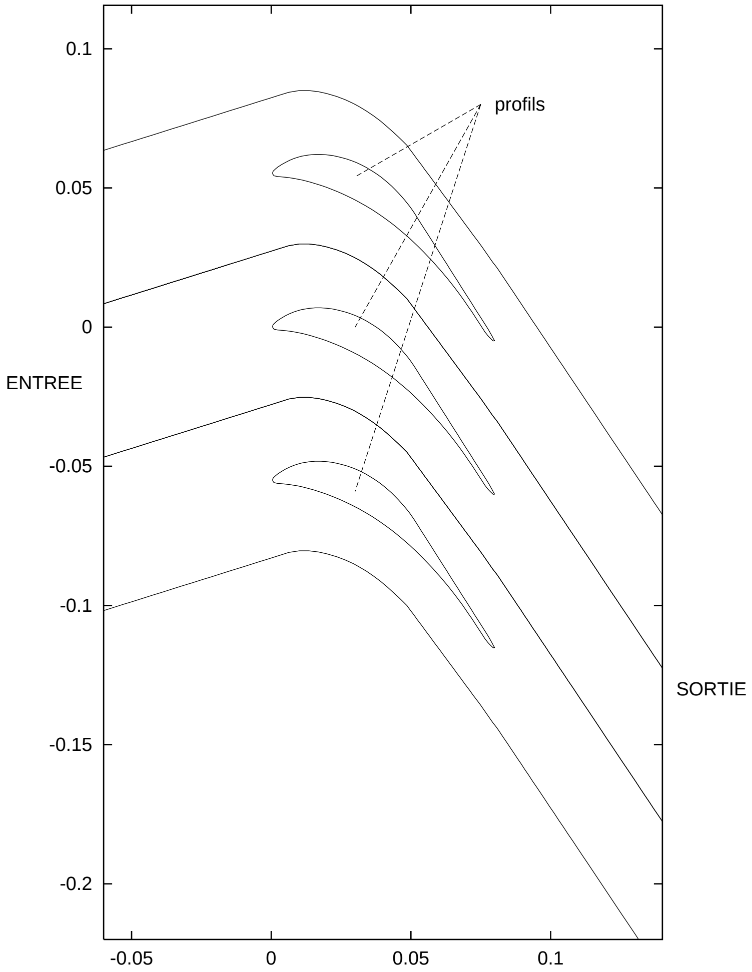


FIG. 4.9: Cascade de profils.

4.3 Cascade de profils

Le troisième exemple, qui est ici présenté, est l'écoulement autour d'une cascade de profils (voir Fig. 4.9). Ce problème simule l'écoulement dans une turbine à vapeur. Pour ce cas on a aussi les résultats expérimentaux, qui ont été réalisés dans l'Institut de Thermomecanique de l'Académie de Sciences de la République Tchèque (voir [59]). On a essayé d'obtenir les résultats numériques en bonne correspondance avec le mesurs. Fig. 4.10 montre un interférogramme avec des lignes de niveau de densité, où quelques ondes du chocs apparaissent.

Les datas expérimentaux et les mêmes pour le calcul sont:

$M_{ent} = 0.32$	nombre de Mach entrée
$\alpha = 19°18'$	l'angle d'incidence
$Re = 1.5 \cdot 10^6$	nombre de Reynolds

Pour ce problème nous avons calculé l'écoulement non visqueux (les équations d'Euler) et aussi visqueux (les équations de Navier–Stokes) et les résultats ont été comparés avec les mesures expérimentales.

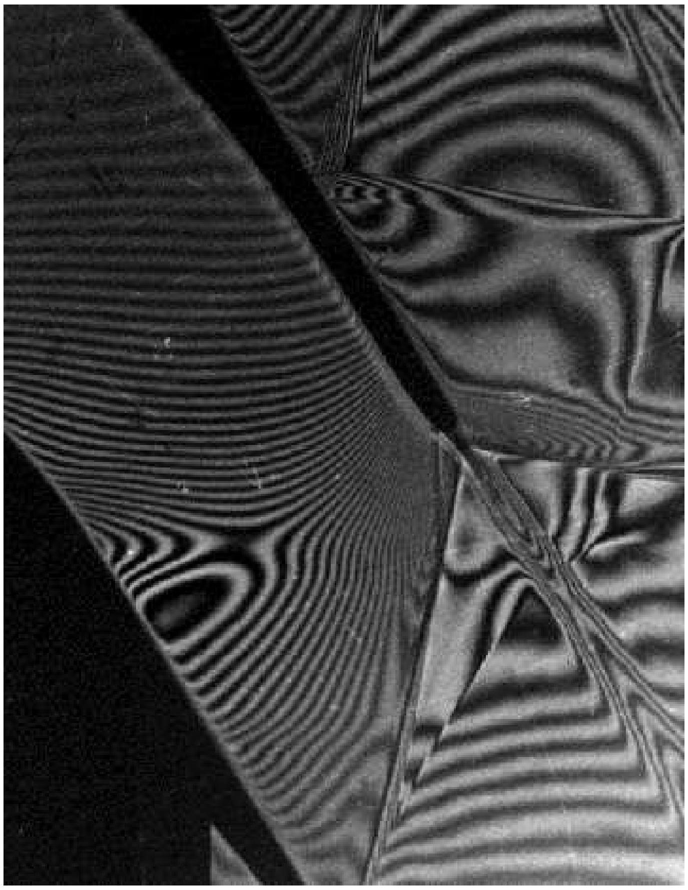


FIG. 4.10: Cascade du profils: l'interférogramme avec des lignes de niveau de la densité.

4.3.1 L'écoulement non visqueux

Pour la solution des équations d'Euler on a testé les volumes finis triangulaires et volumes finis barycentriques, les schémas de base et d'ordre plus élevé. On présente ici le maillage, a distribution de la pression en comparaison avec les datas expérimentaux, les lignes du niveau du nombre de Mach et de la densité. Comme les effets des termes diffusifs dans le cas des cascades de profils sont très petits, on obtient une bonne correspondance avec les mesures même dans le cas d'un calcul non visqueux. L'utilisation des schémas d'ordre plus élevé apporte une amélioration de la solution. Les figures sont classées de la manière suivante:

volumes finis	schéma	figures	numéro d'éléments
triangulaires	de base	Fig. 4.11	10 051
triangulaires	plus élevé	Fig. 4.12	12 248
barycentriques	de base	Fig. 4.13	20 083
barycentriques	plus élevé	Fig. 4.14	23 840

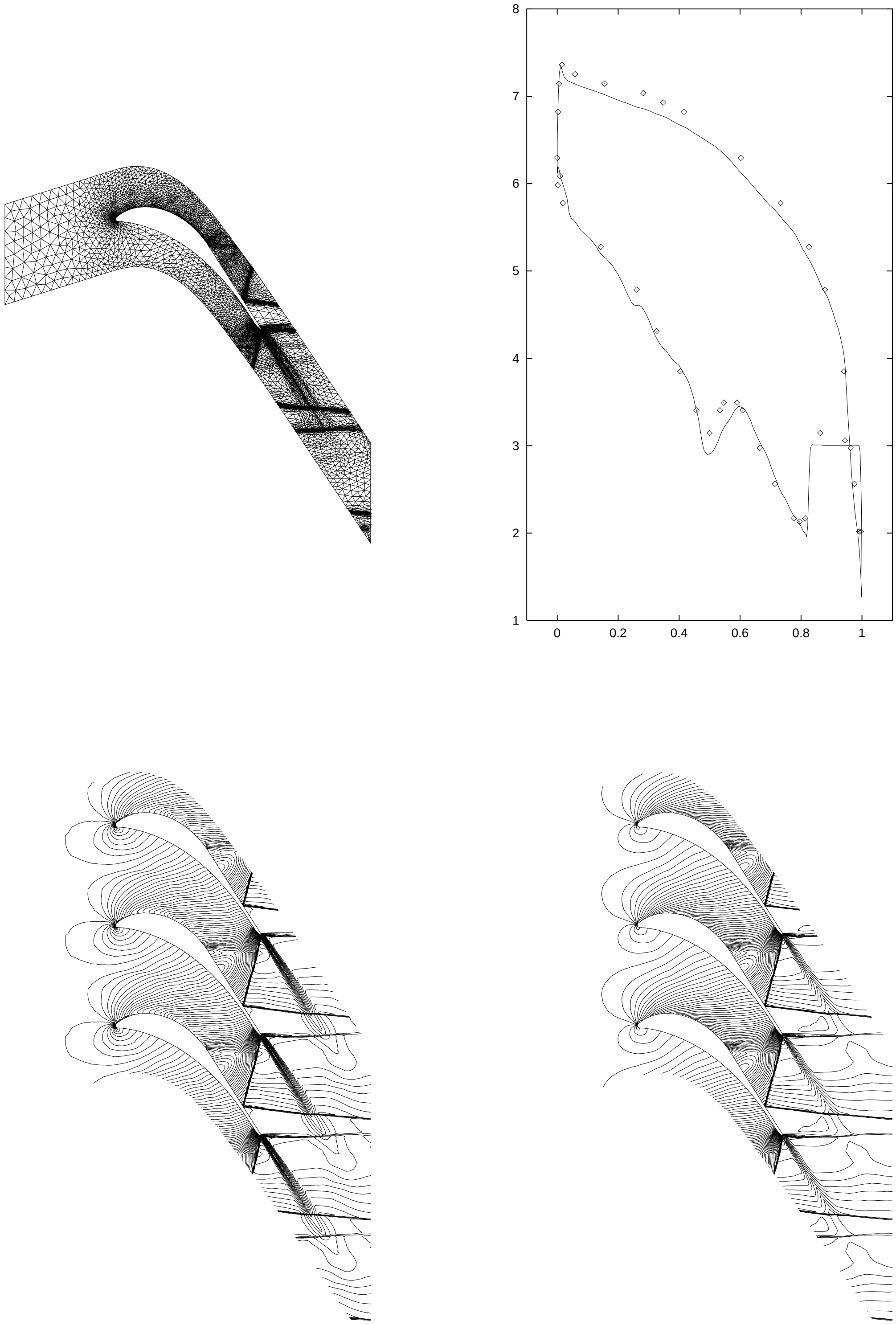


FIG. 4.11: Cas non visqueux, volumes finis triangulaires, schéma de base: maillage (en haut à gauche), distribution de la pression en comparaison avec les mesures (en haut à droite), les lignes de niveau du nombre de Mach(en bas à gauche) et les lignes de niveau de la densité (en bas à droite).

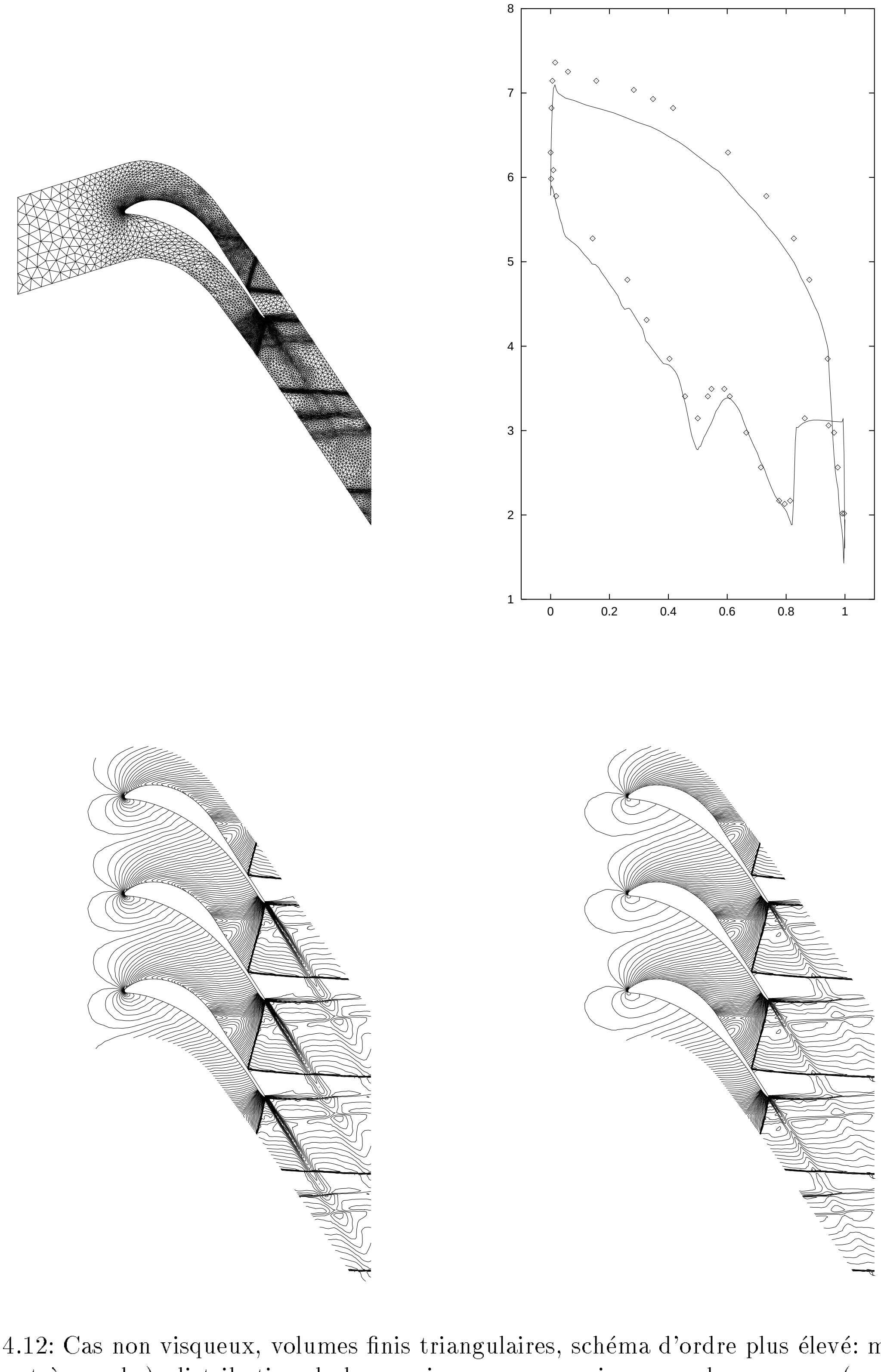


FIG. 4.12: Cas non visqueux, volumes finis triangulaires, schéma d'ordre plus élevé: maillage (en haut à gauche), distribution de la pression en comparaison avec les mesures (en haut à droite),les lignes de niveau du nombre de Mach(en bas à gauche) et les lignes de niveau de la densité (en bas à droite).

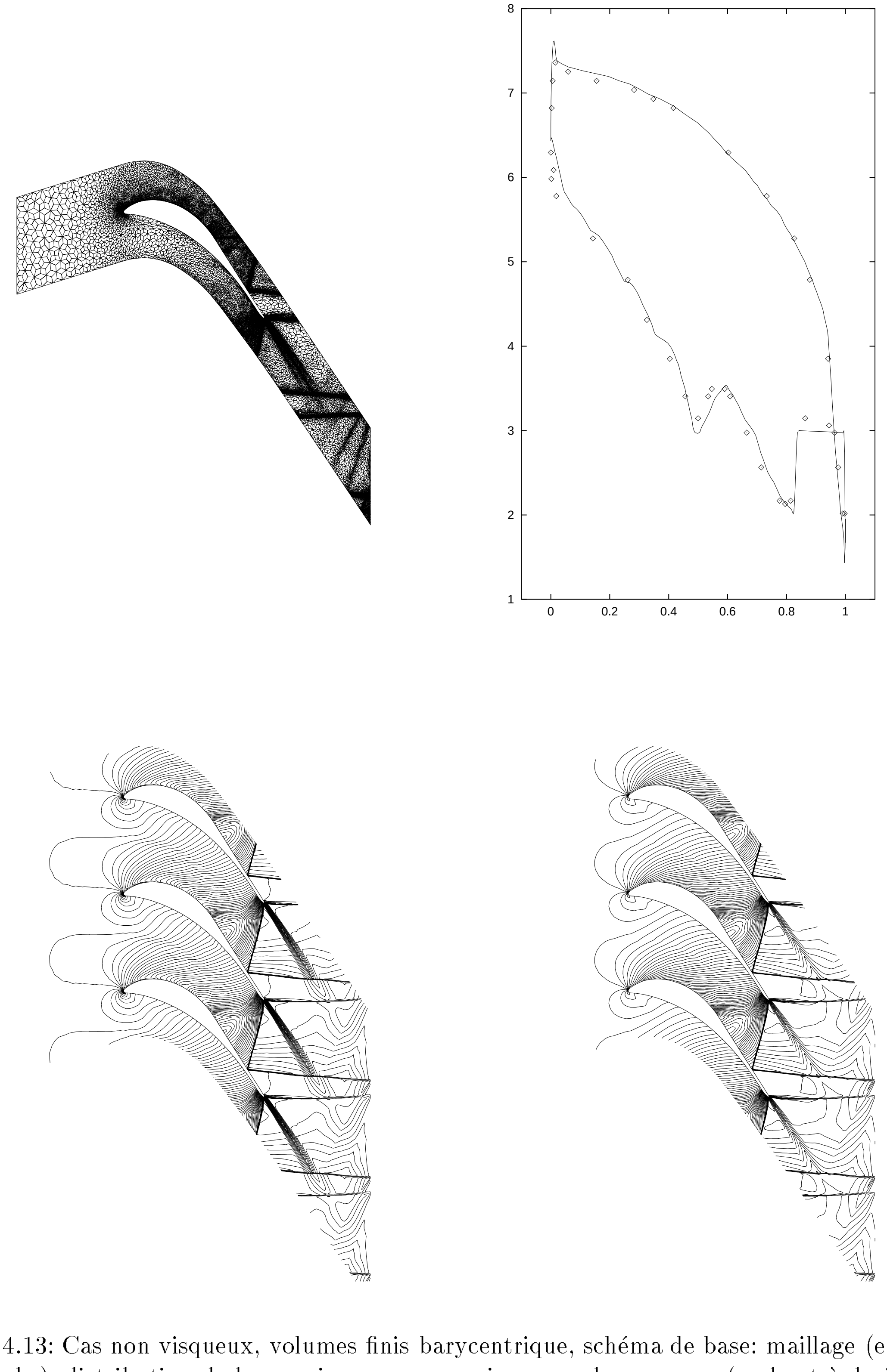


FIG. 4.13: Cas non visqueux, volumes finis barycentrique, schéma de base: maillage (en haut à gauche), distribution de la pression en comparaison avec les mesures (en haut à droite),les lignes de niveau du nombre de Mach(en bas à gauche) et les lignes de niveau de la densité (en bas à droite).

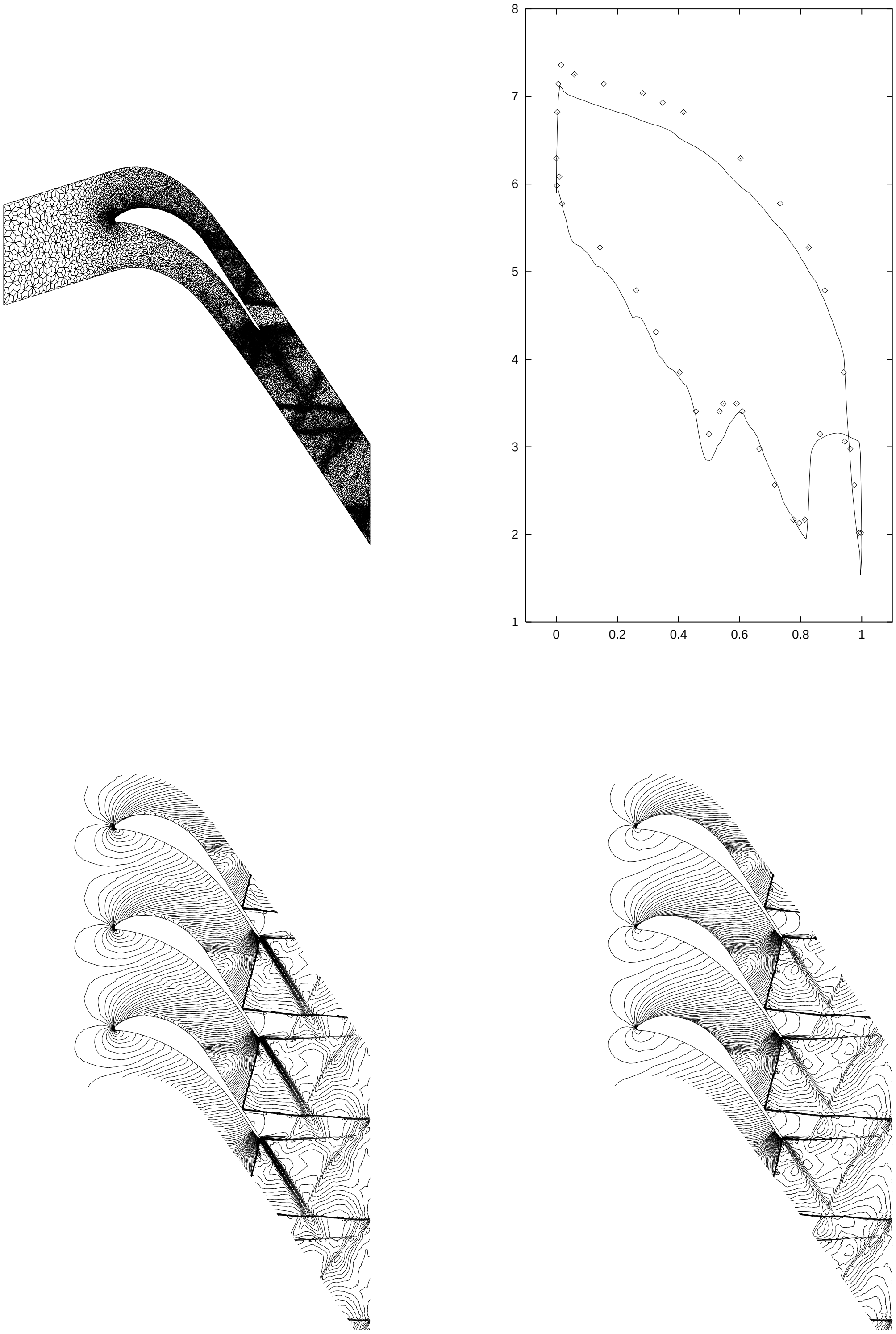


Fig. 4.14: Cas non visqueux, volumes finis barycentrique, schéma d'ordre plus élevé: maillage (en haut à gauche), distribution de la pression en comparaison avec les mesures (en haut à droite), les lignes de niveau du nombre de Mach(en bas à gauche) et les lignes de niveau de la densité (en bas à droite).

4.3.2 L'écoulement visqueux

Le calcul des équations de Navier–Stokes a été exécuté par les deux méthodes combinant des volumes finis et des éléments finis (Sec. 2.2.2):

– la première: volumes finis triangulaire et volumes finis conformes

– la deuxième: volumes finis barycentrique et volumes finis conformes

Pour chacune de ces méthodes le schéma de base et aussi la schéma d'ordre plus élevé a été utilisé. On présente ici le maillage, la distribution de la pression en comparaison avec les datas expérimentaux, les lignes du niveau du nombre de Mach et de la densité.

Les figures sont classées de la manière suivante:

méthode	schéma	figures	nombre d'éléments	
			VF	EF
première	de base	Fig. 4.15	32 721	66 169
première	plus élevé	Fig. 4.16	32 721	66 169
deuxième	de base	Fig. 4.17	49 445	32 721
deuxième	plus élevé	Fig. 4.18	49 445	32 721

Pour la discrétisation des quatre cas, la même triangulation (Fig. 4.15 à gauche haut) a été utilisée, c. à. d. pour la première méthode des termes convectifs le maillage fondamental triangulaire (Fig. 4.15 en haut à gauche 32 721 éléments) et pour les termes diffusifs le maillage correspondant adjoint triangulaire (Fig. 4.16 en haut à gauche 66 169 éléments) et pour la deuxième méthode pour les termes diffusifs le maillage fondamental triangulaire (Fig. 4.15 en haut à gauche) et pour les termes convectifs le maillage correspondant barycentrique (Fig. 4.17 en haut à gauche 49 445 éléments).

Ici les résultats obtenus par la première méthode sont un peu meilleur que pour la deuxième. Au contraire l'utilisation des schémas d'ordre plus élevé n'apporte par d'amélioration des résultats.

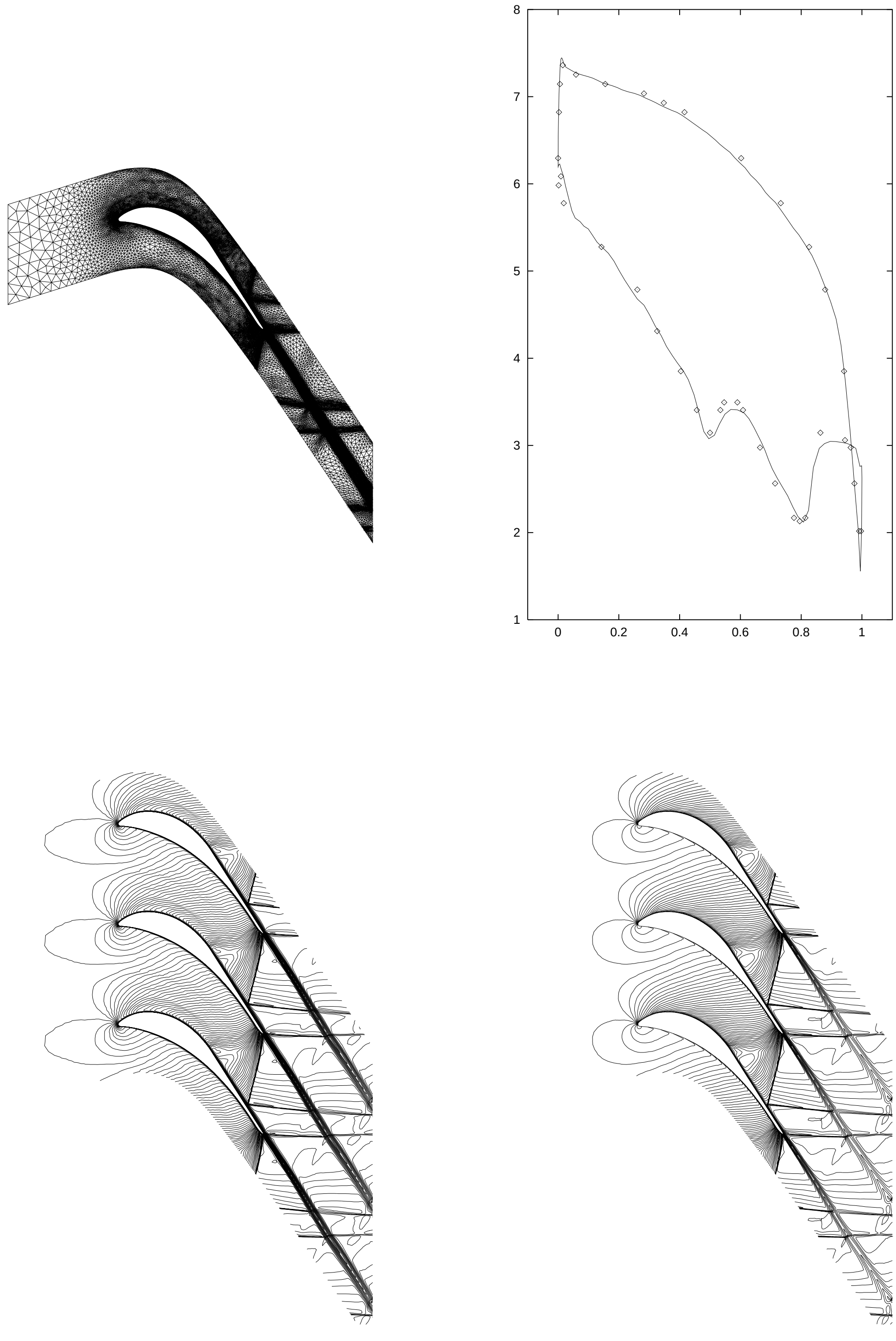


FIG. 4.15: Cas visqueux, maillage triangulaire fondamental (en haut à gauche), la distribution de la pression en comparaison avec les mesures (en haut à droite), les lignes de niveau du nombre de Mach(en bas à gauche) et les lignes de niveau de la densité (en bas à droite) pour première méthode: schéma de base.

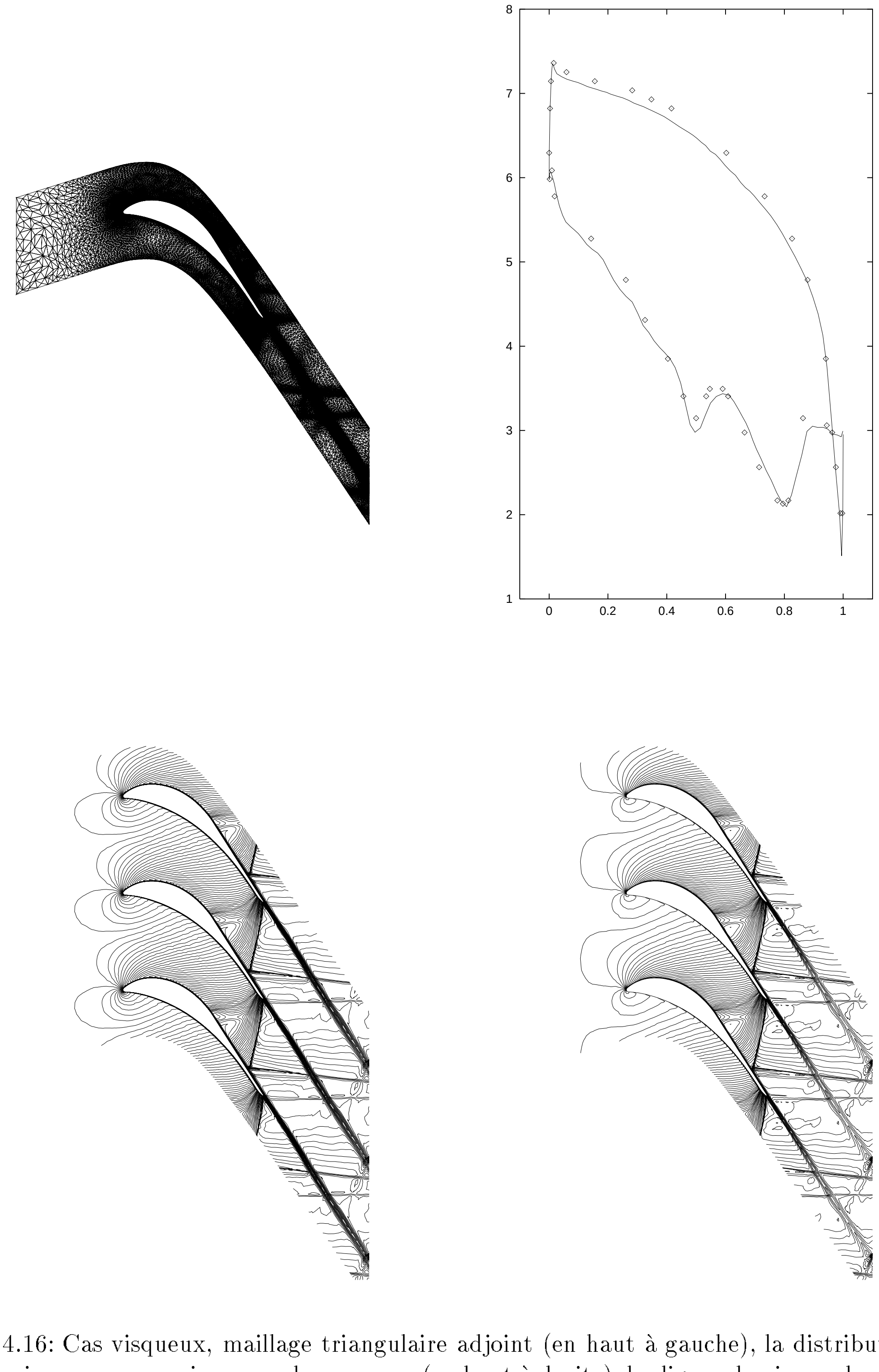


FIG. 4.16: Cas visqueux, maillage triangulaire adjoint (en haut à gauche), la distribution de la pression en comparaison avec les mesures (en haut à droite), les lignes de niveau du nombre de Mach(en bas à gauche) et les lignes de niveau de la densité (en bas à droite) pour première méthode: schéma d'ordre plus élevé.

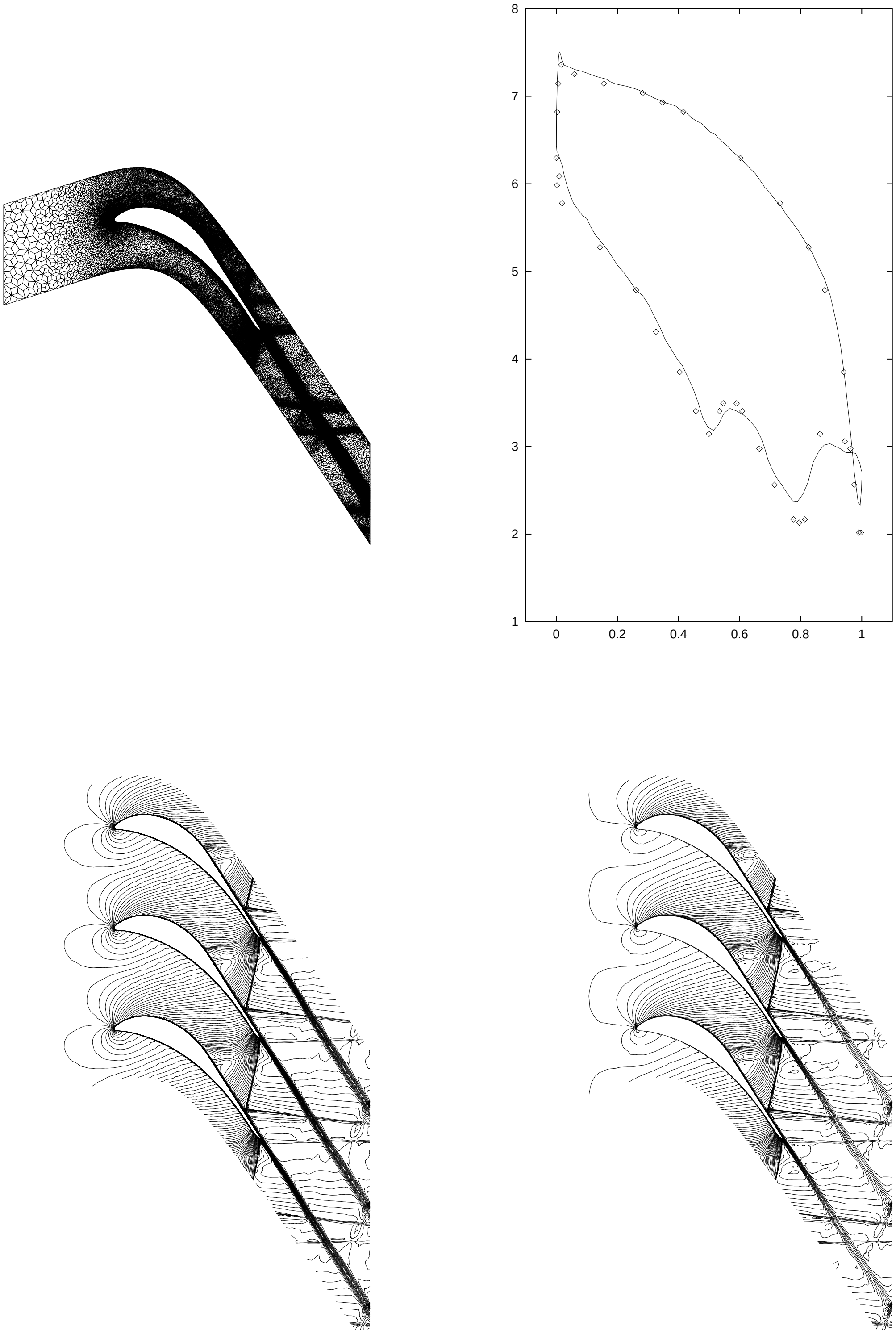


FIG. 4.17: Cas visqueux, maillage barycentrique (en haut à gauche), la distribution de la pression en comparaison avec les mesures (en haut à droite), les lignes de niveau du nombre de Mach(en bas à gauche) et les lignes de niveau de la densité (en bas à droite) pour deuxième méthode: schéma de base.

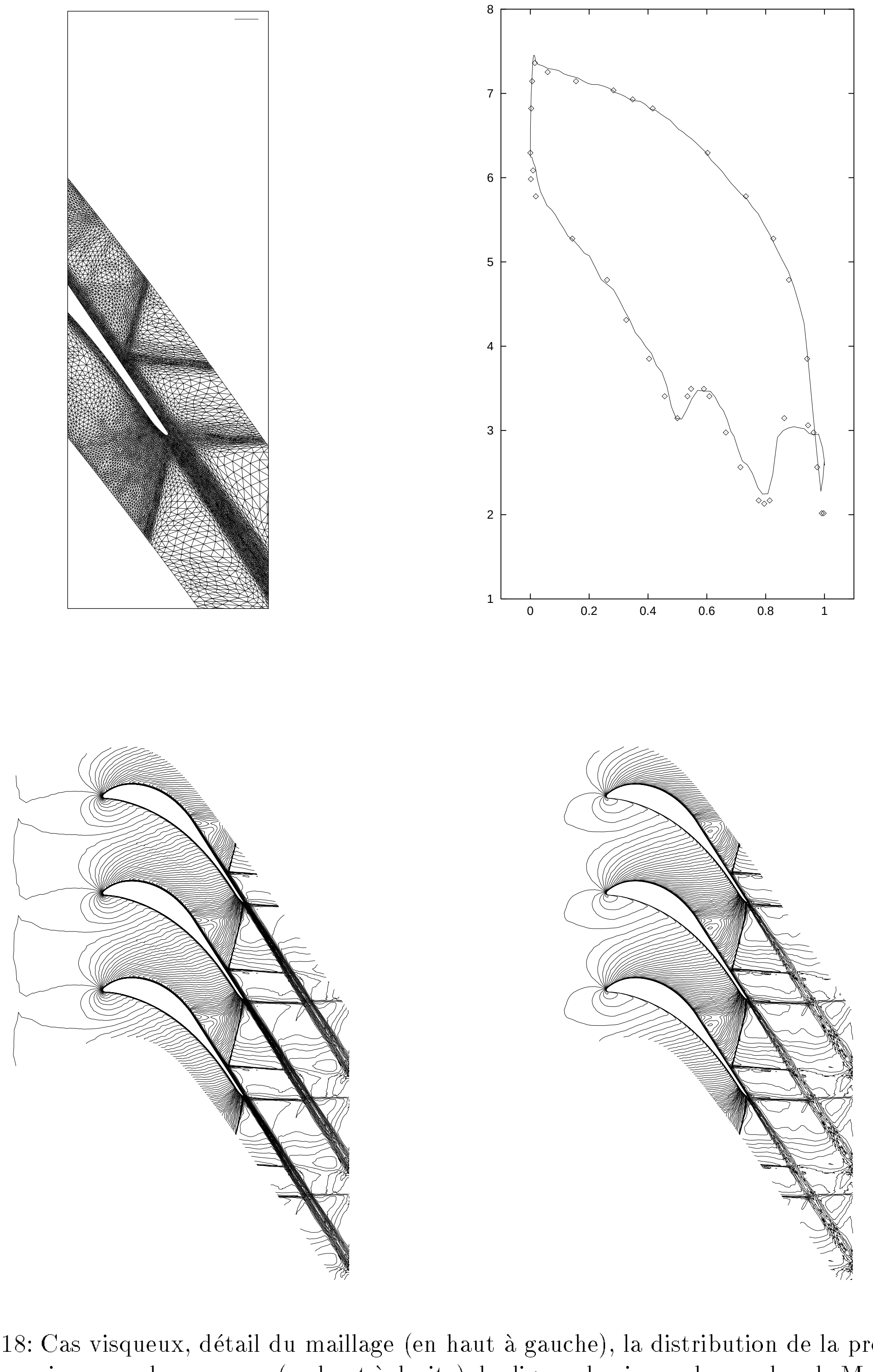


FIG. 4.18: Cas visqueux, détail du maillage (en haut à gauche), la distribution de la pression en comparaison avec les mesures (en haut à droite), les lignes de niveau du nombre de Mach(en bas à gauche) et les lignes de niveau de la densité (en bas à droite) pour deuxième méthode: schéma schéma d'ordre plus élevé.

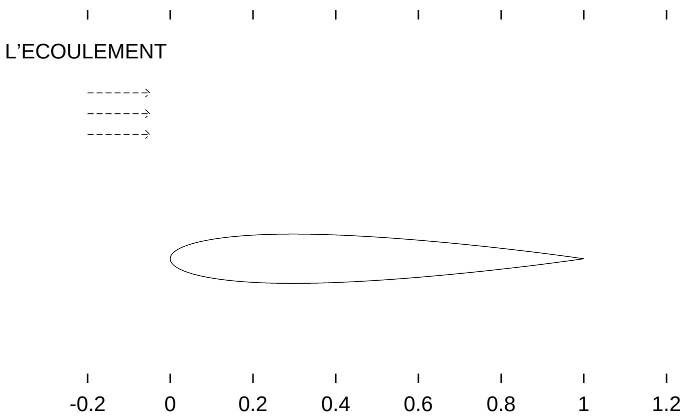


FIG. 4.19: Profile NACA 0012

4.4 NACA 0012 profile

L'exemple suivant est l'écoulement autour du profil NACA 0012 (voir Fig. 4.19). Ce problème est très souvent utilisé pour le calcul visqueux et aussi non visqueux d'écoulement transonique.

L'écoulement non visqueux Ici on présente l'écoulement non visqueux avec les données suivantes:

$$\begin{aligned} M_{ent} &= 0.80 && \text{nombre de Mach entrée} \\ \alpha &= 1.25^\circ && \text{l'angle d'incidence.} \end{aligned}$$

Le calcul a été effectué par les méthodes d'ordre plus élevées sur des maillages triangulaires et barycentriques (Fig. 4.20), qui ont été construits par l'adaptation d'un maillage anisotropique (voir Chap. 6). Les résultats peuvent être comparés avec ceux du Chap. 5, où des maillages sont construits par la méthode de multiniveaux du raffinement.

L'écoulement visqueux Le calcul d'écoulement visqueux a été effectué pour deux cas avec le nombre de Reynolds différents. Les données sont:

$$\begin{aligned} M_{ent} &= 0.85 && \text{nombre de Mach entrée} \\ \alpha &= 1.25^\circ && \text{l'angle d'incidence} \\ Re &= 2 \cdot 10^3 && \text{nombre de Reynolds – le premier cas} \\ Re &= 10^6 && \text{nombre de Reynolds – le deuxiéme cas} \end{aligned}$$

Pour le calcul deux méthodes combinant des VF et EF ont été utilisés. Les résultats pour le cas avec $Re = 2 \cdot 10^3$ sont representés sur la Fig. 4.21 et le cas avec $Re = 10^6$ est representé sur la Fig. 4.22

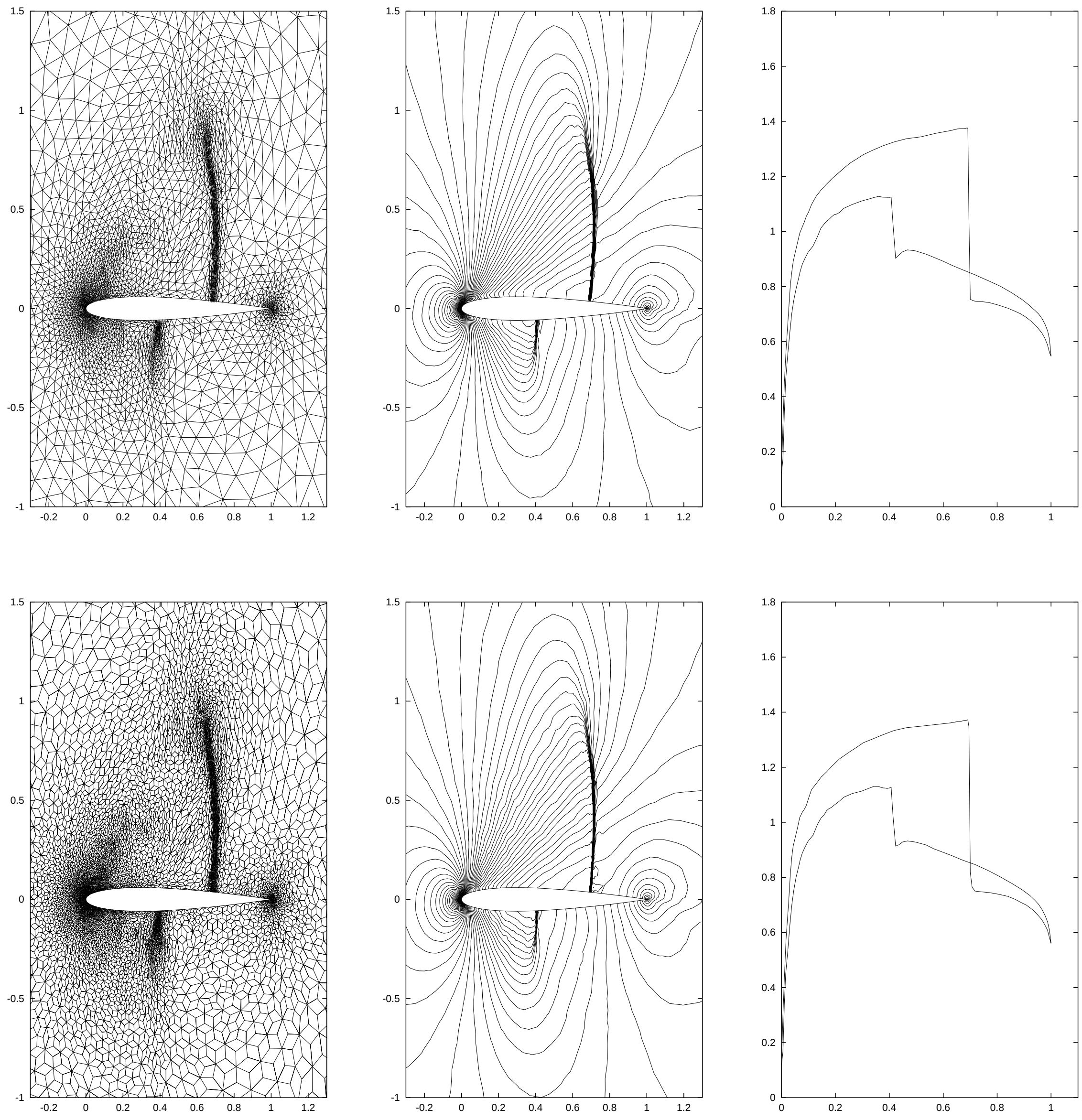


FIG. 4.20: L'écoulement non visqueux: Le maillage, les lignes du niveau du nombre de Mach et la distribution du nombre de Mach sur le profil pour le maillage triangulaire (en haut) et barycentrique (en bas).

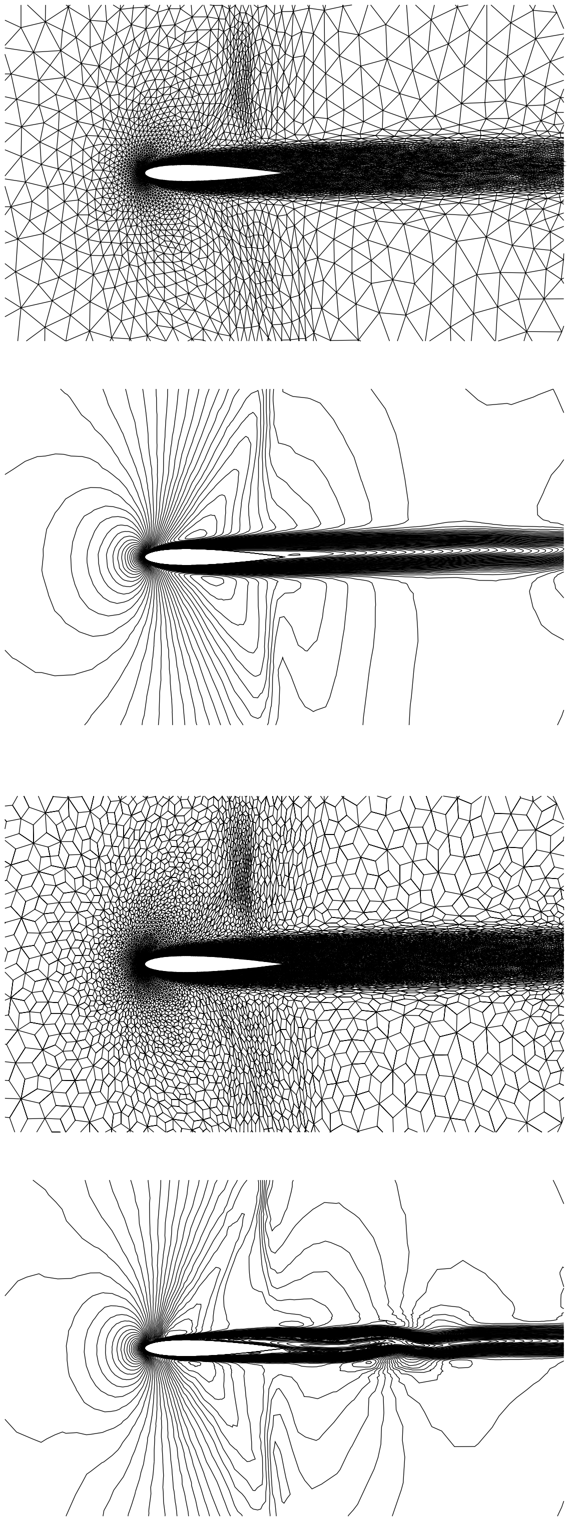


FIG. 4.21: L'écoulement visqueux, $Re = 2 \cdot 10^3$: Le maillage et les lignes du niveau du nombre de Mach pour la méthode de VF triangulaires et EF conformes (en haut) et VF barycentriques et EF non conformes (en bas).

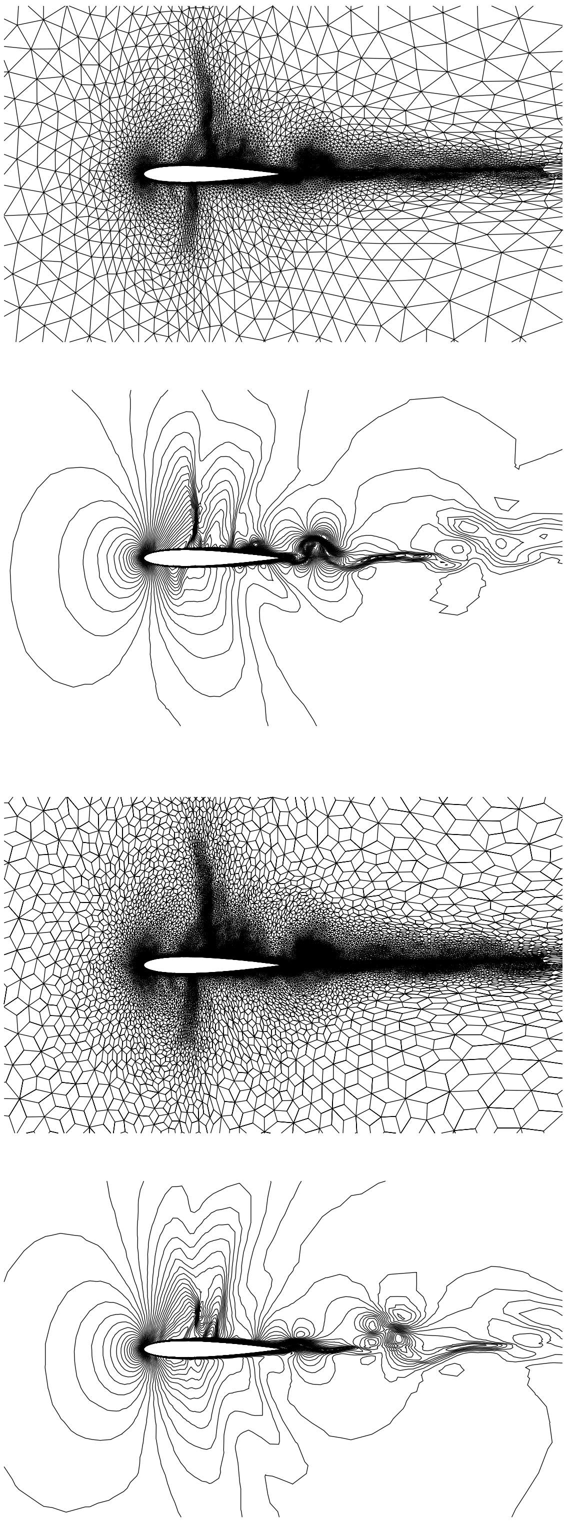


FIG. 4.22: L'écoulement visqueux, $Re = 10^6$: Le maillage et les lignes du niveau du nombre de Mach pour la méthode de VF triangulaires et EF conformes (en haut) et VF barycentriques et EF non conformes (en bas).

Deuxième partie

Adaptation des maillages

Les calculs numériques montrent, que non seulement une méthode numérique mais aussi le choix d'un maillage utilisé pour la discrétisation joue un rôle très important. La limitation d'une mémoire et d'une rapidité des ordinateurs ne nous permet pas d'utiliser un maillage très fin avec un nombre d'éléments très grands. Pour cette raison il est convenable de raffiner un maillage seulement dans certaines parties du domaine du calcul. Un algorithme cherchant ces parties et créant un nouveau maillage s'appelle la *méthode adaptative*.

Les méthodes adaptatives sont des algorithmes itératifs utilisant l'information de la solution approximative originairement calculée pour améliorer sa qualité dans le calcul suivant. De nombreux des algorithmes adaptatifs utilisent le raffinement du maillage, auquel le Chap. 5 est consacré. L'autre approche est l'adaptation complète d'un maillage (c'est-à-dire non seulement raffinement mais aussi deraffinement du maillage et bouge ses noeuds) qui est développée dans le Chap. 6.

Des nombreuses recherches ont été effectuées sur les méthodes adaptatives appliquées aux problèmes elliptiques et paraboliques basés sur les résultats théoriques, voir [5], [4], [16]. Particulièrement dans la mécanique des fluides la résolution correcte des couches limites et des ondes de choc exigent l'utilisation de méthodes adaptatives. De nombreux résultats sont résumés dans [38] et [49].

Chapitre 5

Technique de raffinement du maillage fondamental

Dans ce chapitre nous développons un algorithme de raffinement du maillage. L'outil le plus important pour une méthode adaptative est le choix d'un indicateur d'erreur convenable et, pour l'écoulement avec ondes de choc, aussi d'un indicateur de choc. Le résidu est souvent utilisé comme l'indicateur d'erreur. Cette approche ne paraît pas trop convenable pour des problèmes non linéaires, où il n'existe pas d'opérateur inverse, mais elle a été appliquée dans [35], [33],[58].

5.1 Indicateur de choc

Il existe plusieurs possibilités pour la définition de l'indicateur. Nous utilisons l'approche d'indicateur de choc qui indique les régions où l'on observe de fortes variations des quantités physiques (ondes de choc).

On a testé plusieurs indicateurs de choc qui sont déterminés après la considération sur la forme physique des ondes de choc. Soit $\mathcal{T}_h = \{T_i\}_{i \in J}$ une triangulation. Pour chaque $T_i \in \mathcal{T}_h$ on appelle $s(i)$ l'ensemble des indices $j \in J$ pour lesquels $T_j \in \mathcal{T}_h$ est le voisin de T_i. Soit le vecteur d'état $\boldsymbol{w}$ la fonction constante par chaque $T_i,\ i \in J$. Par $\boldsymbol{w}_i = (w_{i1}, w_{i2}, w_{i3}, w_{i4})^T = (\rho_i, \rho_i u_{i1}, \rho_i u_{i2}, p_i/(\gamma - 1) + \rho_i |\mathbf{u}_i|^2/2)^T$ la valeur du $\boldsymbol{w}$ sur le triangle T_i. On pose

$$
\begin{aligned}
g_{\boldsymbol{w}}(i) &= \max_{j \in s(i)} \max_{k=1,\ldots,4} |w_{ik} - w_{jk}|/h_i, \\
g_{rho}(i) &= \max_{j \in s(i)} \left[-\left(\rho_i - \rho_j\right) \mathbf{u}_i \cdot n_{ij}\right]^+ /h_i, \\
g_v(i) &= \max_{j \in s(i)} \left[\left(|\mathbf{u}_i| - |\mathbf{u}_j|\right) \mathbf{u}_i \cdot n_{ij}\right]^+ /h_i, \\
g_p(i) &= \max_{j \in s(i)} \left[-\left(p_i - p_j\right) \mathbf{u}_i \cdot n_{ij}\right]^+ /h_i, \\
g_{group}(i) &= \max_{j \in s(i)} \{[(|\mathbf{u}_i| - |\mathbf{u}_j|) \mathbf{u}_i \cdot n_{ij}]^+ + [-\left(\rho_i - \rho_j\right) \mathbf{u}_i \cdot n_{ij}]^+ + \\
&+ \ [-\left(p_i - p_j\right) \mathbf{u}_i \cdot n_{ij}]^+\}/h_i, \qquad T_i \in \tau_h,
\end{aligned}
\tag{5.1}
$$

où $a^+ = \max(a, 0)$. L' indicateur g_w contrôle la grandeur du gradient du vecteur d'état $\boldsymbol{w}$ sans observer la direction d'écoulement. Les autres indicateurs sont construits de manière à indiquer les ondes de choc physiquement admissibles.

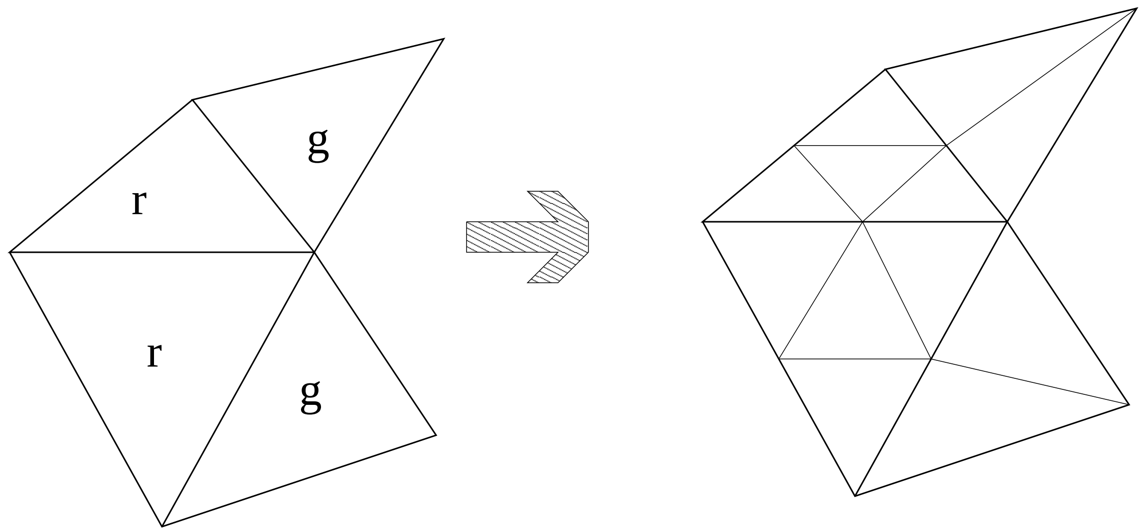


FIG. 5.1: Adaptation du maillage "red–green". Les triangles "red"(**r**) tels que $\eta(i) \geq \eta$ sont tous divisés et quelques autres "green" (**g**) sont divisés pour garantir un maillage conformé.

Chacun de ces indicateurs (5.1) (signifie par $g(i)$) est appliqué de la manière suivante:

1. Pour chaque triangle $T_i \in \mathcal{T}_h, \quad i \in I$ on calcule $g(i)$.
2. Soit η la tolérance donnée. Si pour quelque $T_i \in \mathcal{T}_h$ il est $g(i) \geq \eta$, le triangle "mère" T_i est divisé sur les quatre triangles "filles".
3. Cette division de Ω_h est modifiée de la manière que la triangulation conformée de Ω_h est obtenue (raffinement du maillage "red–green"(voir Fig. 5.1).
4. L'approximation initiale sur le maillage nouveau est définie en prenant les même valeurs sur les triangles filiales comme pour le triangle maternel.

Cet algorithme peut être appliqué aussi pour les volumes finis barycentriques. Il faut interpoler la solution sur le maillage triangulaire (lequel est la base pour la construction du maillage barycentrique), après que le processus décrit est appliqué et à la fin un nouveau maillage barycentrique est créé sur la base de la nouvelle triangulation.

Nous avons testé les indicateurs (5.1) pour le problème d'écoulement non visqueux autour de la cascade des profiles. Fig. 5.2, Fig. 5.3 et Fig. 5.4 montrent le maillage fondamental et les maillages obtenus après raffinement en utilisant les indicateurs (5.1). Les meilleur résultats sont obtenus par l'indicateur g_{rho}. Fig. 5.5 montre les lignes du niveau du nombre de Mach sur le maillage fondamental et le maillage raffiné par g_{rho}. Ici on peut voir qu'après un raffinement les ondes de choc sont plus aigues.

5.2 Technique multiniveau de raffinement

Il est évident qu'après l'application d'un seul raffinement, les ondes du choc n'ont pas encore la netteté suffisante. Il faut appliquer le raffinement plusieurs fois. Le meilleur indicateur qui a été testé paraît

$$g_{rho}(i) = \max_{j \in s(i)} \left[-(\rho_i - \rho_j)\, \boldsymbol{v}_i \cdot \boldsymbol{n}_{ij} \right]^+ / h_i. \tag{5.2}$$

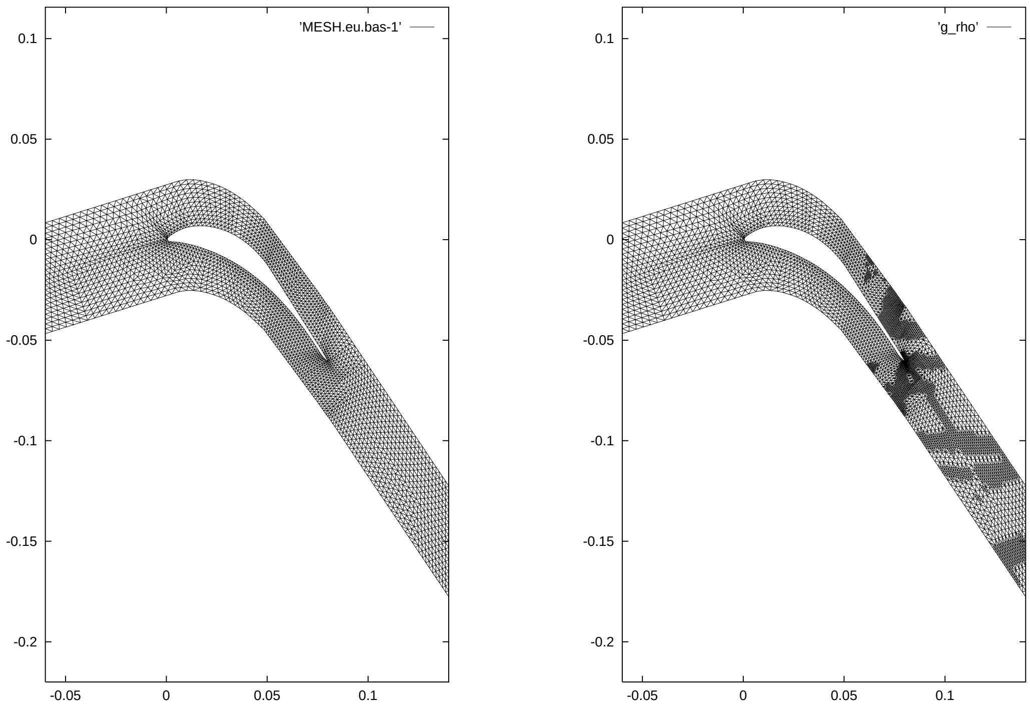


FIG. 5.2: Le maillage fondamental et le maillage raffiné par g_{rho}

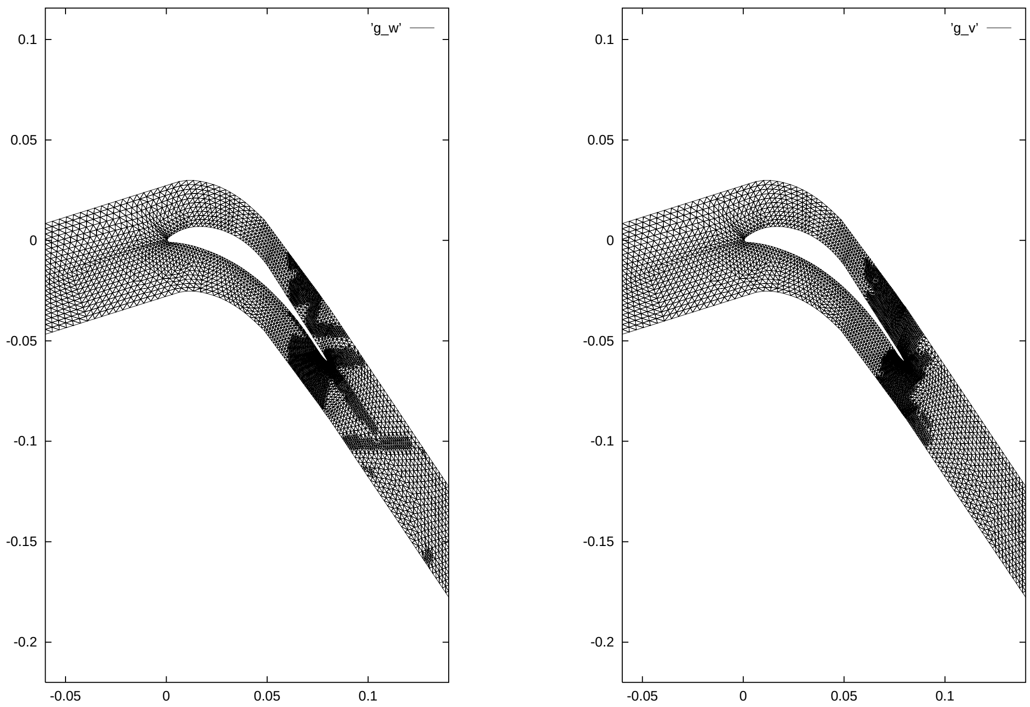


FIG. 5.3: Les maillages raffinés par g_w et g_v.

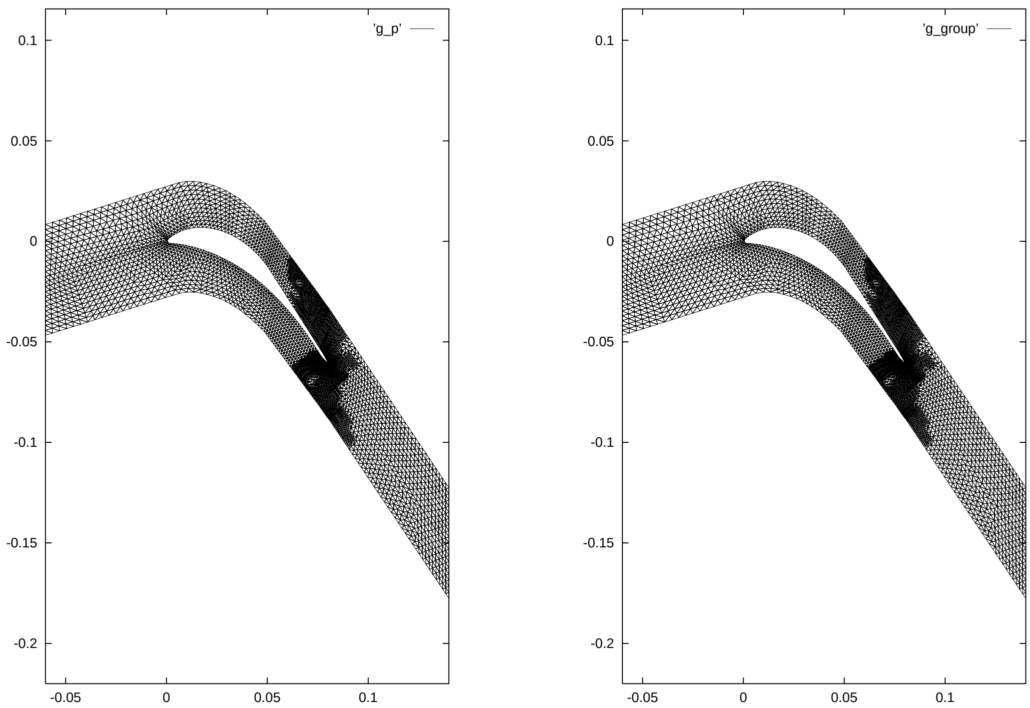


FIG. 5.4: Les maillages raffinés par g_p et g_{group}.

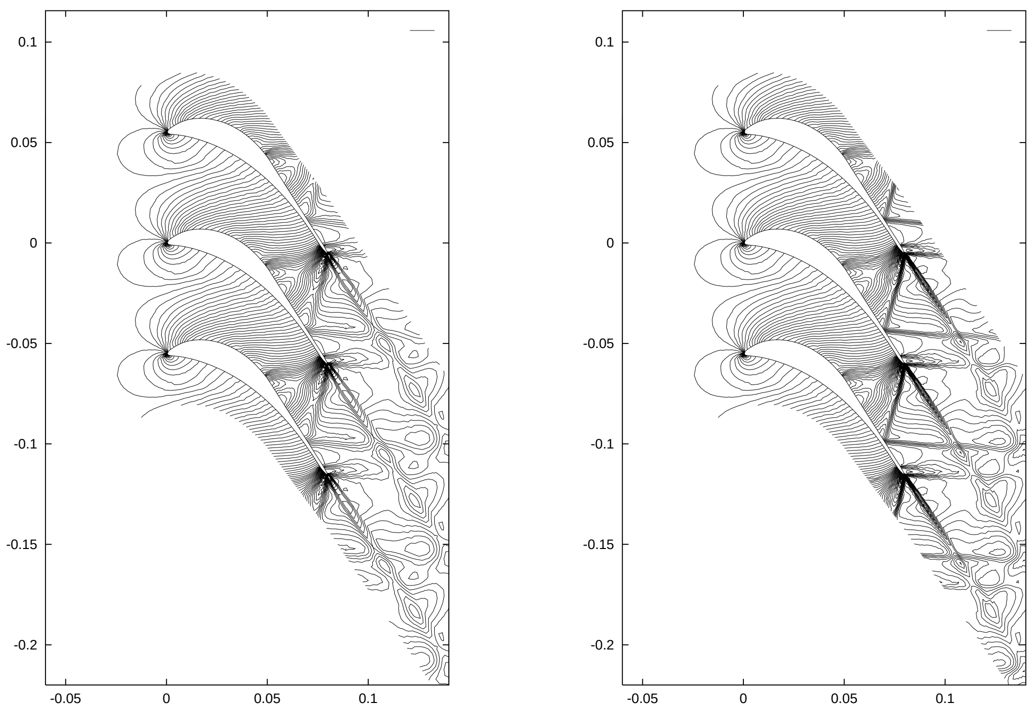


FIG. 5.5: Les lignes du niveau du nombre de Mach sur le maillage fondamental et le maillage refinée

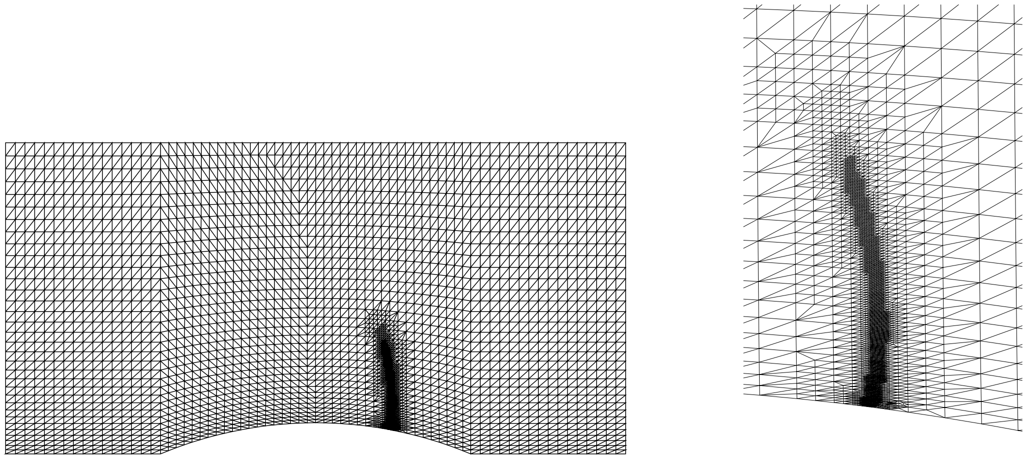


FIG. 5.6: Maillage et son détail après un quatrième raffinement.

Pour le premier raffinement, la tolérance $\eta_1 = \eta$ est choisie expérimentalement et pour $k-$ème raffinement ($k = 2, 3, \dots$) nous supposons que les différences de densité sont les mêmes mais que la grandeur de l'arête h_i est deux fois plus petite. C'est la raison pour laquelle nous imposons

$$\eta_k = \eta_1 \cdot 2^{k-1}, \qquad k = 2, 3, \dots . \tag{5.3}$$

Nous nous attendons à ce que l'indicateur désigne uniquement les volumes qui sont déjà raffinés et que la surface du domaine raffiné diminue lorsque le niveau du raffinement augmente. Fig. 5.6 montre le maillage du GAMM – tube après un quatrième raffinement. Il est aisé de voir que les résultats satisfont notre attente.

Nous présentons l'écoulement autour du profil NACA 0012. On prescrit le nombre de Mach entré $M_{ent} = 0.8$ et l'angle d'incidence $\alpha = 1.25^0$. Nous présentons les résultats obtenus sur la grille fondamentale et sur 4 grilles raffinées successivement par la technique (5.2) – (5.3) avec $\eta_0 = 4$. La méthode de premier ordre sur un maillage triangulaire est utilisée. Figs. 5.7 – 5.11 montrent les lignes du niveau du nombre de Mach et la distribution du nombre de Mach sur le profil pour le maillage fondamental et les maillages après 1ère, 2ème, 3ème et 4ème raffinement. Après le dernier raffinement, la qualité de ondes de choc est très grande.

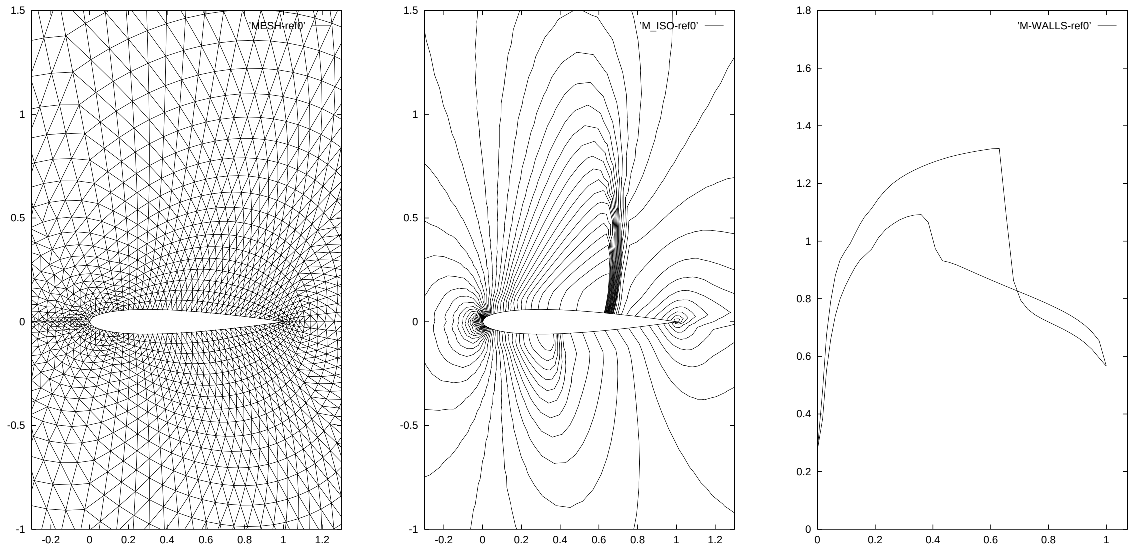


FIG. 5.7: Le maillage, les lignes du niveau du nombre de Mach et la distribution du nombre de Mach sur le profil pour le maillage fondamental.

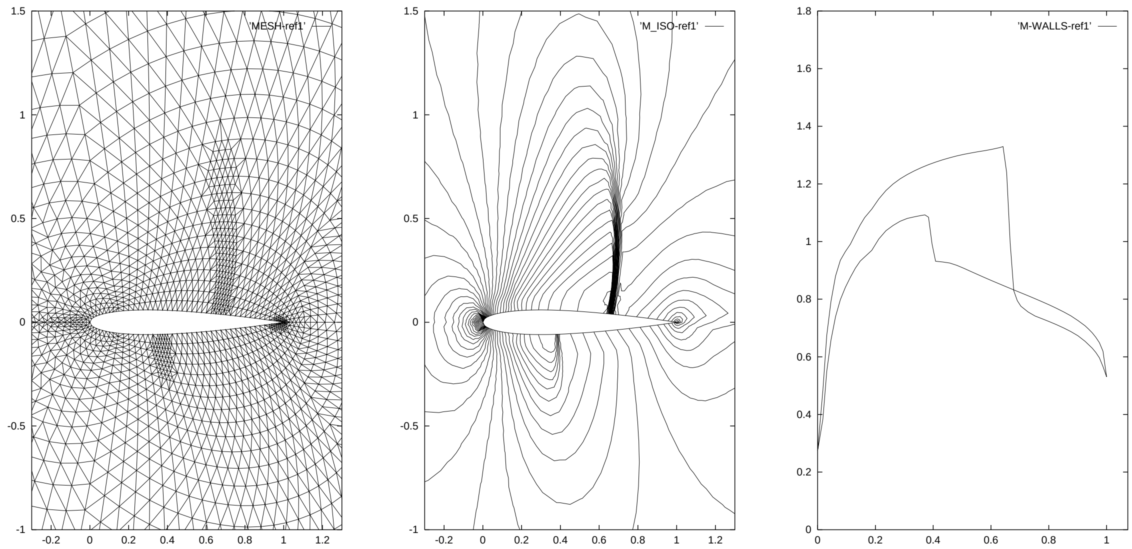


FIG. 5.8: Le maillage, les lignes du niveau du nombre de Mach et la distribution du nombre de Mach sur le profil pour le maillage après un premier raffinement.

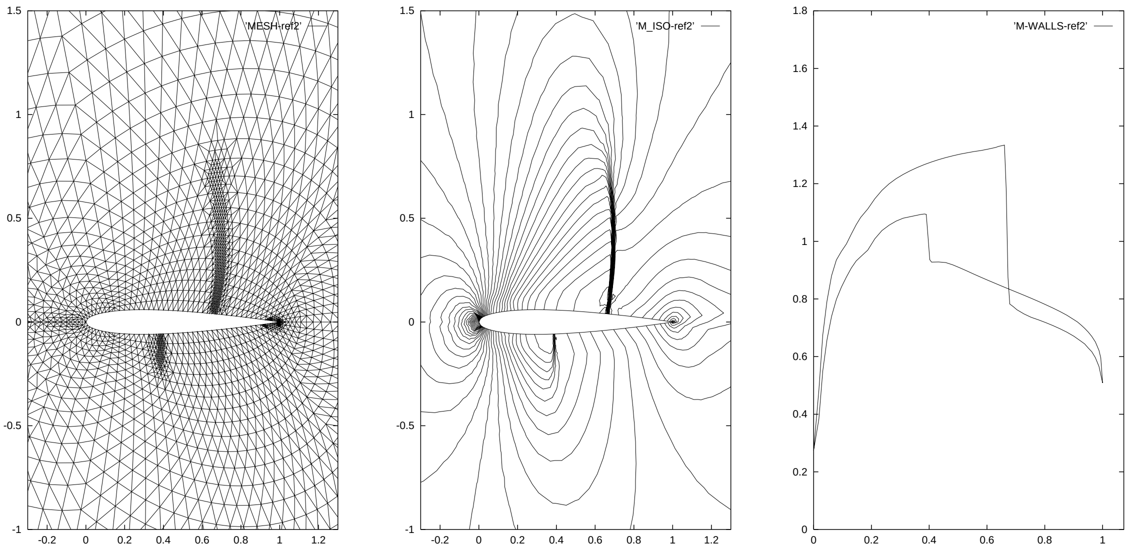


Fig. 5.9: Le maillage, les lignes du niveau du nombre de Mach et la distribution du nombre de Mach sur le profil pour le maillage après un deuxième raffinement.

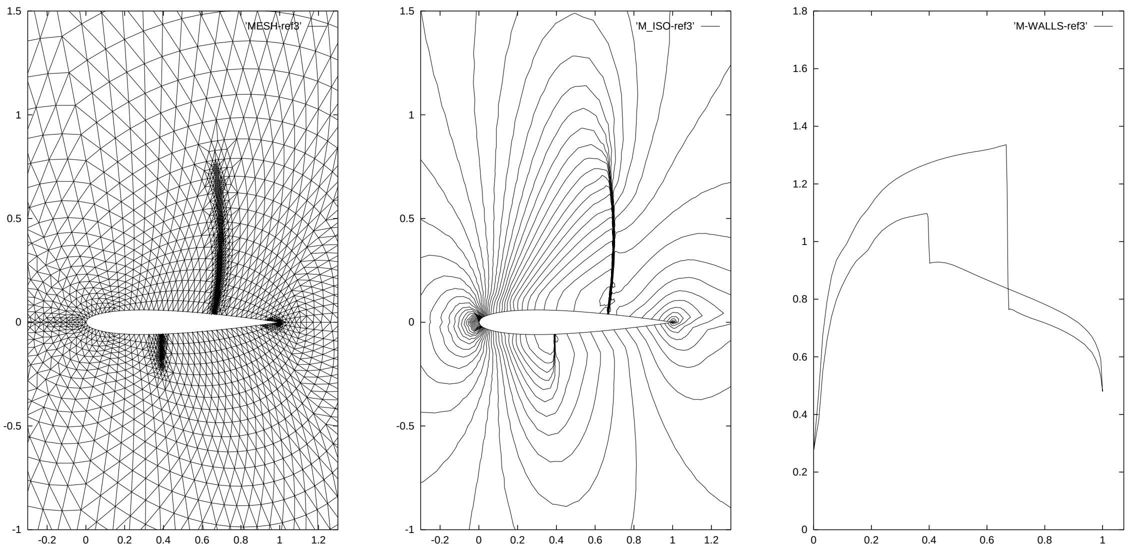


Fig. 5.10: Le maillage, les lignes du niveau du nombre de Mach et la distribution du nombre de Mach sur le profil pour le maillage après un troisième raffinement.

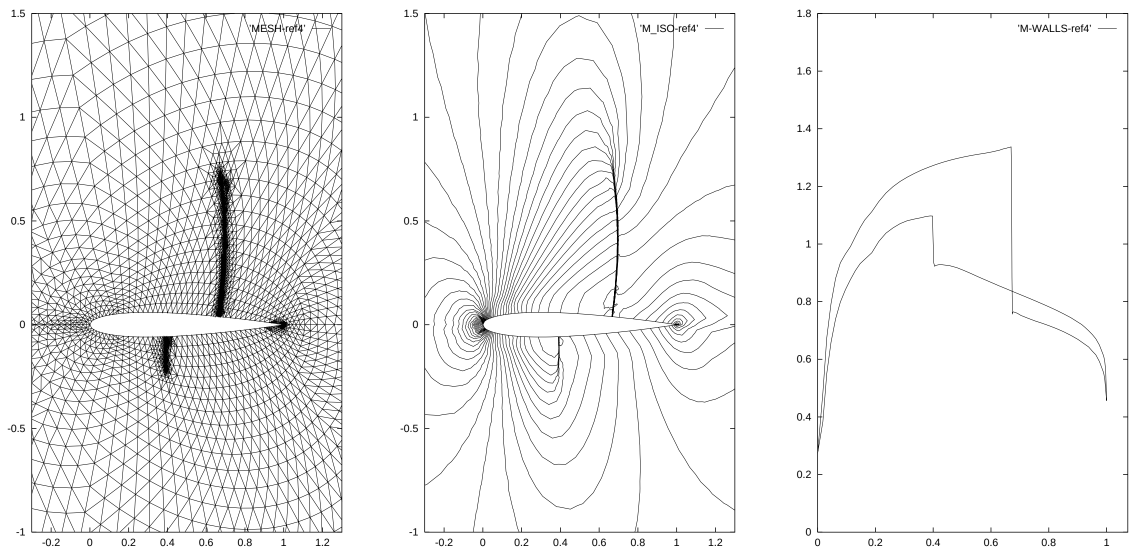


FIG. 5.11: Le maillage, les lignes du niveau du nombre de Mach et la distribution du nombre de Mach sur le profil pour le maillage après un quatrième raffinement.

Chapitre 6

Adaptation des maillages anisotropiques

Le raffinement du maillage décrit ci-dessus donne des résultats satisfaisants, mais après plusieurs niveaux du raffinement le nombre des éléments est très grand. De plus il paraît convenable d'adapter le maillage par rapport à la direction de discontinuité (voir Fig. 6.1). Pour cette raison nous avons développé la stratégie de l'adaptation des maillages anisotropiques (AMA), qui peut être utilisé pour la résolution des problèmes de physique et mécanique. AMA construit un maillage optimal au sens que "l'erreur" de la solution est uniformément distribuée sur tout le maillage.

L'article décrivant l'approche d'AMA a été accepté pour la présentation dans *"Computing and Visualisation in Science"*.

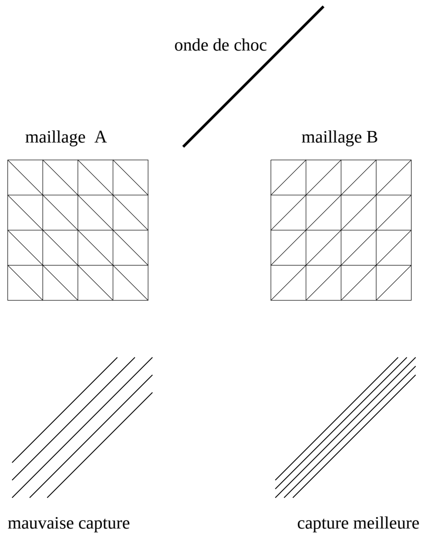


FIG. 6.1: Adaptation du maillage en regard à la direction de discontinuité. B – maillage capture mieux la discontinuité parce que les "diagonales" du maillage sont parallèles à la direction de la discontinuité.

ANISOTROPIC MESH ADAPTATION FOR FINITE VOLUME AND FINITE ELEMENT METHODS ON TRIANGULAR MESHES

Vít Dolejší
Departement of Numerical Mathematics,
Faculty of Mathematics and Physics,
Charles University,
Malostranské nám. 25,
118 00 Prague - Czech Republic

Abstract

The present paper deals with an anisotropic mesh adaptation (AMA) of triangulation which can be employed for the numerical solution of various problems of physics. AMA tries to construct an optimal triangulation of the domain of computation in the sense that an "error" of the solution of the problem considered is uniformly distributed over the whole triangulation. First, we describe the main idea of AMA. We define an optimal triangle and an optimal triangulation. Then we describe the process of optimization of the triangulation and the complete multilevel computational process. We apply AMA to a problem of CFD, namely to inviscid compressible flow. The computational results for a channel flow are presented.

Key words: Mesh adaptation – Riemann metric – unstructured mesh – compressible flows

6.1 Introduction

The *finite volume* and/or *finite element methods* are very often employed for solving various problems of physics (for example fluid mechanics, elasticity, heat conduction, etc.). Here, the computational domain is divided into a finite number of triangles and the system of partial differential equations describing the problem is discretized. To obtain an accurate numerical solution, a fine triangulation has to be used. On the other hand, a very fine triangulation gives a large number of elements and the computation requires a lot of memory and CPU time. This was the reason why adaptative methods were developed. We seek the regions where the numerical solution is inaccurate and we refine the triangulation there. A standard technique is the *red–green refinement*: we define an indicator and triangles with values of the indicator greater than a tolerance prescribed are refined. The red–green strategy gives good results but the number of elements after n refinements (n equals, for example, 3, 4 or 5) is very large. Another possibility is to use *the anisotropic mesh adaptation* (AMA) where we can obtain the same quality of solution, using a lower number of elements. AMA can not only refine the triangulation but also removes some triangles and moves the nodes of the mesh. So we can obtain a triangulation with a not too large number of elements on which a sufficiently accurate solution can be computed.

AMA tries to construct an optimal mesh in the sense that an "error" of the solution of the problem considered is uniformly distributed over the whole triangulation. In distinction to the red–green refinement technique, AMA allows for the refinement of triangles in a special direction. E.g., it is possible to align the refinement along a discontinuity. This requires to have some error indicator which detects not only the value of error but also the changes of error with respect to the direction. Such an indicator can be represented by a set of 2×2

symmetric and positive definite matrices M_i which also represent the Riemann metric in the domain Ω of computation.

The main idea of AMA is given in Section 2. In Section 3 we define *the optimal triangle corresponding to a matrix*, in Section 4 we define *the optimal triangulation*. The construction of the optimal triangulation is described in Section 5. In Section 6 we describe the generation of matrices M and in Section 7 we describe the multilevel computation using AMA. The last section presents some numerical results for CFD problems.

6.2 Interpolation error

Here we describe the main idea of AMA. Let u be an analytical solution of a considered problem and u_h be its piecewise linear approximation. Our aim is to construct a triangulation $\mathcal{T}$ of a computational domain Ω so that the *interpolation error* (see [11]) $e_I(x,y) \equiv |u(x,y) - u_h(x,y)|$, $(x,y) \in \Omega$ is for each $(x,y) \in \Omega$ smaller than given tolerance and on the other hand the number of elements of triangulation is not high.

Let $u \in C^2(\Omega)$. Then, using the Taylor series expansion at $(x_0, y_0) \in \Omega$, we have

$$\begin{aligned} u(x,y) &= u(x_0,y_0) + \frac{\partial u(x_0,y_0)}{\partial x}(x-x_0) + \frac{\partial u(x_0,y_0)}{\partial y}(y-y_0) + \qquad (6.1)\\ &+ \frac{1}{2}\left[\frac{\partial^2 u(x_0,y_0)}{\partial x^2}(x-x_0)^2 + 2\frac{\partial^2 u(x_0,y_0)}{\partial x \partial y}(x-x_0)(y-y_0) + \frac{\partial^2 u(x_0,y_0)}{\partial y^2}(y-y_0)^2\right] + \\ &+ o(|x-x_0, y-y_0|^2) \quad (x,y) \in \Omega. \end{aligned}$$

Now we define the linear approximation of u in some neighbourhood $\mathcal{U}((x_0,y_0)) \subset \Omega$ of (x_0,y_0) by

$$\begin{aligned} u_h|_{\mathcal{U}((x_0,y_0))}(x,y) &:= u(x_0,y_0) + \frac{\partial u(x_0,y_0)}{\partial x}(x-x_0) + \frac{\partial u(x_0,y_0)}{\partial y}(y-y_0), \qquad (6.2)\\ & \qquad (x,y) \in \mathcal{U}((x_0,y_0)). \end{aligned}$$

Then for the interpolation error in $\mathcal{U}((x_0,y_0))$ we obtain (we omit the terms of order $o(|x-x_0, y-y_0|^2)$)

$$\begin{aligned} e_I \ |_{\mathcal{U}((x_0,y_0))}(x,y) &:= \left|u(x,y) - u_h|_{\mathcal{U}((x_0,y_0))}(x,y)\right| \approx \qquad (6.3)\\ &\approx \frac{1}{2}\left|\frac{\partial^2 u(x_0,y_0)}{\partial x^2}(x-x_0)^2 + 2\frac{\partial^2 u(x_0,y_0)}{\partial x \partial y}(x-x_0)(y-y_0) + \frac{\partial^2 u(x_0,y_0)}{\partial y^2}(y-y_0)^2\right|, \end{aligned}$$

which is valid for $(x,y) \in \mathcal{U}((x_0,y_0))$. If $\omega > 0$ is a given tolerance, we require that $e_I(x,y) \leq \omega$. This and (6.3) give

$$\left|\frac{\partial^2 u(x_0,y_0)}{\partial x^2}(x-x_0)^2 + 2\frac{\partial^2 u(x_0,y_0)}{\partial x \partial y}(x-x_0)(y-y_0) + \frac{\partial^2 u(x_0,y_0)}{\partial y^2}(y-y_0)^2\right| \leq 2\omega, \quad (6.4)$$

which can be rewritten as

$$\begin{pmatrix} x-x_0, & y-y_0 \end{pmatrix} \cdot \left.\begin{pmatrix} \frac{\partial^2 u}{\partial x^2} & \frac{\partial^2 u}{\partial x \partial y} \\ \frac{\partial^2 u}{\partial x \partial y} & \frac{\partial^2 u}{\partial y^2} \end{pmatrix}\right|_{(x_0,y_0)} \begin{pmatrix} x-x_0 \\ y-y_0 \end{pmatrix} \leq 2\omega \qquad (6.5)$$

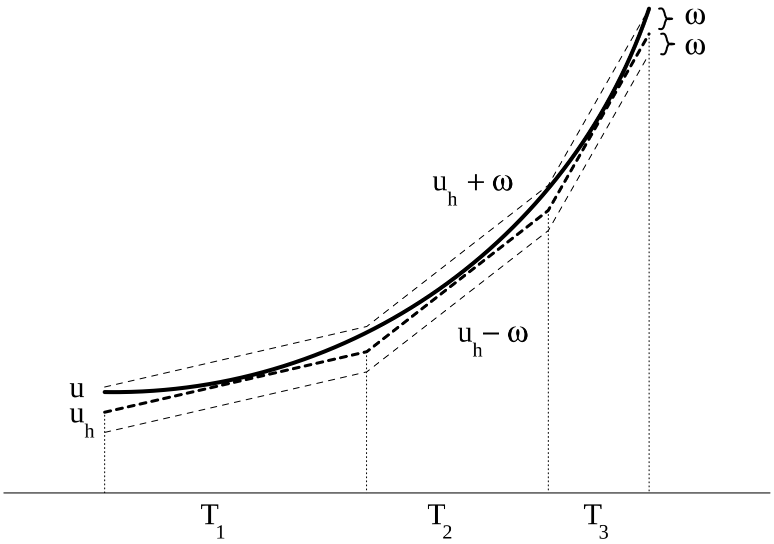


FIG. 6.2: "Triangulation" for 1–D case satisfying the condition (6.6). "Elements" T_1, T_2, T_3, exact solution u (solid line), approximate solution u_h (dashed line).

provided that the Hessian matrix from (6.5) is positive definite. Under this assumption the points (x, y) satisfying the relation (6.5) form an ellipse $\epsilon(\omega) \subset I\!R^2$. From this it follows that the interpolation error $e_I(x, y)$ is less or equal to ω for all $(x, y) \in \epsilon(\omega)$. Furthermore, if T is a triangle such that $T \subset \epsilon(\omega)$, then $e_I(x, y) \leq \omega$ for all $(x, y) \in T$. To have a good triangulation (i.e. with the number of triangles not too high) we choose a triangle which is a subset of $\epsilon(\omega)$ and has the maximal possible area. If the triangulation is formed by the triangles with the above properties (for each triangle we have another matrix) then the required condition will be satisfied:

$$e_{\mathrm{I}}(x, y) = |u(x, y) - u_h(x, y)| \leq \omega \quad \forall (x, y) \in \Omega. \tag{6.6}$$

Here u_h arises as composition of functions of type (6.2). The above considerations are illustrated in Fig. 6.2 for 1–D case. The interpolation error can be locally represented by a matrix from (6.5) which defines the Riemann metric.

There are some problems in the previous considerations. First, the function u is not generally smooth ($u \notin C^2(\Omega)$). Further, the matrix from (6.5) is not positive definite in general. Finally, if the analytical solution is (locally) linear, then the interpolation error given by (6.3) vanishes and the ellipse $\epsilon(\omega)$ does not exist. In the following sections we describe an iterative process for a construction of a triangulation where we must avoid the problems mentioned above.

6.3 Optimal triangle

We will define an optimal triangle corresponding to the matrix M representing the Riemann metric which we have mentioned in the previous section. Here we do not specify the matrix from (6.5) (for the definition of M see Section 6.6). Let M be a symmetric and positive definite 2×2 matrix

$$M = \begin{pmatrix} a & b \\ b & c \end{pmatrix}, \qquad ac - b^2 > 0, \tag{6.7}$$

and let $\boldsymbol{v} = (v_1, v_2)$ be a vector in $I\!R^2$. We define *the norm* of the vector $\boldsymbol{v}$ *corresponding* to the matrix M as

$$\|\boldsymbol{v}\|_M \equiv \left(\boldsymbol{v}M\boldsymbol{v}^T\right)^{\frac{1}{2}} = \left[\begin{pmatrix} v_1 & v_2 \end{pmatrix}\begin{pmatrix} a & b \\ b & c \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}\right]^{\frac{1}{2}} = \left(av_1^2 + 2bv_1v_2 + cv_2^2\right)^{\frac{1}{2}}. \quad (6.8)$$

If $M = I$ (=identical matrix) then $\|\boldsymbol{v}\|_M = |\boldsymbol{v}|$, where $|\boldsymbol{v}| = (v_1^2 + v_2^2)^{1/2}$. The matrix M can be interpreted as *an error matrix* and the norm $\|\boldsymbol{v}\|_M$ as an error of solution over the edge $\boldsymbol{v}$ (see [12], [32]).

We decompose M in the following way:

$$M \equiv \begin{pmatrix} a & b \\ b & c \end{pmatrix} = R^T\Lambda R = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \quad (6.9)$$

where λ_1, λ_2 are the eigenvalues of M and $\theta \in [0, 2\pi)$.

Let Oxy be a Cartesian coordinate system. By ϵ_M we denote an ellipse with the axes lengths $r_1 = 1/\sqrt{\lambda_1}, r_2 = 1/\sqrt{\lambda_2}$. The angle between the axis x and the major axis of ϵ_M is θ. (See Fig. 6.3 .) If P_0 is the center of ϵ_M, then it is easy to see that for any point P lying on the boundary of ϵ_M the norm of the vector $P_0P = (P - P_0)$ corresponding to M is constant and $\|P_0P\|_M = 1$. In comparison to Section 6.2 we can consider $\|P_0P\|_M$ as the interpolation error in P (P_0 corresponds to (x_0, y_0)). Then ϵ_M consists of points with the same interpolation error.

We define *the optimal triangle* T_M *corresponding to* M such that $T_M \subset \epsilon_M$ and $|T_M| \geq |T|$ for any triangle $T \subset \epsilon_M$. (Here $|T|$ means the area of the triangle T). If $M = I$ (= identity matrix) then ϵ_M is an unit circle and T_M is an equilateral triangle with the lengths of edges equal to $\sqrt{3}$.

Now we express analytically the coordinates of the optimal triangle T_M. Let us define a new Cartesian coordinate system $\tilde{O}\tilde{x}\tilde{y}$ with the center at the center of ϵ_M and with the axes parallel with the axes of the ellipse (Fig. 6.4).

The problem of finding a triangle contained in ϵ_M with the maximal area has 4 solutions. These four triangles have the following vertices:

$$\tilde{V}_1^{(1)} = \left[\frac{1}{\sqrt{\lambda_1}}, 0\right], \quad \tilde{V}_2^{(1)} = \left[-\frac{1}{2}\frac{1}{\sqrt{\lambda_1}}, \frac{\sqrt{3}}{2}\frac{1}{\sqrt{\lambda_2}}\right], \quad \tilde{V}_3^{(1)} = \left[-\frac{1}{2}\frac{1}{\sqrt{\lambda_1}}, -\frac{\sqrt{3}}{2}\frac{1}{\sqrt{\lambda_2}}\right], \quad (6.10)$$

$$\tilde{V}_1^{(2)} = \left[-\frac{1}{\sqrt{\lambda_1}}, 0\right], \quad \tilde{V}_2^{(2)} = \left[\frac{1}{2}\frac{1}{\sqrt{\lambda_1}}, -\frac{\sqrt{3}}{2}\frac{1}{\sqrt{\lambda_2}}\right], \quad \tilde{V}_3^{(2)} = \left[\frac{1}{2}\frac{1}{\sqrt{\lambda_1}}, \frac{\sqrt{3}}{2}\frac{1}{\sqrt{\lambda_2}}\right], \quad (6.11)$$

$$\tilde{V}_1^{(3)} = \left[\frac{1}{\sqrt{\lambda_2}}, 0\right], \quad \tilde{V}_2^{(3)} = \left[-\frac{1}{2}\frac{1}{\sqrt{\lambda_2}}, \frac{\sqrt{3}}{2}\frac{1}{\sqrt{\lambda_1}}\right], \quad \tilde{V}_3^{(3)} = \left[-\frac{1}{2}\frac{1}{\sqrt{\lambda_2}}, -\frac{\sqrt{3}}{2}\frac{1}{\sqrt{\lambda_1}}\right], \quad (6.12)$$

$$\tilde{V}_1^{(4)} = \left[-\frac{1}{\sqrt{\lambda_2}}, 0\right], \quad \tilde{V}_2^{(4)} = \left[\frac{1}{2}\frac{1}{\sqrt{\lambda_2}}, -\frac{\sqrt{3}}{2}\frac{1}{\sqrt{\lambda_1}}\right], \quad \tilde{V}_3^{(4)} = \left[\frac{1}{2}\frac{1}{\sqrt{\lambda_2}}, \frac{\sqrt{3}}{2}\frac{1}{\sqrt{\lambda_1}}\right]. \quad (6.13)$$

Then the area of all these triangles $\tilde{V}_1^{(l)}\tilde{V}_2^{(l)}\tilde{V}_3^{(l)}$, $l = 1, \ldots, 4$ is equal to

$$|T_M| = \frac{3\sqrt{3}}{4}\frac{1}{\sqrt{\lambda_1\lambda_2}}. \quad (6.14)$$

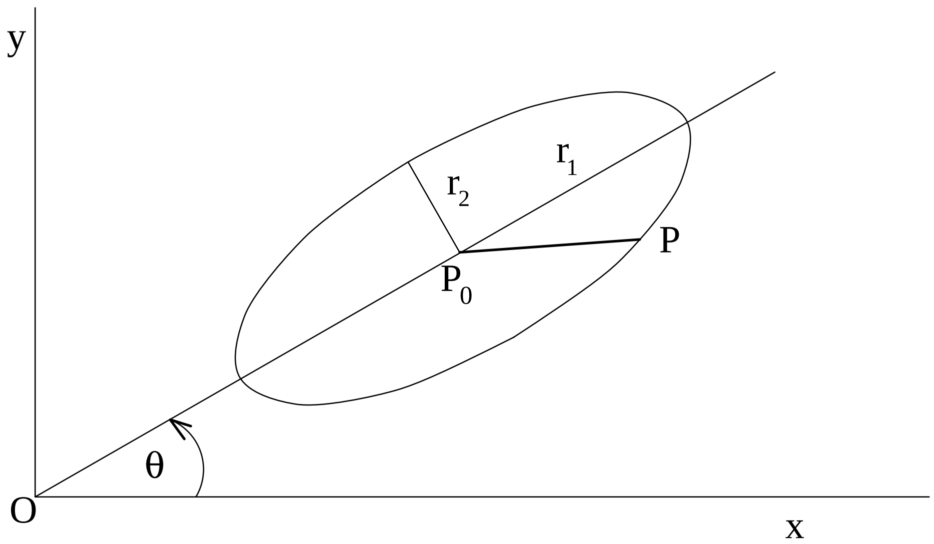


FIG. 6.3: An ellipse ϵ_M corresponding to M with semiaxes length r_1, r_2, P_0 center of ϵ_M, P any point of ϵ_M.

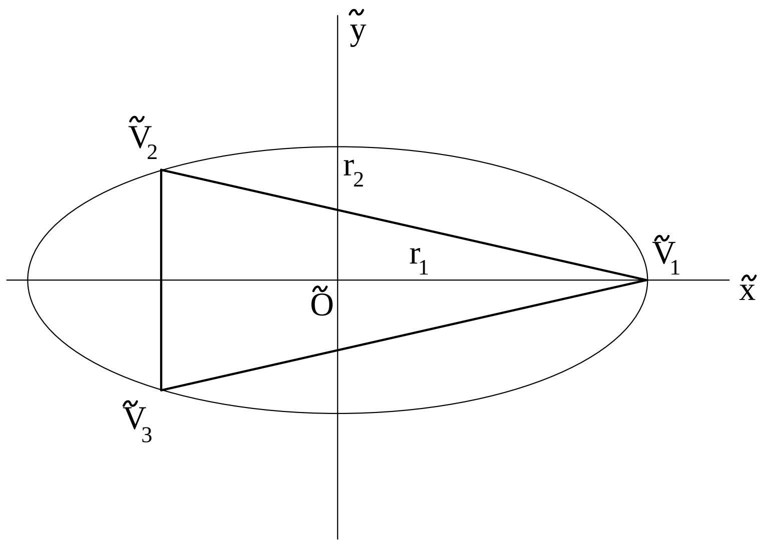


FIG. 6.4: An ellipse ϵ_M corresponding to M in the new Cartesian system and the optimal triangle $T_M = \tilde{V}_1\tilde{V}_2\tilde{V}_3$.

Fig. 6.4 shows the optimal triangle $\tilde{V}_1^{(1)}\tilde{V}_2^{(1)}\tilde{V}_3^{(1)}$. The lengths of edges in the direction $\tilde{y}$ are smaller than the lengths of edges in the direction $\tilde{x}$.

Now we can go back to the original coordinate system by the transformation

$$V = R\tilde{V} + d, \tag{6.15}$$

where V is a point in the system Oxy, $\tilde{V}$ is a point in the system $\tilde{O}\tilde{x}\tilde{y}$, R is the matrix of the rotation from (6.9) and $d = (d_1, d_2)$ is the position of the center of ϵ_M in the coordinate system $\tilde{O}\tilde{x}\tilde{y}$. Applying (6.15) to the points $\tilde{V}_1, \tilde{V}_2, \tilde{V}_3$ (here we use the transformation only for the first triangle given by (6.10) and we omit the superscript $^{(1)}$) we obtain the points V_1, V_2, V_3 which are the vertices of the optimal triangle T_M in the original system:

$$\begin{aligned}
V_1 &= \left[\frac{\cos\theta}{\sqrt{\lambda_1}}, \frac{\sin\theta}{\sqrt{\lambda_2}}\right] + [d_1, d_2] \\
V_2 &= \left[-\frac{\sqrt{3}}{2}\frac{\sqrt{\lambda_2}\cos\theta + \sqrt{\lambda_1}\sin\theta}{\sqrt{\lambda_1\lambda_2}}, \frac{\sqrt{3}}{2}\frac{\sqrt{\lambda_1}\cos\theta - \sqrt{\lambda_2}\sin\theta}{\sqrt{\lambda_1\lambda_2}}\right] + [d_1, d_2], \\
V_3 &= \left[-\frac{\sqrt{3}}{2}\frac{-\sqrt{\lambda_2}\cos\theta + \sqrt{\lambda_1}\sin\theta}{\sqrt{\lambda_1\lambda_2}}, \frac{\sqrt{3}}{2}\frac{\sqrt{\lambda_1}\cos\theta + \sqrt{\lambda_2}\sin\theta}{\sqrt{\lambda_1\lambda_2}}\right] + [d_1, d_2].
\end{aligned}$$

Now we compute the norms corresponding to M of the sides of T_M. We put $\boldsymbol{v}_1 = V_1 - V_2,\ \boldsymbol{v}_2 = V_2 - V_3,\ \boldsymbol{v}_3 = V_3 - V_1$. After some elementary but long calculation we find that

$$\|\boldsymbol{v}_i\|_M = \left(\boldsymbol{v}_i M \boldsymbol{v}_i^T\right)^{\frac{1}{2}} = \sqrt{3}, \qquad i = 1, 2, 3, \tag{6.16}$$

i.e. the norms of all sides of the optimal triangle T_M are constant and equal to $\sqrt{3}$. The relation (6.16) is valid also for triangles which we obtain applying (6.15) to (6.11), (6.12) and (6.13). This is a very interesting and important property in the context of AMA. In the following we will not consider an optimal triangle as a triangle which is subset of ϵ_M and has the maximal possible area but as a triangle with property (6.16).

6.4 Optimal mesh

Here we define an optimal triangulation not as a set of optimal triangles but we use the property (6.16). This approach is more suitable for a numerical realization.

Let $\mathcal{T}$ be a triangulation of the computational domain Ω. Let M_i be a symmetric positive definite 2×2 matrix defined for each node $P_i,\ i \in I(=$ an index set), of the triangulation $\mathcal{T}$. Then for each edge of the triangulation $S_j,\ j \in J(=$ an index set), we put $\bar{M}_j := \frac{1}{2}(M_{j_1} + M_{j_2})$ where P_{j_1}, P_{j_2} are the initial and final points forming S_j, and M_{j_1}, M_{j_2} are the corresponding matrices. Now, we say that the mesh $\mathcal{T}$ is *optimal*, if the norm of S_j corresponding to $\bar{M}_j$ is equal to $\sqrt{3}$ for all $j \in J$, i.e.

$$\|S_j\|_{\bar{M}_j} = \sqrt{3} \qquad \forall j \in J. \tag{6.17}$$

It means that the error (considered as $\|\cdot\|_M$) for each edge of $\mathcal{T}$ is the same, i.e., it is uniformly distributed over the triangulation.

Let $M_i = \bar{c}I,\ i \in I$, where $\bar{c}$ is some constant. Then the optimal mesh is uniform, i.e., not only the norms of all edges of all triangles are the same but also their lengths (in the

Euclidean metric) are the same (because $\|\boldsymbol{v}\|_{M_i} = \sqrt{c}|\boldsymbol{v}|,\ i \in I$) which implies that the inner angles of all triangles are equal to $\pi/3$. It is easy to see that the optimal mesh exists only for special types of domains Ω, for example if Ω is an equilateral triangle. On the other hand, if $\Omega = [0,1] \times [0,1]$ the optimal triangulation does not exist because the angle of triangles at the corners of the domain cannot be equal to $\pi/3$. We will try to construct the mesh which will be close to the optimal one.

We define *the quality parameter of a triangle* $T_n \in \mathcal{T}$ by

$$q_{T_n} = \sum_{l=1}^{3} \left(\|\boldsymbol{v}_{nl}\|_{\bar{M}_{nl}} - \sqrt{3} \right)^2, \tag{6.18}$$

where $\|\boldsymbol{v}_{nl}\|_{\bar{M}_{nl}},\ l = 1,2,3$, are the norms of the sides of the triangle T_n. We define *the quality parameter* $Q_{\mathcal{T}}$ of the mesh $\mathcal{T}$ by

$$Q_{\mathcal{T}} = \frac{1}{nelem} \sum_{T_n \in \mathcal{T}} q_{T_n}, \tag{6.19}$$

where q_{T_n} is the quality parameter of triangle T_n and $nelem$ is a number of triangles in $\mathcal{T}$. It is clear that $Q_{\mathcal{T}}$ is always positive and if the mesh $\mathcal{T}$ is optimal then $Q_{\mathcal{T}} = 0$. The quantity $Q_{\mathcal{T}}$ measures how close the triangulation $\mathcal{T}$ is to the optimal one. We will construct a mesh for which the quality parameter $Q_{\mathcal{T}}$ is minimal.

6.5 Optimization of the mesh

In this section we describe a procedure for the construction of a mesh which will be an approximation of the optimal one (because the optimal triangulation does not exist in general). We start from some initial triangulation $\mathcal{T}_{\text{init}}$ whose quality we want to improve.

Let us suppose that we have a triangulation $\mathcal{T}_{\text{int}}$ with matrices $M_i,\ i \in I$ (M_i symmetric and positive definite). We want to find a new triangulation $\mathcal{T}_{\text{new}}$ such that $Q_{\mathcal{T}_{\text{new}}}$ is minimal. To obtain $\mathcal{T}_{\text{new}}$, an iterative process is used. Each iteration consists of a certain number of the following local operations:

- adding a node at the center of an edge,
- removing an edge,
- swapping the diagonal of the quadrilateral formed by any pair of adjacent elements,
- moving a node.

Now we describe more precisely these local operations.

6.5.1 Adding a node

If the norm of some edge of a triangle is greater than a prescribed constant $L_{\max}$ we put a new node at the center of this edge (Fig. 6.5). Here the norm of the edge AC is greater than $L_{\max}$ and, therefore, the node E is inserted at the center of AC.

We use $L_{\max} = \sqrt{16/3}$. If $\|\boldsymbol{v}\|_M > L_{\max}$ for some $\boldsymbol{v}$ then we have $\left| \|\boldsymbol{v}\|_M - \sqrt{3} \right| > \left| \|\boldsymbol{v}\|_M/2 - \sqrt{3} \right|$, which means that after adding a node, two edges with the norms $\|\boldsymbol{v}\|_M/2$ arise and the difference $\|\boldsymbol{v}\|_M/2$ and $\sqrt{3}$ is smaller than the difference of $\|\boldsymbol{v}\|_M$ and $\sqrt{3}$. The changes in the numbers of nodes, elements, and edges are given in the following table.

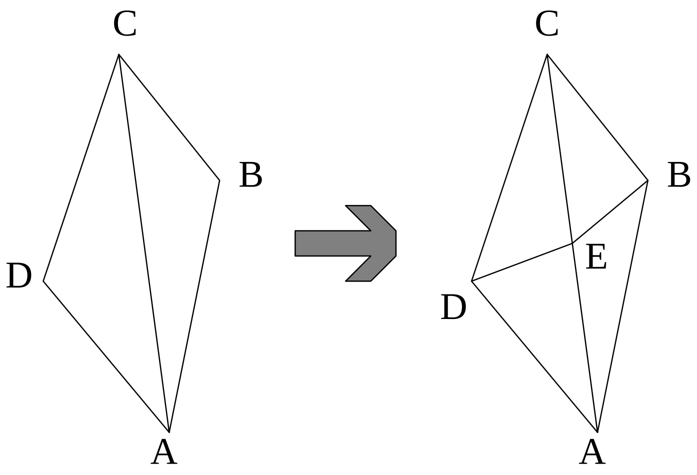


FIG. 6.5: Adding a node E at the center of AC.

type of divided edge	nodes	elements	edges
boundary	+1	+1	+2
non–boundary	+1	+2	+3

We must define the matrix M for the new node. If we insert E on the edge AC we put $M_E = \frac{1}{2}(M_A + M_C)$. We denote by $\mathcal{A}$ one application of the adding of nodes at the centers of all edges with the norms greater than $L_{\max}$ to a given triangulation (i.e., each edge can be divided during $\mathcal{A}$ only once).

6.5.2 Removing an edge

If the norm of some edge of a triangle is smaller than a prescribed constant $L_{\min}$ we remove this edge (Fig. 6.6). Here the norm of the edge AB is smaller than $L_{\min}$ and, therefore, this edge is removed, i.e., the nodes A, B pass into a new node E. The position of the new point E depends on the position of A and B. If A and B are not on the boundary of Ω we can place E at the center of AB. But if A is a boundary node and B is a non–boundary node we must put $E = A$ so that the boundary would not change. Similarly, if A is a non–boundary node and B is a boundary node we put $E = B$.

We use $L_{\min} = \sqrt{4/3}$, which follows from a similar consideration as in Section 6.5.1 and is based heuristically on the implication $\left|\|\boldsymbol{v}\|_M - \sqrt{3}\right| > \left|2\|\boldsymbol{v}\|_M - \sqrt{3}\right| \Rightarrow \|\boldsymbol{v}\|_M < \sqrt{4/3}$. The changes in the numbers of nodes, elements, and edges are given in the following table.

type of divided edge	nodes	elements	edges
boundary	-1	-1	-2
non–boundary	-1	-2	-3

We must define the matrix M for the new node. If we put E at the center of the edge AB

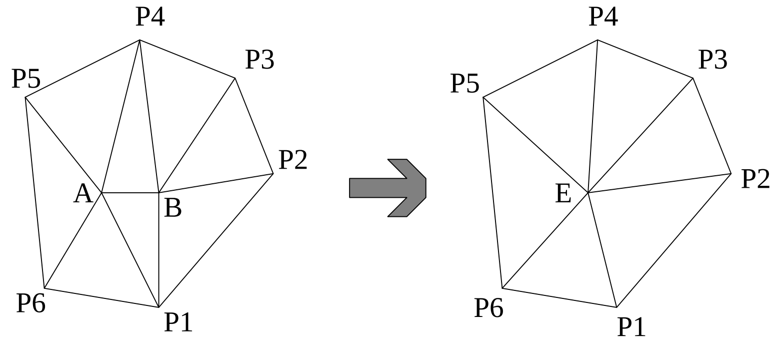


FIG. 6.6: Removing the edge AB and the rise a node E.

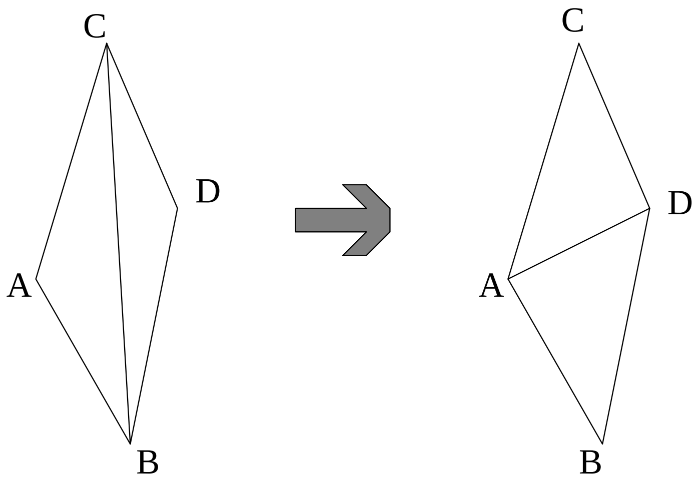


FIG. 6.7: Swapping the diagonal of the quadrilateral formed by pair of adjacent triangles.

then we set $M_E = (M_A + M_B)/2$. If we put $E = A$ (or B) we set $M_E = M_A$ (or $M_E = M_B$). We denote by $\mathcal{R}$ one application of the removing of all edges with norms smaller than $L_{\min}$ to a given triangulation.

6.5.3 Swapping the diagonal

Let ABC and BDC be two adjacent triangles with the common edge BC. Then we can replace the diagonal BC of the quadrilateral $ABCD$ by the diagonal AD and we obtain the triangles ABD and ADC (Fig. 6.7).

The swapping is performed only if the parameter quality $Q_{\mathcal{T}}$ decreases, i.e., $q_{ABD}+q_{ADC} < q_{ABC} + q_{BDC}$ (for the definition of q, see (6.18)). The numbers of nodes, elements, and edges remain the same. We denote by $\mathcal{S}$ the loop over elements consisting of the successive swapping of diagonals as long as $Q_{\mathcal{T}}$ decreases. (The positions of all nodes are fixed.)

6.5.4 Moving a node

The moving of nodes is not a discrete operation. This operation helps us to "smooth" the triangulation. Let P be a node of the triangulation $\mathcal{T}$. Let $P_1, P_2, \ldots, P_m$ be such nodes that PP_iP_{i+1}, $i = 1, \ldots, m$, $(P_{m+1} \equiv P_1)$ are the triangles from $\mathcal{T}$ (m is the number of

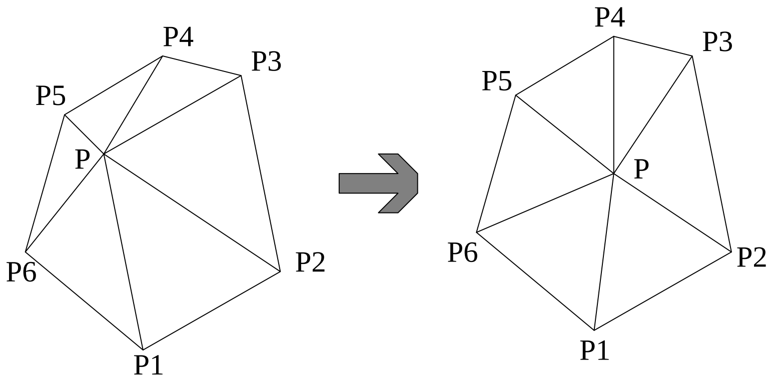


FIG. 6.8: Moving the node P to the new position.

these triangles, see Fig. 6.8). Now the nodes $P_1, P_2, \ldots, P_m$ are fixed and we move the node P. Let P_i^*, $i = 1, \ldots, m$, be such a point that the quality parameter q_{T^i} of the triangle $T^i = P_i^* P_i P_{i+1}$ is minimal for the given position of P_i, P_{i+1} and the given orientation of PP_iP_{i+1}. By the orientation of the triangle ABC we understand the sign of the determinant

$$\begin{pmatrix} x_A & y_A & 1 \\ x_B & y_B & 1 \\ x_C & y_C & 1 \end{pmatrix}, \tag{6.20}$$

where $[x_A, y_A], [x_B, y_B]$, and $[x_C, y_C]$ are the coordinates of A, B, and C. Using the approach given in [7], we have

$$P_i^* = \frac{\begin{pmatrix} d-b & -c \\ a & d+b \end{pmatrix} \cdot (P_{i+1} - P_i)^T}{2d}, \tag{6.21}$$

where $d = \left((ac - b^2)/3\right)^{1/2}$, and a, b, and c are the elements of the matrix M given by (6.7). Then we put

$$P^* = \frac{1}{m} \sum_{i=1}^{m} P_i^*. \tag{6.22}$$

Let $P'(s)$ be a point such that $P'(s) = P + s(P^* - P)$, $s \in [0, 1]$. If we move P into $P'(s)$, we obtain a triangulation $\mathcal{T}(s)$ with the quality parameter $Q_{\mathcal{T}}(s)$. We find $s^* \in [0, 1]$ such that $Q_{\mathcal{T}}(s^*) \leq Q_{\mathcal{T}}(s)$ for all $s \in [0, 1]$ and the new position of P is $P := P'(s^*)$.

The numbers of nodes, elements, and edges remain the same. We denote by $\mathcal{M}$ the application of moving of all nodes in such a way that the quality parameter $Q_{\mathcal{T}}$ is minimal for the given configuration (i.e., during the moving, any swapping, inserting, and removing is not performed).

6.5.5 Boundary nodes

We have to take into account the boundary of the physical domain Ω when adding a node, removing an edge, and moving a node. We have to obey some rules. After each local operation, the boundary of $\mathcal{T}$ must be a polygonal approximation of $\partial\Omega$. We have the list of points lying on $\partial\Omega$ for each problem and we select suitable nodes from this list.

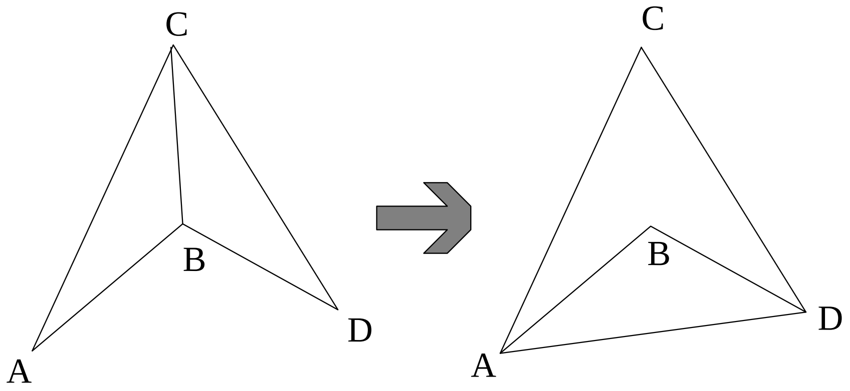


FIG. 6.9: Prohibited case for swapping.

6.5.6 Minimal angle condition

Many theorems concerning FVM and FEM require that the condition, that there exists a positive constant β such that the magnitude of all angles of all triangles is greater than β, be fulfilled. This condition is equivalent (see [8]) to the existence of $c_1 > 0$ such that

$$h_l^2 \leq c_1 |T|, \quad l = 1, 2, 3, \quad T \in \mathcal{T}, \tag{6.23}$$

where h_l, $l = 1, 2, 3$, are the lengths of edges of T. This condition should be satisfied also for the triangulation generated by AMA. Therefore, each of the local operations is performed if for a given $\gamma > 0$, the condition

$$\frac{\det \begin{pmatrix} x_A & y_A & 1 \\ x_B & y_B & 1 \\ x_C & y_C & 1 \end{pmatrix}}{|V_A - V_B|^2 + |V_B - V_C|^2 + |V_C - V_A|^2} \geq \gamma \tag{6.24}$$

is satisfied for each triangle which arises after a local operation. Here, $V_A = (x_A, y_A), V_B = (x_B, y_B)$, and $V_C = (x_C, y_C)$ are the vertices of a triangle from $\mathcal{T}$. The magnitude of the determinant in numerator is the double area of the triangle $V_A V_B V_C$. This condition also verifies whether the orientation of nodes in a triangle is preserved (for the definition of orientation see the text before (6.20)). This can appear in the swapping where we must avoid the case described in Fig. 6.9. The swapping is prohibited in Fig. 6.9 in virtue of (6.24) because the determinant of the triangle ABD is negative in this case. In our code we put $\gamma = 0.12$.

6.5.7 Iterative process

We have tried to find the optimal strategy by the combination of the above operations $\mathcal{A}, \mathcal{R}, \mathcal{S}$, and $\mathcal{M}$. Based on numerical experiments, one of the best strategies seems to be the process represented by the formula

$$\mathcal{S} + m \times [\mathcal{M} + \mathcal{S} + \mathcal{R} + \mathcal{M} + \mathcal{A} + \mathcal{M} + \mathcal{S}] + \mathcal{M} \tag{6.25}$$

(the sequence of operations goes from the left to the right), where m is an integer denoting the number of repetitions of the sequence of operations in the square brackets and its value is chosen in such a way that the sequence in the square bracket is applied until all local operations are performed, i.e., in the last sequence of operations, in the square brackets no edge has been removed, no node has been inserted, and no diagonal has been swapped. We prescribe also a constant $m_{\max}$ and the iterative process is finished if $m \geq m_{\max}$. Our experience gives m equal approximately to 10.

6.6 Generation of the matrix M

The previous section describes the construction of a triangulation $\mathcal{T}_{\text{new}}$ if we know $\mathcal{T}_{\text{old}}$ and the matrices M_i (symmetric and positive definite) are given for each node of $\mathcal{T}_{\text{old}}$. Now we define how to obtain the matrices M_i. We have defined the interpolation error in Section 6.2 which can be represented by a Hessian matrix. This matrix is used for the generation of M_i, but we must avoid the problems mentioned in Section 6.2 (see also [12]).

Let $\boldsymbol{H}_i,\ i \in I$, be a Hessian of some physical quantity η which is characteristic for the problem of consideration (for example, in CFD we put $\eta := \rho$, where ρ is a density of flow). We decompose $\boldsymbol{H}_i$ in the following way:

$$\boldsymbol{H}_i = \left. \begin{pmatrix} \frac{\partial^2 \eta}{\partial x^2} & \frac{\partial^2 \eta}{\partial x \partial y} \\ \frac{\partial^2 \eta}{\partial x \partial y} & \frac{\partial^2 \eta}{\partial y^2} \end{pmatrix} \right|_{P_i} = R_i \begin{pmatrix} \Lambda_{i1} & 0 \\ 0 & \Lambda_{i2} \end{pmatrix} R_i^{-1} \quad i \in I. \tag{6.26}$$

(For the computation of the second order derivatives see Section 6.6.1). The matrices $\boldsymbol{H}_i$ are not suitable as the Riemann metric, because they are not positive definite. Therefore, we put

$$H_i = R_i \begin{pmatrix} |\Lambda_{i1}| & 0 \\ 0 & |\Lambda_{i2}| \end{pmatrix} R_i^{-1} \quad i \in I. \tag{6.27}$$

The matrices $H_i,\ i \in I$ are symmetric and positive definite. It follows from (6.26) and (6.27) , that if η is a polynom of degree less than 2 then $H_i = 0,\ i \in I$. For this case we require to have an uniform triangulation (i.e. $M_i = cI$ for all $i \in I$). This and the numerical experiments lead us to put

$$M_i = c\,[I + \alpha(\|H_i\|)H_i)] \quad i \in I, \tag{6.28}$$

where c is a constant and $\alpha(\|H_i\|)$ is a function of the norms of the matrix H_i. We put $\|H\| = \max_{i,j=1,2} |h_{ij}|$, where $h_{ij},\ i,j = 1,2$ are the elements of H.

Now we determine c and α. For $H_i = 0$ for all $i \in I$ (the uniform mesh) both eigenvalues of M_i are equal to c and therefore the optimal triangulation consists of equilateral triangles with the length edge $\sqrt{3/c}$. Then

$$|T_{M_i}| = \frac{3\sqrt{3}}{4c}. \tag{6.29}$$

Let $|\Omega|$ denote the area of the computational domain and $nelem$ be the number of elements which we prescribe for a uniform mesh. If all triangles of the optimal triangulation are the same then we have

$$nelem \cdot |T_{M_i}| = |\Omega|. \tag{6.30}$$

This and (6.29) imply that

$$nelem \cdot \frac{3\sqrt{3}}{4c} = |\Omega| \tag{6.31}$$

and, hence,

$$c = \frac{3\sqrt{3}}{4} \frac{nelem}{|\Omega|}. \tag{6.32}$$

For the function $\alpha(\|H\|)$ we choose an empiric relation

$$\alpha(\|H_i\|) = \frac{\varepsilon_1}{\varepsilon_2 + \|H_i\|}, \quad i \in I, \tag{6.33}$$

where ε_1 and ε_2 are some constants. The relations (6.28) and (6.33) give

$$\lim_{\|H_i\| \to \infty} \|M_i\| = c(1 + \varepsilon_1) =: d_{\max} \quad \forall i \in I, \tag{6.34}$$

so the norms of M_i, $i \in I$ are bounded and therefore the lengths of edges of $\mathcal{T}$ are bounded from below.

Now let us determine the constants $\varepsilon_1, \varepsilon_2$. Let $H_i = 0$ for some $i \in I$. Then $M_i = cI$ and the corresponding optimal triangle has equal sides with length(in Euclidean norm)

$$l_0 = \sqrt{\frac{3}{c}}. \tag{6.35}$$

Let $H_i = \beta I$ for some $i \in I$ and let $\beta \gg 1$. Then $M_i \to c(1 + \varepsilon_1)I$ for $\beta \to \infty$, as follows from (6.34) and the corresponding optimal triangle has equal sides with length(in Euclidean norm)

$$l_1 = \sqrt{\frac{3}{c(1 + \varepsilon_1)}}. \tag{6.36}$$

It follows from (6.35) and (6.36) that the length l of any edge from the optimal triangulation $\mathcal{T}_{\mathrm{opt}}$ satisfies (provided M_i is defined by (6.28) and (6.33))

$$l_1 = \sqrt{\frac{3}{c(1 + \varepsilon_1)}} \le l \le \sqrt{\frac{3}{c}} = l_0. \tag{6.37}$$

It can be shown that the values l_0 and l_1 are the maximal and minimal lengths of an edge of an optimal triangulation, respectively. The value l_0 is given by (6.35) and (6.30) (where we prescribe $nelem$). It follows from (6.37) that

$$\frac{l_0}{l_1} = \sqrt{c(1 + \varepsilon_1)} \Longrightarrow \varepsilon_1 = \left(\frac{l_0}{l_1}\right)^2 - 1. \tag{6.38}$$

Relation (6.38) shows what role plays the constant ε_1. The quotient $\frac{l_0}{l_1}$ performs the ratio between the longest and the shortest edge admissible in the optimal triangulation. If we prescribe the ratio $\frac{l_0}{l_1}$ (we set in our code $\frac{l_0}{l_1} = 100$), then (6.38) gives the value of ε_1. For ε_2 we use the following empiric relation

$$\varepsilon_2 = \frac{1}{10} |\delta_{\max}| \frac{(\varepsilon_1 + 1)(1 - p)}{p(\varepsilon_1 + 1) - 1}, \tag{6.39}$$

where $|\delta_{\max}| = \frac{1}{l_1^2}$ which estimates the maximal second order derivative for the optimal triangulation and $p \in (0, 1)$. We put $p = 0.8 - 0.95$. This condition prescribes that for $\|H\| = |\delta_{\max}|/10$ we have $\|M\| = pc(1 + \varepsilon_1) = pd_{\max}$. The significance of (6.39) can be seen on Fig. 6.10.

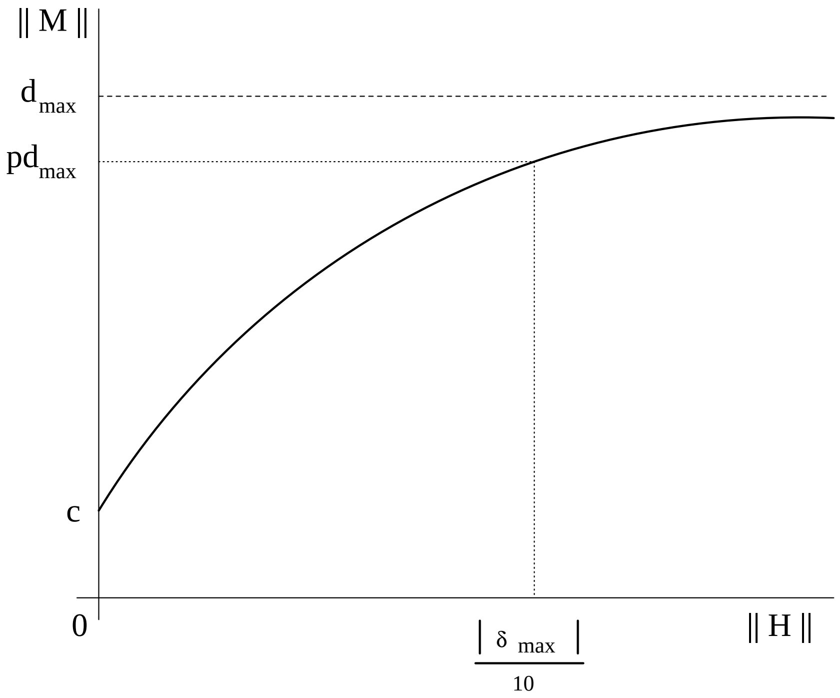


FIG. 6.10: Dependence $\|M\|$ on $\|H\|$ given by (6.28) and (6.33).

6.6.1 Computation of the second order derivatives

We need to know the values of the second order derivative of the function η at each node of $\mathcal{T}$ for the relations (6.26). We approximate η by a piecewise linear function η_h. Let X_h be a space of continuous and piecewise linear functions on each triangle from $\mathcal{T}$ and $V_h = \{\psi,\ \psi \in X_h,\ \psi|_{\partial\Omega} = 0\}$. Let $\eta_h \in X_h$. It is necessary to evaluate the second order derivative of η_h, i.e. $\partial^2\eta_h/\partial x^2$, $\partial^2\eta_h/\partial x\partial y$ and $\partial^2\eta_h/\partial y^2$ which do not exist in the classical sense. We approximate $\partial^2\eta_h/\partial x^2$ by a function $D^2\eta_h \in X_h$ such that

$$\left(\frac{\partial^2\eta_h}{\partial x^2}, \varphi\right) \approx \left(D^2\eta_h, \varphi\right) = -\left(\frac{\partial\eta_h}{\partial x}\frac{\partial\varphi}{\partial x}\right) \quad \forall\varphi \in V_h, \tag{6.40}$$

where $(\cdot,\cdot)$ means the scalar product in $L^2(\Omega_h)$. For boundary nodes we use an extrapolation from the interior part of the computational domain. The R.H.S. of (6.40) has already sense because $\eta_h, \varphi \in V_h$ and therefore the first derivatives exist and are piecewise constant. We choose the special subset of test functions $\{\varphi_i, \varphi_i \in V_h, \varphi_i(P_j) = \delta_{ij}, i,j \in I\}$, where $P_j,\ j \in I$ are nodes from $\mathcal{T}$ and δ_{ij} is the Kronecker delta.

The R.H.S. of (6.40) gives

$$\left(\frac{\partial\eta_h}{\partial x}, \frac{\partial\varphi_i}{\partial x}\right) = \sum_{T\in\mathcal{T}} |T| \left.\frac{\partial\eta_h}{\partial x}\right|_T \left.\frac{\partial\varphi_i}{\partial x}\right|_T = \sum_{T\in d(i)} |T| \left.\frac{\partial\eta_h}{\partial x}\right|_T \left.\frac{\partial\varphi_i}{\partial x}\right|_T \quad i \in I, \tag{6.41}$$

where $d(i),\ i \in I$ is a set of triangles which has P_i as a vertex.

The L.H.S. of (6.40) gives

$$\left(D^2\eta_h, \varphi_i\right) = \sum_{T\in\mathcal{T}} \int_T D^2\eta_h\varphi_i\ dx. \tag{6.42}$$

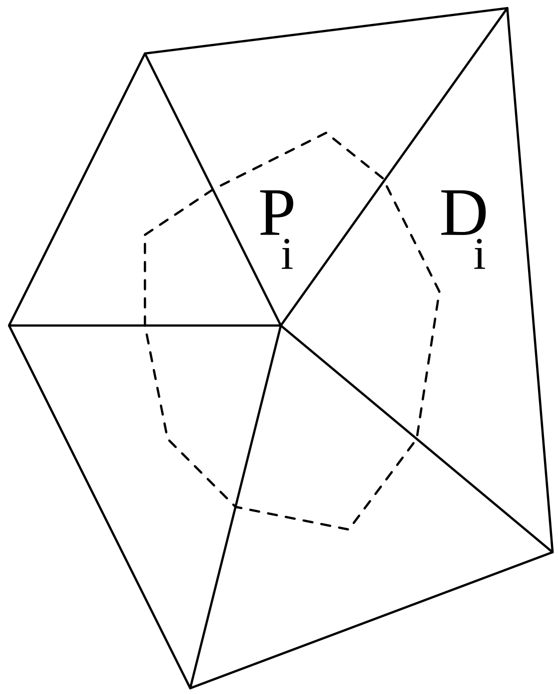


FIG. 6.11: The control volume D_i corresponding to node P_i.

We use the "mass lumping" numerical quadrature

$$\int_T F\ dx \approx \frac{1}{3}|T| \sum_{n=1}^{3} F(P_n^T), \tag{6.43}$$

where $F \in C(T)$ and $P_n^T,\ n = 1,2,3$ are the vertices of T. From (6.42) and (6.43) we have

$$\begin{aligned} \left(D^2\eta_h, \varphi_i\right) &= \sum_{T\in\mathcal{T}} \frac{1}{3}|T| \sum_{m=1}^{3} D^2\eta_h(P_m^T)\varphi_i(P_m^T) = \\ &= \sum_{T\in d(i)} \frac{1}{3}|T| \sum_{m=1}^{3} D^2\eta_h(P_m^T)\varphi_i(P_m^T) = \frac{1}{3}D^2\eta_h(P_i) \sum_{T\in d(i)} |T|. \end{aligned} \tag{6.44}$$

It is easy to show that $\frac{1}{3}\sum_{T\in d(i)} |T| = |D_i|$, where $|D(i)|$ is the area of the so-called dual finite volume (Fig. 6.11) associated with the vertex P_i (see, e.g. [18],[20]).

The relations (6.40), (6.41) and (6.44) yield

$$\frac{\partial^2\eta_h}{\partial x^2}(P_i) \approx D^2\eta_h(P_i) = -\frac{1}{|D_i|} \sum_{T\in d(i)} |T| \frac{\partial\eta_h}{\partial x}\bigg|_T \frac{\partial\varphi_i}{\partial x}\bigg|_T \quad i \in I. \tag{6.45}$$

Similarly we obtain

$$\frac{\partial^2\eta_h}{\partial x\partial y}(P_i) \approx -\frac{1}{|D_i|} \sum_{T\in d(i)} |T| \frac{\partial\eta_h}{\partial x}\bigg|_T \frac{\partial\varphi_i}{\partial y}\bigg|_T \quad i \in I, \tag{6.46}$$

$$\frac{\partial^2 \eta_h}{\partial y \partial x}(P_i) \approx -\frac{1}{|D_i|} \sum_{T \in d(i)} |T| \frac{\partial \eta_h}{\partial y}\bigg|_T \frac{\partial \varphi_i}{\partial x}\bigg|_T \quad i \in I, \tag{6.47}$$

and

$$\frac{\partial^2 \eta_h}{\partial y^2}(P_i) \approx -\frac{1}{|D_i|} \sum_{T \in d(i)} |T| \frac{\partial \eta_h}{\partial y}\bigg|_T \frac{\partial \varphi_i}{\partial y}\bigg|_T \quad i \in I. \tag{6.48}$$

We also need to prove that (6.46) and (6.47) give the same value, i.e. we must prove

$$\begin{aligned} \left(\left(\frac{\partial \eta_h}{\partial x}, \frac{\partial \varphi_i}{\partial y} \right) = \right) \sum_{T \in d(i)} |T| \frac{\partial \eta_h}{\partial x}\bigg|_T \frac{\partial \varphi_i}{\partial y}\bigg|_T &= \\ = \sum_{T \in d(i)} |T| \frac{\partial \eta_h}{\partial y}\bigg|_T \frac{\partial \varphi_i}{\partial x}\bigg|_T & \left(= \left(\frac{\partial \eta_h}{\partial y}, \frac{\partial \varphi_i}{\partial x} \right) \right) \quad \forall i \in I, \end{aligned} \tag{6.49}$$

As $\eta_h \in X_h$ then we can write

$$\eta_h(x) = \sum_{j \in I} \eta_h(P_j) \varphi_j(x). \tag{6.50}$$

Therefore it is enough to prove that

$$\left(\frac{\partial \varphi_i}{\partial x}, \frac{\partial \varphi_j}{\partial y} \right) = \left(\frac{\partial \varphi_i}{\partial y}, \frac{\partial \varphi_j}{\partial x} \right) \qquad i, j \in I. \tag{6.51}$$

For the case $i = j$ the relation (6.51) is evident. If P_i and P_j are not vertices of a common triangle T then the scalar products in (6.51) are equal to 0 because the supports of the functions φ_i and φ_j are disjoint. Now it is enough to prove (6.49) for the case when P_i and P_j form an edge of some $T \in \mathcal{T}$ (see Fig. 6.12).

Let $P_i = (x_i, y_i)$, $P_j = (x_j, y_j)$, $P_k = (x_k, y_k)$, $P_l = (x_l, y_l)$ are nodes and T_1, T_2 are neighbouring triangles with the common edge $P_i P_j$, see Fig. 6.12. Then the relation (6.51) can be rewritten as

$$\int_{T_1} \frac{\partial \varphi_i}{\partial x} \frac{\partial \varphi_j}{\partial y}\, dx + \int_{T_2} \frac{\partial \varphi_i}{\partial x} \frac{\partial \varphi_j}{\partial y}\, dx = \int_{T_1} \frac{\partial \varphi_i}{\partial y} \frac{\partial \varphi_j}{\partial x}\, dx + \int_{T_2} \frac{\partial \varphi_i}{\partial y} \frac{\partial \varphi_j}{\partial x}\, dx. \tag{6.52}$$

Using the standard results (Cf., e.g., [23], Section 4.4) we obtain

$$\begin{aligned} \frac{\partial \varphi_i}{\partial x}\bigg|_{T_1} &= \frac{y_k - y_j}{2|T_1|}, & \frac{\partial \varphi_i}{\partial x}\bigg|_{T_2} &= \frac{y_j - y_l}{2|T_2|}, & \frac{\partial \varphi_i}{\partial y}\bigg|_{T_1} &= \frac{x_j - x_k}{2|T_1|}, & \frac{\partial \varphi_i}{\partial y}\bigg|_{T_2} &= \frac{x_l - x_j}{2|T_2|} \\ \frac{\partial \varphi_j}{\partial x}\bigg|_{T_1} &= \frac{y_i - y_k}{2|T_1|}, & \frac{\partial \varphi_j}{\partial x}\bigg|_{T_2} &= \frac{y_l - y_i}{2|T_2|}, & \frac{\partial \varphi_j}{\partial y}\bigg|_{T_1} &= \frac{x_k - x_i}{2|T_1|}, & \frac{\partial \varphi_j}{\partial y}\bigg|_{T_2} &= \frac{x_i - x_l}{2|T_2|}, \end{aligned} \tag{6.53}$$

where

$$2|T_1| = \begin{vmatrix} x_i & y_i & 1 \\ x_k & y_k & 1 \\ x_j & y_j & 1 \end{vmatrix}, \qquad 2|T_2| = \begin{vmatrix} x_i & y_i & 1 \\ x_j & y_j & 1 \\ x_l & y_l & 1 \end{vmatrix}. \tag{6.54}$$

Now putting (6.53) and (6.54) into (6.52), we arrive after a lengthy calculation at (6.49), which we wanted to prove.

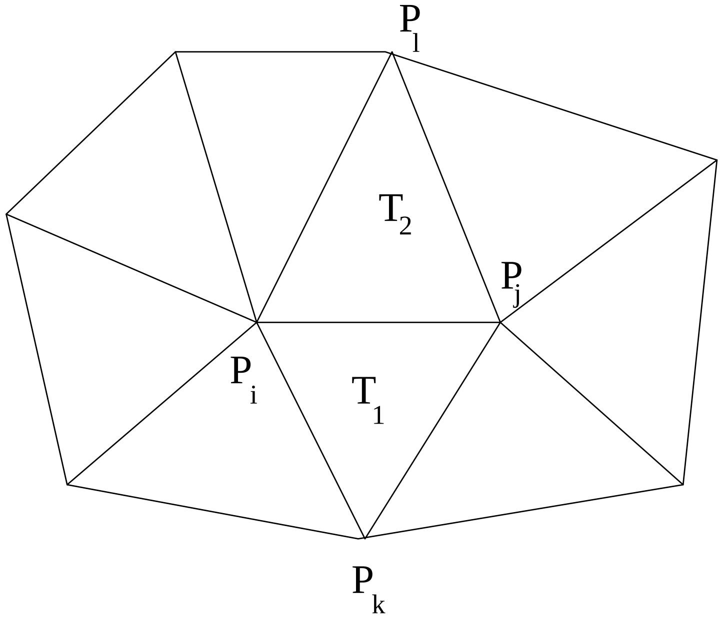


FIG. 6.12: The notation for the prove of (6.49). Triangle T_1 formed by $P_iP_jP_l$ and T_2 by $P_jP_iP_k$.

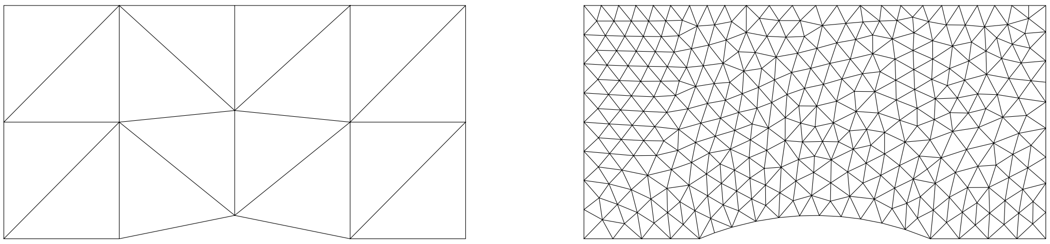


FIG. 6.13: Initial mesh for AMA (left) and the mesh $\mathcal{T}_0$ constructed by AMA (right).

6.7 Multilevel computation

The technique described above is applied in the following way: We start from an initial very coarse triangulation of a suitable polygonal approximation of the computational domain (for example, see Fig. 6.13). Now we can generate a triangulation $\mathcal{T}_0$, using AMA described in Section 6.5, as a uniform triangulation ($M_i = cI,\ i \in I$) with c given by (6.30), where *nelem* represents the number of elements which we prescribe for the triangulation $\mathcal{T}_0$ (for example 500, see Fig. 6.13). In practice, during the generation of a triangulation we do not achieve the prescribed number of elements. The number of triangles is 16 for the initial triangulation and 536 for $\mathcal{T}_0$ (Fig. 6.13).

We compute the solution W_0 on the triangulation $\mathcal{T}_0$, using a numerical method suitable for the problem of consideration (FVM, FEM, ...). With the aid of (6.26) – (6.28), we compute the matrices M_i from W_0 at each node of $\mathcal{T}_0$ and, using the iterative process from Section 6.5, we construct the triangulation $\mathcal{T}_1$. The results from $\mathcal{T}_0$ are transferred onto $\mathcal{T}_1$ and the computation can continue. Then we obtain W_1. (In general we pass from $\mathcal{T}_k$ to $\mathcal{T}_{k+1}$ in the same way.) We obtain, consecutively, $\mathcal{T}_0, \mathcal{T}_1, \mathcal{T}_2, \ldots$ and $W_0, W_1, W_2, \ldots$. There is a natural question when to stop this process. Numerical experiments show that for given $nelem$, ε_1, ε_2, "the computation converges", i.e., there exists a number k for which the triangulations $\mathcal{T}_k$ and $\mathcal{T}_{k+1}$ are (approximately) the same. The number k is approximately between 5 and 8 for our problems. From numerical experiments it follows that the generation of triangulations using AMA needs at most 5% of the total CPU–time of the complete computational process.

6.8 Numerical examples

Here we are concerned with inviscid compressible flow through the GAMM channel with the inlet Mach number $M_{\rm in} = 0.67$. We present the mesh obtained after 4 refinements (Fig. 6.14), the corresponding isolines of the Mach number (Fig. 6.15), and the color map of the Mach number (Fig. 6.16). We use $nelem = 500$, $\varepsilon_1 = 10000$, and $p = 0.98$. Fig. 6.17, Fig. 6.18, and Fig. 6.19 show the results obtained by the standard red–green refinement technique based on a shock indicator [27]. Fig. 6.20 presents the distributions of the Mach number along the solid walls.

The following table compares both of the adaptative methods:

type of the method	AMA	red – green
CPU – time	1570 s	2274 s
total number of levels of mesh adaptations	4	4
number of elements for the last level of adaptation	4601	4731

We see that whereas the numbers of elements for the last level of refinement are approximately the same, the CPU–time is essentially lower for the AMA technique. Moreover, Figs. 6.15, 6.16, 6.18, 6.19 and 6.20 show that the sharpness of shock waves is greater for AMA, i.e., the results obtained by AMA have a better quality.

As the second example we present the simulation of the flow around NACA 0012 profile with the inlet Mach number $M_{\rm in} = 0.8$ and angle of attack $\alpha = 0$. An adaptation was performed of the mesh by AMA with $c = 500$, $\varepsilon_1 = 10000$ and $p = 0.95$. The Figs. 6.21, 6.22, 6.23, 6.24, and 6.25 show the triangulations and the maps of the Mach number of the successive mesh adaptation for the zeroth, first, second, fourth, and sixth level of refinement, respectively. Wee see that from very poor results obtained for the initial triangulation we obtain good results with very sharp shock waves. The results for the fourth and sixth level of refinement (Fig. 6.24 and Fig. 6.25) are almost identical, which corresponds to the assertion that the AMA algorithm "converges" (see Section 6.7). A slight asymmetry of the maps of the Mach number is given by the asymmetry of triangulation, which follows from the fact

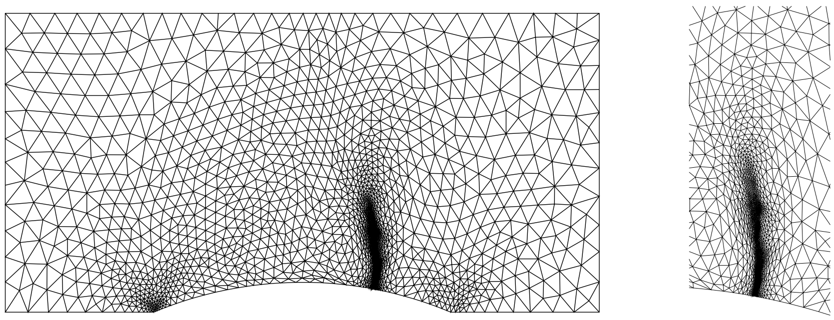


Fig. 6.14: Mesh after fourth refinement and its detail obtained by AMA.

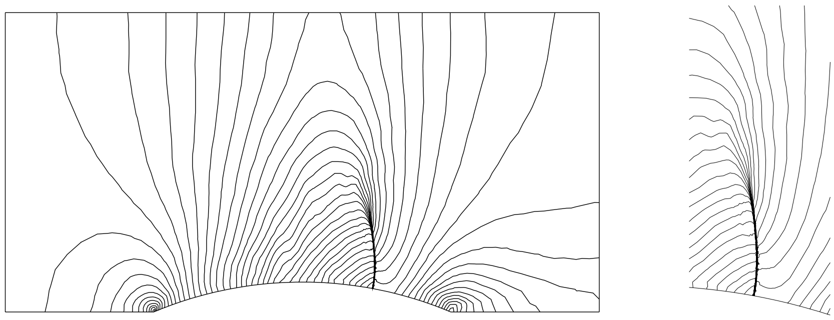


Fig. 6.15: Isolines of Mach number after fourth refinement and its detail obtained by AMA.

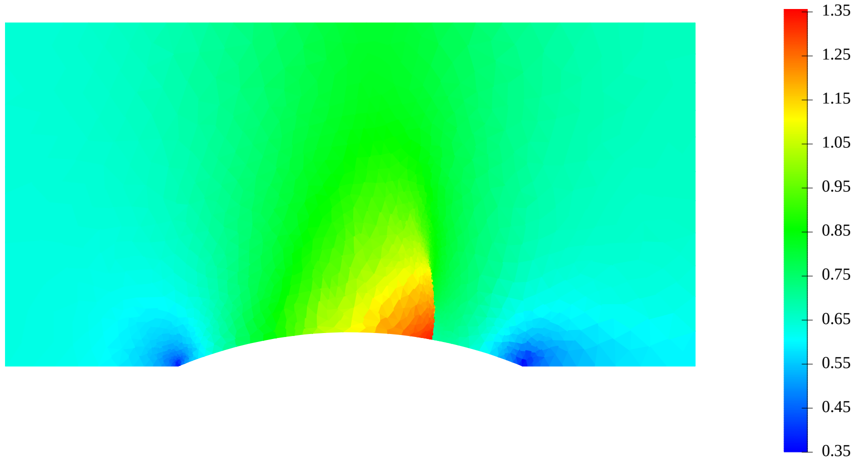


Fig. 6.16: The map of piecewise linear distribution of Mach number after fourth refinement obtained by AMA.

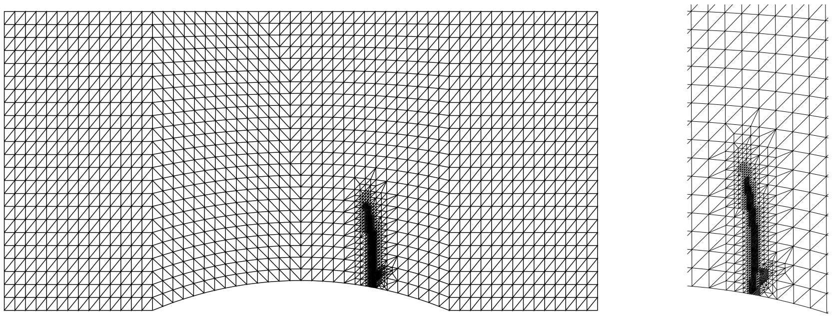


Fig. 6.17: Mesh after fourth refinement and its detail obtained by red–green refinement.

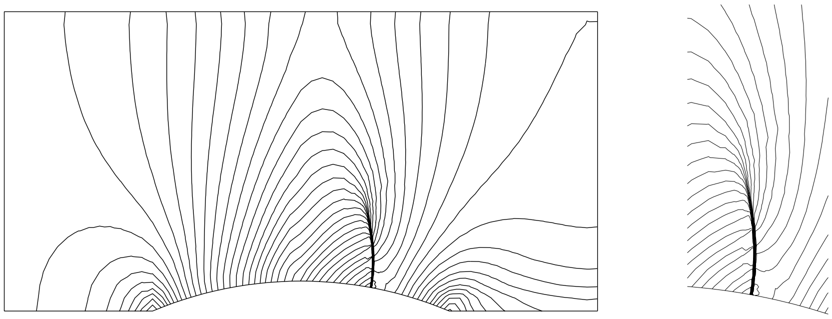


Fig. 6.18: Isolines of Mach number after fourth refinement and its detail obtained by red–green refinement.

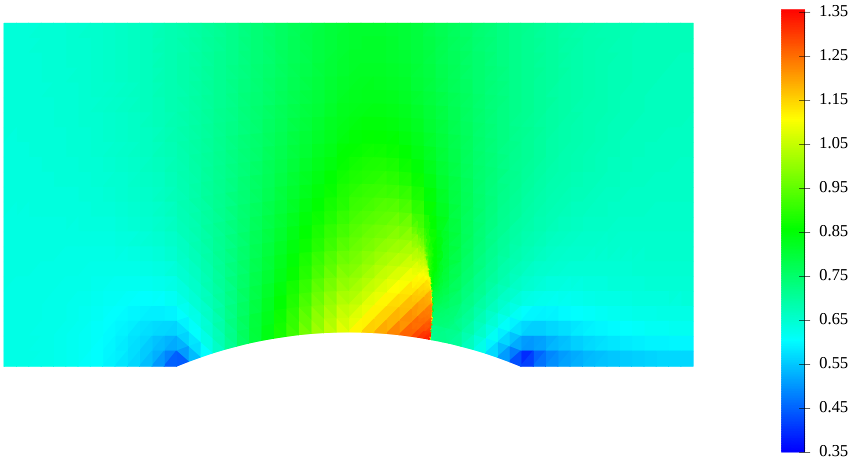


Fig. 6.19: The map of piecewise linear distribution of Mach number after fourth refinement obtained by red–green refinement.

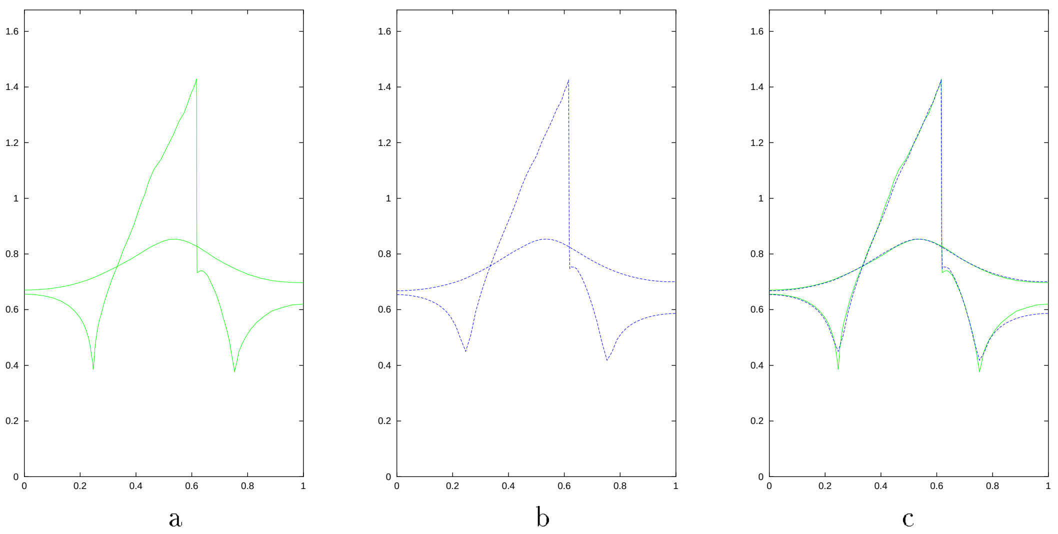


FIG. 6.20: Distribution of Mach number a) obtained by AMA, b) obtained by red – green refinement, c) both of them.

that AMA is an iterative method. To avoid this asymmetry it is necessary to construct the AMA code in such a way that the symmetry prescribed should be preserved.

Acknowledgements

This work was supported by Grant No. 201/96/0313 of the Grant Agency of Czech Republic. The author gratefully acknowledges this support.

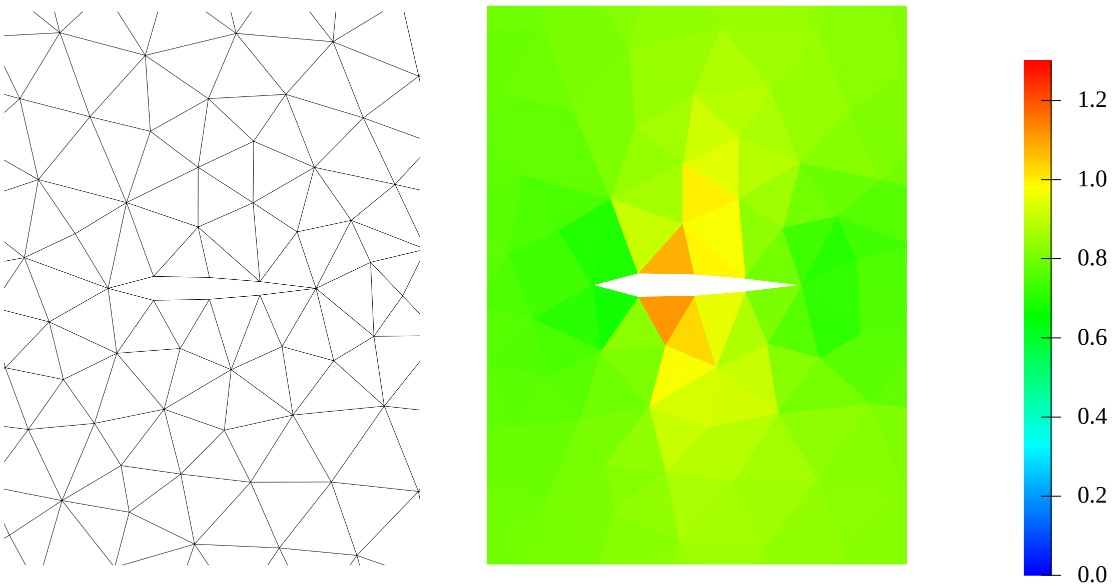


FIG. 6.21: Mesh and map of Mach number around NACA0012 after the zeroth adaptation.

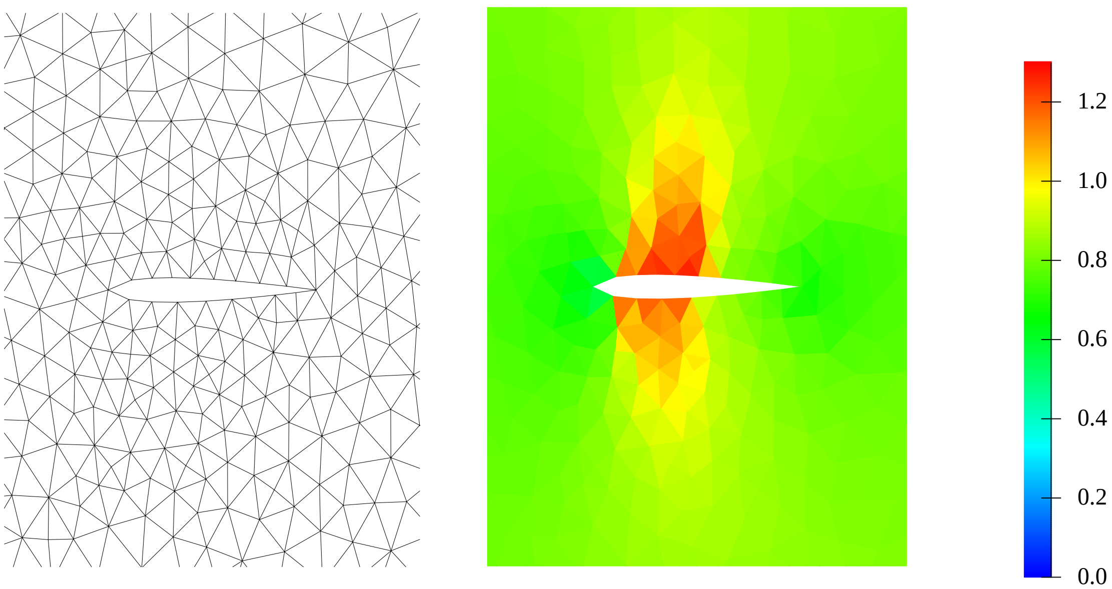


FIG. 6.22: Mesh and map of Mach number around NACA0012 the after first adaptation.

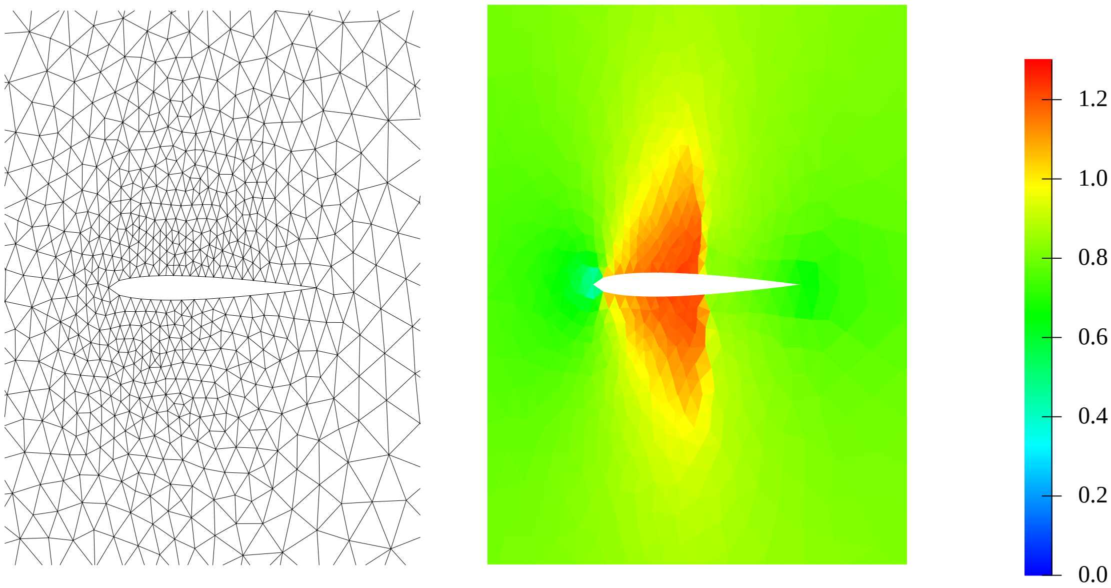


FIG. 6.23: Mesh and map of Mach number around NACA0012 after the second adaptation.

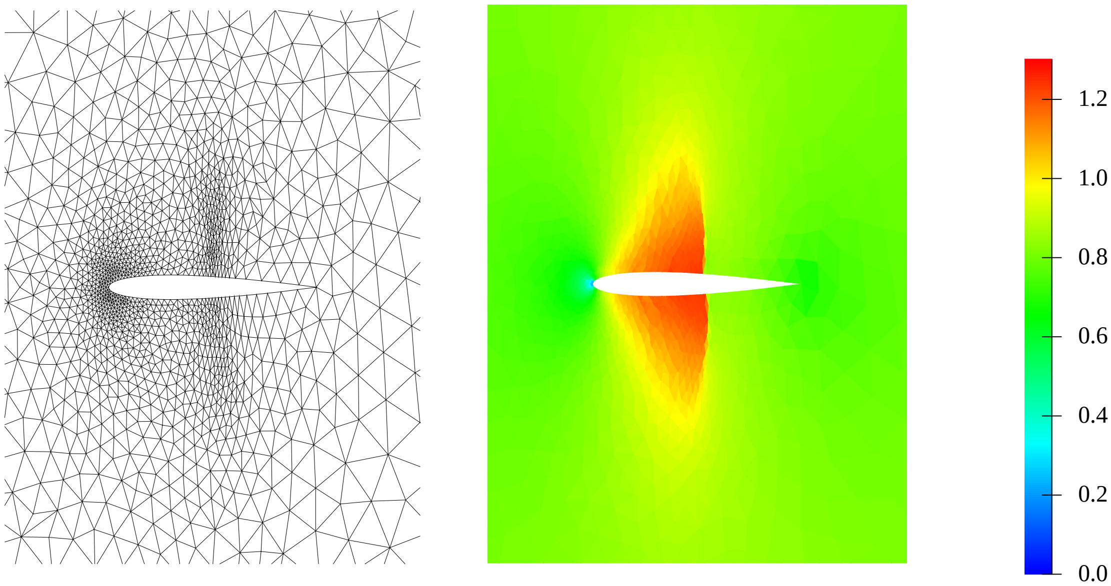


FIG. 6.24: Mesh and map of Mach number around NACA0012 after the fourth adaptation.

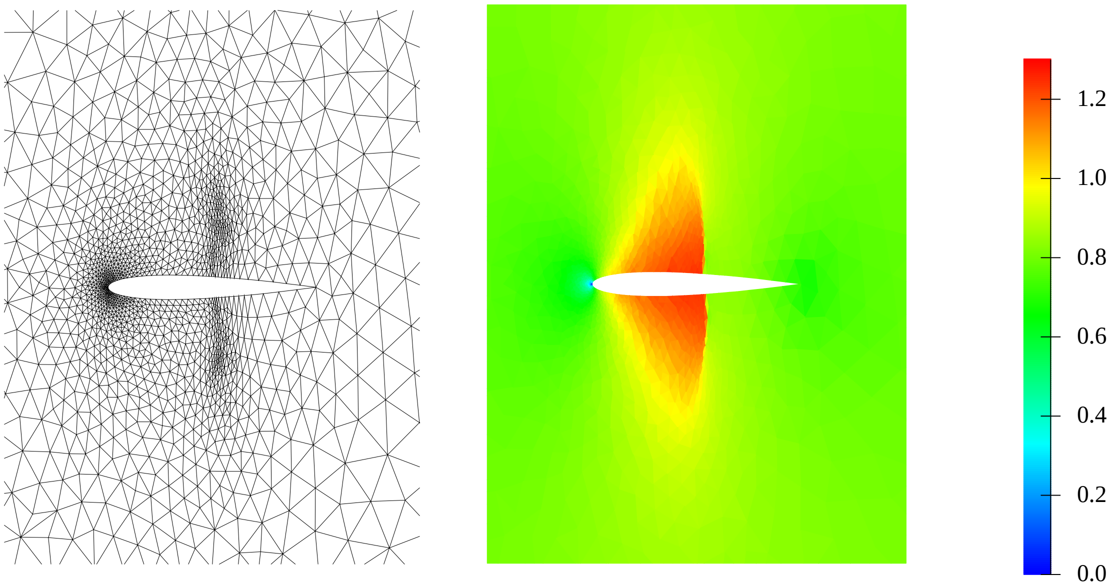


FIG. 6.25: Mesh and map of Mach number around NACA0012 after the sixth adaptation.

Troisième partie

Analyse théorique

Dans la première partie nous avons décrit les méthodes pour la résolution des problèmes de l'écoulement compressible. Une question naturelle est de savoir, si les méthodes sont convergentes. L'absence de l'existence de la solution des équation de Navier–Stokes ne permet pas de répondre avec succès à cette question. Pour cette raison l'analyse numérique des méthodes est effectuée pour le problème convectif–diffusif scalaire: Soit $\Omega \subset I\!R^2$, $T > 0$, on cherche la fonction $u(x,t) : \Omega \times (0,T) \Rightarrow I\!R$, pour laquelle

$$\frac{\partial u}{\partial t} + \sum_{s=1}^{2} \frac{\partial f_s}{\partial x_s} - \nu \Delta u = g \tag{6.55}$$

avec les conditions de frontières et initiales

$$\begin{aligned} & u|_{\partial\Omega \times (0,T)} = 0, \\ & u(x,0) = u_0(x), \quad x \in \Omega, \end{aligned} \tag{6.56}$$

où $\nu > 0$ est une constante donnée et $f_s : I\!R \to I\!R,\ s = 1,2$, $g : \Omega \times (0,T) \to I\!R$, $u_0 : \Omega \to I\!R$ sont les fonctions données. Il est possible de démontrer que la solution faible du problème (6.55) – (6.56) existe et est unique.

Nous avons appliqué notre méthode combinant des volumes finis (VF) barycentriques et des éléments finis (EF) nonconformes au problème (6.55) – (6.56) et on a prouver la convergence de la solution approchée vers la solution précise (Chap. 7). Ceci est un'argument en farent des méthodes de discretisation proposées pour les équations de Navier–Stokes.

Chapitre 7

Analyse de la méthode combinant des volumes finis barycentriques et des éléments finis non conformes pour le problème non linéaire convectif–diffusif

La démonstration de la convergence de la méthode utilisant des volumes finis barycentriques et des éléments finis non conformes a été accepté pour la présentation dans *"Applications of Mathematics"* num. 43/98.

Analysis of a Combined Barycentric Finite Volume – Nonconforming Finite Element Method for Nonlinear Convection – Diffusion Problems

Philippe Angot, Marseille,
Vít Dolejší, Miloslav Feistauer, Jiří Felcman, Praha

(Received December 10, 1997)

Abstract. We present the convergence analysis of an efficient numerical method for the solution of an initial–boundary value problem for a scalar nonlinear conservation law equation with a diffusion term. Nonlinear convective terms are approximated with the aid of a monotone finite volume scheme considered over the finite volume barycentric mesh, whereas the diffusion term is discretized by piecewise linear nonconforming triangular finite elements. Under the assumption that the triangulations are of weakly acute type, with the aid of the discrete maximum principle, a priori estimates and some compactness arguments based on the use of the Fourier transform with respect to time, the convergence of the approximate solutions to the exact solution is proved, provided the mesh size tends to zero.

Key words: nonlinear convection – diffusion problem, barycentric finite volumes, Crouzeix – Raviart nonconforming piecewise linear finite elements, monotone finite volume scheme, discrete maximum principle, a priori estimates, convergence of the method.

AMS Subject Classification: 65M12, 65M50, 35k60, 76M10, 76M25

1. Introduction

Many processes in science and technology are described by convection – diffusion equations with convection dominating over diffusion. We can mention, e.g., processes of fluid dynamics, hydrology and environmental protection. There is an extensive literature on the numerical solution of convection – diffusion problems. Let us mention, e.g., the papers [1], [2], [36], [37], [50], [55], [63], [68], [69], the monographs [48], [54] and the references therein, devoted mainly to linear problems. The main difficulty which must be overcome is the accurate resolution of the so-called boundary layers. If the equation under consideration represents a nonlinear conservation law with a small dissipation, then beside boundary layers also shock waves appear (slightly smeared due to dissipation). This is particularly the case of the system describing the viscous gas flow.

In [13], [18], [19], [20] we developed numerical methods for the solution of high – speed viscous compressible flow in domains with complex geometry. These methods are based on the combination of a finite volume scheme for the discretization of inviscid convective terms and the finite element discretization of viscous terms. Numerical experiments proved the efficiency and robustness of these methods with respect to the precise resolution of boundary layers and shock capturing. (For the finite volume solution of an inviscid gas flow see, e.g., [3], [17], [25], [26], [28], [29], [31], [40], [66]). Since the complete viscous gas flow problem is rather complex, the theoretical analysis of the combined finite volume – finite element method has been carried out for the case of a simplified scalar nonlinear conservation law equation with a small dissipation which is the simplest prototype of the compressible Navier – Stokes equations. Papers [21], [22], [24] are concerned with the convergence and error estimates for the method using dual finite volumes over a triangular mesh combined with conforming piecewise linear triangular finite elements.

Another possibility is the combination of the so-called barycentric finite volumes constructed over a triangular grid with the well–known Crouzeix – Raviart nonconforming piecewise linear finite elements used for the numerical solution of incompressible viscous flow ([9], [17], [62]). The upwind version of the Crouzeix – Raviart finite element method was developed and analyzed in [50] for a linear stationary convection – diffusion equation. This was the inspiration for Schieweck and Tobiska who investigated in [55] upwind schemes for the steady incompressible Navier – Stokes equations.

In the present paper we are concerned with the convergence analysis of the combined barycentric finite volume – nonconforming piecewise linear finite element method for the numerical solution of the nonstationary initial – boundary value problem for a scalar nonlinear conservation law equation with a diffusion term. The main technique used in this paper is based on the discrete maximum principle, a priori estimates and discrete compactness results derived with the aid of the Fourier transform with respect to time.

2. Continuous Problem

Let $\Omega \subset I\!R^2$ be a bounded domain with a Lipschitz–continuous boundary $\partial\Omega$. In the space–time cylinder $Q_T = \Omega \times (0,T)$ $(0 < T < \infty)$ we consider the following initial–boundary value problem:

Find $u : \overline{Q}_T \to I\!R,\ u = u(x,t),\ x \in \Omega,\ t \in [0,T]$, such that

$$\frac{\partial u}{\partial t} + \sum_{s=1}^{2} \frac{\partial f_s(u)}{\partial x_s} - \nu \Delta u = g \quad \text{in } Q_T, \tag{2.1}$$

$$u|_{\partial\Omega \times (0,T)} = 0, \tag{2.2}$$

$$u(x,0) = u^0(x), \quad x \in \Omega, \tag{2.3}$$

where $\nu > 0$ is a given constant and $f_s : I\!R \to I\!R,\ s = 1,2,\ g : Q_T \to I\!R,\ u^0 : \Omega \to I\!R$ are given functions.

We set

$$V = H_0^1(\Omega) = W_0^{1,2}(\Omega). \tag{2.4}$$

In the space $H^1(\Omega)$ besides its norm we will often work with the seminorm

$$|u|_{H^1(\Omega)} = \left(\int_\Omega |\nabla u|^2 \,\mathrm{d}x \right)^{1/2} \tag{2.5}$$

which is an equivalent norm on V: there exist constants $\bar{c}_1, \bar{c}_2 > 0$ such that

$$\bar{c}_1 \|v\|_{H^1(\Omega)} \le |v|_{H^1(\Omega)} \le \bar{c}_2 \|v\|_{H^1(\Omega)}. \tag{2.6}$$

We can write $|u|_{H^1(\Omega)} = ((u,u))^{1/2}$, where

$$((u,v)) = \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x, \quad u,v \in H^1(\Omega), \tag{2.7}$$

is a scalar product on V. Further we set

$$(u, v) = \int_\Omega uv \, \mathrm{d}x, \quad u, v \in L^2(\Omega). \tag{2.8}$$

We will assume that

$$f_s \in C^2(I\!R), \quad f_s(0) = 0, \quad s = 1, 2, \tag{2.9}$$

$$g \in C([0,T]; W^{1,q}(\Omega)) \text{ for some } q > 2, \tag{2.10}$$

$$u^0 \in W^{1,p}(\Omega) \text{ for some } p > 2. \tag{2.11}$$

Now we derive the weak formulation of problem (2.1) – (2.3). Multiplying (2.1) by an arbitrary $v \in V$, integrating over Ω, using Green's theorem we obtain the identity

$$\begin{aligned} \frac{\mathrm{d}}{dt} \int_\Omega u(t) v \, \mathrm{d}x - \int_\Omega \sum_{s=1}^2 f_s(u(t)) \frac{\partial v}{\partial x_s} \, \mathrm{d}x + \nu \int_\Omega \nabla u(t) \cdot \nabla v \, \mathrm{d}x = \\ = \int_\Omega g(t) v \, \mathrm{d}x, \quad \forall v \in V, \ \forall t \in [0,T]. \end{aligned} \tag{2.12}$$

Here, for $t \in [0,T]$, $u(t)$ means the function "$x \in \Omega \to u(t)(x) = u(x,t)$". Let us set

$$b(\varphi, v) = -\int_\Omega \sum_{s=1}^2 f_s(\varphi) \frac{\partial v}{\partial x_s} \, \mathrm{d}x \quad \text{for } \varphi \in L^\infty(\Omega), \quad v \in V. \tag{2.13}$$

Definition 1 We say that a function u is a *weak solution* of problem (2.1) – (2.3), if it satisfies the conditions

$$u \in L^2(0,T;V) \cap L^\infty(Q_T), \tag{2.14}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} (u(t), v) + b\,(u(t), v) + \nu((u(t), v)) = (g(t), v) \quad \forall v \in V, \tag{2.15}$$

in the sense of distributions on $(0,T)$,

$$u(0) = u^0. \tag{2.16}$$

The identity (2.15), which is (2.12) rewritten with the aid of the above notation, means that

$$\begin{aligned} -\int_0^T (u(t), v)\, \psi'(t) \, \mathrm{d}t + \nu \int_0^T ((u(t), v)) \psi(t) \, \mathrm{d}t + \int_0^T b\,(u(t), v)\, \psi(t) \, \mathrm{d}t = \\ = \int_0^T (g(t), v)\, \psi(t) \, \mathrm{d}t \quad \forall v \in V, \ \forall \psi \in C_0^\infty((0,T)). \end{aligned} \tag{2.17}$$

It follows from [21] that problem (2.14) – (2.16) has a unique solution.

3. Discrete Problem

Let Ω_h be a polygonal approximation of the domain Ω. By $\mathcal{T}_h$ we will denote a triangulation of Ω_h with standard properties (see e.g. [8]): $T \in \mathcal{T}_h$ are closed triangles and

$$\overline{\Omega}_h = \bigcup_{T\in\mathcal{T}_h} T, \tag{3.18}$$

$$\text{if } T_1, T_2 \in \mathcal{T}_h, \text{ then } T_1 \cap T_2 = \emptyset \text{ or} \tag{3.19}$$

$$T_1 \cap T_2 \text{ is a common side of } T_1 \text{ and } T_2 \text{ or } T_1 \cap T_2 \text{ is a common vertex of } T_1 \text{ and } T_2,$$

$$P \in \overline{\Omega} \text{ for any vertex } P \text{ of each } T \in \mathcal{T}_h. \tag{3.20}$$

By $\mathcal{S}_h$ we denote the set of all sides of all triangles $T \in \mathcal{T}_h$. We introduce a numbering of triangles $T \in \mathcal{T}_h$ and their sides $S \in \mathcal{S}_h$ in such a way that

$$\begin{aligned} \mathcal{T}_h &= \{T_i;\ i \in I\}, \\ \mathcal{S}_h &= \{S_j;\ j \in J\}, \end{aligned}$$

where I and J are suitable index sets. By Q_j we denote the centre of a side $S_j \in \mathcal{S}_h$ and put $\mathcal{P}_h = \{Q_j;\ j \in J\}$. Moreover, we set

$$J^\circ = \{i \in J;\ Q_i \in \Omega_h\}. \tag{3.21}$$

Sometimes we will use the local notation S_{ij} and Q_{ij}, $j = 1, 2, 3$, for the sides of a triangle $T_i \in \mathcal{T}_h$ and their centres, respectively. Then

$$\begin{aligned} \{Q_j,\ j \in J\} &= \{Q_{ik},\ k = 1,2,3,\ i \in I\}, \\ \{S_j,\ j \in J\} &= \{S_{ik},\ k = 1,2,3,\ i \in I\}. \end{aligned} \tag{3.22}$$

By $h(T)$ and $\theta(T)$ we denote the length of the longest side and the magnitude of the smallest angle, respectively, of the triangle $T \in \mathcal{T}_h$ and put

$$h = \max_{T\in\mathcal{T}_h} h(T), \quad \theta_h = \min_{T\in\mathcal{T}_h} \theta(T). \tag{3.23}$$

Now let us construct the *barycentric mesh* $\mathcal{D}_h = \{D_i; i \in J\}$ over the basic mesh $\mathcal{T}_h$. The *barycentric finite volume* D_i is a closed polygon defined in the following way: We join the barycentre of every triangle $T \in \mathcal{T}_h$ with its vertices. Then around the side S_i, $i \in J^\circ$, we obtain a closed quadrilateral containing S_i. If $S_j \subset \partial\Omega_h$ is a side with vertices P_1, P_2 of a triangle $T \in \mathcal{T}_h$ adjacent to $\partial\Omega_h$, then we denote by D_j the triangle with the sides S_j and segments connecting the barycentre of T with P_1 and P_2. (See Figures 7.1, 7.2.)

It is obvious that

$$\overline{\Omega}_h = \bigcup_{i\in J} D_i. \tag{3.24}$$

If $D_i \neq D_j$ and the set $\partial D_i \cap \partial D_j$ contains more than one point, we call D_i and D_j *neighbours* and set $\Gamma_{ij} = \partial D_i \cap \partial D_j$ (= a common side of D_i and D_j). Further, we define the set $s(i) = \{j \in J;\ D_j \text{ is a neighbour of } D_i\}$. If $Q_i \in \partial\Omega_h$, then we set $S(i) = s(i) \cup \{-1\}$ and $\Gamma_{i,-1} = S_i \subset \partial\Omega_h$, otherwise (for $i \in J^\circ$) we put $S(i) = s(i)$.

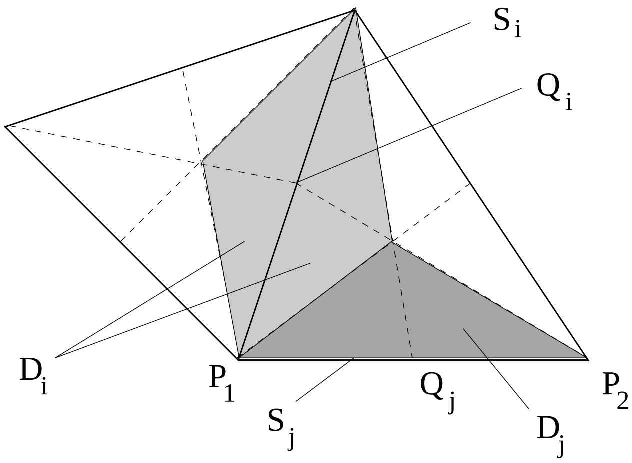


FIG. 7.1: Barycentric finite volumes, $D_i, D_j \in \mathcal{D}_h,\ Q_i, Q_j \in \mathcal{P}_h,\ S_i, S_j \in \mathcal{S}_h,\ S_j \subset \partial\Omega_h$

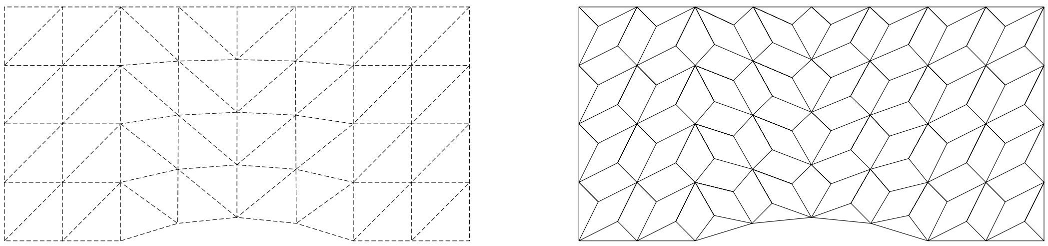


FIG. 7.2: Triangular mesh and the associated barycentric finite volume mesh

In the sequel we use the following notation: $|T|$ = area of $T \in \mathcal{T}_h$, $|D_i|$ = area of $D_i \in \mathcal{D}_h$ (i.e., $i \in J$), l_{ij} = length of the segment Γ_{ij}, $\boldsymbol{n}_{ij} = (n_{ij1}, n_{ij2})$ = unit outer normal to ∂D_i on Γ_{ij} (i.e., $\boldsymbol{n}_{ij}$ points from D_i to D_j). Moreover, let us consider a partition $0 = t_0 < t_1 < \ldots$ of the interval $(0, T)$ and set $\tau_k = t_{k+1} - t_k$ for $k = 0, 1, \ldots$. Obviously, we have

$$\partial D_i = \bigcup_{j \in S(i)} \Gamma_{ij}. \tag{3.25}$$

Let us define the following spaces over the grids $\mathcal{T}_h$ and $\mathcal{D}_h$:

$$\begin{aligned}
X_h &= \left\{ v_h \in L^2(\Omega_h); v_h|_T \text{ is linear } \forall T \in \mathcal{T}_h, v_h \text{ is continuous at } Q_j\ \forall j \in J \right\}, \\
V_h &= \{ v_h \in X_h; v_h(Q_i) = 0\ \forall i \in J - J^\circ \}, \\
Z_h &= \left\{ w_h \in L^2(\Omega_h); w_h|_{D_i} = \text{ const } \forall i \in J \right\}, \\
Y_h &= \{ w_h \in Z_h; w_h = 0 \text{ on } D_i \in \mathcal{D}_h\ \forall i \in J - J^\circ \}.
\end{aligned} \tag{3.26}$$

Let us notice that $X_h \not\subset H^1(\Omega_h)$ and $V_h \not\subset V = H^1_0(\Omega_h)$. Therefore, we speak about *nonconforming, piecewise linear finite elements*. (By G. Strang, the use of nonconforming finite elements belongs to one of the basic *finite element variational crimes*, see [60].)

In the spaces from (3.26) we easily construct *simple bases*: The system $\{w_i;\ i \in J\}$ of functions $w_i \in X_h$ such that $w_i(Q_j) = \delta_{ij} =$ Kronecker delta, $i, j \in J$, forms a basis in X_h. The system $\{w_i,\ i \in J^\circ\}$ is a basis in V_h. Furthermore, denoting by $d_i = \chi_{D_i}$ the characteristic function of $D_i \in \mathcal{D}_h$, we have bases in Z_h and Y_h as the systems $\{d_i;\ i \in J\}$ and $\{d_i;\ i \in J^\circ\}$, respectively.

By I_h we denote the interpolation operator in the space of nonconforming finite elements (see [17], 8.9.79). If $v \in H^1(\Omega)$, then

$$I_h v \in X_h, \quad (I_h v)(Q_{ij}) = \frac{1}{|S_{ij}|} \int_{S_{ij}} v \,\mathrm{d}S, \quad j = 1, 2, 3,\ i \in I. \tag{3.27}$$

This integral exists due to the theorem on traces in the space $H^1(T)$:

$$\|\varphi\|_{L^2(\partial T)} \le c\|\varphi\|_{H^1(T)}, \quad \varphi \in H^1(T),\ T \in \mathcal{T}_h\ \ (c = c(T))\,. \tag{3.28}$$

By $L_h:\ X_h \to Z_h$ we denote the so-called *lumping operator*: if $v:\ \mathcal{P}_h \to I\!R$, then we set

$$L_h v_h = \sum_{i \in J} v_h(Q_i)\, d_i \in Z_h. \tag{3.29}$$

Obviously, $L_h(V_h) = Y_h$.

In order to derive the discrete problem to (2.14) – (2.16) from Definition 1, we put

$$\begin{aligned}
(u, v)_h &= \int_{\Omega_h} (I_h u)(I_h v)\,\mathrm{d}x, \quad u, v \in H^1(\Omega_h), &(3.30)\\
(\!(u, v)\!)_h &= \sum_{i \in I} \int_{T_i} \nabla u \cdot \nabla v \,\mathrm{d}x, \quad u, v \in L^2(\Omega_h), \quad u|_T, v|_T \in H^1(T)\ \forall T \in \mathcal{T}_h,\\
\tilde{b}_h(u, v) &= \sum_{i \in I} \int_{T_i} \sum_{s=1}^{2} \frac{\partial f_s(u)}{\partial x_s} v \,\mathrm{d}x, \quad u \in L^\infty(\Omega_h),\ v \in L^2(\Omega_h), u|_T \in H^1(T)\ \forall T \in \mathcal{T}_h.
\end{aligned}$$

By $|\cdot|_h$ we denote the discrete L^2 – norm induced by $(\cdot,\cdot)_h$. For $v_h \in X_h$ we set $I_h v_h = v_h$ and then

$$(u_h, v_h)_h = (u_h, v_h)_{L^2(\Omega_h)}, |v_h|_h = \|v_h\|_{L^2(\Omega_h)}, \quad u_h, v_h \in X_h. \tag{3.31}$$

If $\Omega_h = \Omega$, then for "regular" functions we have

$$\begin{aligned}
(\!(u, v)\!)_h &= (\!(u, v)\!), && u, v \in H^1(\Omega), &(3.32)\\
\tilde{b}_h(u, v) &= b(u, v), && u \in H^1(\Omega) \cap L^\infty(\Omega),\ v \in V.
\end{aligned}$$

The form $(\!(\cdot,\cdot)\!)_h$ induces the seminorm

$$\|u_h\|_{X_h} = \left(\sum_{i \in I} \int_{T_i} |\nabla u_h|^2 \,\mathrm{d}x\right)^{1/2}, \quad u_h \in X_h. \tag{3.33}$$

Under the notation

$$\|u_h\|_{X_h(T_i)} = \left(\int_{T_i} |\nabla u_h|^2 \,\mathrm{d}x\right)^{1/2}, \quad i \in I,\ u_h \in X_h, \tag{3.34}$$

we have

$$\|u_h\|^2_{X_h} = \sum_{i \in I} \|u_h\|^2_{X_h(T_i)}, \quad u_h \in X_h. \tag{3.35}$$

The following Cauchy inequality holds:

$$((u_h, v_h))_h \le \|u_h\|_{X_h} \|v_h\|_{X_h}, \quad u_h, v_h \in X_h. \tag{3.36}$$

In the case when the diffusion ν is small, it is suitable to modify the discrete "convection" form $\tilde{b}_h$ with the aid of the *finite volume approach*. Let $u \in H^1(\Omega_h), v \in V_h$. Then we write

$$\begin{aligned}
\int_{\Omega_h} \sum_{s=1}^{2} \frac{\partial f_s(u)}{\partial x_s} v \,\mathrm{d}x \quad &\approx \quad \int_{\Omega_h} \sum_{s=1}^{2} \frac{\partial f_s(u)}{\partial x_s} L_h v \,\mathrm{d}x = \\
&= \sum_{i \in J} v(Q_i) \int_{D_i} \sum_{s=1}^{2} \frac{\partial f_s(u)}{\partial x_s} \,\mathrm{d}x = \\
&= \sum_{i \in J} v(Q_i) \int_{\partial D_i} \sum_{s=1}^{2} f_s(u) n_s \,\mathrm{d}S = \\
&= \sum_{i \in J} v(Q_i) \sum_{j \in S(i)} \int_{\Gamma_{ij}} \sum_{s=1}^{2} f_s(u) n_s \,\mathrm{d}S = \\
&= \sum_{i \in J} v(Q_i) \sum_{j \in s(i)} \int_{\Gamma_{ij}} \sum_{s=1}^{2} f_s(u) n_s \,\mathrm{d}S \approx \\
&\approx \sum_{i \in J} v(Q_i) \sum_{j \in s(i)} H\left(u(Q_i), u(Q_j), \boldsymbol{n}_{ij}\right) l_{ij}.
\end{aligned}$$

The function H defined on $I\!R^2 \times \boldsymbol{S}$, where $\boldsymbol{S} = \{\boldsymbol{n} \in I\!R^2; |\boldsymbol{n}| = 1\}$, is called a *numerical flux*.

It is easy to see that the form

$$b_h(u, v) = \sum_{i \in J} v(Q_i) \sum_{j \in s(i)} H\left(u(Q_i), u(Q_j), \boldsymbol{n}_{ij}\right) l_{ij} \tag{3.37}$$

obtained above has sense for all $u, v \in X_h$. We will use it as an approximation of the form $\tilde{b}_h$.

Definition 2 We define the *approximate solution* of problem (2.1) – (2.3) as functions $u_h^k,\ t_k \in [0, T]$, given by the conditions

$$u_h^0 = I_h u^0 \ (\in V_h), \tag{3.38}$$

$$u_h^{k+1} \in V_h, \quad t_k \in [0, T), \tag{3.39}$$

$$\frac{1}{\tau}(u_h^{k+1} - u_h^k, v_h)_h + b_h(u_h^k, v_h) + \nu((u_h^{k+1}, v_h))_h = (g^{k+1}, v_h)_h, \quad \forall v_h \in V_h,\ t_k \in [0, T), \tag{3.40}$$

where $g^k = g(\cdot, t_k)$. The function u_h^k is the approximate solution at time t_k.

Properties of the numerical flux. In what follows we use the following assumptions:
1. $H = H(y, z, \boldsymbol{n})$ is *locally Lipschitz-continuous* with respect to y, z: for any $M^* > 0$ there exists $c(M^*) > 0$ such that

$$\begin{aligned}
|H(y, z, \boldsymbol{n}) - H(y^*, z^*, \boldsymbol{n})| \le c(M^*)(|y - y^*| + |z - z^*|) \\
\forall y, y^*, z, z^* \in [-M^*, M^*],\ \forall \boldsymbol{n} \in \boldsymbol{S}.
\end{aligned} \tag{3.41}$$

2. H is *consistent*:

$$H(u, u, \boldsymbol{n}) = \sum_{s=1}^{2} f_s(u) n_s, \quad \forall u \in I\!R,\ \forall \boldsymbol{n} = (n_1, n_2) \in \mathcal{S}. \tag{3.42}$$

3. H is *conservative*:

$$H(y, z, \boldsymbol{n}) = -H(z, y, -\boldsymbol{n}) \quad \forall y, z \in I\!R,\ \forall \boldsymbol{n} \in \mathcal{S}. \tag{3.43}$$

4. H is *monotone* in the following sense: For a given fixed number $M^* > 0$ the function $H(y, z, \boldsymbol{n})$ is nonincreasing with respect to the second variable z on the set

$$\mathcal{M}_{M^*} = \{(y, z, \boldsymbol{n}); y, z \in [-M^*, M^*], \boldsymbol{n} \in \mathcal{S}\}. \tag{3.44}$$

Lemma 1 *Problem* (3.38) – (3.40) *from Definition* 2 *has the following properties:*
1. *The bilinear forms* $(\cdot,\cdot)_h$ *and* $((\cdot,\cdot))_h$ *defined in* (3.30) *are scalar products on* V_h*.*
2. *For each* $u_h \in X_h$, $b_h(u_h, \cdot)$ *is a linear form on* V_h*.*
3. *If* $i \in J$ *and* $T \in \mathcal{T}_h$ *is a triangle for which* $Q_i \in T$*, then*

$$|T \cap D_i| = \frac{1}{3}|T|. \tag{3.45}$$

4. *The approximation* $(\cdot,\cdot)_h$ *of the* L^2*-scalar product can be defined with the aid of numerical integration using the centres* Q_{ij} *of sides* S_{ij} *of* $T_i \in \mathcal{T}_h$ *as integration points:*

$$(u, v)_h = \frac{1}{3}\sum_{i\in I} |T_i| \sum_{j=1}^{3} u(Q_{ij}) v(Q_{ij}) = \int_\Omega (L_h u)(L_h v)\ \mathrm{d}x, \quad u, v \in X_h. \tag{3.46}$$

5. *We have*

$$(w_i, w_j)_h = \delta_{ij} |D_i|, \quad i, j \in J, \tag{3.47}$$

$$(u, w_i)_h = \frac{1}{3} \sum_{\{T\in\mathcal{T}_h; Q_i \in T\cap\mathcal{P}_h\}} |T| u(Q_i) = |D_i| u(Q_i), \quad i \in J,\ u \in X_h, \tag{3.48}$$

$$(g^k, w_i)_h = |D_i| g(Q_i, t_k), \quad i \in J,\ t_k \in [0, T]. \tag{3.49}$$

6. *Problem* (3.39)–(3.40) *has a unique solution* u_h^{k+1}*.*
7. *Function* $z \in X_h$ *and* $y \in V_h$ *can be expressed in the form*

$$z = \sum_{j\in J} z(Q_j) w_j \ \textit{and}\ y = \sum_{j\in J^\circ} y(Q_j) w_j, \tag{3.50}$$

respectively.
8. *Problem* (3.39)–(3.40) *is equivalent to the system of algebraic equations*

$$|D_i| u_h^{k+1}(Q_i) + \tau\nu \sum_{j\in J^\circ} ((w_i, w_j))_h u_h^{k+1}(Q_j) = \tag{3.51}$$

$$= |D_i| u_h^k(Q_i) - \tau b_h(u_h^k, w_i) + \tau |D_i| g(Q_i, t_k), \quad i \in J^\circ, \tag{3.52}$$

for unknown values $u_h^{k+1}(Q_j)$, $j \in J^\circ$*. This system is uniquely solvable.*

Proof. Assertions 1, 2, 3 and 7 are obvious. By [8], Par. 4.1, the numerical quadrature

$$\int_{T_i} u\ \mathrm{d}x \approx \frac{|T_i|}{3} \sum_{j=1}^{3} u(Q_{ij}), \tag{3.53}$$

is exact for polynomials of degree ≤ 2. This together with 3 implies assertion 4. Assertion 5 is a consequence of 3 and 4. Assertion 6 follows from the Lax–Milgram lemma, 8 is obtained from 5, 6 and 7. □

4. Convergence concept

In what follows, for simplicity *we assume that the domain Ω is polygonal* and, hence, $\Omega_h = \Omega$. Let us consider a system $\{\mathcal{T}_h\}_{h\in(0,h_0)}$ $(h_0 > 0)$ of triangulations of the domain Ω, set $\tau = T/r$ for any integer $r > 1$ and define the partition of the interval $[0,T]$ formed by time instants $t_k = k\tau,\ k = 0,1,\dots,r$.

We define functions $u_{h\tau}, w_{h\tau} : (-\infty,\infty) \to V_h$ associated with an approximate solution $\{u_h^k\}_{k=0}^r$:

$$\begin{aligned} u_{h\tau}(t) &= u_h^0, \quad t \le 0, \\ u_{h\tau}(t) &= u_h^k, \quad t \in (t_{k-1}, t_k], \quad k = 1,\dots,r, \\ u_{h\tau}(t) &= u_h^r, \quad t \ge T; \end{aligned} \tag{4.54}$$

$$\begin{aligned} w_{h\tau} &\quad \text{is a continuous, piecewise linear mapping of } [0,T] \text{ into } V_h, \\ w_{h\tau}(t_k) &= u_h^k, \quad k = 0,\dots,r, \\ w_{h\tau}(t) &= 0 \text{ for } t < 0 \text{ or } t > T. \end{aligned} \tag{4.55}$$

Our goal is to prove that the functions $u_{h\tau}, w_{h\tau}$, constructed from the values of the approximate solution $u_h^k,\ t_k \in [0,T]$ with the aid of scheme (3.38) – (3.40), converge in some sense to the exact solution of problem (2.1) – (2.3) as $h, \tau \to 0$ in a suitable way. In what follows we shall work with a number of constants. By $c, c_1, c_2, \dots, \hat{c}, \hat{c}_1, \dots, \tilde{c}, \tilde{c}_1, \dots$ we will denote constants independent of h, τ, ν, and $C, C_1, \dots$ will denote constants that are independent of h, τ, but depend on ν. Moreover, c will be used as a generic constant attaining in general different values at different places.

Assumptions:

1. Let the system $\{\mathcal{T}_h\}_{h\in(0,h_0)}$ be *regular*, i.e. there exists $\vartheta_0 > 0$ such that

$$\theta_h \ge \vartheta_0 > 0 \quad \forall h \in (0, h_0). \tag{4.56}$$

2.

$$\text{The triangulations } \mathcal{T}_h,\ h \in (0,h_0) \text{ are of } \textit{weakly acute type.} \tag{4.57}$$

This means that the magnitude of all angles of all $T \in \mathcal{T}_h,\ h \in (0,h_0)$, is less than or equal to $\pi/2$.

3. The *inverse assumption* is satisfied: There exists $c_1 > 0$ such that

$$\frac{h}{h(T)} \le c_1 \quad \forall T \in \mathcal{T}_h\ \forall h \in (0,h_0). \tag{4.58}$$

In view of [8], Remark 3.1.3, assumption (4.56) implies the existence of a constant $c_2 > 0$ such that

$$h^2 \le c_2 |T|, \quad T \in \mathcal{T}_h,\ h \in (0,h_0). \tag{4.59}$$

In the next sections, stability, consistency and a priori estimates allows as to construct a limit process and finally to prove the convergence of functions $u_{h\tau}$, $w_{h\tau}$ to a weak solution of the problem (2.1) – (2.3).

5. $L^\infty-$ stability

In virtue of (2.10) and (2.11), $u^0 \in C(\overline{\Omega})$ and $g \in C(\overline{Q}_T)$. Hence, there exist constants $\tilde{M}$ and $\tilde{K}$ such that

$$\|u_h^0\|_{L^\infty(\Omega)} \leq \tilde{M}, \quad \|g\|_{L^\infty(Q_T)} \leq \tilde{K}. \tag{5.60}$$

Let us put

$$M^* = \tilde{M} + T\tilde{K}. \tag{5.61}$$

If $u_h \in X_h$ and $|u_h(Q_i)| \leq M^*$ for all $i \in J$, then $\|u_h\|_{L^\infty(\Omega)} \leq M := 3M^*$. The main tool for proving the L^∞-stability is the *discrete maximum principle* represented by the following results.

Theorem 1 *For $i \in J^\circ$ and $j \in J$ let real numbers a_{ij}, b_{ij}, δ_i, φ_i, u_j, $\tilde{u}_j$, τ satisfy the following conditions:*

$$\begin{aligned}
&\tau > 0,\\
&a_{ii} > 0 \quad \forall i \in J^\circ, \quad a_{ij} \leq 0 \quad \forall i \in J^\circ,\ j \in J,\ i \neq j,\\
&b_{ij} \geq 0 \quad \forall i \in J^\circ,\ j \in J,\\
&\sum_{j\in J} a_{ij} = \sum_{j \in J} b_{ij} = \delta_i > 0\ \forall i \in J^\circ,\\
&\sum_{j\in J} a_{ij}\tilde{u}_j = \sum_{j\in J} b_{ij}u_j + \tau\delta_i\varphi_i\ \forall i \in J^\circ,\\
&\tilde{u}_i = u_i = 0,\ \forall i \in J - J^\circ.
\end{aligned}$$

Then

$$\max_{j\in J} |\tilde{u}_j| \leq \max_{j\in J} |u_j| + \tau \max_{j \in J^\circ} |\varphi_i|. \tag{5.62}$$

Proof follows from [36], Lemma 3.1.1, page 29. □

Lemma 2 *Let w_i, $i \in J$ be the basis functions of X_h defined above. Then under assumption* (4.57) *the following relations are valid:*

$$((w_i, w_i))_h > 0,\ i \in J, \tag{5.63}$$

$$((w_i, w_j))_h \leq 0,\ i, j \in J,\ i \neq j, \tag{5.64}$$

$$\sum_{j\in J} ((w_i, w_j))_h = 0,\ i \in J. \tag{5.65}$$

Proof. By the definition, we have

$$((w_i, w_j))_h = \sum_{T\in\mathcal{T}_h} |T| \nabla w_i|_T \cdot \nabla w_j|_T. \tag{5.66}$$

If $\nabla w_i|_T \cdot \nabla w_j|_T \neq 0$ then Q_i, Q_j must be the midpoints of the sides of the triangle T. So, let T be a triangle with nodes $Q_i = (x_{i1}, x_{i2})$, $Q_j = (x_{j1}, x_{j2})$, $Q_k = (x_{k1}, x_{k2})$. Taking into account that $w_i|_T$ is uniquely determined by its values at the vertices of the triangle $Q_iQ_jQ_k$ and using the standard results (Cf., e.q., [23], Section 4.4), we have

$$\begin{aligned}
\nabla w_i|_T &= \frac{1}{D}(x_{j2} - x_{k2}, x_{k1} - x_{j1}),\\
\nabla w_j|_T &= \frac{1}{D}(x_{k2} - x_{i2}, x_{i1} - x_{k1}),\\
\nabla w_k|_T &= \frac{1}{D}(x_{i2} - x_{j2}, x_{j1} - x_{i1}),
\end{aligned} \tag{5.67}$$

where

$$D = \begin{vmatrix} x_{i1}, & x_{i2}, & 1 \\ x_{j1}, & x_{j2}, & 1 \\ x_{k1}, & x_{k2}, & 1 \end{vmatrix}. \tag{5.68}$$

This implies that

$$|\nabla w_i|_T|^2 = \frac{1}{D^2}\,|Q_j - Q_k|^2 > 0. \tag{5.69}$$

Since $w_i + w_j + w_k = 1$ on T, we have

$$|\nabla w_i|^2 + \nabla w_i \cdot \nabla w_j + \nabla w_i \cdot \nabla w_k = 0 \text{ on } T. \tag{5.70}$$

Further, using (5.67), the well-known expression of the cosine of the angle between two vectors and denoting by α_i the angle in the triangle $Q_iQ_jQ_k$ at the vertex Q_i, we find that

$$\nabla w_i|_T \cdot \nabla w_j|_T = -\frac{1}{D^2}|Q_i - Q_j||Q_i - Q_k| \cos\alpha_i \le 0 \tag{5.71}$$

(similarly for $\nabla w_i{\cdot}\nabla w_k$ and $\nabla w_j{\cdot}\nabla w_k$). The last inequality is a consequence of the assumption (4.57) on the angles of $T \in T_h$, which implies that $\alpha_i \in (0, \pi/2]$. Now we multiply (5.69) – (5.71) by $|T|$, sum over all $T \in T_h$ and use (5.66). As a result we immediately obtain (5.63) – (5.65). □

Theorem 2 *If $\tau > 0$ and $h \in (0, h_0)$ satisfy the condition*

$$\tau c(M^*)|\partial D_i| \le |D_i|, \quad i \in J, \tag{5.72}$$

where $c(M^*)$ is the constant from (3.41), *and if* (5.60) *and* (5.61) *hold, then*

$$\|u_h^k\|_{L^\infty(\Omega)} \le M, \quad t_k \in [0, T], \tag{5.73}$$

$$\|u_{h\tau}\|_{L^\infty(Q_T)},\ \|w_{h\tau}\|_{L^\infty(Q_T)} \le M. \tag{5.74}$$

Proof. In virtue of (3.51) and the fact that $u_h(Q_j) = 0,\ Q_j \in \partial\Omega$, identity (3.40) can be written in the form

$$\begin{aligned} & |D_i|u_h^{k+1}(Q_i) + \tau\nu\sum_{j\in J}((w_i, w_j))_h u_h^{k+1}(Q_j) = \qquad (5.75) \\ = \ & |D_i|u_h^k(Q_i) - \tau b_h(u_h^k, w_i) + \tau|D_i|g(Q_i, t_k), \quad i \in J^\circ,\ k \ge 0. \end{aligned}$$

By induction with respect to k we will prove that

$$\|u_h^k(Q_i)\|_{L^\infty(\Omega)} \le \tilde{M} + k\tau\tilde{K} \le M^*, \quad t_k \in [0, T],\ Q_i \in \mathcal{P}_h. \tag{5.76}$$

Obviously, (5.76) holds for $k = 0$. Let us assume that (5.76) is valid for some $t_k \in [0, T)$.

Let us denote by L_i the left hand side of (5.75) and set $u_i = u_h^k(Q_i)$ and $\varphi_i = g(Q_i, t_k)$ (for simplicity we omit the superscript k). Then (5.75) reads

$$\begin{aligned} L_i &= |D_i|u_i - \tau b_h(u_h, w_i) + \tau|D_i|\varphi_i = \\ &= |D_i|u_i - \tau\sum_{j\in s(i)} H(u_i, u_j, \boldsymbol{n}_{ij})l_{ij} + \tau|D_i|\varphi_i = \\ &= |D_i|u_i + \tau\sum_{j\in s(i)} \Big[H(u_i, u_i, \boldsymbol{n}_{ij}) - H(u_i, u_j, \boldsymbol{n}_{ij}) - H(u_i, u_i, \boldsymbol{n}_{ij})\Big]l_{ij} + \tau|D_i|\varphi_i,\ i \in J^\circ. \end{aligned}$$

In view of the consistency of the numerical flux H (see (3.42)) and Green's theorem, we have

$$\sum_{j\in s(i)} H(u_i, u_i, \boldsymbol{n}_{ij}) l_{ij} = \int_{\partial D_i} \sum_{s=1}^{2} f_s(u_i) n_s \, \mathrm{d}S = 0.$$

Hence, if we set

$$I\!H_{ij} = \begin{cases} 0, & u_i = u_j \\ \frac{H(u_i,u_i,\boldsymbol{n}_{ij})-H(u_i,u_j,\boldsymbol{n}_{ij})}{u_j-u_i} l_{ij}, & u_i \neq u_j, \end{cases} \tag{5.77}$$

we can write

$$L_i = |D_i| u_i + \tau \sum_{j\in s(i)} I\!H_{ij}(u_j - u_i) + \tau |D_i| \varphi_i. \tag{5.78}$$

Due to the monotonicity of the numerical flux,

$$I\!H_{ij} \geq 0, \quad i \in J^\circ, \; j \in s(i). \tag{5.79}$$

In virtue of the induction assumption, $|u_i| \leq \tilde{M} + k\tau\tilde{K} < M^*$ for $i \in J$. This and the local Lipschitz-continuity of H imply that

$$0 \leq I\!H_{ij} \leq c(M^*) l_{ij} = c(M^*)|\Gamma_{ij}|.$$

Using (3.25), we find that

$$0 \leq \sum_{j\in s(i)} I\!H_{ij} \leq c(M^*) \sum_{j\in s(i)} |\Gamma_{ij}| \leq c(M^*)|\partial D_i|$$

and hence, by (5.72),

$$|D_i| - \tau \sum_{j\in s(i)} I\!H_{ij} \geq 0, \quad i \in J^\circ. \tag{5.80}$$

From (5.78) it follows that (5.75) can be written in the form

$$\begin{aligned} &|D_i| u_h^{k+1}(Q_i) + \tau\nu \sum_{j\in J} ((w_i, w_j))_h u_h^{k+1}(Q_j) = \\ = &\left(|D_i| - \tau \sum_{j\in s(i)} I\!H_{ij} \right) u_h^k(Q_i) + \tau \sum_{j\in s(i)} I\!H_{ij} u_h^k(Q_j) + \tau |D_i| \varphi_i^k, \quad i \in J^\circ. \end{aligned} \tag{5.81}$$

Taking into account (5.63) – (5.65), (5.79) and (5.80), we see that Theorem 1 can be applied if we set

$$\begin{aligned} a_{ij} &= |D_i|\delta_{ij} + \tau\nu ((w_i, w_j))_h, \\ b_{ii} &= |D_i| - \tau \sum_{k\in s(i)} I\!H_{ik}, \\ b_{ij} &= \tau I\!H_{ij}, \; i \neq j, \\ u_i &= u_h^k(Q_i), \\ \tilde{u}_i &= u_h^{k+1}(Q_i), \\ \delta_i &= |D_i|. \end{aligned} \tag{5.82}$$

Inequality (5.62) and the fact that $u_h^{k+1}(Q_j) = 0$ for $Q_j \in \partial\Omega_h$ imply that

$$\max_{i\in J} |u_h^{k+1}(Q_i)| \le \max_{i\in J} |u_h^k(Q_i)| + \tau \|g(\cdot, t_k)\|_{L^\infty(\Omega)}.$$

In view of the induction assumption and (5.60), we find that

$$\max_{i\in J} |u_h^{k+1}(Q_i)| \le \tilde{M} + (k+1)\tau\tilde{K} \le M^*.$$

Hence, $\|u_h^{k+1}\|_{L^\infty(\Omega)} \le M = 3M^*$, which we wanted to prove. □

Lemma 3 *Assumption* (4.56) *and its consequence* (4.59) *imply that there exists a constant* $c_3 > 0$ *such that*

$$|D_i|/|\partial D_i| \ge c_3 h, \quad \forall i \in J,\ \forall h \in (0, h_0). \tag{5.83}$$

Proof. From (3.45) we deduce that

$$\forall i \in J\ \exists j_0 \in I \text{ such that } |D_i| \ge \frac{1}{3}|T_{j_0}|, \tag{5.84}$$

which together with (4.59) implies that

$$|D_i| \ge \frac{1}{3c_2}h^2 \qquad \forall i \in J. \tag{5.85}$$

Further, it is easy to see that $|\partial D_i| \le \frac{8}{3}h$, which together with (5.85) gives assertion (5.83). □

As we see, we can consider the *stability condition*

$$0 \le \tau \le c_3 c(M)^{-1} h. \tag{5.86}$$

Obviously, (5.83) and (5.86) yield (5.72).

6. Consistency

Lemma 4 *(Discrete Friedrichs inequality) There exists a constant* $\hat{c}_1$ *independent of* h *such that*

$$\|u_h\|_{L^2(\Omega)} \le \hat{c}_1 \|u_h\|_{X_h}, \quad u_h \in V_h,\ h \in (0, h_0). \tag{6.87}$$

Proof. In [62], Chap. I, Par. 4, Proposition 4.13 or [17], Lemma 8.9.92, this lemma is proved provided Ω is convex. For the case of a general polygonal domain see [14]. □

Definition 3 Let us define the space $L^2(0,T;V_h)$ as the set of all functions $v_h : (0,T) \to V_h$ such that

$$\|v_h\|_{L^2(0,T;V_h)} \equiv \left(\int_0^T \|v_h(t)\|_{X_h}^2 \,\mathrm{d}t\right)^{1/2} = \left(\int_0^T \left(\sum_{i\in I}\int_{T_i} |\nabla v_h(t)|^2 \,\mathrm{d}x\right) \,\mathrm{d}t\right)^{1/2} < \infty. \tag{6.88}$$

Lemma 5 *The interpolation operator I_h defined by* (3.27) *has the following properties:*

$$\text{If } \varphi \in V, \text{ then } I_h\varphi \in V_h. \tag{6.89}$$

Let $\varphi \in H^{k+1}(\Omega)$, where $k = 0$ or 1. Then for $h \in (0, h_0)$ we have

$$\|\varphi - I_h\varphi\|_{X_h} \leq ch^k \|\varphi\|_{H^{k+1}(\Omega)}, \tag{6.90}$$

$$\|\varphi - I_h\varphi\|_{L^2(\Omega)} \leq ch^{k+1} \|\varphi\|_{H^{k+1}(\Omega)}, \tag{6.91}$$

$$\|I_h\varphi\|_{X_h} \leq c\|\varphi\|_{H^1(\Omega)}, \tag{6.92}$$

$$\varphi \in H^1(\Omega) \Rightarrow \|\varphi - I_h\varphi\|_{X_h} \to 0 \text{ as } h \to 0 \tag{6.93}$$

with $c > 0$ independent of φ and h.

Proof. See [17], Lemma 8.9.81. □

Lemma 6 *There exists a constant $c > 0$ such that for any $h \in (0, h_0)$ we have*

$$\|v_h\|_{L^2(\Omega)} = \|L_h v_h\|_{L^2(\Omega)}, \quad v_h \in X_h, \tag{6.94}$$

$$\|v_h - L_h v_h\|_{L^2(\Omega)} \leq ch\|v_h\|_{X_h}, \quad v_h \in X_h, \tag{6.95}$$

$$(u_h, v_h) = (u_h, v_h)_h, \quad u_h, v_h \in X_h, \tag{6.96}$$

$$|(g^k, v_h) - (g^k, v_h)_h| \leq ch\|g^k\|_{W^{1,q}(\Omega)} \|v_h\|_{X_h}, \quad v_h \in V_h. \tag{6.97}$$

If $M > 0$ and $\kappa \in (0,1)$, then there exists a constant $\tilde{c} = \tilde{c}(M, \kappa)$ such that

$$\begin{aligned} |\tilde{b}_h(u_h, v_h) - b_h(u_h, v_h)| \quad &\leq \quad \tilde{c}h^{1-\kappa} \left(\|u_h\|^2_{X_h} + \|u_h\|_{X_h} \right) \|v_h\|_{X_h}, \\ & u_h \in V_h \cap L^\infty(\Omega), \ \|u_h\|_{L^\infty(\Omega)} \leq M, \ v_h \in V_h, \ h \in (0, h_0), \end{aligned} \tag{6.98}$$

where the forms $\tilde{b}_h$ and b_h are defined by (3.30) *and* (3.37), *respectively.*

Proof.
1. Let $v_h \in X_h$. We can write

$$\|v_h\|^2_{L^2(\Omega)} = \int_\Omega |v_h|^2 \, \mathrm{d}x = \sum_{i \in I} \int_{T_i} |v_h|^2 \, \mathrm{d}x. \tag{6.99}$$

By the definition of X_h, $v_h|_{T_i}$ is a linear function. Since the quadrature formula

$$\int_{T_i} \varphi \, \mathrm{d}x \approx \frac{1}{3}|T_i| \sum_{j=1}^{3} \varphi(Q_{ij}) \tag{6.100}$$

is precise for quadratic functions on T_i, we immediately find that

$$\|v_h\|^2_{L^2(\Omega)} = \sum_{i \in I} \frac{1}{3}|T_i| \sum_{j=1}^{3} v_h(Q_{ij})^2. \tag{6.101}$$

On the other hand, from the definition of L_h and (3.45) it follows that

$$\|L_h v_h\|^2_{L^2(\Omega)} = \int_\Omega |L_h v|^2 \, \mathrm{d}x = \sum_{j\in J} \int_{D_j} |L_h v|^2 \, \mathrm{d}x = \tag{6.102}$$
$$= \sum_{j\in J} |D_j| v_h(Q_j)^2 = \sum_{i\in I} \frac{1}{3} |T_i| \sum_{j=1}^{3} v_h(Q_{ij})^2.$$

Now (6.101) and (6.102) yield (6.94).
2. Each $v_h \in X_h$ is linear on $T_i \in \mathcal{T}_h$ and can be expressed in the form

$$v_h(x_1, x_2) = v_h(Q_{ij}) + \left.\frac{\partial v_h}{\partial x_1}\right|_{T_i} (x_1 - x_1(Q_{ij})) + \left.\frac{\partial v_h}{\partial x_2}\right|_{T_i} (x_2 - x_2(Q_{ij})), \quad j = 1, 2, 3, \tag{6.103}$$

where $(x_1(Q_{ij}), x_2(Q_{ij}))$ are the coordinates of Q_{ij}.

Next we have $|\,(x_1 - x_1(Q_{ij}))\,| \le h$, $|\,(x_2 - x_2(Q_{ij}))\,| \le h$ for $(x_1, x_2) \in T_i$. Every triangle $T_i \in \mathcal{T}_h$ can be divided into three subtriangles T_{i1}, T_{i2}, T_{i3} (see Fig. 7.3). Then we have

$$\|v_h - L_h v_h\|^2_{L^2(\Omega)} = \int_\Omega |v_h - L_h v_h|^2 \, \mathrm{d}x = \sum_{i\in I} \sum_{j=1}^{3} \int_{T_{ij}} |v_h - L_h v_h|^2 \, \mathrm{d}x =$$
$$= \sum_{i\in I} \sum_{j=1}^{3} \int_{T_{ij}} \left| \left.\frac{\partial v_h}{\partial x_1}\right|_{T_i} \Big(x_1 - x_1(Q_{ij})\Big) + \left.\frac{\partial v_h}{\partial x_2}\right|_{T_i} \Big(x_2 - x_2(Q_{ij})\Big) \right|^2 \mathrm{d}x \le$$
$$\le h^2 \sum_{i\in I} \sum_{j=1}^{3} \int_{T_{ij}} \left(\left|\frac{\partial v_h}{\partial x_1}\right| + \left|\frac{\partial v_h}{\partial x_2}\right| \right)^2 \mathrm{d}x \le 2h^2 \sum_{i\in I} \int_{T_i} \left(\left|\frac{\partial v_h}{\partial x_1}\right|^2 + \left|\frac{\partial v_h}{\partial x_2}\right|^2 \right) \mathrm{d}x =$$
$$= 2h^2 \|v_h\|^2_{X_h},$$

which proves (6.95).
3. Assertion (6.96) immediately follows from (3.30) and the fact that $u_h = I_h u_h$ for $u_h \in X_h$.
4. Assertion (6.97) follows from relation (3.46), the fact that the quadrature formula (3.53) is exact for polynomials of degree ≤ 2 and [8], Theorem 4.1.5.
5. Let us define the form

$$b_h^*(u_h, v_h) = \sum_{i\in J} v_h(Q_i) \sum_{j\in s(i)} \int_{\Gamma_{ij}} \sum_{s=1}^{2} f_s(u_h) n_s \, \mathrm{d}S, \;\; u_h, v_h \in V_h. \tag{6.104}$$

We write

$$\tilde b_h(u_h, v_h) - b_h(u_h, v_h) = \Big[\tilde b_h(u_h, v_h) - \tilde b_h(u_h, L_h v_h)\Big] + \tag{6.105}$$
$$+ \;\; \Big[\tilde b_h(u_h, L_h v_h) - b_h^*(u_h, v_h)\Big] + \Big[b_h^*(u_h, v_h) - b_h(u_h, v_h)\Big]$$

and estimate the expressions in square brackets separately. Obviously, due to (6.95) and the bound $\|u_h\|_{L^\infty(\Omega)} \le M$,

$$\Big|\tilde b_h(u_h, v_h) - \tilde b_h(u_h, L_h v_h)\Big| = \left| \sum_{i\in I} \int_{T_i} \sum_{s=1}^{2} f_s'(u_h) \frac{\partial u_h}{\partial x_s} (v_h - L_h v_h) \, \mathrm{d}x \right| \le \tag{6.106}$$
$$\le \max_{|\xi|\le M} \max_{s=1,2} |f_s'(\xi)| \, \|\nabla u_h\|_{L^2(\Omega)} \|v_h - L_h v_h\|_{L^2(\Omega)} \le \tilde c h \|u_h\|_{X_h} \|v_h\|_{X_h}, \quad \tilde c = \tilde c(M).$$

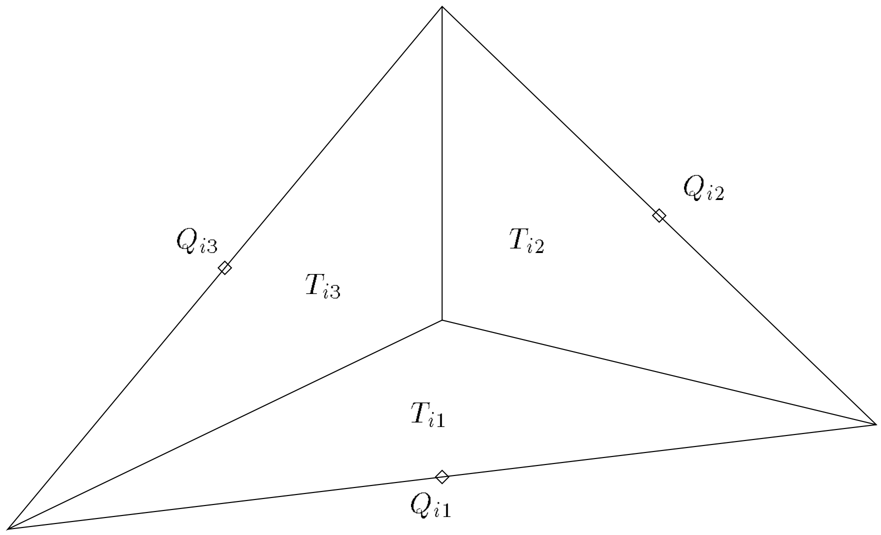


FIG. 7.3: Partition of a triangle T_i into subtriangles T_{i1}, T_{i2}, T_{i3}

Using notation from Fig. 7.3, we have

$$\tilde{b}_h(u_h, L_h v_h) = \sum_{i\in I} \int_{T_i} \sum_{s=1}^{2} \frac{\partial f_s(u_h)}{\partial x_s} L_h v_h \, \mathrm{d}x = \tag{6.107}$$

$$= \sum_{i\in I} \sum_{j=1}^{3} v_h(Q_{ij}) \int_{T_{ij}} \sum_{s=1}^{2} \frac{\partial f_s(u_h)}{\partial x_s} \, \mathrm{d}x = \sum_{i\in I} \sum_{j=1}^{3} v_h(Q_{ij}) \int_{\partial T_{ij}} \sum_{s=1}^{2} f_s(u_h) n_s \, \mathrm{d}S.$$

It is evident (see Fig. 7.4) that for each $k \in J$ there exist $i, i^* \in I$ and $j, j^* \in \{1,2,3\}$ such that

$$D_k = T_{ij} \cup T_{i^*j^*}, \quad S_k = T_{ij} \cap T_{i^*j^*}, \quad Q_{ij} = Q_{i^*j^*} = Q_k. \tag{6.108}$$

If D_k is a boundary finite volume then $i = i^*$, $j = j^*$ and $S_k = T_{ij} \cap \partial\Omega_h$.

The function $u_h \in X_h$ is in general discontinuous on $S_k - \{Q_k\}$. We denote $u_h|_{T_i} = (u_h)_k^p$ and $u_h|_{T_i^*} = (u_h)_k^n$. We denote the outer unit normals to T_i, T_{i^*} on S_k by $\boldsymbol{n}_k^p, \boldsymbol{n}_k^n$ and their components by n_{ks}^p, n_{ks}^n, $s = 1, 2$, respectively. (Obviously $\boldsymbol{n}_k^p = -\boldsymbol{n}_k^n$.) We have

$$\int_{\partial T_{ij}} \sum_{s=1}^{2} f_s(u_h) n_s \, \mathrm{d}S + \int_{\partial T_{i^*j^*}} \sum_{s=1}^{2} f_s(u_h) n_s \, \mathrm{d}S = \tag{6.109}$$

$$= \int_{\partial D_k} \sum_{s=1}^{2} f_s(u_h) n_s \, \mathrm{d}S + \int_{S_k} \sum_{s=1}^{2} f_s\left((u_h)_k^p\right) n_{ks}^p \, \mathrm{d}S + \int_{S_k} \sum_{s=1}^{2} f_s\left((u_h)_k^n\right) n_{ks}^n \, \mathrm{d}S =$$

$$= \int_{\partial D_k} \sum_{s=1}^{2} f_s(u_h) n_s \, \mathrm{d}S + \int_{S_k} \sum_{s=1}^{2} \left[f_s\left((u_h)_k^p\right) - f_s\left((u_h)_k^n\right)\right] n_{ks}^p \, \mathrm{d}S.$$

Now from (6.107), (6.109) and the definition of the forms b_h^*, $\tilde{b}_h$ we have

$$\tilde{b}_h(u_h, L_h v_h) - b_h^*(u_h, v_h) = \sum_{k\in J} v_h(Q_k) \int_{S_k} \sum_{s=1}^{2} \left[f_s\left((u_h)_k^p\right) - f_s\left((u_h)_k^n\right)\right] n_{ks}^p \, \mathrm{d}S. \tag{6.110}$$

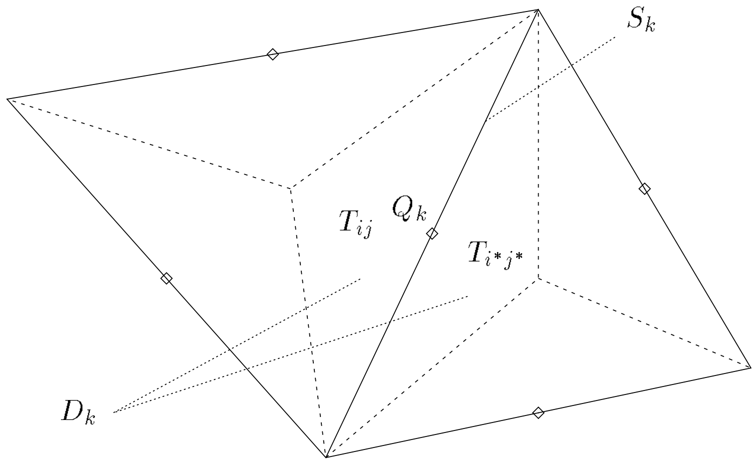


FIG. 7.4: Line of discontinuity S_k of a finite volume D_k

In the following we omit for simplicity the subscript k and write $u_h^p = (u_h)_k^p$, $u_h^n = (u_h)_k^n$, $n_s = n_{ks}^p$. By assumption (2.9) and the Taylor formula we can write

$$\begin{aligned} f_s(u_h^p) &= f_s(u_K) + f_s'(u_K)(u_h^p - u_K) + \frac{1}{2} f_s''(\eta_{sp})(u_h^p - u_K)^2, \quad s = 1,2\,, \qquad (6.111)\\ f_s(u_h^n) &= f_s(u_K) + f_s'(u_K)(u_h^n - u_K) + \frac{1}{2} f_s''(\eta_{sn})(u_h^n - u_K)^2, \quad s = 1,2\,, \end{aligned}$$

where $u_K = u_h^p(Q_k) = u_h^n(Q_k)$, η_{sp} lies between u_h^p and u_K, η_{sn} lies between u_h^n and u_K. Using the above notation for $s = 1, 2$, we have

$$f_s(u_h^p) - f_s(u_h^n) = f_s'(u_K)(u_h^p - u_h^n) + \frac{1}{2}\left[f_s''(\eta_{s1})(u_h^p - u_K)^2 - f_s''(\eta_{s2})(u_h^n - u_K)^2\right]. \qquad (6.112)$$

Since u_h^p, u_h^n are linear functions, then ∇u_h^p and ∇u_h^n are constant and

$$\begin{aligned} u_h^p(x) - u_h^n(x) &= (\nabla u_h^p - \nabla u_h^n)\cdot(x - Q_k)\,, \quad x \in S_k,\\ u_h^p(x) - u_K &= \nabla u_h^p \cdot (x - Q_k)\,, \quad x \in S_k, \qquad (6.113)\\ u_h^n(x) - u_K &= \nabla u_h^n \cdot (x - Q_k)\,, \quad x \in S_k,\\ |x - Q_k| &\le \frac{h}{2}, \quad x \in S_k. \end{aligned}$$

Now from the assumptions on u_h, (6.112) and (6.113) we have

$$\left| \int_{S_k} \sum_{s=1}^{2} \left(f_s(u_h^p) - f_s(u_h^n)\right) n_s \, \mathrm{d}S \right| \le \qquad (6.114)$$

$$\le \left| \int_{S_k} \sum_{s=1}^{2} f_s'(u_K)\left(u_h^p - u_h^n\right) \mathrm{d}S \right| + \max_{|\xi|\le M} \max_{s=1,2} |f_s''(\xi)| \frac{h^2}{4} \int_{S_k} \left| |\nabla u_h^p|^2 - |\nabla u_h^n|^2 \right| \mathrm{d}S \le$$

$$\leq \max_{|\xi|\leq M}\max_{s=1,2}|f_s''(\xi)|\frac{h^3}{4}\left(|\nabla u_h^p|^2+|\nabla u_h^n|^2\right),$$

since $\int_{S_k}(u_h^p-u_h^n)\mathrm{d}S=0$, as one can easily show. Putting (6.114) into (6.110) and taking into account that $u(Q_k)=0$ for $Q_k\in\partial\Omega_h$, we obtain

$$\begin{aligned}
&\left|\tilde{b}_h(u_h,L_hv_h)-b_h^*(u_h,v_h)\right|\leq\\
&\leq \max_{|\xi|\leq M}\max_{s=1,2}|f_s''(\xi)|\frac{h^3}{4}\sum_{k\in J^\circ}|v_h(Q_k)|\left(|\nabla(u_h)_k^p|^2+|\nabla(u_h)_k^n|^2\right)\leq\\
&\leq \max_{|\xi|\leq M}\max_{s=1,2}|f_s''(\xi)|\frac{h^3}{2}\sum_{i\in I}\sum_{j=1}^{3}\left|\nabla u_h|_{T_i}\right|^2\left|L_hv_h|_{T_{ij}}\right|\leq \qquad (6.115)\\
&\leq \hat{c}\max_{|\xi|\leq M}\max_{s=1,2}|f_s''(\xi)|\frac{h}{2}\sum_{i\in I}\sum_{j=1}^{3}|T_i|\left|\nabla u_h|_{T_i}\right|^2\left|L_hv_h|_{T_{ij}}\right|\leq ch\|v_h\|_{L^\infty(\Omega)}\|u_h\|_{X_h}^2.
\end{aligned}$$

Let us put $p=2/\kappa$ ($\in(2,\infty)$). Similarly as in [55] we have

$$\|v_h\|_{L^p(\Omega)}\leq c(p)\|v_h\|_{X_h},\quad v_h\in V_h,\ h\in(0,h_0). \qquad (6.116)$$

(Cf. also [62], Chap. II, Par. 23.) Then, using the inverse assumption (4.58), with the aid of the inverse inequality ([8], Theorem 3.2.6), we obtain

$$\|v_h\|_{L^\infty(\Omega)}\leq \tilde{c}(p)h^{-\frac{2}{p}}\|v_h\|_{L^p(\Omega)},\quad v_h\in V_h,\ h\in(0,h_0). \qquad (6.117)$$

Now (6.115), (6.116) and (6.117) imply that

$$\left|\tilde{b}_h(u_h,L_hv_h)-b_h^*(u_h,v_h)\right|\leq c(M,\kappa)h^{1-\kappa}\|u_h\|_{X_h}^2\|v_h\|_{X_h},\quad v_h\in V_h. \qquad (6.118)$$

Using (3.37), (6.104), the conservativity of the numerical flux H and the relations $\Gamma_{ij}=\Gamma_{ji}, l_{ij}=l_{ji}, \boldsymbol{n}_{ij}=-\boldsymbol{n}_{ji}$, we arrive at

$$\begin{aligned}
&b_h^*(u_h,v_h)-b_h(u_h,v_h)= \qquad (6.119)\\
&=\sum_{i\in J}v_h(Q_i)\sum_{j\in s(i)}\left\{\int_{\Gamma_{ij}}\sum_{s=1}^{2}f_s(u_h)n_s\,\mathrm{d}S-H\left(u_h(Q_i),u_h(Q_j),\boldsymbol{n}_{ij}\right)l_{ij}\right\}=\\
&=\frac{1}{2}\sum_{i\in J}\sum_{j\in s(i)}\left[\int_{\Gamma_{ij}}\sum_{s=1}^{2}f_s(u_h)n_s\,\mathrm{d}S-H\left(u_h(Q_i),u_h(Q_j),\boldsymbol{n}_{ij}\right)l_{ij}\right](v_h(Q_i)-v_h(Q_j)).
\end{aligned}$$

If $i\in J$ and $j\in s(i)$ then we denote by T^{ij} the triangle from $\mathcal{T}_h$ such that $\Gamma_{ij}\subset T^{ij}$. It is easy to see that

$$\begin{aligned}
&|Q_i-Q_j|\leq\frac{h}{2},\quad |x-Q_i|\leq\frac{h}{2}\text{ for }x\in\Gamma_{ij},\ l_{ij}\leq\frac{2}{3}h, \qquad (6.120)\\
&|u_h(Q_i)-u_h(Q_j)|\leq\frac{h}{2}|\nabla u_h|_{T^{ij}}|,\\
&|u_h(x)-u_h(Q_i)|\leq\frac{h}{2}|\nabla u_h|_{T^{ij}}|\ \text{ for }x\in\Gamma_{ij},\\
&|v_h(Q_i)-v_h(Q_j)|\leq\frac{h}{2}|\nabla v_h|_{T^{ij}}|.
\end{aligned}$$

In virtue of the consistency and local Lipschitz-continuity of H, the bound $\|u_h\|_{L^\infty(\Omega)} \leq M$ and (6.120), we conclude that

$$\begin{aligned}
& \left| \int_{\Gamma_{ij}} \sum_{s=1}^{2} f_s(u_h) n_s \, \mathrm{d}S - H\left(u_h(Q_i), u_h(Q_j), \boldsymbol{n}_{ij}\right) l_{ij} \right| \leq \\
\leq & \left| \int_{\Gamma_{ij}} \sum_{s=1}^{2} \left(f_s(u_h) - f_s(u_h(Q_i))\right) n_s \, \mathrm{d}S \right| + \\
& + \left| \sum_{s=1}^{2} f_s(u_h(Q_i)) - H\left(u_h(Q_i), u_h(Q_j), \boldsymbol{n}_{ij}\right) \right| l_{ij} = \\
= & \left| \int_{\Gamma_{ij}} \left[H\left(u_h(x), u_h(x), \boldsymbol{n}_{ij}\right) - H\left(u_h(Q_i), u_h(Q_i), \boldsymbol{n}_{ij}\right)\right] \mathrm{d}S \right| + \\
& + \left| H\left(u_h(Q_i), u_h(Q_i), \boldsymbol{n}_{ij}\right) - H\left(u_h(Q_i), u_h(Q_j), \boldsymbol{n}_{ij}\right) \right| l_{ij} \leq \\
\leq & \; 2c(M) \max_{x \in \Gamma_{ij}} |u_h(x) - u_h(Q_i)| \, l_{ij} + c(M) \, |u_h(Q_i) - u_h(Q_j)| \, l_{ij} \leq \\
\leq & \; c(M) h^2 \left| \nabla u_h|_{T^{ij}} \right|. \qquad (6.121)
\end{aligned}$$

This, (6.119) and (6.120) immediately yield the estimate

$$|b_h^*(u_h, v_h) - b_h(u_h, v_h)| \leq \frac{1}{2} \sum_{i \in J} \sum_{j \in s(i)} c(M) h^3 \left| \nabla u_h|_{T^{ij}} \right| \left| \nabla v_h|_{T^{ij}} \right|. \qquad (6.122)$$

Taking into account that each triangle $T \in \mathcal{T}_h$ appears in the sum in (6.122) as some T^{ij} at most six times and using (4.59), we find that

$$\begin{aligned}
|b_h^*(u_h, v_h) - b_h(u_h, v_h)| & \leq 3c_2 c(M) h \sum_{i \in I} |T_i| \left| \nabla u_h|_{T_i} \right| \left| \nabla v_h|_{T_i} \right| = \\
& = ch \sum_{i \in I} \int_{T_i} |\nabla u_h| \, |\nabla v_h| \, \mathrm{d}x \leq ch \|u_h\|_{X_h} \|v_h\|_{X_h}.
\end{aligned}$$

This, (6.105), (6.106) and (6.118) finally yield (6.98). □

Lemma 7 *If $M > 0$, then there exists a constant $c^* = c^*(M)$ such that*

$$\begin{aligned}
|b_h(u_h, v_h)| \leq & \; c^* \|u_h\|_{L^\infty(\Omega)} \|v_h\|_{X_h}, \qquad (6.123) \\
& u_h \in V_h \cap L^\infty(\Omega), \; \|u_h\|_{L^\infty(\Omega)} \leq M, \; v_h \in V_h, \; h \in (0, h_0).
\end{aligned}$$

Proof. Let $u_h, v_h \in V_h$ and $\|u_h\|_{L^\infty(\Omega)} \leq M$. Using (3.37), the conservativity of the numerical flux and the relations $\Gamma_{ij} = \Gamma_{ji}$, $l_{ij} = l_{ji}$, $\boldsymbol{n}_{ij} = -\boldsymbol{n}_{ji}$, we find that

$$\begin{aligned}
b_h(u_h, v_h) & = \sum_{i \in J} v_h(Q_i) \sum_{j \in s(i)} H\left(u(Q_i), u(Q_j), \boldsymbol{n}_{ij}\right) l_{ij} = \qquad (6.124) \\
& = \frac{1}{2} \sum_{i \in J} \sum_{j \in s(i)} H\left(u(Q_i), u(Q_j), \boldsymbol{n}_{ij}\right) \left(v_h(Q_i) - v_h(Q_j)\right) l_{ij}.
\end{aligned}$$

Let us use the symbol T^{ij} in the same way as in the proof of Lemma 6. Then (6.124), (2.9), (6.120), the consistency and local Lipschitz–continuity of H imply that

$$|b_h(u_h, v_h)| \leq \frac{1}{2} c(M) \max_{\overline{\Omega}} |u_h| \sum_{i \in J} \sum_{j \in s(i)} h^2 \left| \nabla v_h|_{T^{ij}} \right|. \qquad (6.125)$$

From (6.125), (4.59), the fact that each $T \in \mathcal{T}_h$ appears in the above sum as some T^{ij} at most six times and from the Cauchy inequality we conclude that

$$
\begin{aligned}
|b_h(u_h, v_h)| &\leq 3c_2 c(M) \|u_h\|_{L^\infty(\Omega)} \sum_{i \in I} |T_i| \left| \nabla v_h|_{T_i} \right| = 3c_2 c(M) \|u_h\|_{L^\infty(\Omega)} \int_\Omega |\nabla v_h| \, \mathrm{d}x \leq \\
&\leq 3c_2 c(M) \left(\mathrm{meas}(\Omega)\right)^{1/2} \|u_h\|_{L^\infty(\Omega)} \|v_h\|_{X_h}, \qquad (6.126)
\end{aligned}
$$

which is (6.123) with $c^* := 3c_2 c(M)(\mathrm{meas}(\Omega))^{1/2}$. □

The above results imply the following

Theorem 3 *Let* (5.60), (5.61) and (5.72) *hold. Then the solution* u_h^{k+1} *of the discrete problem* (3.38) – (3.40) *satisfies the relation*

$$
\begin{aligned}
(u_h^{k+1} - u_h^k, v_h) + \tau \tilde{b}_h(u_h^k, v_h) + \tau\nu((u_h^{k+1}, v_h))_h = \tau(g^{k+1}, v_h) + l_h^k(v_h), \qquad (6.127)\\
v_h \in V_h, \ t_k \in [0, T), \ h \in (0, h_0),
\end{aligned}
$$

where

$$l_h^k(v_h) = l_{1h}^k(v_h) + l_{2h}^k(v_h), \qquad (6.128)$$

$$l_{1h}^k(v_h) = \tau\left(\tilde{b}_h(u_h^k, v_h) - b_h(u_h^k, v_h)\right), \qquad (6.129)$$

$$l_{2h}^k(v_h) = -\tau\left((g^k, v_h) - (g^k, v_h)_h\right). \qquad (6.130)$$

Moreover, for any $\kappa \in (0,1)$ *there exists a constant* $c > 0$ *independent of* v_h, k, τ *and* h *(but dependent on* κ *and* M*) such that*

$$\left| l_{1h}^k(v_h) \right| \leq c\tau h^{1-\kappa} \left(\|u_h^k\|_{X_h}^2 + \|u_h^k\|_{X_h} \right) \|v_h\|_{X_h}. \qquad (6.131)$$

There exists a constant $\hat{c}$ *independent of* v_h, k, τ *and* h *such that*

$$\left| l_{2h}^k(v_h) \right| \leq \hat{c}\tau h \|g^k\|_{W^{1,q}(\Omega)} \|v_h\|_{X_h}. \qquad (6.132)$$

Proof is an immediate consequence of Theorem 2 and Lemma 6. □

7. A Priori Estimates

Theorem 4 *Let* (5.60) *and* (5.61) *hold. Then there exists a constant* $\hat{C} > 0$ *independent of* h, τ *and* m *(but dependent on* ν*) such that*

$$\max_{t_k \in [0,T]} \|u_h^k\|_{L^2(\Omega)} \leq \hat{C}, \qquad (7.133)$$

$$\sum_{k=1}^m \|u_h^k - u_h^{k-1}\|_{L^2(\Omega)}^2 \leq \hat{C}, \quad t_m \in (0, T], \qquad (7.134)$$

$$\nu\tau \sum_{k=0}^m \|u_h^k\|_{X_h}^2 \leq \hat{C}, \quad t_m \in [0, T], \qquad (7.135)$$

for all τ, $h > 0$ *satisfying the conditions* $h \in (0, h_0)$ *and* (5.72).

Proof. In view of Lemma 3 and Theorem 2, conditions (5.60), (5.61) and (5.86) imply (5.73). If we set $v_h := u_k^{k+1}$ in (3.40) and use the relation

$$(y-z,y) = \frac{1}{2}\left(\|y\|^2_{L^2(\Omega)} - \|z\|^2_{L^2(\Omega)} + \|y-z\|^2_{L^2(\Omega)}\right) \tag{7.136}$$

valid for $y, z \in L^2(\Omega)$, we get

$$\begin{aligned} &\|u_h^{k+1}\|^2_{L^2(\Omega)} - \|u_h^k\|^2_{L^2(\Omega)} + \|u_h^{k+1} - u_h^k\|^2_{L^2(\Omega)} + 2\tau\nu\|u_h^{k+1}\|^2_{X_h} = \\ = \;\; & 2\tau(g^k, u_h^{k+1})_h - 2\tau b_h(u_h^k, u_h^{k+1}). \end{aligned} \tag{7.137}$$

In virtue of Theorem 2, Lemma 7 and Young's inequality for $\varepsilon > 0$, we have

$$2\left|b_h(u_h^k, u_h^{k+1})\right| \le (c^* M)^2/\varepsilon + \varepsilon\|u_h^{k+1}\|^2_{X_h}. \tag{7.138}$$

By (3.46) and the Cauchy inequality,

$$\left|(g^k, u_h^{k+1})_h\right| = \left|(L_h g^k, L_h u_h^{k+1})\right| \le \|L_h g^k\|_{L^2(\Omega)}\|L_h u_h^{k+1}\|_{L^2(\Omega)}. \tag{7.139}$$

With the aid of the definition of the operator L_h and the continuous imbedding $W^{1,q}(\Omega) \hookrightarrow C(\overline{\Omega})$ it is easy to find that

$$\|L_h g^k\|_{L^2(\Omega)} \le c\|g\|_{C(0,T;W^{1,q}(\Omega))} \tag{7.140}$$

with c independent of h and k. Further, from (3.31) and (3.46) we have

$$\|v_h\|_{L^2(\Omega)} = \|v_h\|_h = \|L_h v_h\|_{L^2(\Omega)}, \quad \forall v_h \in X_h. \tag{7.141}$$

This, (7.139), (7.140), Lemma 4 and Young's inequality yield the estimate

$$\begin{aligned} 2\left|(g^k, u_h^{k+1})_h\right| \;&\le\; 2\hat{c}_1 c\|g\|_{C(0,T;W^{1,q}(\Omega))}\|u_h\|_{X_h} \le \\ &\le\; \hat{c}_1^2 c^2\|g\|^2_{C(0,T;W^{1,q}(\Omega))}/\varepsilon + \varepsilon\|u_h\|^2_{X_h}, \quad \varepsilon > 0. \end{aligned} \tag{7.142}$$

Now choosing $\varepsilon = \nu/2$, from (7.137), (7.138) and (7.142) we get

$$\begin{aligned} &\|u_h^{k+1}\|^2_{L^2(\Omega)} - \|u_h^k\|^2_{L^2(\Omega)} + \|u_h^{k+1} - u_h^k\|^2_{L^2(\Omega)} + \nu\tau\|u_h^{k+1}\|^2_{X_h} \le \bar{C}\tau, \\ &\quad \bar{C} = 2(\hat{c}_1^2 c^2\|g\|^2_{C(0,T;W^{1,q}(\Omega))} + (c^* M)^2)/\nu. \end{aligned} \tag{7.143}$$

Summation over $k = 0, \dots, m-1$ ($t_m \in (0,T]$) and the use of (7.141), (3.38), (6.92) and Lemma 4 yield

$$\begin{aligned} &\|u_h^m\|^2_{L^2(\Omega)} + \sum_{k=1}^m \|u_h^k - u_h^{k-1}\|^2_{L^2(\Omega)} + \tau\nu\sum_{k=1}^m \|u_h^k\|^2_{X_h} \le \\ \le \;\; & \bar{C}T + \|u_h^0\|^2_{L^2(\Omega)} \le \bar{C}T + \hat{c}_1^2\|u_h^0\|^2_{X_h} \le \bar{C}T + \hat{c}_1^2 c\|u^0\|^2_{H^1(\Omega)} \le \hat{C}, \quad t_m \in (0,T]. \end{aligned} \tag{7.144}$$

Now, estimates (7.144) immediately imply (7.133) – (7.135). □

Theorem 5 *Let (5.60) and (5.61) hold. Then there exists a constant $C > 0$ such that functions $u_{h\tau}$ and $w_{h\tau}$ defined by (4.54) and (4.55) satisfy the estimates*

$$\|u_{h\tau}\|_{L^2(-1,T;L^2(\Omega))} \leq C, \tag{7.145}$$

$$\|w_{h\tau}\|_{L^2(0,T;L^2(\Omega))} \leq C, \tag{7.146}$$

$$\|u_{h\tau}\|_{L^2(-1,T;V_h)} \leq C, \tag{7.147}$$

$$\|w_{h\tau}\|_{L^2(0,T;V_h)} \leq C \tag{7.148}$$

for all $h \in (0, h_0)$ and $\tau > 0$ satisfying condition (5.72). Moreover, there exists a constant $\tilde{C} > 0$ such that

$$\|u_{h\tau} - w_{h\tau}\|_{L^2(Q_T)} \leq \tilde{C}\sqrt{\tau} \tag{7.149}$$

for all h and τ with the above properties.

Proof. Assertions (7.145) and (7.147) immediately follow from (7.133) and (7.135), respectively.

Now let us prove (7.149). We have

$$\|u_{h\tau} - w_{h\tau}\|^2_{L^2(Q_T)} = \sum_{k=1}^{r} \int_{t_{k-1}}^{t_k} \left\| \frac{t-t_k}{\tau} \left(u_h^k - u_h^{k-1}\right) \right\|^2_{L^2(\Omega)} \mathrm{d}t \leq$$

$$\leq \sum_{k=1}^{r} \|u_h^k - u_h^{k-1}\|^2_{L^2(\Omega)} \int_{t_{k-1}}^{t_k} \left(\frac{t-t_k}{\tau}\right)^2 \mathrm{d}t \leq \hat{C}\frac{\tau}{3},$$

as follows from (7.134).

Assertion (7.146) is a consequence of (7.145) and (7.149). Finally,

$$\|w_{h\tau}\|^2_{L^2(0,T;V_h)} = \sum_{k=1}^{r} \int_{t_{k-1}}^{t_k} \|w_{h\tau}(t)\|^2_{X_h} \,\mathrm{d}t \tag{7.150}$$

and for $t \in (t_{k-1}, t_k)$, using the convexity of the function "$u \to \|u\|^2_{X_h}$", we get

$$\|w_{h\tau}(t)\|^2_{X_h} = \left\| u_h^{k-1} + \frac{t-t_k}{\tau}\left(u_h^k - u_h^{k-1}\right) \right\|^2_{X_h} \leq \|u_h^{k-1}\|^2_{X_h} + \|u_h^k\|^2_{X_h}. \tag{7.151}$$

This and (7.135) already yield (7.148). □

8. Passage to Limit

We rewrite scheme (3.40) in the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(w_{h\tau}(t), v_h\right)_h + \nu(\!(u_{h\tau}, v_h)\!)_h + b_h(u_{h\tau}(t-\tau), v_h) = (g_{h\tau}(t), v_h)_h \tag{8.152}$$

$$\text{for a. e. } t \in (0,T),\ v_h \in V_h,$$

where

$$g_{h\tau}(t) = g^{k+1} \quad \text{for} \quad t \in (t_k, t_{k+1}). \tag{8.153}$$

The weak solution of the continuous problem (2.1) – (2.3) satisfies the condition $u(\cdot, t) \in V$ for a.e. $t \in (0,T)$ and the approximate solution $u_h^k \in V_h$ for $t_k \in [0,T]$. Since we use nonconforming FEM and thus $V_h \not\subset V$, the convergence analysis is more complex than in the conforming case investigated in [21]. Our further considerations will be based on results from [62] and [17], Section 8.9.

If $v_h \in V_h$, then the distributional derivatives are not elements of $L^2(\Omega)$. Therefore, we will define the discrete analogue $d_{ih}v_h$ of the derivatives $\frac{\partial v_h}{\partial x_i}$, $i = 1, 2$:

$$(d_{ih}v_h)(x) = \left(\frac{\partial v_h}{\partial x_i}\right)(x), \quad x \in T,\ T \in \mathcal{T}_h. \tag{8.154}$$

Obviously, $d_{ih}v \in L^2(\Omega)$.

We introduce the space $F = \left[L^2(\Omega)\right]^3$ and the mapping $\omega : V \to F$ defined by

$$u \in V \longmapsto \omega u = \left(u, \frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}\right) \in F. \tag{8.155}$$

The space F is equipped with the norm

$$\|\varphi\|_F = \left(\sum_{i=0}^{2} \|\varphi_i\|^2_{L^2(\Omega)}\right)^{1/2} \quad \text{for } \varphi = (\varphi_0, \varphi_1, \varphi_2) \in F. \tag{8.156}$$

We define a scalar product in F by

$$(\varphi, \psi)_F = \sum_{i=0}^{2} (\varphi_i, \psi_i), \quad \varphi = (\varphi_0, \varphi_1, \varphi_2),\ \psi = (\psi_0, \psi_1, \psi_2) \in F. \tag{8.157}$$

Further, we define the imbedding operator $J_h : V_h \to F$ by

$$v_h \in V_h \longmapsto J_h v_h = (v_h, d_{1h}v_h, d_{2h}v_h) \in F. \tag{8.158}$$

From (8.156) and the discrete Friedrichs inequality (6.87) we have

$$\|J_h v_h\|_F^2 = \|v_h\|^2_{L^2(\Omega)} + \|v_h\|^2_{X_h} \le (c_1^2 + 1)\|v_h\|^2_{X_h} \text{ for } v_h \in V_h, \tag{8.159}$$

which leads to

$$\|J_h v_h\|_F \le c\|v_h\|_{X_h} \text{ for } v_h \in V_h. \tag{8.160}$$

This implies that the operators J_h, $h \in (0, h_0)$, are *uniformly bounded*:

$$\|J_h\| = \sup_{0 \ne v_h \in V_h} \frac{\|J_h v_h\|_F}{\|v_h\|_{X_h}} \le c, \quad h \in (0, h_0). \tag{8.161}$$

We will also work with the operator $I_h : V \to V_h$ defined by (3.27). Let us prove several auxiliary results.

Lemma 8 *1. For each $v \in V$,*

$$\lim_{h \to 0} J_h(I_h v) = \omega v \text{ strongly in } F. \tag{8.162}$$

2. If for a sequence $h_n \in (0, h_0)$, $n = 1, 2, \ldots$ we have $h = h_n \to 0$ as $n \to \infty$, $v_h \in V_h$ and

$$\lim_{h \to 0} J_h v_h = \phi \text{ weakly in } F, \tag{8.163}$$

then there exists $v \in V$ such that $\phi = \omega v$.

Proof.
1. Let $v \in V$. In view of (8.155) and (8.158) we have

$$\|J_h(I_h v) - \omega v\|_F^2 = \|I_h v - v\|_{X_h}^2 + \|I_h v - v\|_{L^2(\Omega)}^2 \to 0 \text{ for } h \to 0 \tag{8.164}$$

as follows from Lemma 5.
2. To establish assertion 2, see the 2nd and 3rd part of the proof of 8.9.118 from [17] or Chap. 1, Sec. 5 from [34]. □

Remark 1 The family of triplets $\{V_h, J_h, I_h\}_{h\in(0,h_0)}$ together with $\{V, F, \omega\}$ is called *the external approximation* of the space V. If (8.161) holds, the external approximation of V is called *stable*. If the operators I_h, J_h have properties (8.162) and (8.163), then the external approximation of V is called *convergent* (cf. [34], Chap. I, Sec 5. or [62], Chap. I, Par. 3).

Lemma 9 *There exists a constant $c > 0$ such that*

$$\left|\sum_{T\in\mathcal{T}_h} \int_{\partial T} v_h \varphi n_i \,\mathrm{d}S\right| \le ch\|\varphi\|_{H^1(\Omega)}\|v_h\|_{X_h}, \quad h \in (0, h_0), \tag{8.165}$$

for any $\varphi \in H^1(\Omega)$ and $v_h \in V_h$. Here n_i denotes the i-th component of the unit outer normal to ∂T.

Proof. See [17], Lemma 8.9.85 and Lemma 4. □

Now let us return to the systems of functions $u_{h\tau}$ and $w_{h\tau}$ defined in (4.54) and (4.55), respectively, for $h \in (0, h_0)$ and $\tau > 0$ satisfying the stability condition (5.72). Then $u_{h\tau}$ and $w_{h\tau}$ satisfy estimates (7.145), (7.147) and (7.146), (7.148), respectively.

Lemma 10 *There exist sequences $h = h_n,\ \tau = \tau_n \to 0$ as $n \to \infty$ satisfying (5.72) and functions $u \in L^2(-1, T; V),\ \phi \in L_2(-1, T; F)$ such that*

$$u_{h\tau} \to u \text{ weakly in } L^2(-1, T; L^2(\Omega)), \tag{8.166}$$
$$J_h u_{h\tau} \to \phi \text{ weakly in } L^2(-1, T; F), \tag{8.167}$$

and $\phi = \omega u$.

Proof. In view of (7.147) and (8.160) we have

$$\|J_h u_{h\tau}\|_{L^2(-1,T;F)} \le c\|u_{h\tau}\|_{L^2(-1,T;V_h)} \le C, \quad h \in (0, h_0),\ \tau > 0, \tag{8.168}$$

where $C > 0$ is a constant independent of h and τ. Since the spaces $L^2(-1, T; L^2(\Omega))$ and $L^2(-1, T; F)$ are reflexive, we obtain sequences $h = h_n,\ \tau = \tau_n \to 0$ and functions u, ϕ such that (8.166) and (8.167) hold.

Further, we prove that $\phi = \omega u$. If $\phi = (\phi_0, \phi_1, \phi_2)$, then obviously $u = \phi_0$. We want to show that $\frac{\partial u}{\partial x_s} = \phi_s,\ s = 1, 2$ in the sense of distributions on $\tilde{Q}_T = \Omega \times (-1, T)$. We can proceed similarly as in the proof of 8.9.81 from [17]. Let $\varphi \in C^\infty(\overline{\tilde{Q}}_T)$. Then (8.166) and (8.167) imply that

$$\int_{\tilde{Q}_T} u_{h\tau} \frac{\partial \varphi}{\partial x_s} \,\mathrm{d}x \,\mathrm{d}t \quad \to \quad \int_{\tilde{Q}_T} u \frac{\partial \varphi}{\partial x_s} \,\mathrm{d}x \,\mathrm{d}t, \tag{8.169}$$
$$\int_{\tilde{Q}_T} d_{sh} u_{h\tau} \varphi \,\mathrm{d}x \,\mathrm{d}t \quad \to \quad \int_{\tilde{Q}_T} \phi_s \varphi \,\mathrm{d}x \,\mathrm{d}t. \tag{8.170}$$

Using Fubini's and Green's theorems, we get

$$\int_{\tilde{Q}_T} u_{h\tau} \frac{\partial \varphi}{\partial x_s} \,\mathrm{d}x\,\mathrm{d}t = \int_{-1}^{T} \left(\int_{\Omega} u_{h\tau} \frac{\partial \varphi}{\partial x_s} \,\mathrm{d}x \right) \mathrm{d}t = \tag{8.171}$$
$$= -\int_{-1}^{T} \left(\sum_{i\in I} \int_{T_i} \frac{\partial u_{h\tau}}{\partial x_s} \varphi \,\mathrm{d}x \right) \mathrm{d}t + \int_{-1}^{T} \left(\sum_{i\in I} \int_{\partial T_i} u_{h\tau} \varphi n_s \,\mathrm{d}S \right) \mathrm{d}t.$$

Since $\varphi \in C^\infty(\overline{\tilde{Q}}_T)$ and $u_{h\tau} : (-1,T) \to V_h$, we conclude from Lemma 9 and the Cauchy inequality that

$$\left| \int_{-1}^{T} \left(\sum_{i\in I} \int_{\partial T_i} u_{h\tau} \varphi n_s \,\mathrm{d}S \right) \mathrm{d}t \right| \le ch \int_{-1}^{T} \|u_{h\tau}(t)\|_{X_h} \|\varphi(t)\|_{H^1(\Omega)} \,\mathrm{d}t \le \tag{8.172}$$
$$\le ch \|u_{h\tau}\|_{L^2(-1,T;V_h)} \|\varphi\|_{L^2(-1,T;H^1(\Omega))} \le \tilde{C}h \to 0 \text{ as } h,\tau \to 0.$$

This, (8.169) – (8.171) and the relation

$$\int_{-1}^{T} \left(\sum_{i\in I} \int_{T_i} \frac{\partial u_{h\tau}}{\partial x_s} \varphi \,\mathrm{d}x \right) \mathrm{d}t = \int_{\tilde{Q}_T} d_{sh} u_{h\tau} \varphi \,\mathrm{d}x\,\mathrm{d}t \tag{8.173}$$

imply that

$$\int_{\tilde{Q}_T} u \frac{\partial \varphi}{\partial x_s} \,\mathrm{d}x\,\mathrm{d}t = -\int_{\tilde{Q}_T} \phi_s \varphi \,\mathrm{d}x\,\mathrm{d}t. \tag{8.174}$$

Taking here $\varphi \in C_0^\infty(\tilde{Q}_T) \subset C^\infty(\overline{\tilde{Q}}_T)$, we find that $\frac{\partial u}{\partial x_s} = \phi_s \in L^2(-1,T;L^2(\Omega)),\ s = 1,2$. Hence, $u \in L^2(-1,T;H^1(\Omega))$ and $\phi = \omega u$. As we see, we have

$$\int_{\tilde{Q}_T} u \frac{\partial \varphi}{\partial x_s} \,\mathrm{d}x\,\mathrm{d}t = -\int_{\tilde{Q}_T} \frac{\partial u}{\partial x_s} \varphi \,\mathrm{d}x\,\mathrm{d}t \qquad \forall \varphi \in C^\infty(\overline{\tilde{Q}}_T),\ s = 1,2. \tag{8.175}$$

The application of Green's theorem yields the identity

$$\int_{-1}^{T} \left(\int_{\partial\Omega} u \varphi n_s \,\mathrm{d}S \right) \mathrm{d}t = 0 \qquad \forall \varphi \in C^\infty(\overline{\tilde{Q}}_T),\ s = 1,2, \tag{8.176}$$

which implies that $u(t) = 0$ on $\partial\Omega$ for a.e. $t \in (-1,T)$. Thus, $u \in L^2(-1,T;V)$. □

Lemma 11 *If $h = h_n$ and $\tau = \tau_n$ are sequences from* Lemma 10*, then*

$$w_{h\tau} \to u \text{ weakly in } L^2(0,T;L^2(\Omega)) \tag{8.177}$$

$$J_h w_{h\tau} \to \omega u \text{ weakly in } L^2(0,T;F). \tag{8.178}$$

Proof. We use Lemma 10 and (7.149). □

Lemma 12 *Let $h = h_n \to 0$ as $n \to \infty$, $v_h \in X_h$, $v \in V$, $J_h v_h \to \omega v$ weakly in F. Then $v_h \to v$ strongly in $L^2(\Omega)$.*

Proof. See part 5) of the proof of Theorem 8.9.118 from [17]. □

In the sequel we will use the compactness criterion based on the Fourier transform $\hat{w}_{h\tau}$ of the function $w_{h\tau}$ with respect to time:

$$\hat{w}_{h\tau}(s) = \int_{\mathbb{R}} w_{h\tau}(t) e^{-2\pi i s t} \,\mathrm{d}t. \tag{8.179}$$

Lemma 13 *We have*

$$\int_{\mathbb{R}} |s|^{2\gamma} \|\hat{w}_{h\tau}(s)\|^2_{L^2(\Omega)} \, \mathrm{d}s \le c \quad \textit{for } 0 < \gamma < 1/4 \tag{8.180}$$

with a constant c independent of h, τ.

Proof. For a.e. $t \in (0, T)$ we define $r_{h\tau}(t) \in V_h$ by the identity

$$((r_{h\tau}(t), v_h))_h = (g_{h\tau}(t), v_h)_h - \nu((u_{h\tau}(t), v_h))_h - b_h(u_{h\tau}(t-\tau), v_h) \quad \forall v_h \in V_h. \tag{8.181}$$

Hence, by (8.152),

$$\frac{\mathrm{d}}{\mathrm{d}t}(w_{h\tau}(t), v_h)_h = ((r_{h\tau}(t), v_h))_h, \quad v_h \in V_h, \text{ a.e. } t \in (0, T). \tag{8.182}$$

Substituting $v_h := r_{h\tau}(t)$ in (8.181) and using (6.123), (6.87), (2.10) and (5.74), we obtain

$$\begin{aligned} \|r_{h\tau}(t)\|^2_{X_h} &\le \|r_{h\tau}(t)\|_{L^2(\Omega)} \|g_{h\tau}(t)\|_{L^2(\Omega)} + \nu \|u_{h\tau}(t)\|_{X_h} \|r_{h\tau}(t)\|_{X_h} + \\ &+ c^* \|u_{h\tau}(t)\|_{L^\infty(\Omega)} \|r_{h\tau}(t)\|_{X_h} \le c \|r_{h\tau}(t)\|_{X_h} \left(1 + \|u_{h\tau}(t)\|_{X_h}\right) \end{aligned}$$

and thus, in view of (4.55),

$$\|r_{h\tau}(t)\|_{X_h} \le c\left(1 + \|u_h^k\|_{X_h}\right), \quad t_{k-1} \le t \le t_k. \tag{8.183}$$

This, (7.135) and the Cauchy inequality imply that

$$\int_0^T \|r_{h\tau}(t)\|_{X_h} \, \mathrm{d}t \le \sqrt{T}\left(\int_0^T \|r_{h\tau}(t)\|^2_{X_h} \, \mathrm{d}t\right)^{1/2} \le c\sqrt{T}\left(T + \tau \sum_{k=1}^r \|u_h^k\|^2_{X_h}\right)^{1/2} \le \text{const.} \tag{8.184}$$

Now we put

$$\bar{r}_{h\tau}(t) = \begin{cases} r_{h\tau}(t), & t \in (0, T), \\ 0, & t \notin (0, T). \end{cases} \tag{8.185}$$

Then the Fourier transform $\hat{\bar{r}}_{h\tau}$ of $\bar{r}_{h\tau}$ satisfies the relations

$$\|\hat{\bar{r}}_{h\tau}(s)\|_{X_h} = \left\| \int_{-\infty}^{\infty} e^{-2\pi i s t} \bar{r}_{h\tau}(t) \, \mathrm{d}t \right\|_{X_h} \le \int_0^T \|r_{h\tau}(t)\|_{X_h} \, \mathrm{d}t \le c, \quad s \in \mathbb{R}. \tag{8.186}$$

The distribution derivative of the function $(w_{h\tau}(t), v_h)_h$ over $\mathbb{R}$ has the form

$$\frac{\mathrm{d}}{\mathrm{d}t}(w_{h\tau}(t), v_h)_h = ((\bar{r}_{h\tau}(t), v_h))_h + (u_h^0, v_h)_h \delta_0 - (u_h^r, v_h)_h \delta_T, \quad v_h \in V_h, \tag{8.187}$$

where δ_0 and δ_T are the Dirac distributions concentrated at $t = 0$ and $t = T$, respectively. The Fourier transform yields

$$2\pi i s(\hat{w}_{h\tau}(s), v_h)_h = ((\hat{\bar{r}}_{h\tau}(s), v_h))_h + (u_h^0, v_h)_h - (u_h^r, v_h)_h \exp(-2\pi i s T), \quad s \in \mathbb{R}. \tag{8.188}$$

Putting here $v_h := \hat{w}_{h\tau}(s)$, we have

$$\begin{aligned} & 2\pi i s(\hat{w}_{h\tau}(s), \hat{w}_{h\tau}(s))_h = \qquad (8.189) \\ = \; & ((\hat{\tilde{r}}_{h\tau}(s), \hat{w}_{h\tau}(s)))_h + (u_h^0, \hat{w}_{h\tau}(s))_h - (u_h^r, \hat{w}_{h\tau}(s))_h \exp(-2\pi i s T), \quad c \in I\!R. \end{aligned}$$

From (6.96), (8.189) and the Cauchy inequality we find that

$$\begin{aligned} 2\pi|s|\|\hat{w}_{h\tau}(s)\|^2_{L^2(\Omega)} \; & \leq \; \|\hat{\tilde{r}}_{h\tau}(s)\|_{X_h}\|\hat{w}_{h\tau}(s)\|_{X_h} + \qquad (8.190) \\ & + \; \|u_h^0\|_{L^2(\Omega)}\|\hat{w}_{h\tau}(s)\|_{L^2(\Omega)} + \|u_h^r\|_{L^2(\Omega)}\|\hat{w}_{h\tau}(s)\|_{L^2(\Omega)}. \end{aligned}$$

In view of (8.190), (8.186), (7.133) and (6.87), we obtain

$$|s|\|\hat{w}_{h\tau}(s)\|^2_{L^2(\Omega)} \leq c\|\hat{w}_{h\tau}(s)\|_{X_h}, \quad s \in I\!R. \qquad (8.191)$$

Let $0 < \gamma < 1/4$. Obviously, there exists a constant $c(\gamma)$ such that

$$|s|^{2\gamma} \leq c(\gamma)\frac{1+|s|}{1+|s|^{1-2\gamma}}, \; s \in I\!R, \qquad (8.192)$$

and

$$\int_{-\infty}^{\infty} \frac{\mathrm{d}s}{(1+|s|^{1-2\gamma})^2} < \infty. \qquad (8.193)$$

Using (8.191) – (8.193), the Cauchy inequality and (6.87), we find that

$$\begin{aligned} & \int_{-\infty}^{\infty} |s|^{2\gamma}\|\hat{w}_{h\tau}(s)\|^2_{L^2(\Omega)} \,\mathrm{d}s \leq c\int_{-\infty}^{\infty} \|\hat{w}_{h\tau}(s)\|^2_{L^2(\Omega)} \,\mathrm{d}s + c\int_{-\infty}^{\infty} \frac{\|\hat{w}_{h\tau}(s)\|_{X_h}}{1+|s|^{1-2\gamma}} \,\mathrm{d}s \leq \\ \leq \; & c\int_{-\infty}^{\infty} \|\hat{w}_{h\tau}(s)\|^2_{X_h} \,\mathrm{d}s \left\{1 + \left(\int_{-\infty}^{\infty} \frac{\mathrm{d}s}{(1+|s|^{1-2\gamma})^2}\right)^{\frac{1}{2}}\right\} \leq c\int_{-\infty}^{\infty} \|\hat{w}_{h\tau}(s)\|^2_{X_h} \,\mathrm{d}s. \; (8.194) \end{aligned}$$

With the aid of (3.33), Fubini's theorem, the differentiation of the integral with respect to a parameter, Parseval's equality and (7.148), we obtain

$$\begin{aligned} & \int_{-\infty}^{\infty} \|\hat{w}_{h\tau}(s)\|^2_{X_h} \,\mathrm{d}s = \qquad (8.195) \\ = \; & \int_{-\infty}^{\infty} \sum_{i\in I} \int_{T_i} \left|\nabla \int_{-\infty}^{\infty} w_{h\tau}(t) e^{-2\pi i t s} \,\mathrm{d}t\right|^2 \mathrm{d}x \,\mathrm{d}s = \\ = \; & \sum_{i\in I} \int_{T_i} \left(\int_{-\infty}^{\infty} \left|\int_{-\infty}^{\infty} \nabla w_{h\tau}(t) e^{-2\pi i t s} \,\mathrm{d}t\right|^2 \,\mathrm{d}s\right) \,\mathrm{d}x \overset{\text{Parseval's equality}}{=} \\ = \; & \sum_{i\in I} \int_{T_i} \left(\int_{-\infty}^{\infty} |\nabla w_{h\tau}(t)|^2 \,\mathrm{d}t\right) \,\mathrm{d}x = \int_0^T \left(\sum_{i\in I} \int_{T_i} |\nabla w_{h\tau}(t)|^2 \,\mathrm{d}x\right) \,\mathrm{d}t = \\ = \; & \|w_{h\tau}\|^2_{L^2(0,T;V_h)} \leq C. \qquad (8.196) \end{aligned}$$

$\square$

Now we prove the *fundamental compactness result*.

Lemma 14 *Let us consider the sequences* $h = h_n$, $\tau = \tau_n \to 0$ *and* $w_{h\tau} = w_{h_n\tau_n}$ *from Lemma* 11 *satisfying* (7.148) *and* (8.177)*. Then* (8.178) *holds and*

$$w_{h\tau} \to u \textit{ strongly in } L^2(Q_T), \tag{8.197}$$

where u *is the limit function from Lemma* 11*.*

Proof. Let us set
$w(t) = u(t), \quad t \in (0,T),$
$w(t) = 0, \qquad t < 0 \text{ or } t > T$.
Then, in virtue of (8.177) and (8.178),

$$J_h w_{h\tau} \to \omega w \text{ weakly in } L^2(I\!R, F), \tag{8.198}$$

$$w_{h\tau} \to w \text{ weakly in } L^2(I\!R, L^2(\Omega)). \tag{8.199}$$

Our goal is to prove that

$$\mathcal{F}_{h\tau} = \int_{-\infty}^{\infty} \|w_{h\tau}(t) - w(t)\|^2_{L^2(\Omega)} \, \mathrm{d}t \to 0. \tag{8.200}$$

In virtue of (7.146), $\mathcal{F}_{h\tau}$ is uniformly bounded for $h \in (0, h_0)$ and $\tau > 0$ satisfying (5.72). By Parseval's equality,

$$\mathcal{F}_{h\tau} = \int_{-\infty}^{\infty} \|\hat{w}_{h\tau}(s) - \hat{w}(s)\|^2_{L^2(\Omega)} \, \mathrm{d}s, \tag{8.201}$$

where $\hat{w}$ is the Fourier transform of w.

For $\gamma > 0$ we define the space

$$\mathcal{H}^\gamma = \left\{ v; \omega v \in L^2(-1,1;F), \int_{I\!R} |s|^{2\gamma} \|\hat{v}(s)\|^2_{L^2(\Omega)} \, \mathrm{d}s < \infty \right\} \tag{8.202}$$

equipped with the scalar product

$$(v,w)_{\mathcal{H}^\gamma} = \int_{-1}^{T} (\omega v(t), \omega w(t))_F \; \mathrm{d}t + \int_{I\!R} |s|^{2\gamma} \left(\hat{v}(s), \hat{w}(s)\right) \; \mathrm{d}s. \tag{8.203}$$

It can be proved that $\mathcal{H}^\gamma$ is a Hilbert space. In virtue of Theorem 5 and Lemma 13, the system $\{w_{h\tau}\}$ is uniformly bounded in $\mathcal{H}^\gamma$ for all $h \in (0, h_0)$ and $\tau > 0$ satisfying condition (5.72). Then, taking into account (8.199), we can write

$$w_{h\tau} \to w \text{ weakly in } \mathcal{H}^\gamma \tag{8.204}$$

and thus in view of the boundedness of $\mathcal{F}_{h\tau}$ and relation (8.201),

$$\int_{-\infty}^{\infty} (1 + |s|^{2\gamma}) \|\hat{w}_{h\tau}(s) - \hat{w}(s)\|^2_{L^2(\Omega)} \, \mathrm{d}s \le C < \infty. \tag{8.205}$$

Now we write

$$\begin{aligned} \mathcal{F}_{h\tau} &= \int\limits_{|s|\le M} \|\hat{w}_{h\tau}(s) - \hat{w}(s)\|^2_{L^2(\Omega)} \, \mathrm{d}s + \int\limits_{|s|>M} \left(1 + |s|^{2\gamma}\right) \|\hat{w}_{h\tau}(s) - \hat{w}(s)\|^2_{L^2(\Omega)} \frac{\mathrm{d}s}{1 + |s|^{2\gamma}} \le \\ &\le \mathcal{J}_{h\tau} + \frac{1}{1 + M^{2\gamma}} \int\limits_{-\infty}^{\infty} \left(1 + |s|^{2\gamma}\right) \|\hat{w}_{h\tau}(s) - \hat{w}(s)\|^2_{L^2(\Omega)} \, \mathrm{d}s \le \mathcal{J}_{h\tau} + \frac{C}{1 + M^{2\gamma}}, \end{aligned} \tag{8.206}$$

where

$$\mathcal{J}_{h\tau} = \int_{|s|\le M} \|\hat{w}_{h\tau}(s) - \hat{w}(s)\|^2_{L^2(\Omega)}\, \mathrm{d}s. \tag{8.207}$$

For a given $\varepsilon > 0$ we choose $M > 0$ such that

$$\frac{C}{1+M^{2\gamma}} \le \frac{\varepsilon}{2}. \tag{8.208}$$

Hence,

$$\mathcal{F}_{h\tau} \le \mathcal{J}_{h\tau} + \frac{\varepsilon}{2}. \tag{8.209}$$

Now we want to prove that

$$\mathcal{J}_{h\tau} \to 0 \text{ as } h, \tau \to 0. \tag{8.210}$$

As we will show, this is a consequence of the Lebesgue theorem. We have

$$\hat{w}_{h\tau}(s) = \int_{-\infty}^{\infty} w_{h\tau}(t) e^{-2\pi its}\, \mathrm{d}t = \int_{-\infty}^{\infty} w_{h\tau}(t)\chi(t) e^{-2\pi its}\, \mathrm{d}t, \quad \forall s \in I\!R, \tag{8.211}$$

where χ is the characteristic function of the interval $[0,T]$ (hence $w_{h\tau} = \chi w_{h\tau}$). Then, using (7.146), we have

$$\begin{aligned} \|\hat{w}_{h\tau}(s)\|_{L^2(\Omega)} &= \| \int_{-\infty}^{\infty} w_{h\tau}(t)\chi(t) e^{-2\pi its}\, \mathrm{d}t\|_{L^2(\Omega)} \le \\ \le \quad & \|w_{h\tau}(t)\|_{L^2(I\!R, L^2(\Omega))} \|\chi(t) e^{-2\pi its}\|_{L^2(I\!R)} \le C \end{aligned} \tag{8.212}$$

and

$$\|\hat{w}_{h\tau}(s) - \hat{w}(s)\|^2_{L^2(\Omega)} \le 2\left(C^2 + \|\hat{w}(s)\|^2_{L^2(\Omega)}\right) \ \forall s \in I\!R. \tag{8.213}$$

The function on the right-hand side of (8.213) is integrable over the interval $(-M, M)$. By definition, (8.177) is equivalent to

$$\int_{I\!R\times\Omega} (w_{h\tau} - w)\varphi\, \mathrm{d}x\, \mathrm{d}t \to 0 \quad \forall \varphi \in L^2(I\!R, L^2(\Omega)). \tag{8.214}$$

For $\vartheta \in L^2(\Omega)$ we have $\varphi(x,t) = \vartheta(x)\chi(t)e^{-2\pi its} \in L^2(I\!R, L^2(\Omega))$ for any fixed $s \in I\!R$. Then, by the definition of the Fourier transform, Fubini's theorem and (8.214),

$$\begin{aligned} &\int_\Omega (\hat{w}_{h\tau}(s) - \hat{w}(s))\,\vartheta\, \mathrm{d}x = \int_\Omega \left(\int_{I\!R} (w_{h\tau}(x,t) - w(x,t))\, e^{-2\pi its}\chi(t)\, \mathrm{d}t \right) \vartheta(x)\, \mathrm{d}x = \\ = \quad & \int_{I\!R\times\Omega} (w_{h\tau}(x,t) - w(x,t))\,\vartheta(x)\chi(t)e^{-2\pi its}\, \mathrm{d}t\, \mathrm{d}x \to 0 \text{ as } h, \tau \to 0, \end{aligned}$$

which means that

$$\hat{w}_{h\tau}(s) \to \hat{w}(s) = \int_{-\infty}^{\infty} w(t)\chi(t)e^{-2\pi its}\, \mathrm{d}t \qquad \text{weakly in } L^2(\Omega)\ \forall s \in I\!R. \tag{8.215}$$

Due to (8.211), the Cauchy inequality and (7.148), we have

$$\begin{aligned} \|\hat{w}_{h\tau}(s)\|_{X_h} &= \| \int_{-\infty}^{\infty} w_{h\tau}(t)\chi(t) e^{-2\pi its}\, \mathrm{d}t\|_{X_h} \le \\ \le \quad & \|w_{h\tau}(t)\|_{L^2(0,T;V_h)} \|\chi(t) e^{-2\pi its}\|_{L^2(I\!R)} \le C. \end{aligned} \tag{8.216}$$

Hence

$$\|J_h \hat{w}_{h\tau}(s)\|_F \leq C \text{ for all } s \in I\!R. \tag{8.217}$$

Now (8.215), reflexivity of the space F, (8.217) and assertion 2 of Lemma 8 imply that

$$J_h \hat{w}_{h\tau}(s) \to \omega \hat{w}(s) \text{ weakly in } F \text{ for each } s \in I\!R. \tag{8.218}$$

Since $\hat{w}_{h\tau}(s) \in X_h$ ($h = h_n$, $\tau = \tau_n \to 0$ as $n \to \infty$), the application of Lemma 12 implies that

$$\hat{w}_{h\tau}(s) \to \hat{w}(s) \text{ strongly in } L^2(\Omega) \text{ for all } s \in I\!R. \tag{8.219}$$

Hence,

$$\|\hat{w}_{h\tau}(s) - \hat{w}(s)\|^2_{L^2(\Omega)} \to 0 \ \forall s \in I\!R. \tag{8.220}$$

From (8.220), the bound (8.213) and the Lebesgue theorem we obtain (8.210). This proves the lemma.

□

From Lemmas 10, 11 and 14 and assertion (7.149) we can conclude that there exist sequences $h = h_n \to 0$, $\tau = \tau_n \to 0$ as $n \to \infty$ satisfying (5.72) and a function u such that

$$\begin{aligned} &J_h u_{h\tau} \to \omega u \quad \text{weakly in } L^2(-1,T;F), \\ &J_h w_{h\tau} \to \omega u \quad \text{weakly in } L^2(-1,T;F), \\ &u_{h\tau} \to u \quad \text{strongly in } L^2(\tilde{Q}_T), \\ &w_{h\tau} \to u \quad \text{strongly in } L^2(\tilde{Q}_T) \end{aligned} \tag{8.221}$$

as $n \to \infty$. Since $L^\infty(Q_T)$ is the dual to the separable Banach space $L^1(Q_T)$, the above results and (5.74) imply that

$$\begin{aligned} &u_{h\tau} \to u \text{ weak-* in } L^\infty(Q_T), \\ &w_{h\tau} \to u \text{ weak-* in } L^\infty(Q_T). \end{aligned} \tag{8.222}$$

9. Limit Process

Let us consider sequences $h = h_n$, $\tau = \tau_n \to 0$ satisfying (5.72) and assume that the corresponding approximate solutions $u_{h\tau}$, $w_{h\tau}$ satisfy conditions (5.74) and (8.221). Our goal is to show that the limit function u is a weak solution of problem (2.1)–(2.3), i.e. u satisfies (2.14)–(2.16).

Multiplying (8.152) by any $\psi \in C_0^\infty([0,T)) := \{\varphi \in C^\infty([0,T]); \varphi(T) = 0\}$, integrating over $(0,T)$, applying the integration by parts in the first term and using (4.55), which implies that $w_{h\tau}(0) = u_h^0$, we find that

$$\begin{aligned} &- \int_0^T \left(w_{h\tau}(t), \psi'(t) v_h\right)_h \, dt + \nu \int_0^T ((u_{h\tau}(t), \psi(t) v_h))_h \, dt + \\ &+ \int_0^T b_h(u_{h\tau}(t-\tau), \psi(t) v_h) \, dt = \\ &= \int_0^T \left(g_{h\tau}(t), \psi(t) v_h\right)_h \, dt + (u_h^0, v_h)\psi(0), \quad v_h \in V_h, \ \psi \in C_0^\infty([0,T)). \end{aligned} \tag{9.223}$$

For $t \in [0,T], v_h \in V_h,\ \psi \in C_0^\infty([0,T))$ we set

$$\vartheta_{h\tau}(t;\psi,v_h) = (g_{h\tau}(t), \psi(t)v_h)_h - (g_{h\tau}(t), \psi(t)v_h)\,. \tag{9.224}$$

Taking into account (6.96), we see that (9.223) is equivalent to

$$\begin{aligned} &- \int_0^T (w_{h\tau}(t), \psi'(t)v_h)\ \mathrm{d}t + \nu \int_0^T ((u_{h\tau}(t), \psi(t)v_h))_h\ \mathrm{d}t + \\ &+ \int_0^T b_h(u_{h\tau}(t-\tau), \psi(t)v_h)\ \mathrm{d}t = \\ &= \int_0^T (g_{h\tau}(t), \psi(t)v_h)\ \mathrm{d}t + (u_h^0, v_h)\psi(0) + \int_0^T \vartheta_{h\tau}(t;\psi,v_h)\ \mathrm{d}t. \end{aligned} \tag{9.225}$$

In virtue of (6.97) and (2.10), we obtain

$$\left| \int_0^T \vartheta_{h\tau}(t;\psi,v_h)\ \mathrm{d}t \right| \le ch\|v_h\|_{X_h}. \tag{9.226}$$

Let $v \in C_0^\infty(\Omega), v_h = I_h v$. From (6.87), (6.90) and (6.92) we have

$$\begin{aligned} &\|v_h - v\|_{L^2(\Omega)} \le \hat{c}_1\|v_h - v\|_{X_h} \le ch\|v\|_{H^2(\Omega)} \\ &\|v_h\|_{X_h} \le c, \quad h \in (0,h_0). \end{aligned} \tag{9.227}$$

This implies that

$$J_h v_h \to \omega v \text{ strongly in } F. \tag{9.228}$$

Hence,

$$\psi' v_h \to \psi' v \quad \text{strongly in } L^2(Q_T) \tag{9.229}$$

and

$$J_h \psi v_h \to \omega \psi v \quad \text{strongly in } L^2(0,T;F). \tag{9.230}$$

The analysis of the limit process will be divided into several lemmas. In what follows we consider sequences $h = h_n \to 0$, $\tau = \tau_n \to 0$ as $n \to \infty$ satisfying (5.72), such that (8.221) and (8.222) hold.

Lemma 15 *Let $\psi(t) \in C_0^\infty([0,T))$ and let $w_{h\tau}, v_{h\tau}$ be two sequences satisfying*

$$w_{h\tau} \to u \qquad \textit{strongly in } L^2(Q_T), \tag{9.231}$$

$$v_{h\tau} \to v \qquad \textit{strongly in } L^2(Q_T). \tag{9.232}$$

Then

$$\int_0^T (w_{h\tau}(t), \psi'(t)v_{h\tau})\ \mathrm{d}t \to \int_0^T (u(t), \psi'(t)v)\ \mathrm{d}t \tag{9.233}$$

as $h = h_n \to 0$ and $\tau = \tau_n \to 0$.

Proof is evident. □

Lemma 16 *Let $\psi(t) \in C_0^\infty([0,T))$ and let $u_{h\tau}, v_{h\tau}$ be two sequences satisfying*

$$J_h u_{h\tau} \to \omega u \qquad \text{weakly in } L^2(0,T;F), \tag{9.234}$$

$$J_h v_h \to \omega v \qquad \text{strongly in } F. \tag{9.235}$$

Then

$$\int_0^T ((u_{h\tau}(t), \psi(t) v_{h\tau}))_h \,\mathrm{d}t \to \int_0^T ((u(t), \psi(t) v)) \,\mathrm{d}t \tag{9.236}$$

as $h = h_n \to 0$ and $\tau = \tau_n \to 0$.

Proof. From (9.235) it follows that

$$J_h \psi(t) v_h \to \psi(t) v \text{ strongly in } L^2(0,T;F). \tag{9.237}$$

We have

$$\begin{aligned}
\int_0^T ((u_{h\tau}(t), \psi(t) v_h))_h \,\mathrm{d}t &= \int_0^T (J_h u_{h\tau}(t), J_h \psi(t) v_h)_F \,\mathrm{d}t - \int_0^T (u_{h\tau}(t), \psi(t) v_h) \,\mathrm{d}t, \\
\int_0^T ((u(t), \psi(t) v))_h \,\mathrm{d}t &= \int_0^T (\omega u(t), \psi(t) \omega v)_F \,\mathrm{d}t - \int_0^T (u(t), \psi(t) v) \,\mathrm{d}t.
\end{aligned}$$

By the definition, (9.234) means that

$$\int_0^T (J_h u_{h\tau}(t), \vartheta(t)) \,\mathrm{d}t \to \int_0^T (\omega u(t), \vartheta(t)) \,\mathrm{d}t \quad \forall \vartheta \in L^2(0,T;F). \tag{9.238}$$

We can write

$$\left| \int_0^T ((u_{h\tau}(t), \psi(t) v_{h\tau}))_h \mathrm{d}t - \int_0^T ((u(t), \psi(t) v)) \mathrm{d}t \right| \le \tag{9.239}$$

$$\begin{aligned}
\le\ & \left| \int_0^T (J_h u_{h\tau}(t), J_h \psi(t) v_h)_F \,\mathrm{d}t - \int_0^T (\omega u(t), \psi(t) \omega v)_F \,\mathrm{d}t \right| + \\
& + \left| \int_0^T (u_{h\tau}(t), \psi(t) v_h) \,\mathrm{d}t - \int_0^T (u(t), \psi(t) v) \,\mathrm{d}t \right| \le
\end{aligned}$$

$$\begin{aligned}
\le\ & \int_0^T \|J_h u_{h\tau}(t)\|_F \|\psi(t)(J_h v_h - \omega v)\|_F \mathrm{d}t + \\
& + \left| \int_0^T (J_h u_{h\tau}(t), \psi(t) \omega v)_F \,\mathrm{d}t - \int_0^T (\omega u(t), \psi(t) \omega v) \,\mathrm{d}t \right| + \\
& + \left| \int_0^T (u_{h\tau}(t), \psi(t) v_h) \,\mathrm{d}t - \int_0^T (u_{h_\tau}(t), \psi(t) v) \,\mathrm{d}t \right| + \\
& + \left| \int_0^T (u_{h\tau}(t), \psi(t) v) \,\mathrm{d}t - \int_0^T (u(t), \psi(t) v) \,\mathrm{d}t \right| \to 0,
\end{aligned} \tag{9.240}$$

as follows from the boundedness of the sequence $\{J_h u_{h\tau}\}$ in $L^2(0,T;F)$, (9.234), (9.237) and (9.238), where we substitute $\vartheta(t) = \psi(t) v$. □

Lemma 17 *Let $\psi(t) \in C_0^\infty([0,T))$ and let $u_{h\tau}, v_h$ be two sequences satisfying*

$$u_{h\tau} \to u \qquad \text{strongly in } L^2(\tilde{Q}_T), \quad \tilde{Q}_T = \Omega \times (-1,T), \tag{9.241}$$

$$J_h u_{h\tau} \to \omega u \qquad \text{weakly in } L^2(-1,T;F), \tag{9.242}$$

$$J_h v_h \to \omega v \qquad \text{strongly in } F, v \in C_0^\infty(\Omega). \tag{9.243}$$

Then

$$\int_0^T b_h(u_{h\tau}(t-\tau), v_h)\psi(t)\,\mathrm{d}t \to \int_0^T b(u(t),v)\psi(t)\,\mathrm{d}t \tag{9.244}$$

as $h = h_n \to 0,\ \tau = \tau_n \to 0$.

Proof. We write

$$\begin{aligned}
b_h(u_{h\tau}(t-\tau), v_h) - b(u(t),v) &= \\
= b_h(u_{h\tau}(t-\tau), v_h) - \tilde{b}_h(u_{h\tau}(t-\tau), v_h) &+ \qquad (=: \sigma(1)\ ) \\
+\tilde{b}_h(u_{h\tau}(t-\tau), v_h) - \tilde{b}_h(u_{h\tau}(t-\tau), v) &+ \qquad (=: \sigma(2)\ ) \\
+\tilde{b}_h(u_{h\tau}(t-\tau), v) - b(u(t-\tau), v) &+ \qquad (=: \sigma(3)\ ) \\
+b(u(t-\tau), v) - b(u(t), v) & \qquad\ \ (=: \sigma(4)\ )
\end{aligned}$$

($\tilde{b}_h$ is defined in (3.30)) and successively estimate the terms $\sigma(1) - \sigma(4)$:
In virtue of (6.98),

$$|\sigma(1)| \leq ch^{1-\kappa}(\|u_{h\tau}(t-\tau)\|_{X_h}^2 + \|u_{h\tau}(t-\tau)\|_{X_h})\|v_h\|_{X_h}$$

with $\kappa \in (0,1)$. Hence,

$$\left|\int_0^T \left[b_h(u_{h\tau}(t-\tau), v_h) - \tilde{b}_h(u_{h\tau}(t-\tau), v_h)\right] \ \psi(t)\ \mathrm{d}t\right| \leq \tag{9.245}$$

$$\leq \ ch^{1-\kappa}\|v_h\|_{X_h}\left[\int_0^T \|u_{h\tau}(t-\tau)\|_{X_h}^2 \mathrm{d}t + \left(\int_0^T \|u_{h\tau}(t-\tau)\|_{X_h}^2\ \mathrm{d}t\right)^{1/2}\right] \leq ch^{1-\kappa},$$

as follows from (7.147), (9.242) and (9.243). Similarly, using the Cauchy inequality and (9.227), we obtain

$$\left|\int_0^T \left[\tilde{b}_h(u_{h\tau}(t-\tau), v_h) - \tilde{b}_h(u_{h\tau}(t-\tau), v)\right] \psi(t)\ \mathrm{d}t\right| \leq \tag{9.246}$$

$$\leq 2 \max_{|\xi|\leq M} \max_{s=1,2} |f_s'(\xi)| \int_0^T \|u_{h\tau}(t-\tau)\|_{X_h}^2 \mathrm{d}t \|v_h - v\|_{L^2(\Omega)} \leq$$

$$\leq c\|J_h v_h - \omega v\|_F \to 0.$$

Further,

$$\sigma(3) \ = \ \sum_{i\in I}\int_{T_i}\sum_{s=1}^2 \left(\frac{\partial f_s(u_{h\tau}(t-\tau))}{\partial x_s} - \frac{\partial f_s(u(t-\tau))}{\partial x_s}\right) v\,\mathrm{d}x =$$

$$
\begin{aligned}
&= \sum_{i\in I}\int_{T_i}\sum_{s=1}^{2}\left(f_s'(u_{h\tau}(t-\tau))\frac{\partial u_{h\tau}(t-\tau)}{\partial x_s} - f_s'(u(t-\tau))\,\frac{\partial u(t-\tau)}{\partial x_s}\right) v\,\mathrm{d}x = \\
&= \overbrace{\sum_{i\in I}\int_{T_i}\sum_{s=1}^{2}\left(f_s'(u_{h\tau}(t-\tau)) - f_s'(u(t-\tau))\right)\frac{\partial u_{h\tau}(t-\tau)}{\partial x_s}}^{\sigma^*(3)} v\,\mathrm{d}x + \\
&\quad + \sum_{i\in I}\int_{T_i}\sum_{s=1}^{2} f_s'(u(t-\tau))\left(\frac{\partial u_{h\tau}(t-\tau)}{\partial x_s} - \frac{\partial u(t-\tau)}{\partial x_s}\right) v\,\mathrm{d}x.
\end{aligned}
$$

Using the mean value theorem in the integral form, we find that

$$
\begin{aligned}
|\sigma^*(3)| &\le \max_{\xi\in[-M,M]}\max_{s=1,2}|f_s''(\xi)|\sum_{i\in I}\int_{T_i}|u_{h\tau}(t-\tau)-u(t-\tau)|\,|\nabla u_{h\tau}(t-\tau)|\;|v|\,\mathrm{d}x \le \\
&\le c\,\|u_{h\tau}(t-\tau)-u(t-\tau)\|_{L^2(\Omega)}\,\|u_{h\tau}(t-\tau)\|_{X_h}\,.
\end{aligned}
$$

Since $\psi(t)=0$ for $t\ge T$, the substitution $t:=t-\tau$ yields

$$
\begin{aligned}
&\left|\int_0^T\left[\tilde{b}_h(u_{h\tau}(t-\tau),\,v) - b(u(t-\tau),\,v)\right]\psi(t)\,\mathrm{d}t\right| \le \qquad (9.247)\\
&\le c\left(\|u_{h\tau}-u\|^2_{L^2(Q_T)} + \tau\|u_h^0-u^0\|^2_{L^2(\Omega)}\right)^{1/2}\left(\|J_h\,u_{h\tau}\|_F^2 + \tau\|u_h^0\|^2_{X_h}\right)^{1/2} + \\
&\quad + \left|\int_0^T\psi(t+\tau)\sum_{i\in I}\int_{T_i}\sum_{s=1}^{2}f_s'(u(t))\left(\frac{\partial u_{h\tau}(t)}{\partial x_s} - \frac{\partial u(t)}{\partial x_s}\right)v\,\mathrm{d}x\mathrm{d}t + \right.\\
&\quad \left. + \tau\sum_{i\in I}\int_{T_i}\sum_{s=1}^{2}f_s'(u^0)\left(\frac{\partial u_h^0}{\partial x_s} - \frac{\partial u^0}{\partial x_s}\right)v\,\mathrm{d}x\right| \longrightarrow 0 \quad \text{as } h,\,\tau\to 0
\end{aligned}
$$

due to (9.241), (9.242) and (6.93) valid for $\varphi = u_0$.

Similar calculation yields the estimate

$$
\begin{aligned}
&\left|\int_0^T(b(u(t-\tau),v) - b(u(t),v))\,\psi(t)\;\mathrm{d}t\right| \le \qquad (9.248)\\
&\le c\int_0^T\int_\Omega|u(x,t-\tau)-u(x,t)|^2\;\mathrm{d}x\,\mathrm{d}t\;\|v\|_{H^1(\Omega)} \to 0 \quad \text{as } h,\tau\to 0,
\end{aligned}
$$

which is a consequence of the continuity in the mean of $u\in L^2(\tilde{Q}_T)$. (Cf., e.g., [43], Theorem 2.4.2.)

Now (9.245) – (9.248) immediately imply (9.244). □

Lemma 18 *Let $\psi(t)\in C_0^\infty([0,T))$ and let $v_{h\tau}, g_{h\tau}$ be two sequences such that*

$$
\begin{aligned}
&v_h \to v \qquad \textit{strongly in } L^2(\Omega), && (9.249)\\
&g_{h\tau}(t) = g^k = g(\cdot,t_k),\ \forall t\in[t_k,t_{k+1}), && (9.250)
\end{aligned}
$$

where g satisfies (2.10). *Then*

$$
\int_0^T(g_{h\tau}(t),v_h)\,\psi(t)\;\mathrm{d}t \to \int_0^T(g(t),v)\,\psi(t)\;\mathrm{d}t \quad \textit{as } h,\tau\to 0. \qquad (9.251)
$$

Proof.

Obviously, by (2.10) and (9.249),

$$\left|\int_0^T ((g_{h\tau}(t), v_h) - (g(t), v))\, \psi(t)\ \mathrm{d}t\right| \leq \tag{9.252}$$

$$\leq\ c\int_0^T \|g_{h\tau}(t)\|_{L^2(\Omega)}\|v_h - v\|_{L^2(\Omega)}\ \mathrm{d}t + c\int_0^T \|g_{h\tau}(t) - g(t)\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)}\ \mathrm{d}t \leq$$

$$\leq\ c\|v_h - v\|_{L^2(\Omega)} + c\left(\int_0^T \|g_{h\tau}(t) - g(t)\|^2_{L^2(\Omega)}\ \mathrm{d}t\right)^{1/2}.$$

In virtue of the uniform continuity of the mapping $g : [0,T] \to L^2(\Omega)$, we have

$$\int_0^T \|g_{h\tau}(t) - g(t)\|^2_{L^2(\Omega)}\ \mathrm{d}t = \sum_{k=0}^{r-1}\int_{t_k}^{t_{k+1}} \|g(t_k) - g(t)\|^2_{L^2(\Omega)}\ \mathrm{d}t \to 0 \quad \text{as } \tau \to 0.$$

This, (9.249) and (9.252) yield (9.251). □

Finally, using (6.87), (6.93) and (9.227) we conclude that

$$(u_h^0, v_h) \to (u^0, v) \quad \text{as } h \to 0. \tag{9.253}$$

Now, summarizing (9.223), (9.224), (9.225), (9.226), (9.233), (9.236), (9.244), (9.251) and (9.253), we see that the limit function $u \in L^2(0,T;V) \cap L^\infty(Q_T)$ satisfies the identity

$$-\int_0^T (u(t), v)\, \psi'(t)\ \mathrm{d}t + \nu\int_0^T ((u(t), v))\psi(t)\ \mathrm{d}t + \int_0^T b\,(u(t), v)\, \psi(t)\ \mathrm{d}t = \tag{9.254}$$

$$=\ \int_0^T (g(t), v)\, \psi(t)\ \mathrm{d}t + (u^0, v)\psi(0), \quad v \in C_0^\infty(\Omega),\ \ \psi \in C_0^\infty([0,T)).$$

Since the space $C_0^\infty(\Omega)$ is dense in V, (9.254) holds for all $v \in V$. Moreover, $C_0^\infty((0,T)) \subset C_0^\infty([0,T))$ and identity (9.254) implies (2.17). Hence, u satisfies (2.14) – (2.15).

It is possible to show that $u' \in L^2(0,T;V^*)$ and

$$\langle u'(t), v\rangle + \nu((u(t), v)) + b(u(t), v) = (g(t), v), \quad v \in V, \quad \text{a.e. } t \in (0,T), \tag{9.255}$$

where $\langle\cdot,\cdot\rangle$ is the duality between V^* and V. (Cf., e.g., [62], Sec. 8.6.)

If we multiply (9.255) by any $\psi \in C_0^\infty([0,T))$, integrate over $(0,T)$ and transform the first term with the aid of integration by parts, we obtain the identity

$$-\int_0^T (u(t), v)\, \psi'(t)\ \mathrm{d}t + \nu\int_0^T ((u(t), v))\psi(t)\ \mathrm{d}t + \int_0^T b\,(u(t), v)\, \psi(t)\ \mathrm{d}t = \tag{9.256}$$

$$=\ \int_0^T (g(t), v)\, \psi(t)\ \mathrm{d}t + (u(0), v)\psi(0), \quad v \in V,\ \ \psi \in C_0^\infty([0,T)).$$

The comparison of (9.254) (with $v \in V$) and (9.256) immediately implies that $u(0) = u^0$. Hence, we have proved that u is a solution of problem (2.14) – (2.16). □

On the basis of the above considerations we come to the following conclusion:

Let us consider approximate solutions of problem (2.14) – (2.16) obtained from (3.38) – (3.40) with $\tau,\ h > 0$ satisfying condition (5.72). Then the system of functions $u_{h\tau},\ w_{h\tau}$ defined by (4.54) and (4.55) can be split into sequences converging in the sense of (8.221) and (8.222). Every limit function of such a sequence is a solution of problem (2.14) – (2.16). (As we see, we have proved the existence of a weak solution of (2.1) – (2.3).) Taking into account the uniqueness of the solution of (2.14) – (2.16) we obtain the convergence of the whole systems $\{u_{h\tau}\},\ \{w_{h\tau}\}$ to the weak solution u of problem (2.1) – (2.3). Thus, we come to the **main result** of this paper:

Theorem 6 *Let us assume that the domain* $\Omega \subset I\!R^2$ *is bounded and polygonal and that conditions* (2.9) – (2.11)*,* (3.18) – (3.20)*,* (3.41) – (3.44)*,* (4.56) – (4.58)*,* (5.60) *and* (5.61) *are satisfied. For* $h \in (0, h_0)$ *and* $\tau \in (0, T)$ *let us construct approximate solutions with the aid of the finite volume – finite element scheme* (3.38) – (3.40) *and define functions* $u_{h\tau}$ *and* $w_{h\tau}$ *by* (4.54) *and* (4.55)*. Then the systems* $\{u_{h\tau}\},\ \{w_{h\tau}\}$ *with* $h \in (0, h_0),\ \tau \in (0, T)$ *satisfying the "stability condition"* (5.72) *fulfil estimates* (5.73) *and* (7.145) – (7.148)*. Moreover,*

$$
\begin{aligned}
&J_h u_{h\tau},\ J_h w_{h\tau} \to \omega u \quad \textit{weakly in } L^2(0,T;F),\\
&u_{h\tau},\ w_{h\tau} \to u \quad \textit{weak-* in } L^\infty(Q_T),\\
&u_{h\tau},\ w_{h\tau} \to u \quad \textit{strongly in } L^2(Q_T), \quad \textit{as } h,\ \tau \to 0,\ \ h,\tau \textit{ satisfy } (5.72),
\end{aligned}
$$

where u is the unique weak solution of problem (2.1) – (2.3) *(i.e., u satisfies* (2.14) – (2.16)*).*

Remark 2 There are several unsolved problems connected with the above results:

- error estimates and a posteriori error estimates,
- analysis of the problem in a nonpolygonal domain, i.e., the effect of the approximation of a curved boundary,
- analysis of the problem with nonhomogeneous Dirichlet boundary conditions and/or mixed Dirichlet - Neumann boundary conditions.

Acknowledgements. The research of M. Feistauer, J. Felcman and V. Dolejší has been supported under the Grant No. 3/1998/B MAT/MFF of the Grant Agency of the Charles University. The authors gratefully acknowledge this support.

Bibliographie

[1] D. Adam, A. Felgenhauer, H.-G. Ross, and M. Stynes. A nonconforming finite element method for a singularly perturbed boundary value problem. Preprint, 1992–10, Department of Mathematics, University College, Cork, Ireland, 1982.

[2] D. Adam and H.-G. Ross. A nonconforming exponentially field fitted finite element method I: The interpolation error. Preprint MATH-nm-06-1993, Technische Universität Dresden.

[3] P. Arminjou, A. Dervieux, L. Fezoui, H. Steve, and B. Stoufflet. Non-oscillatory schemes for multidimensional Euler calculations with unstructured grids. Technical report, Vieweg, Braunschweig–Wiesbaden, 1989. In Ballman J. and Jeltsch R. (eds) *Notes on Numerical Fluid Mechanics Volume 24 (Nonlinear Hyperbolic Equations – Theory, Computation Methods and Applications).*

[4] I. Babuška and W. C. Rheinboldt. A posteriori error estimators for the finite element method. *Internat. J. Numer. Methods Engrg.*, 12:1597–1615, 1978.

[5] R. Bank and A. Weiser. Some a posteriori erroe estimators for elliptic partial differential equations. *Math. Comp.*, 44:283–301, 1985.

[6] T. J. Barth and D. C. Jesperson. The design and application of upwinding schemes on unstructured meshes. *AIAA journal*, page 366, 1989.

[7] H. Borouchaki, P. L. George, F. Hecht, P. Laugh, and E. Saltel. Mailleur bidimensionnel de Delaunay gouverné par une carte de métriques. Partie I: Algorithmes. Technical Report 2741, INRIA, 1995.

[8] P. G. Ciarlet. *The Finite Elements Method for Elliptic Problems.* North-Holland, Amsterodam, 1979.

[9] M. Crouzeix and P.-A. Raviart. Conforming and nonconforming finite element methods for solving the stationary Stokes equations. *RAIRO, Anal. Numér.*, 7:33–76, 1973.

[10] H. Beirao da Veiga. Stationary motions and the incompressible limit for compressible viscous fluid. *Houston J. Math.*, 13:327–341, 1987.

[11] E. F. D'Azevedo and R. B. Simpson. On optimal triangular meshes for minimizing the gradient error. *Numer. Math.*, 59:321–348, 1991.

[12] M. J. Castro Díaz, H. Borouchaki, P. L. George, F. Hecht, and B. Mohammadi. Anisotropic Adaptive Mesch Generation in Two Dimensions for CFD. In *Proceeding of ECCOMAS96 – Computational Fluid Dynamics, Paris*, pages 181–186, 1996.

[13] V. Dolejší and P. Angot. Finite Volume Methods on Unstructured Meshes for Compressible Flows. In F. Benkhaldoun and R. Vilsmeier, editors, *Finite Volumes for Complex Applications (Problems and Perspectives)*, pages 667–674. Hermes, Rouen, 1996. ISBN 2-86601-556-8.

[14] V. Dolejší and M. Feistauer. On the discrete Friedrichs inequality for nonconforming finite elements. Preprint, Faculty of Mathematics and Physics, Charles University, Prague 1997.

[15] L. J. Durlovsky, B. Enquist, and S. Osher. Triangle based adaptive stencils for the solution of hyperbolic conservation laws. *J. of Comput. Phys.*, 98:64–73, 1992.

[16] K. Eriksson and C. Johnson. Adaptive finite element methods for parabolic problems I–IV. Preprints, Chalmers University of Technology, 1990-1992.

[17] M. Feistauer. *Mathematical methods in fluid dynamics.* Longman Scientific & Technical,Harlow, 1993.

[18] M. Feistauer and J. Felcman. Convection-Diffusion Problems and Compressible Navier–Stokes Equations. In J. R. Whitemen, editor, *The Mathematics of Finite Elements and Applications*, pages 175–194. John Wiley & Sons, 1996.

[19] M. Feistauer, J. Felcman, and V. Dolejší. Numerical Simulation of Compresssible Viscous Flow through Cascades of Profiles. *ZAMM*, 76:297–300, 1996.

[20] M. Feistauer, J. Felcman, and M. Lukáčová. Combined finite elements–finite volume solution of compressible flow. *J. Comput. Appl. Math.*, 63:179–199, 1995.

[21] M. Feistauer, J. Felcman, and M. Lukáčová. On the convergence of a combined finite volume – finite element method for nonlinear convection – diffusion problems. *Numer. Methods for PDE's*, 13:163–190, 1997.

[22] M. Feistauer, J. Felcman, M. Lukáčová, and G. Warnecke. Error estimates of a combined finite volume – finite element method for nonlinear convection – diffusion problems. Preprint, Faculty of Mathematics and Physics, Charles Univ., Prague, 1996.

[23] M. Feistauer, J. Felcman, M. Rokyta, and Z. Vlášek. Finite element solution of flow problems with trailing conditions. *J. Comput. Appl. Math*, 44:131–145, 1992.

[24] M. Feistauer and P. Stupka. On the convergence of the combined finite volume – finite element method for nonlinear convection – diffusion problems. ii Explicit schemes. Preprint, Faculty of Mathematics and Physics, Charles Univ., Prague, 1997.

[25] J. Felcman. Finite volume solution of the inviscid compressible fluid flow. *ZAMM*, 72:513–516, 1992.

[26] J. Felcman and V. Dolejší. Adaptive Methods for the Solution of the Euler Equations in Elements of the Blade Machines. *ZAMM*, 76:301–304, 1996.

[27] J. Felcman and V. Dolejší. Adaptive Methods in Internal and External Flow Computations. In M. P. Bendsøe M. Brøns and M. P. Sørensen, editors, *Progress in Industrial Mathematics at ECMI 96*, pages 424–431. B. G. Teubner, Stuttgart, 1997. ISBN 3-519-02607-4.

[28] J. Felcman, V. Dolejší and M. Feistauer. Adaptive Finite Volume Method for the Numerical Solution of the Compressible Euler Equations. In J. Périaux S. Wagner, E. H. Hirschel and R. Piva, editors, *Computational Fluid Dynamics '94*, pages 894–901. John Wiley and Sons, Stuttgart, 1994. ISBN 0-471-950637.

[29] J. Felcman and G. Warnecke. Adaptive computational methods for gas flow. In *Proceedings of the Prague Mathematical Conference, Prague*, pages 99–104. ICARIS, 1996.

[30] L. Flandrin, P. Charrier, and B. Dubroca. A robust finite volume method for computations of two-dimensional unstructured hybrid meshes. In S. Wagner, E. H. Hirschel,

J. Périaux, and R. Piva, editors, *Computational Fluid Dynamics '94*, pages 301–308, ISBN 0-471-950637, 1994. John Wiley and Sons, Stuttgart.

[31] J. Fořt, M. Huněk, K. Kozel, and M. Vavřincová. Numerical simulation of steady and unsteady flows through plane cascades. In R. Ranacher M. Feistauer and K. Kozel, editors, *Numerical Modeling in Continuum Mechanics II*, pages 95–102. Faculty of Mathematics and Physics, Charles Univ., Prague, 1995.

[32] M. Fortin, M.-G. Vallet, J. Dompierre, Y. Bourgault, and W. G. Habashi. Anisotropic Mesch Adaptation: Theory, Validation and Applications. In *Proceeding of ECCOMAS96 – Computational Fluid Dynamics, Paris*, pages 174–180, 1996.

[33] M. Geiben. Numerische Simulation von dreidimensionalen instationären kompressiblen Strömungen in komplexen Geometrien. In *Strömungen mit Ablösung, Proc. of 8. DGLR–Fach–Symposium*, pages 351–360. DGLR–Bericht 92–07, Bonn, Nov. 1992.

[34] R. Glowinski, J. L. Lions, and R. Trémalières. *Analyse numérique des inéquations variationnelles*. Dunod, Paris, 1976.

[35] V. Hannemann, D. Hemple, and T. Sonar. Adaptive computation of compressible flow. In G. Wittum W. Hackbusch, editor, *Proceedings of the 9th GAMM Seminar on Adaptive Methods*. Vieweg, Braunschweig – Wiesbaden, 1992.

[36] T. Ikeda. Maximum principle in finite element models for convection–diffusion phenomena. In *Mathematics Studies 76, Lecture Notes in Numerical and Applied Analysis Vol. 4, North-Holland, Amsterdam – New York – Oxford*, 1983.

[37] C. Johnson. Finite element methods for convection – diffusion problems. In R. Glowinski and J. L. Lions, editors, *Computing Methods in Engineering and Applied Sciences V*. North–Holland, Amsterodam, 1981.

[38] C. Johnson and P. Hansbo. Adaptive finite element methods in computational mechanics. *Comput. Methods Appl. Mech. Engrg.*, 101:143–181, 1992.

[39] P. Knobloch. Numerical modelling of viscous compressible flows. Master's thesis, Charles University, 1993. (en tchèque).

[40] D. Kröner. *Numerical Schemes for Conservation Laws*. Wiley – Teubner, Stuttgart, 1997.

[41] D. Kröner and M. Rokyta. Convergence of upwind finite volume schemes for scalar conservation laws in 2D. Preprint no **208**, SFB, Univ. Bonn, 1992.

[42] S. N. Kruzhkov. First order quasilinear equations in several independent variables. *Maath. USSR Sbornik*, 10:217–243, 1970.

[43] A. Kufner, O. John, and S. Fučík. *Function Spaces*. Academia, Prague, 1977.

[44] O. A. Ladyzhenskaya. *The Mathematical Theory of Viscous Incompressible Flow*. Nauka, Mosow, 1970. (en russe).

[45] G. Lukaszewicz. An existence theorem for compressible viscous and heat conducting fluids. *Math. Methd. Appl. Sci.*, 6:234–247, 1984.

[46] A. Matsumura and T. Nishida. Initial boundary value problems for the equations of motion of general fluids. In Lions J. L. Glowinski R., editor, *Computing Methods in Applied Sciences and Engineering*. North – Holland, Amsterodam, 1982.

[47] M. Lukáčová Medviďová. *Combined finite element–finite volume solution of compressible flow*. PhD thesis, Charles University of Prague, 1994.

[48] K. W. Morton. *Numerical Solution of Convection – Diffusion Problems.* Chapman & Hall, London, 1996.

[49] T. Oden. Progress in adaptive methods in computational fluid dynamics. In Flaherty et al., editor, *Adaptive Methods in PDE's, Proc. of International Workshop*, pages 206–252. North – Holland, Amsterodam, 1988.

[50] K. Ohmori and T. Ushijima. A technique of upstream type applied to a linear nonconforming finite element approximation of convective diffusion equations. *RAIRO, Anal. Numér.*, 18:309–322, 1984.

[51] S. Osher and F. Solomon. Upwind difference schemes for hyperbolic systems of conservation laws. *Math. Comp.*, 38:339–374, 1982.

[52] M. Padula. Existence and uniqueness for viscous steady compressible motions. *Arch. Rational Mach., Anal.*, 97:89–102, 1987.

[53] O. Pironneau. *Finite Element Methods for Fluids.* Wiley, Chichester – New York – Brisbane – Toronto – Singapore, 1989.

[54] H.-G. Ross, M. Stynes, and L. Tobiska. Numerical Methods for Singularly Perturbed Differential Equation. *Springer Series in Computational Mathematics*, 24, 1996. Springer-Verlag, Berlin.

[55] F. Schieweck and L. Tobiska. A nonconforming finite element method of upstream type applied to the stationary Navier – Stokes equation. M^2AN, 23:627–647, 1989.

[56] S. Schochet. The compressible Euler equations in a bounded domain: Existence of solutions and incompressible limit. *Comm. Math. Phys.*, 104:49–75, 1986.

[57] G. A. Sod. A survey of several finite difference methods for system of nonlinear hyperbolic conservation laws. *J. of Comput. Phys.*, 27:1–31, 1978.

[58] T. Sonar, J. Felcman, G. Warnecke, and W. Wendland. Adaptive Berechnung kompressible Strömungsfelder. In *Strömungen mit Ablösung, Proc. of 8. DGLR–Fach–Symposium*, pages 345–350. DGLR–Bericht 92–07, Bonn, Nov. 1992.

[59] M. Šťastný and P. Šafařík. Experimental analysis data on the transonic flow past a plane turbine. Technical report, ASME Paper 90-GT-313, New York, 1990.

[60] G. Strang. Variational crimes in the finite element method. In *The Mathematical Foundations of the Finite Element Method (A. K. Aziz, ed.)*, pages 689–710. Academic Press, New York, 1972.

[61] J. Tani. On the first initial – boundary value problem of compressible viscous fluid montion. *Publ. RIMS Kyoto Univ.*, 13:193–253, 1977.

[62] R. Teman. *Navier – Stokes Equations.* North – Holland, Amsterodam – New York – Oxford, 1977.

[63] L. Tobiska. Full and weighted upwind finite element methods. In J. W. Schmidt and H. Spath, editors, *Splines in Numerical Analysis Mathematical Research Volume*, volume 32. Akademie-Verlag, Berlin, 1989.

[64] A. Valli. Mathematical results for compressible flows. Preprint Univ. degli Studi di Trento, 1992.

[65] A. Valli and W. M. Zajaczkowski. Navier – S tokes equations for compressible fluids: Global existence and qualitative properties of the solution in the general case. *Comm. Math. Phys.*, 103:259–296, 1986.

[66] G. Vijayasundaram. Transonic flow simulation using upstream centered scheme of Godunov type in finite elements. *J. Comp. Phys.*, 63:416–433, 1986.

[67] P. Woodward and P. Colella. The numerical simulation of two-dimensional fluid flow with strong shocks. *J. of Comput. Phys.*, 54:115–173, 1984.

[68] G. Zhou. A local L^2 – error analysis of the streamline diffusion method for nonstationary convection – diffusion systems. M^2AN, 29:577–603, 1995.

[69] G. Zhou and R. Rannacher. Pointwise superconvergence of streamline diffusion finite – element method. *Numer. Methods for PDE's*, 12:123–145, 1996.