Energy Methods
in 3D Spline Approximations of the Navier-Stokes equations

by

Gerard M. Awanou

(Under the direction of Ming-Jun Lai)

Abstract

In this work, we use splines of arbitrary degree and arbitrary smoothness to find approximations of the 3D Navier-Stokes equations in velocity-pressure formulation. We start by showing how to use splines to solve the 3D Poisson equation, the 3D biharmonic equation and the 3D Stokes equations. For these problems the equations are put in variational form and the associated energy functional minimized over a subset of a space of splines. For the Navier-Stokes equations, we also derived based on energy arguments the discrete equations satisfied by the solution. Numerous numerical results are provided.

Index words: Energy, Lagrange multipliers, Matlab, Navier-Stokes equations, Trivariate Splines, Refinement

Energy Methods
in 3D Spline Approximations of the Navier-Stokes equations

by

Gerard M. Awanou

Licence, Université Nationale du Bénin, Bénin, 1995

Maitrise, Université Nationale du Bénin, Bénin, 1996

ICTP Diploma, Abdus Salam International Centre of Theoretical Physics, Italy, 1998

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2003

Energy Methods
in 3D Spline Approximations of the Navier-Stokes equations

by

Gerard M. Awanou

Approved:

Major Professor: Ming-Jun Lai

Committee: Malcolm Adams
Edward Azoff
Kenneth Johnson
Paul Wenston

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2003

Acknowledgments

I would like to thank my advisor Ming-Jun Lai for having supervised this dissertation. His 2D codes have helped speed and debug my 3D codes. I also thank him for discussions that helped improve this work. I would also like to thank him to have challenged me by not showing me many of the things he understood leading me to dig deeper. Finally I would like to thank the members of my advisory committee who at different levels have supported me through this dissertation.

Table of Contents

Chapter 1

Introduction

We consider in this dissertation numerical solutions of Partial Differential Equations using splines of arbitrary degree and arbitrary smoothness over an arbitrary tetrahedral partition of a bounded domain in $\mathbf{R}^3$. The basic strategy used here is to minimize a suitable functional over a subset of a spline space. We have traced back the idea of minimizing functionals to as early as 1972 where the first theoretical study of the method of Lagrange multipliers appeared [Babuska'72]. The book [Fortin and Glowinski'83] gives an account of such methods. The main contribution of this dissertation is that we have implemented approximations of PDE's on bounded domains in $\mathbf{R}^3$ using splines of arbitrary degree and arbitrary smoothness. The approach we take here using the B-form of splines over tetrahedral partitions falls within the framework of the so-called Generalized Finite Elements [Babuska'96]. We notice that extension to arbitrary smoothness has been announced in [Babuska'96] but has not been yet implemented tested and experimented. However in the framework of [Babuska'96], the basis functions are assumed to already have some smoothness properties. In [Babuska'72], Lagrange multipliers are only used to enforce the boundary conditions, the idea to use them to systematically enforce smoothness conditions does not seem to have attracted attention. Dorr, [Dorr'88] used Lagrange multipliers to enforce continuity compatibility constraints in domain decomposition methods. Let's also mention that there's a current trend to use penalty terms to weakly enforce interelement continuity of solutions and of their normal flux in the

so-called discontinuous Galerkin methods, [Cockburn, Karniadakis and Shu'00]. We have been able to construct numerical solutions of arbitrary smoothness across interior triangular facets which do not share a face with the boundary. To summarize, our approach is like the finite element method using piecewise polynomials over tetrahedral partitions. The main features are: no local basis is constructed, smoothness can be imposed in a flexible way across the domain at places where the solution is expected to be smoo th, and we can use polynomials of arbitrary degrees. There are still questions about the convergence of the method for higher order smoothness we have not settled. This has to do with the fact that not much is known about the approximation properties of trivariate spline spaces. We first review the question of how to refine a tetrahedral partition while retaining basic properties that guaranty the quality of the numerical solution. We identify a refinement strategy and indicate some of its properties. Then we introduce the spline spaces which will be used in this dissertation. The B-form representation of splines [cf. de Boor'87] is discussed in that chapter and we review what is known about the approximation properties of trivariate spline spaces. We continue by exposing our method of solving numerically partial differential equations. The basic idea to use splines was first suggested by Ming-Jun Lai then we framed the problem in a minimization setting. We have in an extensive way benefited from his implementation techniques for 2D problems. The programs have been implemented in MATLAB. To compute least squares solutions of equations involving square singular matrices, we added a row of zeros making the matrices non square. When the matrices become very large, we have found useful to use an algorithm that reduces the problem to solve systems of equations of smaller size. We consider how the method is applied to the Poisson equation with both Dirichlet and Neumann boundary conditions then we study the biharmonic equation. The motivation to study the biharmonic equation is that it would be useful to approximate the solution of the Navier-Stokes equations in stream

function formulation. In any case it gives an example of how our techniques can be applied to a high order partial differential equation. The next chapter is devoted to the Navier-Stokes equations. We provide a heuristic derivation of the equations. Then using energy arguments, we derive the discrete equations satisfied by the solution. We then consider the numerical approximations of the Stokes equations and the Navier-Stokes equations. We give a uniqueness result for the discrete equations and prove convergence of our numerical schemes. Finally we conclude with remarks about the strengths and weaknesses of our techniques and some additional topics that will constitute a sequel to this work.

Chapter 2

Tetrahedral Partition of Polygonal Domains of $\mathbf{R}^3$

To increase the accuracy of numerical computations, one typically subdivides the domains of numerical computation into smaller elements taking care that the elements do not degenerate, (a measure of degeneracy is introduced below). This is because the error bounds for numerical solutions involve a constant which depends on the degeneracy of elements and the size of these elements. So the idea is to assure that the degeneracy of the elements does not exceed a threshold, i.e. that one has a quasi-uniform partition.

In $\mathbf{R}^2$, by connecting the middle points of the three edges of a triangle, one obtains four sub-triangles which are all similar to the original one. Proceeding this way, one can get a uniform triangulation. The situation for tetrahedral elements is completely different. It is not difficult to convince oneself that it is impossible to subdivide a regular tetrahedron into eight identical sub-tetrahedra. We start by showing that without careful subdivisions, one can at each level of refinement introduce tetrahedra which are more degenerate than the ones at the previous level. Then we show that it is possible to subdivide a model tetrahedron (which is non-degenerate) in such a way that only one type of tetrahedron arises upon successive refinements. Then we show how to extend this procedure to an arbitrary tetrahedral partition to get a quasi-uniform refinement.

We start with a few notations and definitions.

2.1 Notations and definitions

A tetrahedron is said to be non degenerate if

$$\sigma_T = \frac{h_T}{\rho_T} < \infty$$

where h_T is the diameter of T, i.e its longest edge and ρ_T is the diameter of the largest sphere inscribed in T. A tetrahedral partition is said to be quasi-uniform or shape regular if there is a constant σ such that

$$\sigma_T = \frac{h_T}{\rho_T} \leq \sigma < \infty$$

for each tetrahedron T in the partition, [Braess'92].

We will deal in this dissertation with a polygonal domain Ω of $\mathbf{R}^3$ with piecewise planar boundary. This hypothesis is made to make sure that the domain can be subdivided into tetrahedra. Some results hold for more general domains. At times we will indicate other assumptions on the boundary. A tetrahedral partition $\mathcal{T} = \{t_1, \ldots, t_M\}$ of Ω is said to be admissible provided that

1. $t \subset \overline{\Omega}\ \forall t \in \mathcal{T},\ \bigcup_{t\in\mathcal{T}} t = \overline{\Omega}$
2. $t_i^\circ \cap t_j^\circ = \emptyset,\ \forall\ t_i, t_j \in \mathcal{T}$ and $t_i \neq t_j$, where t_i° denotes the interior of t_i.
3. $\forall\ t_i, t_j \in \mathcal{T}$, exactly one of the following holds:
   - $t_i \cap t_j = \emptyset$
   - t_i and t_j have a common vertex, a common edge, a common face or their intersection is empty.

2.2 A global refinement scheme

It is desirable when refining a tetrahedral partition to keep the partition admissible as well as getting a quasi-uniform partition as explained above. We distinguish two basic

strategies to refine a tetrahedron. A refinement into two tetrahedra and a refinement into eight tetrahedra. We'll discuss extensively the latter since it leads to a 'more stable' refinement in the sense that the sub-tetrahedra are never more degenerate than the original tetrahedron. Throughout this dissertation, by tetrahedral partition, we shall mean an admissible one.

Given a tetrahedron T with four vertices a_1, a_2, a_3 and a_4, we let $a_{12}, a_{13}, a_{14}, a_{23}, a_{24}$ and a_{34} be the midpoints of the edges, where a_{ij} is the midpoint of the edge joining the vertex a_i to the vertex a_j. Using these points we form four tetrahedra in the corners of T and an octahedron. The four corner tetrahedra are $\langle a_1, a_{12}, a_{13}, a_{14}\rangle$, $\langle a_2, a_{12}, a_{23}, a_{24}\rangle$, $\langle a_3, a_{23}, a_{13}, a_{34}\rangle$ and $\langle a_4, a_{14}, a_{24}, a_{34}\rangle$. There are three diagonals of the octahedron: $\langle a_{12}, a_{34}\rangle$, $\langle a_{14}, a_{23}\rangle$ and $\langle a_{24}, a_{13}\rangle$. The choice of a diagonal of the octahedron determine the other sub-tetrahedra.

If the diagonal $\langle a_{12}, a_{34}\rangle$ is chosen we get the following tetrahedra: $\langle a_{14}, a_{24}, a_{12}, a_{34}\rangle$, $\langle a_{12}, a_{34}, a_{13}, a_{14}\rangle$, $\langle a_{12}, a_{34}, a_{24}, a_{23}\rangle$ and $\langle a_{12}, a_{34}, a_{13}, a_{23}\rangle$. The octahedron can be seen as formed with two pyramids which are then each divided into two tetrahedra.

We have then three possible refinements $\mathcal{T}_1$, $\mathcal{T}_2$, and $\mathcal{T}_3$ depending on the choice of the diagonal. Let $\sigma_i = \max\{\sigma_{t_k}, t_k \in \mathcal{T}_i\}$, $i = 1, 2, 3$. There are various rules to pick the diagonal. We suggest to simply pick the one that yields the smallest σ. It appears that for an arbitrary tetrahedron, one of the $\sigma_i's$ is always bigger than the others so it is possible to always pick a diagonal generating a sequence of degenerate tetrahedral partitions. The proof of this result can be found in [Zhang'88]. We show that for the model tetrahedron introduced in [Ong'94], the smallest σ corresponds to the one that measures the degeneracy of the original tetrahedron. This assures quasi-uniformity of the refinement. The extension of this algorithm to a tetrahedral partition is straightforward. For each tetrahedron in the partition, we simply apply the above refinement strategy. The only inconvenience for this refinement strategy is that it is time consuming to compare all the choices before picking one.

To proceed we'll need a more computer tractable formula for the shape measure σ. We have for any tetrahedron T,

$$\rho_T = \frac{6V_T}{S_T}, \tag{2.1}$$

where V_T is the volume of T and S_T its surface area.

To see this, notice that the four tetrahedra formed by connecting the vertices of T to its center have the same height, the radius r of the largest sphere inscribed in T. The volume V_T is given by the sum of the volume of the four sub-tetrahedra. So

$$V_T = \frac{1}{3}S_1 r + \frac{1}{3}S_2 r + \frac{1}{3}S_3 r + \frac{1}{3}S_4 r,$$

where the S_i's are the area of the faces. It follows that $V_T = \frac{1}{3}S_T r$ where S_T is the total surface area. So $r = \frac{3V_T}{S_T}$ and finally we get (2.1).

Using (2.1) as a shape measure we have made some experiments which we now report. The type of a tetrahedron is defined by the lengths of its edges.

We refine a model tetrahedron T_0 with vertices $a_1 = (0,0,0)$, $a_2 = (1,0,0)$, $a_3 = (0,1,0)$ and $a_4 = (0,0,1)$ such that the diagonal that produces the most degenerate sub-tetrahedron is chosen. We list in the following table the number of tetrahedra, the maximum shape measure σ for the refinement and the number of types of tetrahedra.

Tetrahedra	Sigma	Types
1	4.1815	1
8	6.8102	2
64	10.1948	4
512	16.6254	10
4096	26.8399	24

We now choose the diagonal such that the shape measure σ is the minimum among all three choices. This leads to a uniform refinement in the sense that we have tetrahedra of the same shape at all levels of refinement. This rule seems to satisfy Ong's rule of uniformly refining a cube [Ong'94]. We therefore conjecture that when applied to an arbitrary tetrahedron, we'll get a quasi-uniform refinement and that at most six types of tetrahedra will arise upon iterative refinements.

Tetrahedra	Sigma	Types
1	4.1815	1
8	4.1815	1
64	4.1815	1
512	4.1815	1
4096	4.1815	1

Notice that for the model tetrahedron, only one type of sub-tetrahedron arises after refinements. This is not the case for an arbitrary tetrahedron. For example if we uniformly refine the tetrahedron with vertices $a_1 = (1, 0, 0)$, $a_2 = (2, 2, 0)$, $a_3 = (0, 1, 0)$ and $a_4 = (0, 0, 1)$, we get the following table:

Tetrahedra	Sigma	Types
1	5.3660	1
8	5.4495	3
64	5.4495	4
512	5.4495	4
4096	5.4495	4

2.3 Local refinement

In the previous section, we described a strategy to refine a single tetrahedron into eight sub-tetrahedra and indicated that this strategy could be applied to all tetrahedra in a partition to get an admissible one. Let's point out that in some situations it is desirable instead of refining all tetrahedra in the partition, to refine only a few. For example one may want to avoid solving a very large system of equations as a result of having too many tetrahedra and at the same time want to reduce the approximation errors in some regions where it is unacceptable. In that case it is more delicate to get an admissible partition. Our refinement scheme for a single tetrahedron could be combined with ideas of [Liu and Joe'96] to get a local refinement strategy. We refer to their paper for additional details.

2.4 Conclusion

We have presented a refinement strategy which is basically a concatenation of previous results. The numerical solution of three dimensional partial differential equations is expensive in terms of CPU time and computer memory so in general we have not been able to show several levels of refinement. If one simply subdivides a tetrahedron into two sub-tetrahedra, the accuracy of the numerical solution does not decrease fast through refinements. On the other hand it does not lead to a uniform refinement which will definitely help show numerical evidence of convergence of our algorithms. The bisection method for tetrahedra has been theoretically studied in [Liu and Joe'94]. Basically, given a tetrahedron $T = \langle a_1, a_2, a_3, a_4 \rangle$, using the midpoint t of its longest edge say $\langle a_1, a_2 \rangle$, it is bisected into two sub-tetrahedra $\langle a_1, t, a_3, a_4 \rangle$ and $\langle t, a_2, a_3, a_4 \rangle$. We show here for a cube of volume one subdivided into six tetrahedra the shape measure and the number of type of tetrahedra which appear at different level of refinements. We have numerically checked that for this

situation, simply bisecting each tetrahedron in the cube with its longest edge leads to an admissible tetrahedral partition.

Tetrahedra	Sigma	Types
6	4.1815	1
12	4.4142	1
24	3.8284	1
48	4.1815	1
96	4.4142	1
192	3.8284	1
384	4.1815	1
768	4.4142	1
1536	3.8284	1

These jumps in the shape measure σ, as we shall see, will have an effect on the quality of the numerical solutions. For this reason, the bisection method has not been used extensively in this study.

Chapter 3

Trivariate Spline Spaces

In this chapter, we introduce trivariate spline spaces. We generalize to the 3D setting some results in [Lai and Schumaker'00] which will be needed. [de Boor'87] presents a different approach in a multivariate setting.

Given a bounded domain Ω of $\mathbf{R}^3$ with piecewise planar boundary, let $d \geq 0$ be a fixed integer and let $\mathcal{T}$ be a tetrahedral partition of Ω. We are going to use the spline spaces

$$S_d^r(\Omega) = \{p \in C^r(\Omega),\ p_{|t} \in P_d,\ \forall t \in \mathcal{T}\}$$

to approximate solutions of PDE's defined on Ω. P_d denotes the space of trivariate polynomials of degree d. The B-form representation of splines on tetrahedra will be used, [de Boor'87]. This enables us to efficiently handle the condition that $p \in C^r(\Omega)$.

We are interested in trivariate polynomials of degree d. Those are functions defined on $\mathbf{R}^3$ of the form

$$p(x, y, z) = \sum_{0 \leq i+j+k \leq d} \alpha_{ijk} x^i y^j z^k, \tag{3.1}$$

where the α_{ijk} are real numbers.

3.1 The B-form of polynomials on tetrahedra

We will derive a representation of a polynomial p on a tetrahedron equivalent to (3.1) using barycentric coordinates.

3.1.1 Barycentric coordinates

Let $T = \langle v_1, v_2, v_3, v_4 \rangle$ be a non-degenerate tetrahedron with v_i having coordinates (x_i, y_i, z_i), $i = 1, 2, 3, 4$. We have

Lemma 3.1.1 *Every point $v = (x, y, z)$ can be written uniquely in the form*

$$v = b_1 v_1 + b_2 v_2 + b_3 v_3 + b_4 v_4, \tag{3.2}$$

with

$$b_1 + b_2 + b_3 + b_4 = 1. \tag{3.3}$$

The $b_i's, i = 1, 2, 3, 4$ are called the barycentric coordinates of the point $v = (x, y, z)$ relative to the tetrahedron T. Moreover each b_i is a linear polynomial in x, y, z.

Proof: (3.3) can be written $\begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} = 1$ and (3.2) can be written

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = b_1 \begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} + b_2 \begin{pmatrix} x_2 \\ y_2 \\ z_2 \end{pmatrix} + b_3 \begin{pmatrix} x_3 \\ y_3 \\ z_3 \end{pmatrix} + b_4 \begin{pmatrix} x_4 \\ y_4 \\ z_4 \end{pmatrix}.$$

This written in matrix form combined with (3.3) in matrix form gives

$$V \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} = \begin{pmatrix} 1 \\ x \\ y \\ z \end{pmatrix}$$

where $V = \begin{pmatrix} 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \\ z_1 & z_2 & z_3 & z_4 \end{pmatrix}$.

We have det $V = 6$ volume(T), which shows that the equation is nondegenerate proving existence and uniqueness. Using Kramer's rule, we have

$$b_1 = \frac{1}{\det V}\det \begin{pmatrix} 1 & 1 & 1 & 1 \\ x & x_2 & x_3 & x_4 \\ y & y_2 & y_3 & y_4 \\ z & z_2 & z_3 & z_4 \end{pmatrix}$$

and similar formulas for the other $b_i's, i = 2, 3, 4$. This shows that they are linear polynomials in x, y, z.

Remark: In the interior of the tetrahedron $\langle v_1, v_2, v_3, v_4\rangle$, we have $b_i > 0,\ i = 1, 2, 3$. This can be seen by noticing that b_i is a linear polynomial with value 1 at the vertex v_i which vanishes on the face opposite to v_i.

3.1.2 The Bernstein form of polynomials

For $v = (x, y, z) \in \mathbf{R}^3$, let (b_1, b_2, b_3, b_4) be the barycentric coordinates of v with respect to $T = \langle v_1, v_2, v_3, v_4\rangle$. We introduce the Bernstein polynomials of degree d as follows

$$B^d_{ijkl}(v) = \frac{d!}{i!j!k!l!} b_1^i b_2^j b_3^k b_4^l, \qquad i + j + k + l = d.$$

They are polynomials of degree d since each b_i is a linear polynomial. We have

Theorem 3.1.2 *The set $\mathcal{B}^d = \{B^d_{ijkl}(x, y, z), i + j + k + l = d\}$ is a basis for the space of polynomials P_d.*

Proof: We recall that the dimension of P_d is $\binom{d+3}{3}$ and notice that the number of Bernstein basis polynomials is equal to the dimension of P_d. It is therefore enough to show that any polynomial $x^\nu y^\mu z^\kappa, 0 \le \nu+\mu+\kappa \le d$ can be written as a combination

of elements of $\mathcal{B}_d$.

$$\begin{aligned} 1 &= (b_1 + b_2 + b_3 + b_4)^d \\ &= \sum_{i+j+k+l=d} \frac{d!}{i!j!k!l!} b_1^i b_2^j b_3^k b_4^l \\ &= \sum_{i+j+k+l=d} B_{ijkl}^d(v), \quad \forall v \in \mathbf{R}^3, \end{aligned}$$

so 1 belongs to the linear span of $\mathcal{B}^d$. And this is valid for any $d \geq 0$. Now

$$\begin{aligned} x &= b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 \\ &= (b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4) \sum_{i+j+k+l=d-1} B_{ijkl}^{d-1}(v), \end{aligned}$$

and we have

$$\begin{aligned} b_1 B_{ijkl}^{d-1} &= b_1 \frac{(d-1)!}{i!j!k!l!} b_1^i b_2^j b_3^k b_4^l \\ &= \frac{1}{d} \frac{d!}{i!j!k!l!} b_1^{i+1} b_2^j b_3^k b_4^l \\ &= \frac{i+1}{d} \frac{d!}{(i+1)!j!k!l!} b_1^{i+1} b_2^j b_3^k b_4^l \\ &= \frac{i+1}{d} B_{i+1,j,k,l}^d, \end{aligned}$$

etc, so

$$\begin{aligned} x = &\sum_{i+j+k+l=d-1} \frac{1}{d} \left[(i+1) x_1 B_{i+1,j,k,l}^d + (j+1) x_2 B_{i,j+1,k,l}^d \right] \\ &\frac{1}{d} \left[+(k+1) x_3 B_{i,j,k+1,l}^d + (l+1) x_4 B_{i,j,k,l+1}^d \right] \\ = &\sum_{i+j+k+l=d} \frac{1}{d} (i x_1 + j x_2 + k x_3 + l x_4) B_{i,j,k,l}^d. \end{aligned}$$

and analogous relations for y and z. This shows that x, y and z belong to the linear span of $\mathcal{B}^d$.

Now for $1 \leq \mu + \nu + \kappa \leq d$, we have

$$\begin{aligned} x^\mu y^\nu z^\kappa = &(b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4)^\mu (b_1 y_1 + b_2 y_2 + b_3 y_3 + b_4 y_4)^\nu \\ &\times (b_1 z_1 + b_2 z_2 + b_3 z_3 + b_4 z_4)^\kappa (b_1 + b_2 + b_3 + b_4)^{d-\mu-\nu-\kappa} \end{aligned}$$

and

$$(b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4)^\mu = \sum_{i+j+k+l=\mu} \frac{d!}{i!j!k!l!}(b_1x_1)^i(b_2x_2)^j \times (b_3x_3)^k(b_4x_4)^l = \sum_{i+j+k+l=\mu} c(i,j,k,l,x_1,x_2,x_3,x_4)B^\mu_{ijkl},$$

with $c(i,j,k,l,x_1,x_2,x_3,x_4) = x_1^i x_2^j x_3^k x_4^l$. Finally

$$\begin{aligned} x^\mu y^\nu z^\kappa = & \sum_{i+j+k+l=\mu} c(i,j,k,l,x_1,x_2,x_3,x_4)B^\mu_{ijkl} \\ & \times \sum_{i+j+k+l=\nu} c(i,j,k,l,y_1,y_2,y_3,y_4)B^\mu_{ijkl} \\ & \times \sum_{i+j+k+l=\kappa} c(i,j,k,l,z_1,z_2,z_3,z_4)B^\mu_{ijkl} \\ & \times \sum_{i+j+k+l=d-\mu-\nu-\kappa} B^{d-\mu-\nu-\kappa}_{ijkl}. \end{aligned}$$

Collecting terms, we get

$$x^\mu y^\nu z^\kappa = \sum_{i+j+k+l=d} c(\mu,\nu,\kappa)_{ijkl} B^d_{ijkl}(x,y,z)$$

for some constants $c(\mu,\nu,\kappa)_{ijkl}$ which depends on (x_1,x_2,x_3), (y_1,y_2,y_3) and (z_1,z_2,z_3). This completes the proof.

As a consequence any polynomial p of degree d on T can be written uniquely

$$p = \sum_{i+j+k+l=d} c_{ijkl} B^d_{ijkl}. \tag{3.4}$$

The representation (3.4) is referred to as the B-form of p. The c_{ijkl} are the B-coefficients of p. The B-net of p is a vector of all B-coefficients for all the tetrahedra in the tetrahedral partition. For later use we define the set of domain points in T to be

$$\mathcal{D}_{d,T} = \{\xi_{ijkl} = \frac{iv_1 + jv_2 + kv_3 + lv_4}{d}, i+j+k+l = d\}. \tag{3.5}$$

It is common practice to associate the coefficients c_{ijkl} with the domain points ξ_{ijkl}.

3.1.3 Interpolation by trivariate polynomials on Tetrahedra

The following will be an essential tool when we study smoothness conditions, i.e. conditions on the B-coefficients that assure that a piecewise polynomial on a tetrahedral partition is smooth across tetrahedra.

Theorem 3.1.3 *There is a unique polynomial p of degree d that interpolates any given function f on a tetrahedron $T = \langle v_1, v_2, v_3, v_4 \rangle$ at the points in (3.5).*

Proof: Let m=dimP_d and $(\gamma_\alpha)_{\alpha=1}^m$ be a set of basis functions of P_d. We can choose for $(\gamma_\alpha)_{\alpha=1}^m$ the Bernstein polynomials we have just introduced. The problem is to find

$$p = \sum_{\alpha=1}^{m} c_\alpha \gamma_\alpha$$

such that

$$p(a_\beta) = f(a_\beta), \quad \beta = 1, \ldots, m,$$

where the a_β's denote the domain points in (3.5). This amounts to finding $(c_\alpha)_{\alpha=1}^m$ such that

$$\sum_{\alpha=1}^{m} c_\alpha \gamma_\alpha(a_\beta) = f(a_\beta), \quad \beta = 1, \ldots, m.$$

Let $G = (\gamma_\alpha(a_\beta))_{\alpha,\beta=1}^m$, $c = (c_1, \ldots, c_m)^T$ and $b = (f(a_1), \ldots, f(a_m))^T$. The problem is then to solve $Gc = b$. Since G is square it is sufficient to show that G is nonsingular. Let c such that $Gc = 0$. That is, there is a polynomial $p = \sum_{\alpha=1}^m c_\alpha \gamma_\alpha$ such that

$$p(a_\beta) = 0, \quad \beta = 1, \ldots, m.$$

We show that p is the zero polynomial which will imply that $c = 0$. We use the following lemma whose proof can be found in [Lai and Schumaker'00].

Lemma 3.1.4 *A bivariate polynomial p of degree d is uniquely determined on a triangle $\langle v_1, v_2, v_3\rangle$ by its values at the domain points $\xi_{ijk} = \frac{iv_1+jv_2+kv_3}{d}$. Note that here the domain points have three indices.*

For each $l = 0, \ldots, d$, let $\alpha_l x + \beta_l y + \gamma_l z = \delta_l$ be the equation of the plane P_l containing the domain points $\{\xi_{ijkl}, i+j+k+l = d\}$. For each l the domain points on that plane can be considered as domain points on a suitable triangle. Let's assume that the theorem is true for $d-1$, we show that it is true for d. Since p reduces to a bivariate polynomial of degree d on P_0, by the previous lemma p vanishes identically on P_0 and we write

$$p(x, y, z) = (\alpha_0 x + \beta_0 y + \gamma_0 z - \delta_0) q(x, y, z),$$

with q of degree $d-1$. This implies that q of degree $d-1$ is identically zero on the domain points $\{\xi^d_{ijkl}, l \neq 0\} = \mathcal{D}_{d-1,T}$. By induction hypothesis q is the zero polynomial. This completes the proof.

Numerical example: For computational purpose, we need to order the $c_{ijkl}, i+j+k+l = d$. The order we choose for the domain points on the tetrahedron $\langle v_1, v_2, v_3, v_4\rangle$ can be described as follows: First we list the coefficients on the triangle $\langle v_1, v_2, v_3\rangle$ the ones on the edge $\langle v_1, v_2\rangle$ first starting with the coefficient associated with the vertex v_1, then we continue with the other coefficients towards the coefficient associated with v_4. Specifically for d=2, the order is

$$c_{2000}, c_{1100}, c_{0200}, c_{1010}, c_{0110}, c_{0020}, c_{1001}, c_{0101}, c_{0011}, c_{0002}.$$

Explicitly, for $t = \langle v_1, v_2, v_3, v_4\rangle$ with $v_1 = (0,0,0)$, $v_2 = (1,0,0)$ $v_3 = (0,1,0)$ and $v_4 = (0,0,1)$ and $p(x,y,z) = x^2 + 3xz + y^2$ of degree 2, the B-net is $(0,0,1,0,0,1,0,1.5,0,1)$. This result was obtained by running a Matlab code but it can be computed by hand.

3.1.4 De Casteljau algorithm

There's a recurrence relation about Bernstein polynomials that has algorithmic consequences (De Casteljau algorithm). Notice that

$$\begin{aligned} B^d_{ijkl} &= \frac{d!}{i!j!k!l!} b_1^i b_2^j b_3^k b_4^l \\ &= d\frac{(d-1)!}{i!j!k!l!} b_1^i b_2^j b_3^k b_4^l \\ &= (i+j+k+l)\frac{(d-1)!}{i!j!k!l!} b_1^i b_2^j b_3^k b_4^l, \end{aligned}$$

since $i+j+k+l = d$, so that

$$B^d_{ijkl} = b_1 B^{d-1}_{i-1,j,k,l} + b_2 B^{d-1}_{i,j-1,k,l} + b_3 B^{d-1}_{i,j,k-1,l} + b_4 B^{d-1}_{i,j,k,l-1}.$$

Now, let

$$p = \sum_{i+j+k+l=d} c_{ijkl} B^d_{ijkl}$$

and let's write $c_{ijkl} = c^{(0)}_{ijkl}$. Using the recurrence relation for B^d_{ijkl}, we get,

$$\begin{aligned} p = \sum_{i+j+k+l=d} & b_1 c^{(0)}_{ijkl} B^{d-1}_{i-1,j,k,l} + b_2 c^{(0)}_{ijkl} B^{d-1}_{i,j-1,k,l} + b_3 c^{(0)}_{ijkl} B^{d-1}_{i,j,k-1,l} \\ & + b_4 c^{(0)}_{ijkl} B_{i,j,k,l-1}. \end{aligned} \tag{3.6}$$

Each term in the sum (3.6) can be treated the same way. For example,

$$\sum_{i+j+k+l=d} b_1 c^{(0)}_{ijkl} B^{d-1}_{i-1,j,k,l} = \sum_{i+j+k+l=d-1} b_1 c^{(0)}_{i+1,j,k,l} B^{d-1}_{i,j,k,l}.$$

We can therefore write, noticing that $B^{d-1}_{i,j,k,l}$ is a common factor in all 4 terms,

$$p = \sum_{i+j+k+l=d-1} (b_1 c_{i+1,j,k,l} + b_2 c_{i,j+1,k,l} + b_3 c_{i,j,k+1,l} + b_4 c_{i,j,kl+1}) B^{d-1}_{i,j,k,l}. \tag{3.7}$$

We therefore define, for a positive integer r

$$c^{(r)}_{ijkl}(b) = b_1 c^{(r-1)}_{i+1,j,k,l} + b_2 c^{(r-1)}_{i,j+1,k,l} + b_3 c^{(r-1)}_{i,j,k+1,l} + b_4 c^{(r-1)}_{i,j,k,l+1},$$

so that

$$p = \sum_{i+j+k+l=d-1} c^{(1)}_{ijkl}(b) B^{d-1}_{ijkl},$$

with

$$c^{(0)}_{ijkl} = c_{ijkl}.$$

And one can easily verify, using induction that

$$p = \sum_{i+j+k+l=d-r} c^{(r)}_{ijkl}(\mathbf{b}) B^{d-r}_{ijkl},$$

where $\mathbf{b} = (b_1, b_2, b_3, b_4)$ denotes the barycentric coordinates of the evaluation point. As a consequence

$$p = c^{(d)}_{0000}.$$

This is the so-called de Casteljau algorithm.

It is possible to write $c^{(r)}_{ijkl}(b), i+j+k+l = d-r$ in terms of the $c_{ijkl}, i+j+k+l = d$. We have

Theorem 3.1.5

$$c^{(m)}_{ijkl}(\mathbf{b}) = \sum_{\alpha+\beta+\gamma+\delta=m} c_{i+\alpha,j+\beta,k+\gamma,l+\delta} B^m_{\alpha,\beta,\gamma,\delta}(v), \quad i+j+k+l = d-m. \tag{3.8}$$

Proof: By induction on m. For $m = 1$, we have by definition

$$c^{(1)}_{ijkl}(v) = b_1 c_{i+1,j,k,l} + b_2 c_{i,j+1,k,l} + b_3 c_{i,j,k+1,l} + b_4 c_{i,j,k,l+1},$$

since $B^1_{1000}(b) = b_1$ and similar formulas for B^1_{0100}, B^1_{0010} and B^1_{0001}.

We assume the result true for $s-1$. We have

$$c^{(s)}_{ijkl}(v) = b_1 c^{s-1}_{i+1,j,k,l} + b_2 c^{s-1}_{i,j+1,k,l} + b_3 c^{s-1}_{i,j,k+1,l} + b_4 c^{s-1}_{i,j,k,l+1};$$

but by the induction hypothesis

$$c^{(s-1)}_{i+1,j,k,l}(\mathbf{b}) = \sum_{\alpha+\beta+\gamma+\delta=s-1} c_{i+1+\alpha,j+\beta,k+\gamma,l+\delta} B^{s-1}_{\alpha,\beta,\gamma,\delta}(v), \quad i+j+k+l = d-s+1,$$

and similar formulas for the other terms. We also have

$$\begin{aligned} b_1 B^{m-1}_{\alpha\beta\gamma\delta} &= \frac{(m-1)!}{\alpha!\beta!\gamma!\delta!} b_1^{\alpha+1} b_2^{\beta} b_3^{\gamma} b_4^{\delta} \\ &= \frac{(\alpha+1)}{m} B^m_{\alpha+1,\beta,\gamma,\delta}, \end{aligned}$$

and similar relations for $b_i B^{m-1}_{\alpha\beta\gamma\delta}, i = 2, 3, 4$. So

$$\begin{aligned} c^m_{i+1,j,k,l} = & \sum_{\alpha+\beta+\gamma+\delta=m-1} \frac{(\alpha+1)}{m} c_{i+1+\alpha,j+\beta,k+\gamma,l+\delta} B^m_{\alpha+1,\beta,\gamma,\delta} \\ & + \frac{(\beta+1)}{m} c_{i+\alpha,j+1+\beta,k+\gamma,l+\delta} B^m_{\alpha,\beta+1,\gamma,\delta} \\ & + \frac{(\gamma+1)}{m} c_{i+\alpha,j+\beta,k+1+\gamma,l+\delta} B^m_{\alpha,\beta,\gamma+1,\delta} \\ & + \frac{(\delta+1)}{m} c_{i+\alpha,j+\beta,k+\gamma,l+1+\delta} B^m_{\alpha,\beta,\gamma,\delta+1} \\ = & \sum_{\alpha+\beta+\gamma+\delta=m} \frac{[\alpha c_{i+\alpha,j+\beta,k+\gamma,l+\delta} + \beta c_{i+\alpha,j+\beta,k+\gamma,l+\delta}]}{m} B^m_{\alpha,\beta,\gamma,\delta} \\ & + \frac{[\gamma c_{i+\alpha,j+\beta,k+\gamma,l+\delta} + \delta c_{i+\alpha,j+\beta,k+\gamma,l+\delta}]}{m} B^m_{\alpha,\beta,\gamma,\delta} \\ = & \sum_{\alpha+\beta+\gamma+\delta=m} c_{i+\alpha,j+\beta,k+\gamma,l+\delta} B^m_{\alpha,\beta,\gamma,\delta}, \end{aligned}$$

since $\alpha + \beta + \gamma + \delta = m$. This completes the proof.

3.1.5 Directional derivatives of polynomials in B-form

We want to give formulas for the directional derivatives of p in a direction defined by a vector $\mathbf{u}$. For

$$\begin{aligned} p &= \sum_{i+j+k+l=d} c_{ijkl}\ B^d_{ijkl}, \\ D_{\mathbf{u}} p &= \sum_{i+j+k+l=d} c_{ijkl}\ D_{\mathbf{u}} B^d_{ijkl}, \end{aligned}$$

noticing that B^d_{ijkl} is a product of 4 terms, we get

$$\begin{aligned} D_{\mathbf{u}}B^d_{ijkl} &= \frac{d!}{i!j!k!l!}(iD_{\mathbf{u}}b_1\ b_1^{i-1}b_2^j b_3^k b_4^l + jD_{\mathbf{u}}b_2\ b_1^i b_2^{j-1} b_3^k b_4^l + kD_{\mathbf{u}}b_3\ b_1^i b_2^j b_3^{k-1} b_4^l \\ &\quad + lD_{\mathbf{u}}b_4\ b_1^i b_2^j b_3^k b_4^{l-1}) \\ &= d(D_{\mathbf{u}}b_1\ B^{d-1}_{i-1,j,k,l} + D_{\mathbf{u}}b_2\ B_{i,j-1,k,l} + D_{\mathbf{u}}b_3\ B_{i,j,k-1,l} + D_{\mathbf{u}}b_4\ B_{i,j,k,l-1}). \end{aligned}$$

We write $D_{\mathbf{u}}b_t = a_t, \quad t = 1, \ldots, 4$, with $\mathbf{a} = (a_1, a_2, a_3, a_4)$ so

$$D_{\mathbf{u}}B^d_{ijkl} = d(a_1\ B^{d-1}_{i-1,j,k,l} + a_2\ B^{d-1}_{i,j-1,k,l} + a_3\ B^{d-1}_{i,j,k-1,l} + a_4\ B^{d-1}_{i,j,k,l-1}),$$

and collecting terms we can write

$$D_{\mathbf{u}}p = d \sum_{i+j+k+l=d-1} c^{(1)}_{ijkl}(\mathbf{a})\ B^{d-1}_{ijkl}.$$

The $a_t, \quad t = 1, \ldots, 4$ turn out to be what are called T-coordinates of $\mathbf{u}$. If $\mathbf{u} = y - x$ with $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ and $(\beta_1, \beta_2, \beta_3, \beta_4)$ being the barycentric coordinates of x and y respectively, the T-coordinates of $\mathbf{u}$ are defined to be $\beta_k - \alpha_k$ for all k. Using the definition of directional derivative, we have

$$\begin{aligned} D_{\mathbf{u}}b_t(v) &= \lim_{s\to 0} \frac{b_t(v + s\mathbf{u}) - b_t(v)}{s} \\ &= \frac{d}{ds} b_t(v + s\mathbf{u})_{|s=0}. \end{aligned}$$

Recall that (b_1, b_2, b_3, b_4) denote the barycentric coordinates of v. Those of $v + s\mathbf{u}$ are $b_t + s(\beta_t - \alpha_t)$ so that we get $D_{\mathbf{u}}b_t = \beta_t - \alpha_t = a_t$ for all t.

For a polynomial of degree d, and for $1 \leq m \leq d$, if $\mathbf{u_i}$, $i = 1, \ldots, m$ are m directions defined by T-coordinates $\mathbf{a}^{(i)}$, we immediately get

$$D_{\mathbf{u_m}} \cdots D_{\mathbf{u_1}}p(v) = \frac{d!}{(d-m)!} \sum_{i+j+k+l=d-m} c^{(m)}_{ijkl}(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m)}) B^{d-m}_{ijkl}(v),$$

with the obvious notation

$$c^{(m)}_{ijkl}(\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(m)}) = c^{(m)}_{ijkl}(\mathbf{a}^{(1)}) \ldots (\mathbf{a}^{(m)}).$$

For example, if we put $d_{ijkl} = c^{(1)}_{ijkl}(\mathbf{a_1})$, then $c^{(2)}_{ijkl}(\mathbf{a_1}, \mathbf{a_2}) = d^{(1)}_{ijkl}(\mathbf{a_2})$. In particular

$$D^m_{\mathbf{u}} p(v) = \frac{d!}{(d-m)!} \sum_{i+j+k+l=d-m} c^{(m)}_{ijkl}(\mathbf{a}) B^{d-m}_{ijkl}(v).$$

Example: For $\mathbf{u} = v_2 - v_1$, $\mathbf{a} = (-1, 1, 0, 0)$ we have

$$c^{(1)}_{ijkl}(\mathbf{a}) = -c_{i+1,j,k,l} + c_{i,j+1,k,l}.$$

3.1.6 Integrals and inner product of polynomials in B-form

There are precise formulas for the inner product and integrals of polynomials in B-form. The inner product we use here is the integral of the product of two polynomials. We begin with

Theorem 3.1.6 *Let p have B-net c_{ijkl}, $i+j+k+l = d$ on a tetrahedron t. We have*

$$\int_t p(x, y, z)\ \mathrm{d}x\mathrm{d}y\mathrm{d}z = \frac{\mathrm{volume}(t)}{\binom{d+3}{3}} \sum_{i+j+k+l=d} c_{ijkl}.$$

Proof: Using an integral formula for the multi gamma function, we get

$$\int_t B^d_{ijkl}(x, y, z)\ \mathrm{d}x\ \mathrm{d}y\ \mathrm{d}z = \frac{\mathrm{volume}(t)}{\binom{d+3}{3}}.$$

Therefore

$$\int_t p(x, y, z)\ \mathrm{d}x\mathrm{d}y\mathrm{d}z = \sum_{i+j+k+l=d} c_{ijkl} \int_t B^d_{ijkl} = \frac{\mathrm{volume}(t)}{\binom{d+3}{3}} \sum_{i+j+k+l=d} c_{ijkl}.$$

This proves the theorem. Now

$$\begin{aligned} B^d_{ijkl} B^d_{rstu} &= \frac{d!}{i!j!k!l!} \frac{d!}{r!s!t!u!} b_1^i b_2^j b_3^k b_4^l b_1^r b_2^s b_3^t b_4^u \\ &= \frac{d!}{i!j!k!l!} \frac{d!}{r!s!t!u!} b_1^{i+r} b_2^{j+s} b_3^{k+t} b_4^{l+u} \end{aligned}$$

and

$$B^{2d}_{i+r,j+s,k+t,l+u} = \frac{(2d)!}{(i+r)!(j+s)!(k+t)!(l+u)!} b_1^{i+r} b_2^{j+s} b_3^{k+t} b_4^{l+u},$$

so

$$\begin{aligned} B^d_{ijkl}B^d_{rstu} &= \frac{d!}{i!j!k!l!}\frac{d!}{r!s!t!u!}\frac{(i+r)!(j+s)!(k+t)!(l+u)!}{(2d)!} \\ &\qquad \times B^{2d}_{i+r,j+s,k+t,l+u} \\ &= \frac{(i+r)!}{i!r!}\frac{(j+s)!}{j!s!}\frac{(k+t)!}{k!t!}\frac{(l+u)!}{l!u!}\frac{d!d!}{(2d)!}B^{2d}_{i+r,j+s,k+t,l+u} \\ &= \frac{\binom{i+r}{i}\binom{j+s}{j}\binom{k+t}{k}\binom{l+u}{l}}{\binom{2d}{d}}B^{2d}_{i+r,j+s,k+t,l+u}. \end{aligned}$$

It follows that

$$\int_t B^d_{ijkl}(x,y,z)B^d_{rstu}(x,y,z)\ \mathrm{d}x\mathrm{d}y\mathrm{d}z = \frac{\binom{i+r}{i}\binom{j+s}{j}\binom{k+t}{k}\binom{l+u}{l}\mathrm{volume}(t)}{\binom{2d}{d}\binom{2d+3}{3}}.$$

As a consequence for two polynomials

$$p = \sum_{i+j+k+l=d} c_{ijkl}B^d_{ijkl} \text{ and } q = \sum_{r+s+t+u=d} \widetilde{c}_{rstu}B^d_{rstu}, \tag{3.9}$$

their inner product is given by

$$\begin{aligned} \int p(x,y,z)q(x,y,z)\ \mathrm{d}x\mathrm{d}y\mathrm{d}z = \frac{\mathrm{volume}(t)}{\binom{2d}{d}\binom{2d+3}{3}} &\sum_{\substack{i+j+k+l=d \\ r+s+t+u=d}} c_{ijkl}\widetilde{c}_{rstu} \\ &\times \binom{i+r}{i}\binom{j+s}{j}\binom{k+t}{k}\binom{l+u}{l}. \end{aligned} \tag{3.10}$$

This can also be written in the form

$$\int p(x,y,z)q(x,y,z)\mathrm{d}x\mathrm{d}y\mathrm{d}z = \frac{\mathrm{volume}(t)}{\binom{2d}{d}\binom{2d+3}{3}}C^TG\widetilde{C}, \tag{3.11}$$

where C and $\widetilde{C}$ encode respectively the B-net of p and q respectively and G is a (m,m) square matrix with m=dimP_d.

This process can be carried out for the product of three polynomials of degree d_1, d_2

and d_3. We have

$$\begin{aligned}
B^{d_1}_{ijkl}B^{d_2}_{rstu}B^{d_3}_{\mu\nu\kappa\delta} &= \frac{d_1!}{i!j!k!l!}\frac{d_2!}{r!s!t!u!}\frac{d_3!}{\mu!\nu!\kappa!\delta!}b_1^i b_2^j b_3^k b_4^l b_1^r b_2^s b_3^t b_4^u b_1^\mu b_2^\nu b_3^\kappa b_4^\delta \\
&= \frac{d_1!}{i!j!k!l!}\frac{d_2!}{r!s!t!u!}\frac{d_3!}{\mu!\nu!\kappa!\delta!}b_1^{i+r+\mu}b_2^{j+s+\nu}b_3^{k+t+\kappa}b_4^{l+u+\delta} \\
&= \frac{d_1!}{i!j!k!l!}\frac{d_2!}{r!s!t!u!}\frac{d_3!}{\mu!\nu!\kappa!\delta!} \\
&\quad \times \frac{(i+r+\mu)!(j+s+\nu)!(k+t+\kappa)!(l+u+\delta)!}{(d_1+d_2+d_3)!} \\
&\quad \times B^{d_1+d_2+d_3}_{i+r+\mu,j+s+\nu,k+t+\kappa,l+u+\delta} \\
&= \frac{(i+r+\mu)!}{i!r!\mu!}\frac{(j+s+\nu)!}{j!s!\nu!}\frac{(k+t+\kappa)!}{k!t!\kappa!}\frac{(l+u+\delta)!}{l!u!\delta!} \\
&\quad \times \frac{d_1!d_2!d_3!}{(d_1+d_2+d_3)!}B^{d_1+d_2+d_3}_{i+r+\mu,j+s+\nu,k+t+\kappa,l+u+\delta} \\
&= \binom{i+r}{i}\binom{i+r+\mu}{i+r}\binom{j+s}{j}\binom{j+s+\nu}{j+s}\binom{k+t}{k} \\
&\quad \times \binom{k+t+\kappa}{k+t}\binom{l+u}{l}\binom{l+u}{l+u}\frac{d_1!d_2!d_3!}{(d_1+d_2+d_3)!} \\
&\quad \times B^{d_1+d_2+d_3}_{i+r+\mu,j+s+\nu,k+t+\kappa,l+u+\delta}
\end{aligned}$$

so if $(C^1_{ijkl})_{i+j+k+l=d_1}$, $(C^2_{rstu})_{r+s+t+u=d_2}$ and $(C^3_{\mu\nu\kappa\delta})_{\mu+\nu+\kappa+\delta=d_3}$ encode the B-nets of p_1, p_2 and p_3 respectively, we have

$$\begin{aligned}
\int_t p_1(x,y,z)p_2(x,y,z)p_3(x,y,z)\,\mathrm{d}x\mathrm{d}y\mathrm{d}z &= \frac{\mathrm{volume}(t)(d_1+d_2+d_3)!}{d_1!d_2!d_3!\binom{d_1+d_2+d_3+3}{3}} \\
&\quad \times \sum_{\mu+\nu+\kappa+\delta=d_3} C^3_{\mu\nu\kappa\delta}(C^1)^T G_{\mu\nu\kappa\delta}C^2,
\end{aligned} \tag{3.12}$$

where $G_{\mu\nu\kappa\delta}$ is a (m_1, m_2) matrix with m_1=dimP_{d_1} and m_2=dimP_{d_2}.

3.2 Smoothness Conditions

Let $\mathcal{T}$ be a subdivision of a domain Ω into tetrahedra and let p be a spline of degree d defined over Ω, i.e. $p_{|t} \in P_d\ \forall t \in \mathcal{T}$. It is assumed that we have the B-form of each

polynomial piece

$$p_t = \sum_{i+j+k+l=d} c^t_{ijkl} B^d_{ijkl}. \tag{3.13}$$

We say that we have the B-form representation of the spline p. We would like to give conditions on the c^t_{ijkl} that will assure that p has certain global smoothness properties.

Theorem 3.2.1 *Let $t = \langle v_1, v_2, v_3, v_4\rangle$ and $t' = \langle v_1, v_2, v_3, v_5\rangle$ be two tetrahedra with common face $\langle v_1, v_2, v_3\rangle$. Then p is of class C^r on $t \cup t'$ if and only if*

$$c^{t'}_{ijkm} = \sum_{\mu+\nu+\kappa+\delta=m} c^t_{i+\mu,j+\nu,\gamma+\kappa,\delta} B^l_{\mu,\nu,\kappa,\delta}(v_5), \text{ for } m = 0, \ldots, r$$

and i+j+k=d-m.

The proof will follow from several lemmas.

Lemma 3.2.2 *If u_1, u_2 and u_3 are 3 independent directions in $\mathbf{R}^3$, f is of class C^r at v if and only if $D^{\alpha_1}_{u_1} D^{\alpha_2}_{u_2} D^{\alpha}_{u} f(v)$ is continuous at v for all α, $|\alpha| = \alpha_1+\alpha_2+\alpha_3 \le r$, where D_{u_i} is the directional derivative operator in the direction u_i.*

Proof: Let u_i have coordinates (u^1_i, u^2_i, u^3_i), $i = 1, 2, 3$. We have

$$\begin{pmatrix} D_{u_1} \\ D_{u_2} \\ D_{u_3} \end{pmatrix} = \begin{pmatrix} u^1_1 & u^2_1 & u^3_1 \\ u^1_2 & u^2_2 & u^3_2 \\ u^1_3 & u^2_3 & u^3_3 \end{pmatrix} \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{pmatrix}.$$

Because u_1, u_2 and u_3 are linearly independent, the transformation matrix is invertible so that the continuity of $\frac{\partial}{\partial x}, \frac{\partial}{\partial y}$ $\frac{\partial}{\partial z}$ is equivalent to the continuity of the $D_{u_i}, i = 1, 2, 3$. This proves the lemma for $r = 1$. Iterating we can express the $D^{\alpha_1}_{u_1} D^{\alpha_2}_{u_2} D^{\alpha}_{u} f(v)$ in terms of the $\frac{\partial^{|\alpha|} f}{\partial x^{\alpha_1} \partial y^{\alpha_2} \partial z^{\alpha_3}}$ and conversely. This shows the result.

Lemma 3.2.3 *$D^m_u p(v)$ at $v \in \langle v_1, v_2, v_3\rangle$ is uniquely determined by the $c_{ijkl}, 0 \le l \le m$.*

Proof: We look at the B-net of $D_u^m p(v)$.

$$D_u^m p(v) = \frac{d!}{(d-m)!} \sum_{i+j+k+l=d-m} c_{ijkl}^{(m)}(\mathbf{a}) B_{ijkl}^{d-m}(v),$$
$$B_{ijkl}^{d-m}(v) = \frac{(d-m)!}{i!j!k!l!} b_1^i b_2^j b_3^k b_4^l$$

for $v \in \langle v_1, v_2, v_3 \rangle$, $b_4 = 0$ so for $l \neq 0$, $B_{ijk0}^{d-m}(v) = 0$. Therefore

$$D_u^m p(v) = \frac{d!}{(d-m)!} \sum_{i+j+k=d-m} c_{ijk0}^{(m)}(\mathbf{a}) B_{ijkl}^{d-m}(v),$$

which shows the result taking into account (3.8)

$$c_{ijk0}^{(m)}(\mathbf{a}) = \sum_{\alpha+\beta+\gamma+\delta=m} c_{i+\alpha,j+\beta,k+\gamma,\delta} B_{\alpha,\beta,\gamma,\delta}^m(\mathbf{a}).$$

Lemma 3.2.4 *If p and q are defined on two adjacent tetrahedra which share a common face $\langle v_1, v_2, v_3 \rangle$, then p and q are joined in a C^r fashion if and only if*

$$D_u^m p(v) = D_u^m q(v), \quad v \in \langle v_1, v_2, v_3 \rangle, \quad m = 0, \ldots, r$$

where u is a direction not in the face $\langle v_1, v_2, v_3 \rangle$.

Proof: We note that $u_1 = v_2 - v_1$, $u_2 = v_3 - v_1$ and u are 3 independent directions and we need only to show that the condition of the lemma is equivalent to

$$D_{u_1}^{\alpha_1} D_{u_2}^{\alpha_2} D_u^{\alpha} p(v) = D_{u_1}^{\alpha_1} D_{u_2}^{\alpha_2} D_u^{\alpha} q(v), \quad \alpha_1 + \alpha_2 + \alpha = m,$$

$m = 0, \ldots, r$ in view of Lemma (3.2.2). This will be done by looking at $D_u^{\alpha} p(v)$. Let

$$D_u^{\alpha} p(v) = \sum_{i+j+k+l=s} c_{ijkl} B_{ijkl}^d, \qquad D_u^{\alpha} q(v) = \sum_{i+j+k+l=s} d_{ijkl} B_{ijkl}^d$$

Since $\alpha \leq r$, by the lemma $c_{ijkl} = d_{ijkl}, i+j+k+l = s$. We then show that $D_{u_1}^{\alpha_1} D_{u_2}^{\alpha_2} D_u^{\alpha} p(v)$ depends only on the c_{ijkl} and this forces continuity. But this is immediate since these derivatives are determined by the values of $D_u^{\alpha} p(v)$ on $\langle v_1, v_2, v_3 \rangle$. And these values are completely determined by the $c_{ijkl} = d_{ijkl}, i+j+k+l = s$.

Lemma 3.2.5 *Let $t = \langle v_1, v_2, v_3, v_4\rangle$ be a tetrahedron and w another point in $\mathbf{R}^3$ with barycentric coordinates $\mathbf{a} = (a_1, a_2, a_3, a_4)$. On $\langle v_1, v_2, v_3, w\rangle$ p can be written*

$$p = \sum_{i+j+k+l=d} d_{ijkl} B^d_{ijkl},$$

with

$$d_{ijkl} = c^{(l)}_{ijk0}(\mathbf{a}) = \sum_{\mu+\nu+\kappa+\delta=m} c^t_{i+\mu,j+\nu,\gamma+\kappa,\delta} B^l_{\mu,\nu,\kappa,\delta}(w).$$

Proof: For each $v \in t$, let $\widetilde{b_1}(v), \widetilde{b_2}(v), \widetilde{b_3}(v)$ and $\widetilde{b_4}(v)$ be the barycentric coordinates of v relative to $\langle v_1, v_2, v_3, w\rangle$. We substitute $w = a_1v_1 + a_2v_2 + a_3v_3 + a_4v_4$ in $v = \widetilde{b_1}v_1 + \widetilde{b_2}v_2 + \widetilde{b_3}v_3 + \widetilde{b_4}w$ and get $v = (\widetilde{b_1} + \widetilde{b_4}a_1)v_1 + (\widetilde{b_2} + \widetilde{b_4}a_2)v_2 + (\widetilde{b_3} + \widetilde{b_4}a_3)v_3 + \widetilde{b_4}a_4v_4$. Thus

$$\begin{aligned}
B^d_{\alpha,\beta,\gamma,\delta}(v) &= \frac{d!}{\alpha!\beta!\gamma!\delta!}(\widetilde{b_1} + \widetilde{b_4}a_1)^\alpha(\widetilde{b_2} + \widetilde{b_4}a_2)^\beta(\widetilde{b_3} + \widetilde{b_4}a_3)^\gamma(\widetilde{b_4}a_4)^\delta \\
&= \sum_{\mu=0}^{\alpha}\sum_{\nu=0}^{\beta}\sum_{\kappa=0}^{\gamma} \frac{d!}{\alpha!\beta!\gamma!\delta!}\binom{\alpha}{\mu}\widetilde{b_1}^{\alpha-\mu}\widetilde{b_4}^{\mu}a_1^\mu\binom{\beta}{\nu}\widetilde{b_1}^{\beta-\nu}\widetilde{b_4}^{\nu}a_1^\nu \\
&\quad \times \binom{\gamma}{\kappa}\widetilde{b_1}^{\gamma-\kappa}\widetilde{b_4}^{\kappa}a_1^\kappa\widetilde{b_4}^{\delta}a_4^\delta \\
&= \sum_{\mu=0}^{\alpha}\sum_{\nu=0}^{\beta}\sum_{\kappa=0}^{\gamma} \frac{d!}{\alpha!\beta!\gamma!\delta!}\frac{\alpha!}{\mu!(\alpha-\mu)!}\frac{\beta!}{\nu!(\beta-\nu)!}\frac{\gamma!}{\kappa!(\gamma-\kappa)!} \\
&\quad \times \widetilde{b_1}^{\alpha-\mu}\widetilde{b_2}^{\beta-\nu}\widetilde{b_3}^{\gamma-\kappa}\widetilde{b_4}^{\mu+\nu+\kappa+\delta}a_1^\mu a_2^\nu a_3^\kappa a_4^\delta.
\end{aligned}$$

We have

$$B^{\mu+\nu+\kappa+\delta}_{\mu,\nu,\kappa,\delta}(w) = \frac{(\mu+\nu+\kappa+\delta)!}{\mu!\nu!\kappa!\delta!}a_1^\mu a_2^\nu a_3^\kappa a_4^\delta$$

and

$$\begin{aligned}
\widetilde{B}^d_{\alpha-\mu,\beta-\nu,\gamma-\kappa,\mu+\nu+\kappa+\delta}(v) &= \frac{d!}{(\alpha-\mu)!(\beta-\nu)!(\gamma-\kappa)!(\mu+\nu+\kappa+\delta)!} \\
&\quad \times \widetilde{b_1}^{\alpha-\mu}\widetilde{b_2}^{\beta-\nu}\widetilde{b_3}^{\gamma-\kappa}\widetilde{b_4}^{\mu+\nu+\kappa+\delta},
\end{aligned}$$

so

$$B^d_{\alpha,\beta,\gamma,\delta}(v) = \sum_{\mu=0}^{\alpha}\sum_{\nu=0}^{\beta}\sum_{\kappa=0}^{\gamma} \widetilde{B}^d_{\alpha-\mu,\beta-\nu,\gamma-\kappa,\mu+\nu+\kappa+\delta}(v) B^{\mu+\nu+\kappa+\delta}_{\mu,\nu,\kappa,\delta}(w).$$

We substitute this in

$$p(v) = \sum_{\alpha+\beta+\gamma+\delta=d} c_{\alpha,\beta,\gamma,\delta} B^d_{\alpha,\beta,\gamma,\delta}(v)$$

and get

$$p(v) = \sum_{\alpha+\beta+\gamma+\delta=d} \sum_{\mu=0}^{\alpha} \sum_{\nu=0}^{\beta} \sum_{\kappa=0}^{\gamma} c_{\alpha,\beta,\gamma,\delta} B^{\mu+\nu+\kappa+\delta}_{\mu,\nu,\kappa,\delta}(w)$$
$$\times \widetilde{B}^d_{\alpha-\mu,\beta-\nu,\gamma-\kappa,\mu+\nu+\kappa+\delta}(v).$$

We put

$$l = \mu+\nu+\kappa+\delta, \quad k = \gamma-\kappa,$$
$$j = \beta-\nu, \qquad\qquad i = \alpha-\mu,$$

so $\widetilde{B}^d_{\alpha-\mu,\beta-\nu,\gamma-\kappa,\mu+\nu+\kappa+\delta} = \widetilde{B}^d_{ijkl}$. It appears that the coefficient of $\widetilde{B}^d_{ijkl}$ is

$$\widetilde{c}_{ijkl} = \sum_{\mu+\nu+\kappa+\delta=l} c_{i+\mu,j+\nu,k+\gamma,\delta} B^l_{\mu,\nu,\kappa,\gamma}.$$

And this proves the lemma.

Proof of the theorem: We can now give a proof of the theorem. Let u with T-coordinates $\mathbf{a} = (a_1, a_2, a_3, a_4)$ be a direction not parallel to the face $\langle v_1, v_2, v_3\rangle$. We need only to show that the conditions of the lemma is equivalent to

$$D^m_u p_t(v) = D^m_u p_{t'}(v), \quad v \in \langle v_1, v_2, v_3\rangle \quad m = 0, \ldots, r.$$

Let $p = \sum_{i+j+k+l=d} \widetilde{c}^t_{ijkl} B^d_{ijkl}$ on $\langle v_1, v_2, v_3, v_5\rangle$. The previous lemma tells how to relate $\widetilde{c}^t_{ijkl}$ and c^t_{ijkl}. To say that p_t and $p_{t'}$ agree on $\langle v_1, v_2, v_3\rangle$ gives $\widetilde{c}^t_{ijk0} = c^{t'}_{ijk0}, i+j+k = d$. Moreover $D_u p_t(v) = D_u p_{t'}(v)$ gives $\widetilde{c}^{t(1)}_{ijk0} = c^{t'(1)}_{ijk0}, i+j+k = d$ which written explicitly is

$$a_1\widetilde{c}^t_{i+1,j,k,0} + a_2\widetilde{c}^t_{i,j+1,k,0} + a_3\widetilde{c}^t_{i,j,k+1,0} + a_4\widetilde{c}^t_{i,j,k,1} =$$
$$a_1 c^{t'}_{i+1,j,k,0} + a_2 c^{t'}_{i,j+1,k,0} + a_3 c^{t'}_{i,j,k+1,0} + a_4 c^{t'}_{i,j,k,1}.$$

This immediately gives $\widetilde{c}^t_{ijk1} = c^{t'}_{ijk1}, i+j+k = d-1$. We now assume that

$$D^m_u p_t(v) = D^m_u p_{t'}(v), \quad v \in \langle v_1, v_2, v_3\rangle, \quad m = 0, \ldots, s-1;\ s \leq r$$

is equivalent to

$$\widetilde{c}^t_{ijkm} = c^{t'}_{ijkm},\ i+j+k = d-m,\ m = 0, \ldots, s-1;\ s \leq r.$$

$D^s_u p_t(v) = D^s_u p_{t'}(v), \quad v \in \langle v_1, v_2, v_3\rangle$ gives $\widetilde{c}^{t(s)}_{ijk0} = c^{t'(s)}_{ijk0}, i+j+k = d-s$ which written out in terms of $c^{t(s-1)}_{ijk0}$ gives $\widetilde{c}^{t'}_{ijks} = c^t_{ijks},\ i+j+k = d-s$. This completes the proof of the theorem using the expression of $\widetilde{c}^t_{ijkm}$ in terms of c^t_{ijkm}.

Example: We give the conditions of C^0 and C^1 continuity for $\langle v_1, v_2, v_3, v_4\rangle$ and $\langle v_1, v_2, v_3, v_5\rangle$. Let v_5 have B-coordinates (b_1, b_2, b_3, b_4), (c^t_{ijkl}) and $c^{t'}_{ijkl}$ encode the B-net of p_t and $p_{t'}$ respectively. For C^0 continuity we have

$$c^{t'}_{ijk0} = c^t_{ijk0},\ i+j+k = d.$$

For C^1 continuity, we add

$$\begin{aligned} c^{t'}_{ijk1} = {} & c^t_{i+1,j,k,0} B^1_{1000}(v_5) + c^t_{i,j+1,k,0} B^1_{0100}(v_5) + c^t_{i,j,k+1,0} B^1_{0010}(v_5) \\ & + c^t_{i,j,k,1} B^1_{0001}(v_5), \end{aligned}$$

but $B^1_{1000} = b_1$ and similar formulas for the other Bernstein polynomials. So

$$c^{t'}_{ijk1} = b_1 c^t_{i+1,j,k,0} + b_2 c^t_{i,j+1,k,0} + b_3 c^t_{i,j,k+1,0} + b_4 c^t_{i,j,k,1}.$$

Chapter 4

Weak formulation of PDE's and Spline Approximations

In this chapter, we recall basic properties of Sobolev spaces then we introduced the main ideas of this dissertation on an abstract variational problem. We review the approximation properties of spline spaces and give classical error estimates.

4.1 Sobolev spaces

We recall in this section the main results about Sobolev spaces which we shall use later. A reference for this section is [Girault and Raviart' 86]. Let Ω denote an open subset of $\mathbf{R}^3$ with boundary $\partial\Omega$. $\mathcal{D}(\Omega)$ is the linear space of infinitely differentiable functions with compact support on Ω. Let $\mathcal{D}'(\Omega)$ denote the dual space of $\mathcal{D}(\Omega)$ also called space of distributions. Let $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ and set

$$|\alpha| = \alpha_1 + \alpha_2 + \alpha_3.$$

For $u \in \mathcal{D}'(\Omega)$, we define $\partial^\alpha u$ in $\mathcal{D}'(\Omega)$ by

$$\langle \partial^\alpha u, \phi \rangle = (-1)^{|\alpha|} \langle u, \partial^\alpha \phi \rangle, \quad \forall u \in \mathcal{D}'(\Omega)$$

when u is $|\alpha|$ times differentiable, $\partial^\alpha u$ is the usual notion of derivative:

$$\partial^\alpha u = \frac{\partial^{|\alpha|} u}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \partial x_3^{\alpha_3}}.$$

For each integer $m \geq 0$ and real p, $1 \leq p \leq \infty$, we define the Sobolev space

$$W^{m,p}(\Omega) = \{v \in L^p(\Omega), \partial^\alpha u \in L^p(\Omega), \ \forall |\alpha| \leq m\}.$$

This is a Banach space when endowed with the norm

$$||u||_{m,p,\Omega} = (\sum_{|\alpha|\leq m} \int_\Omega |\partial^\alpha u(x)|^p \mathrm{d}x)^{\frac{1}{p}}, \quad p < \infty,$$

or

$$||u||_{m,\infty,\Omega} = \max_{|\alpha|\leq m}(\mathrm{esssup}_{x\in\Omega}|\partial^\alpha u(x)|), \quad p = \infty.$$

We also equip $W^{m,p}(\Omega)$ with the following semi-norm

$$|u|_{m,p,\Omega} = (\sum_{|\alpha|= m} \int_\Omega |\partial^\alpha u(x)|^p \mathrm{d}x)^{\frac{1}{p}}, \quad p < \infty$$

and the modification above for $p = \infty$.

When $p = 2$, $W^{m,2}(\Omega)$ is denoted $H^2(\Omega)$, and its norm and semi-norm are simply referred as $||u||_{m,\Omega}$ and $|u|_{m,\Omega}$. Also $H^m(\Omega)$ is a Hilbert space for the scalar product

$$(u, v)_{m,\Omega} = \sum_{|\alpha|\leq m} \int_\Omega \partial^\alpha u(x)\partial^\alpha v(x)\mathrm{d}x.$$

Let $W_0^{m,p}(\Omega)$ be the closure of $\mathcal{D}(\Omega)$ in $W^{m,p}(\Omega)$ for the norm $||.||_{m,p,\Omega}$. $W_0^{m,2}(\Omega)$ is denoted $H_0^m(\Omega)$. The following Poincare-Friedrichs lemma says that on $H_0^m(\Omega)$, $||u||_{m,\Omega}$ and $|u|_{m,\Omega}$ are two equivalent norms, [cf. Girault and Raviart'86].

Lemma 4.1.1 *If Ω is connected and bounded in at least one direction, i.e. there exists $\mathbf{n}$ such that $\{|\mathbf{x}\cdot\mathbf{n}|, \mathbf{x}\in\Omega\}$ is bounded, then for each $m \geq 0$, there exists a constant $K = K(m,\Omega) > 0$ such that*

$$||u||_{m,\Omega} \leq K|u|_{m,\Omega}, \quad \forall v \in H_0^m(\Omega).$$

For $1 \leq p \leq \infty$, we denote by $W^{-m,p'}(\Omega)$ the dual space of $W^{m,p}(\Omega)$, with p' satisfying

$$\frac{1}{p} + \frac{1}{p'} = 1. \tag{4.1}$$

It is equipped with the norm

$$||f||_{-m,p',\Omega} = \sup_{v \in W_0^{m,p}(\Omega),\ v \neq 0} \frac{\langle f, v \rangle}{||v||_{m,p,\Omega}}.$$

As usual $W^{-m,2}(\Omega) = H^{-m}(\Omega)$. The following lemma, [cf. Girault and Raviart'86], describes elements of $W^{-m,p'}(\Omega)$.

Lemma 4.1.2 *For p and p' satisfying (4.1), f belongs to $W^{-m,p'}(\Omega)$ if and only if there exist functions $f_\alpha \in L^{p'}(\Omega)$, for $|\alpha| \leq m$, such that*

$$f = \sum_{|\alpha| \leq m} \partial^\alpha f_\alpha.$$

Definition 4.1.3 *Ω is said to have a Lipschitz continuous boundary if for each $x_0 \in \partial\Omega$ there is a ball B of center x_0 and radius r and a Lipschitz continuous function ϕ defined on a domain $D \subset \mathbf{R}^2$ such that in a system of coordinates with the origin at x_0:*

1. *The set $\partial\Omega \cap B$ can be represented by an equation of type $x_3 = \phi(x_1, x_2)$*
2. *Each $x \in \Omega \cap B$ satisfies $x_3 < \phi(x_1, x_2)$.*

Although in this dissertation, we limit ourselves to domains with piecewise planar boundaries, it is interesting to note that some results hold for more general domains. We also recall the Sobolev embedding theorem. For our purposes it says that elements of $H^2(\Omega)$ are globally continuous.

Lemma 4.1.4 *Let Ω be an open subset of $\mathbf{R}^3$ with a Lipschitz continuous boundary and let $p \in \mathbf{R}$ with $1 \leq p < \infty$ and $m, n \in \mathbf{N}$ with $n \leq m$. We have*

$$W^{m,p}(\Omega) \subset \mathcal{C}^n(\Omega) \text{ provided } \frac{1}{p} < \frac{m-n}{3}.$$

For example, $H^2(\Omega) = W^{2,2}(\Omega) \subset \mathcal{C}^0(\Omega)$. We therefore have conditions under which point values for functions in a Sobolev space are well defined. To give a meaning for their values on the boundary, we'll need the following trace theorem, [cf. Brenner and Scott'94].

Theorem 4.1.5 *Suppose that Ω has a Lipschitz boundary, and that p is a real number in the range $1 \leq p \leq \infty$. Then there is a constant, C, such that*

$$||v||_{L^p(\partial\Omega)} \leq C||v||^{1-\frac{1}{p}}_{L^p(\Omega)}||v||^{\frac{1}{p}}_{W^{1,p}(\Omega)} \quad \text{for all } v \in W^{1,p}(\Omega).$$

For $v \in W^{1,p}(\Omega)$, we will call trace of v on $\partial\Omega$, its restriction on the boundary interpreted as an element of $L^2(\partial\Omega)$. By definition,

$$H^{\frac{1}{2}}(\partial\Omega) = \{\tau(u), u \in H^1(\Omega)\},$$

where $\tau(u)$ stands for trace of u. $H^{-\frac{1}{2}}(\partial\Omega)$ will denote the dual of $H^{\frac{1}{2}}(\partial\Omega)$. Finally we give a few Green's formulas, [cf. Brenner and Scott'94]. Here

$$\frac{\partial v}{\partial \nu} = \nu \cdot \nabla v.$$

Lemma 4.1.6 *Let Ω be a bounded open subset of $\mathbf{R}^3$ with a Lipschitz continuous boundary and let ν denote the unit outward normal on $\partial\Omega$, defined almost everywhere and by assumption is in $L^\infty(\partial\Omega)$. We have:*

- *For $\mathbf{u} \in W^{1,1}(\Omega)^3$,*

$$\int_\Omega \nabla \cdot \mathbf{u} = \int_{\partial\Omega} \mathbf{u} \cdot \nu.$$

- *Let $v, w \in H^1(\Omega)$. Then, for i=1,2,3,*

$$\int_\Omega (\frac{\partial v}{\partial x_i}) w \ \mathrm{d}x = -\int_\Omega v(\frac{\partial w}{\partial x_i}) \ \mathrm{d}x + \int_{\partial\Omega} vw \ \nu_i. \tag{4.2}$$

- *For $u \in H^2(\Omega)$ and $v \in H^1(\Omega)$, we have*

$$\int_{\partial\Omega} (-\Delta u) v \ \mathrm{d}x = \int_{\partial\Omega} \nabla u \cdot \nabla v \ \mathrm{d}x - \int_{\partial\Omega} \frac{\partial u}{\partial \nu} v. \tag{4.3}$$

4.2 Abstract variational problems

We have the following theorem due to Lax and Milgram [Lax'54]:

Theorem 4.2.1 *Let V be a real Hilbert space with norm $||.||$ and let $(u,v) \to a(u,v)$ be a real bilinear form on $V \times V$ and $f : V \to \mathbf{R}$ be a continuous linear form. We assume that a is continuous and elliptic on V, i.e. there exist two constants M and $\alpha > 0$ such that*

$$|a(u,v)| \leq M||u||||v||, \qquad \forall u, v \in V \tag{4.4}$$

$$a(v,v) \geq \alpha||v||^2, \qquad \forall v \in V. \tag{4.5}$$

Then the problem: Find $u \in V$ such that

$$a(u,v) = f(v), \qquad \forall v \in V, \tag{4.6}$$

has one and only one solution.

We have the following corollary which will be extensively used.

Corollary 4.2.2 *When a is symmetric, i.e. $a(u,v) = a(v,u)$ $\forall u, v \in V$, then the solution of (4.6) is also the only element of V that minimizes the following functional also called energy functional*

$$J(v) = \frac{1}{2}a(v,v) - f(v).$$

Proof: We show that there's only one element in V which minimizes J and that it is also a solution of (4.6). By the V-ellipticity, $a(v,v) \geq 0, \forall v \in V$ and $a(v,v) = 0$ only if $v = 0$. The bilinear form a therefore defines an inner product over V. (4.4) and (4.5) give

$$\alpha||v||^2 \leq a(v,v) \leq M||v||^2.$$

This shows that the associated norm to a is equivalent to $||.||$ and so when V is equipped with the inner product $a(.,.)$, it is a Hilbert space. By the Riesz representation theorem there exists $u \in V$ unique such that

$$a(u,v) = f(v). \tag{4.7}$$

We have $a(v-u, v-u) = a(v,v) - 2a(u,v) + a(u,u)$ using the symmetry of a so

$$\begin{aligned} J(v) &= \frac{1}{2}a(v,v) - f(v) = \frac{1}{2}a(v,v) - a(u,v) \\ &= \frac{1}{2}a(v-u, v-u) - \frac{1}{2}a(u,u). \end{aligned}$$

And this shows that minimizing J amounts to minimize $a(v-u, v-u)$. We conclude that J has a unique minimizer u solution of (4.7).

Remark: We have also thus proved Theorem (4.2.1) in the case a symmetric.

4.3 Spline Approximations by Energy Minimization

We are going to seek approximate solutions of boundary value problems in finite dimensional subspaces of Sobolev spaces. Specifically we shall use the spline space $S_d^r(\Omega)$. We define

$$S_d^r(\Omega) = \{p \in C^r(\Omega), p_t \in P_d \ \forall t \in \mathcal{T}\}.$$

We have $S_d^r(\Omega) \subset H^{r+1}(\Omega)$. This follows from the following lemma, [Braess'92,p 60].

Lemma 4.3.1 *Let $k \geq 1$ and suppose Ω is bounded. Then a piecewise infinitely differentiable function $v : \overline{\Omega} \to \mathbf{R}$ belongs to $H^k(\Omega)$ if and only if $v \in C^{k-1}(\overline{\Omega})$.*

4.3.1 Approximation properties of spline spaces

For $r = 0$, $S_d^r(\Omega)$ is the classical Lagrangian finite element space for arbitrary d. There are classical approximation results for this finite element space which depend on the following interpolation operator which we now introduce.

Let $\mathcal{T}$ be a tetrahedral partition of Ω and let $h = \max\, h_T,\ T \in \mathcal{T}$ where h_T is a measure of the size of a tetrahedron T in $\mathcal{T}$ and denote by σ the maximum shape measure associated with $\mathcal{T}$.

Let $T = \langle v_1, v_2, v_3, v_4 \rangle \in \mathcal{T}$, for v continuous on T let $\Pi^d(v_T)$, $d \geq 1$ be the unique polynomial on T which interpolates v at the domain points $\psi_{ijkl} = \frac{iv_1+jv_2+kv_3+lv_4}{d}$. We define $\Pi^d(v)$ globally by

$$\Pi^d(v)_{|T} = \Pi^d(v_{|T}).$$

The interpolant $\Pi^d(v)$ is therefore automatically continuous since the value of a polynomial at the domain points $\psi_{ijk} = \frac{iv_1+jv_2+kv_3}{d}$ of a face $\langle v_1, v_2, v_3 \rangle$ uniquely defines it on that face. For $m = 0, 1$, we have the following inequality [Quarteroni and Valli'97, p.91]

$$|v - \Pi^d(v)|_{m,\Omega} \leq Ch^{d+1-m}|v|_{d+1,\Omega}, \qquad \forall v \ \in H^{d+1}(\Omega). \tag{4.8}$$

The constant C depends on Ω, σ and d and the restriction on the index m is due to the fact that $\Pi^d(v)$ is merely continuous. It is the lack of a suitable interpolation operator and estimates as the above for $S_d^r(\Omega)$ for $r \geq 1$ that makes it difficult to derive error estimates for arbitrary smoothness and an arbitrary tetrahedral partition. Ming-Jun Lai has proved in his dissertation these kind of estimates for $S_d^r(\Omega)$ when $d \geq 6r + 3$ in the L_∞ norm.

4.3.2 Approximate solutions

We assume that the hypotheses of the Lax-Milgram lemma (Theorem 4.2.1) hold, and a is symmetric. We approximate the problem

$$(P) \qquad \text{Find } u \in V \text{ such that } a(u, v) = f(v), \quad \forall\ v \in V,$$

by a similar finite dimensional one

$$(P_1) \qquad \text{Find } u \in S \text{ such that } a(u, v) = f(v), \quad \forall\ v \in S,$$

where S is a finite dimensional subset of V. Typically, S will be a subset of $S_d^r(\Omega)$.

We now describe $S_d^r(\Omega)$ in a way more suitable for our purposes. Recall the following B-form representation of a spline u with B-net $\mathbf{c} \in \mathbf{R}^N$ where N is equal to the product of the number of tetrahedra and the dimension of P_d,

$$u = \sum_t \sum_{i+j+k+l=d} c_{ijkl}^t B_{ijkl}^d.$$

We shall identify S_d^r with

$$\{\mathbf{c} \in \mathbf{R}^N, H\mathbf{c} = 0\},$$

where H is a smoothness matrix of order r and degree d that encodes the smoothness conditions that ensure that a spline is in $S_d^r(\Omega)$.

In addition since $S \subset S_d^r(\Omega)$, we shall describe it by additional constraints. These constraints take the form $B\mathbf{c} = \mathbf{G}$ for some matrices B and $\mathbf{G}$ to impose boundary conditions for PDE's or $D\mathbf{c} = 0$ to impose the divergence free constraint for the Stokes and Navier-Stokes problems. Here D is a discrete divergence matrix. For the following abstract discussion, we shall assume that

$$S = \{\mathbf{c} \in \mathbf{R}^N, H\mathbf{c} = 0, U\mathbf{c} = \mathbf{d}\},$$

where $U\mathbf{c} = \mathbf{d}$ encode constraints in the approximating space for some matrix U and a vector $\mathbf{d}$.

Problem (P_1) is equivalent to

$$(P_2) \qquad \text{Find } u \in S \text{ which minimizes, } J(u) = \frac{1}{2}a(u,u) - f(u) \text{ over } S.$$

For $u \in S$, we can write for simplicity

$$u = \sum_{i=1}^N c_i \psi_i,$$

where ψ_i is one of the Bernstein polynomials B^d_{ijkl} on a tetrahedron of the partition. We have

$$J(u) = \frac{1}{2}a(\sum_{i=1}^{N} c_i\psi_i, \sum_{i=1}^{N} c_i\psi_i) - f(\sum_{i=1}^{N} c_i\psi_i)$$
$$= \frac{1}{2}\sum_{i,j=1}^{N} c_i c_j a(\psi_i, \psi_j) - \sum_{i=1}^{N} c_i f(\psi_i).$$

Denote by $\mathbf{F}$ the vector $(f(\psi_1), \cdots, f(\psi_N))^T$ and by A the matrix with entries $a_{ij} = a(\psi_i, \psi_j)$. A will be called the stiffness matrix and $\mathbf{F}$ the load vector. We can write

$$J(u) = J(c) = \frac{1}{2}\mathbf{c}^T A\mathbf{c} - \mathbf{c}^T\mathbf{b}.$$

We note that A is symmetric and that problem (P_2), is equivalent to the following constrained optimization problem.

(P_3) Minimize $J(c) = \frac{1}{2}\mathbf{c}^T A\mathbf{c} - \mathbf{c}^T\mathbf{F}$ over $\mathbf{R}^N$ under the constraints

$$H\mathbf{c} = 0 \text{ and } U\mathbf{c} = \mathbf{d}.$$

By the theory of Lagrange multipliers there exist two vectors λ_1 and λ_2 such that

$$\begin{aligned} \mathbf{c}^T A + \lambda_1^T H + \lambda_2^T U &= \mathbf{F}^T, \\ H\mathbf{c} &= 0, \\ U\mathbf{c} &= \mathbf{d}. \end{aligned} \tag{4.9}$$

Note that the conditions encoded in U are not linearly independent since in general U will be a (N, m) matrix with $m > N$. The Lagrange multipliers corresponding to the redundant equations are zero.

(4.9) can be written

$$\begin{aligned} A\mathbf{c} + H^T\lambda_1 + U^T\lambda_2 &= \mathbf{F}, \\ H\mathbf{c} &= 0, \\ U\mathbf{c} &= \mathbf{d}, \end{aligned}$$

or in matrix form

$$\begin{pmatrix} A & H^T & U^T \\ H & 0 & 0 \\ U & 0 & 0 \end{pmatrix} \begin{bmatrix} \mathbf{c} \\ \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} \mathbf{F} \\ 0 \\ \mathbf{d} \end{bmatrix}. \tag{4.10}$$

Since we have assumed that the conditions of the Lax-Milgram theorem hold, problem (P) and similarly (P_1), (P_2) and (P_3) have a unique solution. But the situation is not so clear for (4.10). We have existence of the component $\mathbf{c}$ of the unknown in (4.10) by the existence theorem for (P_3). However there is no evidence that solving (4.10) will give the unique solution of (P_3). Indeed the matrix

$$R = \begin{pmatrix} A & H^T & U^T \\ H & 0 & 0 \\ U & 0 & 0 \end{pmatrix}$$

will be, in general, singular. Let $Z = \left[\mathbf{F}^T, 0, \mathbf{d}^T\right]^T$; We have

$$R \ \left[\mathbf{c}^T, \lambda_1^T, \lambda_2^T\right]^T - Z = 0.$$

It turns out that it is very convenient to find a least squares solution of the equation $Rx = Z$ with MATLAB. Since R does not have full rank, there is not a unique least squares solution to this equation. This can be proved using the projection theorem, [Ciarlet'89, p.272]. However, for such a least squares solution $(\mathbf{e}^T, \beta_1^T, \beta_2^T)^T$ we have

$$||R \ \left[\mathbf{e}^T, \beta_1^T, \beta_2^T\right]^T - Z|| = 0,$$

so

$$R \ \left[\mathbf{e}^T, \beta_1^T, \beta_2^T\right]^T = Z.$$

This means that $\left[\mathbf{e}^T, \beta_1^T, \beta_2^T\right]^T$ satisfies the necessary conditions for $\mathbf{e}$ being a minimizer of J in S. Those conditions are also sufficient since the functional J is convex

[Ciarlet'89, p 246]. The convexity of J is proven by using the fact that A is symmetric, [Ciarlet'89]. By the unicity of a solution to the problem P_3, $\mathbf{e} = \mathbf{c}$. It is such a $\mathbf{c}$ that we compute. We'll now indicate two issues that arise from the discretization process. Since we are going to interpolate functions on the boundary with an interpolation operator which is merely continuous it will be impossible to satisfy smoothness conditions near the boundary. So we relaxed the condition of smoothness and construct approximations which are smooth across tetrahedra which do not share a face with the boundary, i.e. if we let $\widetilde{\Omega}$ denote the union of these tetrahedra

$$\{p \ \in \ C^r(\widetilde{\Omega}),\ p_t \in P_d, \quad \forall t \in \mathcal{T}\}. \tag{4.11}$$

The second issue is that the matrix R is quite large, so it is desirable to reduce its size. We now describe an algorithm which will be referred to later as the matrix iterative algorithm. It is a variant of the augmented Lagrangian algorithm. It can be used to cope with matrices of large size. The trade off is that as an iterative method, it is less accurate than the least squares method. On the other hand, we prove below the convergence of this algorithm only in the symmetric case. A convergence proof for the nonsymmetric case can be found in [Awanou and Lai'03]. Oscillations in the numerical error when the tetrahedral partition is refined makes it unappealing in the non symmetric case.

4.4 A matrix iterative Algorithm

With obvious notations let's write the equation

$$\begin{pmatrix} A & H^T & U^T \\ H & 0 & 0 \\ U & 0 & 0 \end{pmatrix} \begin{bmatrix} \mathbf{c} \\ \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} \mathbf{F} \\ 0 \\ \mathbf{d} \end{bmatrix}$$

as

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \mathbf{c} \\ \lambda \end{pmatrix} = \begin{bmatrix} \mathbf{F} \\ \mathbf{G} \end{bmatrix}. \tag{4.12}$$

We consider for l= $0, 1, 2, \ldots$, the sequence of problems

$$\begin{pmatrix} A & B^T \\ B & -\epsilon I \end{pmatrix} \begin{bmatrix} \mathbf{c}^{l+1} \\ \lambda^{l+1} \end{bmatrix} = \begin{bmatrix} \mathbf{F} \\ \mathbf{G} - \epsilon\lambda^l \end{bmatrix}, \tag{4.13}$$

where λ^0 is a suitable initial guess, for example $\lambda^0 = 0$, and I is the identity matrix. (4.13) reads

$$A\mathbf{c}^{l+1} + B^T\lambda^{l+1} = \mathbf{F} \tag{4.14}$$

$$B\mathbf{c}^{l+1} - \epsilon\lambda^{l+1} = \mathbf{G} - \epsilon\lambda^l. \tag{4.15}$$

Multiplying (4.15) on the left by B^T we get

$$B^TB\mathbf{c}^{l+1} - \epsilon B^T\lambda^{l+1} = B^T\mathbf{G} - \epsilon B^T\lambda^l$$

or

$$B^T\lambda^{l+1} = -\frac{1}{\epsilon}B^T\mathbf{G} + B^T\lambda^l + \frac{1}{\epsilon}B^TB\mathbf{c}^{l+1}.$$

Substituting this last relation back into (4.14), we get

$$(A + \frac{1}{\epsilon}B^TB)\mathbf{c}^{l+1} = \mathbf{F} + \frac{1}{\epsilon}B^T\mathbf{G} - B^T\lambda^l. \tag{4.16}$$

which reads for $l = 0$

$$(A + \frac{1}{\epsilon}B^TB)\mathbf{c}^1 = \mathbf{F} + \frac{1}{\epsilon}B^T\mathbf{G} - B^T\lambda^0.$$

Using $A\mathbf{c}^l = \mathbf{F} - B^T\lambda^l$, we have

$$(A + \frac{1}{\epsilon}B^TB)\mathbf{c}^{l+1} = A\mathbf{c}^l + \frac{1}{\epsilon}B^T\mathbf{G}, \qquad \text{for } l = 1, 2, \ldots \tag{4.17}$$

This suggests the following algorithm.

ALGORITHM:

Given λ^0, choose $\epsilon > 0$ small enough and define $\mathbf{c}^1$ by

$$\mathbf{c}^1 = (A + \frac{1}{\epsilon}B^TB)^{-1}(\mathbf{F} + \frac{1}{\epsilon}B^T\mathbf{G} - B^T\lambda^0)$$

and iteratively define

$$\mathbf{c}^{l+1} = (A + \frac{1}{\epsilon}B^TB)^{-1}(A\mathbf{c}^l + \frac{1}{\epsilon}B^T\mathbf{G}),\ l = 1, 2, \ldots$$

We have the following theorem:

Theorem 4.4.1 *Assume that A is symmetric positive definite with respect to B, i.e. $x^TAx \geq 0$, and $x^TAx = 0$ with $Bx = 0$ implies $x = 0$. Then, the sequence $(\mathbf{c}^{l+1})$ converges to the solution $\mathbf{c}$ of (4.12).*

Remark: We prove below not only the convergence of the sequence, but also that the convergence factor tends to 0 as $\epsilon \to 0$. We refer to [Fortin and Glowinski'83], or [Awanou and Lai'03] for another proof of the convergence of the algorithm in the context of augmented lagrangian algorithms.

Proof of the theorem.

First, we need to show that $A + \frac{1}{\epsilon}B^TB$ is invertible. Since A is a square matrix, it is enough to show that

$$(A + \frac{1}{\epsilon}B^TB)x = 0 \Rightarrow x = 0.$$

$$x^T(A + \frac{1}{\epsilon}B^TB)x = x^TAx + \frac{1}{\epsilon}(Bx)^T(Bx).$$

Since $x^TAx \geq 0$ and $x^TB^TBx \geq 0$,

$$x^T(A + \frac{1}{\epsilon}B^TB)x = 0 \Rightarrow x^TAx = 0 \text{ and } (Bx)^T(Bx) = 0,$$

so $x^T Ax = 0$ and $Bx = 0$. Since A is assumed to be symmetric positive definite with respect to B, we get $x = 0$. The sequence $\mathbf{c}^{l+1}$ is therefore well-defined. Let's write

$$E = (A + \frac{1}{\epsilon} B^T B).$$

By (4.16),

$$E\mathbf{c}^{l+1} = \mathbf{F} + \frac{1}{\epsilon} B^T \mathbf{G} - B^T \lambda^l.$$

The same way, using (4.12), we have

$$\mathbf{F} = A\mathbf{c} + B^T \lambda$$

$$G = B\mathbf{c},$$

so

$$\begin{aligned} \mathbf{F} + \frac{1}{\epsilon} B^T \mathbf{G} &= A\mathbf{c} + B^T \lambda + \frac{1}{\epsilon} B^T B\mathbf{c} \\ &= (A + \frac{1}{\epsilon} B^T B)\mathbf{c} + B^T \lambda \\ &= E\mathbf{c} + B^T \lambda. \end{aligned}$$

It follows that

$$E\mathbf{c} = \mathbf{F} + \frac{1}{\epsilon} B^T \mathbf{G} - B^T \lambda. \tag{4.18}$$

Therefore

$$\mathbf{c}^{l+1} - \mathbf{c} = E^{-1} B^T (\lambda - \lambda^l).$$

We'll show convergence of the sequence $(\lambda^l),\ l = 1, 2, \ldots$ to λ. This will prove the result. Using (4.15), we get

$$-\epsilon(\lambda^{l+1} - \lambda) = -\epsilon(\lambda^l - \lambda) + \mathbf{G} - B\mathbf{c}^{l+1}.$$

But

$$\mathbf{c}^{l+1} = E^{-1}\mathbf{F} + \frac{1}{\epsilon} E^{-1} B^T \mathbf{G} - E^{-1} B^T \lambda^l,$$

so

$$\begin{aligned}\epsilon(\lambda^{l+1}-\lambda) &= \epsilon(\lambda^l-\lambda)-\mathbf{G}+BE^{-1}\mathbf{F}\\ &\quad+\frac{1}{\epsilon}BE^{-1}B^T\mathbf{G}-BE^{-1}B^T\lambda^l\\ &= \epsilon(\lambda^l-\lambda)-\mathbf{G}+BE^{-1}(\mathbf{F}+\frac{1}{\epsilon}B^T\mathbf{G}-B^T\lambda^l).\end{aligned}$$

Using (4.18), we have

$$\begin{aligned}\epsilon(\lambda^{l+1}-\lambda) &= \epsilon(\lambda^l-\lambda)-\mathbf{G}+BE^{-1}(E\mathbf{c}+B^T\lambda-B^T\lambda^l)\\ &= \epsilon(\lambda^l-\lambda)-D(\lambda^l-\lambda),\end{aligned}$$

where $D=BE^{-1}B^T$ and we used $B\mathbf{c}=\mathbf{G}$. Finally

$$\lambda^{l+1}-\lambda=(I-\frac{1}{\epsilon}D)(\lambda^l-\lambda). \tag{4.19}$$

From (4.12) and (4.13),

$$B(\mathbf{c}-\mathbf{c}^l)=\epsilon(\lambda^{l-1}-\lambda^l),$$

so $\lambda^{l-1}-\lambda^l$ is in the range of B and we may assume that the same is true of $\lambda^{l-1}-\lambda$ by writing the later in terms $\lambda^{k-1}-\lambda^k,\ k\le l$ and choosing λ^0 such that $\lambda^0-\lambda$ is in the range of B. This suggests that we regard D as a mapping from $\mathrm{Im}(B)$ to $\mathrm{Im}(B)$, where $\mathrm{Im}(B)$ denotes the range of B. We claim that D is an invertible mapping from $\mathrm{Im}(B)$ to $\mathrm{Im}(B)$. Since D is symmetric, it is enough to show that

$$R(y)=\frac{y^TDy}{y^Ty}>0,\ \forall\ y\ \in\ \mathrm{Im}(B),\ y\neq 0,$$

where $R(y)$ is the Raleigh quotient. We have

$$\begin{aligned}R(y)=\frac{y^TBE^{-1}B^Ty}{y^Ty}&=\frac{y^TBE^{-1}(EE^{-1})B^Ty}{y^Ty}\\ &=\frac{(y^TBE^{-1})E(E^{-1}B^Ty)}{y^Ty}.\end{aligned}$$

So

$$R(y)=\frac{||E^{-1}B^Ty||_E^2}{y^Ty}, \tag{4.20}$$

where we have defined a norm $||\ ||_E$ associated with the positive definite matrix E;

$$||u||_E^2 = b(u,u) \text{ with } b(u,v) = v^T E u.$$

We have

$$||E^{-1}B^T y||_E = \sup_{v\in\mathbf{R}^N} \frac{b(E^{-1}B^T y, v)}{||v||_E},$$

with

$$\frac{b(E^{-1}B^T y, v)}{||v||_E} = \frac{v^T E E^{-1} B^T y}{||v||_E} = \frac{v^T B^T y}{||v||_E} = \frac{y^T B v}{||v||_E}, \quad \forall\ v \neq 0\ \in\ \mathbf{R}^N.$$

So

$$||E^{-1}B^T y||_E = \sup_{v\in\mathbf{R}^N} \frac{y^T B v}{||v||_E}. \tag{4.21}$$

We claim that B is an invertible mapping from $\mathrm{Im}(B^T)$ to $\mathrm{Im}(B)$, [Segal'79]. Since $y \in \mathrm{Im}B$, there's v in $\mathrm{Im}(B^T)$ such that $y = Bv$. We write $||v||^2 = v^T v$. Then,

$$\begin{aligned} R(y) &\geq \frac{1}{||Bv||^2}\left(\frac{(Bv)^T(Bv)}{||v||_E}\right)^2 = \frac{||Bv||^2}{||v||_E^2} \\ &= \frac{||Bv||^2}{v^T A v + \frac{1}{\epsilon} v^T B^T B v} = \frac{||Bv||^2}{||v||_A^2 + \frac{1}{\epsilon}||Bv||^2}. \end{aligned}$$

For an operator X, $\mu_{X\min}$ and $\mu_{X\max}$ denote respectively the smallest and greatest eigenvalues of X. We have, using Raleigh's principle,

$$||v||_A^2 \leq \mu_{A\max}||v||^2 \quad \text{and} \quad ||v||^2\mu_{A\min} \leq ||v||_A^2.$$

On the other hand, since $y = Bv \neq 0$, $||Bv|| \geq \frac{||v||}{||B^{-1}||}$. We therefore have

$$R(y) \geq \frac{\frac{1}{||B^{-1}||^2}}{\mu_{A\max} + \frac{1}{\epsilon}||B||^2} > 0,$$

which shows that D is invertible. We want to estimate the spectral radius of $I - \frac{1}{\epsilon}D$. We have

$$\begin{aligned} \mu_{\frac{1}{\epsilon}D\min} &\geq \frac{1}{\epsilon}R(y) \\ &\geq \frac{\frac{1}{\epsilon}||Bv||^2}{||v||_A^2 + \frac{1}{\epsilon}||Bv||^2}, \qquad y = Bv,\ v\ \in\ \mathrm{Im}B \end{aligned}$$

Taking $\epsilon \to 0$, we get

$$\mu_{\frac{1}{\epsilon}D\min} \geq 1.$$

On the other hand,

$$\begin{aligned}
\mu_{\frac{1}{\epsilon}D\max} &= \sup_{y\in \mathrm{Im}(B)} \frac{1}{\epsilon}\frac{y^T D y}{y^T y} \\
&= \sup_{y\in \mathrm{Im}(B)} \frac{1}{\epsilon}\frac{||E^{-1}B^T y||_E^2}{y^T y} \quad \text{using (4.20)} \\
&= \sup_{y\in \mathrm{Im}(B)} \sup_{v\in \mathbf{R}^N} \frac{1}{\epsilon y^T y}\frac{(y^T B v)^2}{||v||_E^2} \quad \text{using (4.21)} \\
&\leq \sup_{v\in \mathbf{R}^N} \frac{1}{\epsilon}\frac{||Bv||^2}{||v||_E^2} \\
&= \sup_{v\in \mathbf{R}^N} \frac{\frac{1}{\epsilon}||Bv||^2}{||v||_A^2 + \frac{1}{\epsilon}||Bv||^2} \\
&\leq 1.
\end{aligned}$$

These results show that the convergence factor in (4.19) goes to zero as $\epsilon \to 0$ since if μ is an eigenvalue of $\frac{1}{\epsilon}D$, $1-\mu$ is an eigenvalue of $I - \frac{1}{\epsilon}D$. This completes the proof.

To finish let's point out a classical error estimate for finite dimensional approximations, Cea's lemma, [Brenner and Scott'94]. With the notations of this section, let u_S be the approximate solution in S. Then

$$||u - u_S||_V \leq C \min_{v\in S} ||u - v||_V$$

for a constant C. Using the inequality (4.8), one shows for specific problems inequalities of this type with right hand side depending on the discretization parameter h.

Chapter 5

Spline Approximations of the 3D Poisson Equation and the 3D Biharmonic equation

We now apply the techniques we have developed to a simple second order elliptic equation, the Poisson equation and to a fourth order elliptic equation, the biharmonic equation.

5.1 The Case of the Poisson Equation

We consider numerical approximations of the 3D Poisson equation by splines of arbitrary degree and arbitrary smoothness over an arbitrary tetrahedral partition of a polygonal domain of $\mathbf{R}^3$. The equation is put in variational form and the associated energy functional is minimized over a subset of a space of splines. The domain Ω here is a bounded open subset of $\mathbf{R}^3$ with piecewise planar boundary. We consider both the Dirichlet and Neumann boundary conditions and in the later case, the domain will be assumed connected.

5.1.1 Existence and Uniqueness

We first consider the Dirichlet problem

$$\begin{cases} -\Delta u &= f \text{ in } \Omega \\ u &= g \text{ on } \partial\Omega \end{cases}$$

where $\partial\Omega$ will denote the boundary of Ω and Δ denotes the Laplace operator, $\Delta = \sum_{i=1}^{3} \frac{\partial^2}{\partial x_i^2}$. Multiplying $-\Delta u = f$ by v sufficiently smooth which vanishes on the

boundary, integrating over Ω we get after using Green's formula,

$$\int_\Omega \nabla u \nabla v = \int_\Omega fv. \tag{5.1}$$

Let $a(u,v) = \int_\Omega \nabla u \nabla v$. We have the following existence and uniqueness results of a weak solution of the homogeneous Poisson, equation. [Girault and Raviart'86].

Theorem 5.1.1 *For f in $H^{-1}(\Omega)$, the problem: Find u in $H_0^1(\Omega)$, such that*

$$a(u,v) = \int_\Omega fv, \qquad \forall v \in H_0^1(\Omega) \tag{5.2}$$

has a unique solution u in $H_0^1(\Omega)$. Moreover, since the bilinear form a is symmetric, the solution u minimizes the functional

$$J : v \longrightarrow J(v) = \frac{1}{2}a(v,v) - \int_\Omega fv \text{ over } H_0^1(\Omega). \tag{5.3}$$

We would like to also describe the solution of the non homogeneous Dirichlet problem as a solution of a minimization problem.

Let $V = \{w \in H^1(\Omega), w = g \text{ on } \partial\Omega\}$ and u_0 an element of V. We may then assume that g is in $H^{\frac{1}{2}}(\partial\Omega)$. We notice that $u_0 + H_0^1(\Omega) = V$. For $u \in V$, we write $u = u_0 + w$ and multiply $-\Delta(u_0 + w) = f$ by v an element of $H_0^1(\Omega)$. We get

$$a(w,v) = \int_\Omega fv - a(u_0, v), \tag{5.4}$$

so the problem: Find w in $H_0^1(\Omega)$, such that

$$a(w,v) = \int_\Omega fv - a(u_0,v), \quad \forall v \in H_0^1(\Omega)$$

has a unique solution w_0 in $H_0^1(\Omega)$ by the Lax-Milgram theorem. By its corollary (4.2.2), w_0 is a minimizer of

$$K(w) = \frac{1}{2}a(w,w) - \int_\Omega fw + a(u_0, w), \tag{5.5}$$

A simple substitution shows that $u_0 + w_0$ solves the non homogeneous problem

$$a(u, v) = \int_\Omega fv, \qquad \forall v \in H_0^1(\Omega), u \in V.$$

Let

$$L(w) = J(u_0 + w) = \frac{1}{2}a(u_0 + w, u_0 + w) - \int_\Omega f(u_0 + w),$$

we have

$$L(w) = K(w) + J(u_0).$$

Indeed for w in $H_0^1(\Omega)$,

$$\begin{aligned} L(w) &= \frac{1}{2}\int_\Omega \nabla(u_0 + w)\nabla(u_0 + w) - \int_\Omega f(u_0 + w) \\ &= \frac{1}{2}\int_\Omega |\nabla w|^2 - \int_\Omega fw + \int_\Omega \nabla u_0 \nabla w + \frac{1}{2}\int_\Omega |\nabla u_0|^2 - \int_\Omega fu_0 \\ &= K(w) + J(u_0). \end{aligned}$$

So

$$L(w_0) = K(w_0) + J(u_0) \leq K(w) + J(u_0) = L(w)$$

since w_0 is a minimizer of K. As $u_0 + H_0^1(\Omega) = V$, this shows that $u_0 + w_0$ minimizes $L(w)$ over V. We have then proved the following well-known existence and uniqueness result for the non-homogeneous Poisson equation.

Theorem 5.1.2 *For f in $H^{-1}(\Omega)$ and g in $H^{\frac{1}{2}}(\partial\Omega)$, the functional*

$$J : v \longrightarrow J(v) = \frac{1}{2}a(v, v) - \int_\Omega fv$$

has a unique minimizer in V which is the unique solution of the non-homogeneous Poisson problem.

We now consider the Neumann problem

$$\begin{cases} -\Delta u &= f \text{ in } \Omega \\ \frac{\partial u}{\partial n} &= g \text{ on } \partial\Omega. \end{cases} \tag{5.6}$$

f and g are given. Assuming that u is smooth, multiplying the first of (5.6) by v smooth, we have

$$\int_\Omega (-\Delta u)v = \int_\Omega \nabla u \cdot \nabla v - \int_{\partial\Omega} \frac{\partial u}{\partial \mathbf{n}} v.$$

So

$$\int_\Omega \nabla u \cdot \nabla v = \int_\Omega fv + \int_{\partial\Omega} \frac{\partial u}{\partial \mathbf{n}} v. \tag{5.7}$$

If u is a solution of (5.6), then $u + c$ is also a solution for a constant c. Let's assume that $f \in L^2(\Omega)$, $g \in H^{-\frac{1}{2}}(\partial\Omega)$. We define on $H^1(\Omega)$ the equivalence relation

$$u \simeq v \text{ if and only if } u - v \in \mathbf{R}$$

and seek $u \in H^1(\Omega)/\mathbf{R}$ such that (5.7) holds. For $\dot{u} \in H^1(\Omega)/\mathbf{R}$ and $\dot{v} \in H^1(\Omega)/\mathbf{R}$, let

$$a(\dot{u}, \dot{v}) = \int_\Omega \nabla u \cdot \nabla v, \ \forall u \in \dot{u}, \forall v \in \dot{v}, \forall\ \dot{u}, \dot{v} \in H^1(\Omega)/\mathbf{R}.$$

and consider the variational problem:

$$\text{Find } \ \dot{u} \in H^1(\Omega)/\mathbf{R} \text{ such that } a(\dot{u}, \dot{v}) = \int_\Omega fv + \langle g, v\rangle, \quad \forall\ \dot{v} \in H^1(\Omega)/\mathbf{R}. \tag{5.8}$$

Here $\langle , \rangle$ denotes the duality between $H^{-\frac{1}{2}}(\partial\Omega)$ and $H^{\frac{1}{2}}(\partial\Omega)$. We need to find out under which conditions the right-hand side in (5.8) is independent of v in $\dot{v}$. Let then

$$F(v) = \int_\Omega fv + \langle g, v\rangle.$$

For $v, w \in \dot{v}$, $v - w \in \mathbf{R}$ and

$$F(v) - F(w) = \int_\Omega f(v - w) + \langle g, v - w\rangle = (v - w)\Big(\int_\Omega f + \langle g, 1\rangle\Big).$$

We therefore require

$$\int_\Omega f + \langle g, 1\rangle = 0.$$

Notice that for $g \in L^2(\partial\Omega)$, this reads $\int_\Omega f + \int_{\partial\Omega} g = 0$. On the other hand,

$$a(\dot{u}, \dot{v}) = \int_\Omega \nabla u \cdot \nabla v, \ \forall\ \dot{u}, \dot{v} \in H^1(\Omega)/\mathbf{R}$$

defines an inner product on $H^1(\Omega)/\mathbf{R}$ which makes it a Hilbert space provided Ω is connected. We only need to check that if $a(\dot{u}, \dot{u}) = 0$, then $\dot{u} = 0$.

$$a(\dot{u}, \dot{u}) = \int_\Omega |\nabla u|^2 = 0$$

implies that u is constant i.e. $\dot{u} = 0$. As a consequence

$$a(\dot{u}, \dot{v}) = || \ \dot{u} \ ||_{H^1(\Omega)/\mathbf{R}}$$

is elliptic on $H^1(\Omega)/\mathbf{R}$. We conclude that:

Theorem 5.1.3 *Let Ω be connected with a Lipschitz continuous boundary. For $f \in L^2(\Omega)$, $g \in H^{-\frac{1}{2}}(\partial\Omega)$ satisfying*

$$\int_\Omega f + \langle g, 1\rangle = 0$$

the problem (5.8) has a unique solution in $H^1(\Omega)/\mathbf{R}$.

By the corollary of the Lax-Milgram lemma, $\dot{u}$ is the unique minimizer in $H^1(\Omega)/\mathbf{R}$ of

$$J(\dot{u}) = \frac{1}{2} a(\dot{u}, \dot{u}) - \int_\Omega fu - \langle g, u\rangle.$$

It is useful to notice that algebraically and topologically, $H^1(\Omega)/\mathbf{R}$ and

$$\Big\{u \in H^1(\Omega), \int_\Omega u = 0\Big\}$$

are the same. This can be seen by considering the surjective mapping from $H^1(\Omega)$ to $\{u \in H^1(\Omega), \int_\Omega u = 0\}$, defined by

$$u \mapsto u - \frac{1}{\text{meas}(\Omega)} \int_\Omega u.$$

The kernel of this mapping is $\mathbf{R}$ and the result follows.

5.1.2 Spline approximations of the Dirichlet problem

To approximate the right hand side f in the Poisson equation, we approximate f on each tetrahedron $t = \langle v_1, v_2, v_3, v_4 \rangle$ by the unique polynomial of degree d on t which interpolates f at the domain points ψ_{ijkl}. This gives rise to a global interpolation operator Π^d.

Let nth be the number of tetrahedra and nbf the number of boundary faces. We introduce the boundary interpolation operator Π_b^d which is defined on each boundary face $f = \langle v_1, v_2, v_3 \rangle$, as the B-net of the unique polynomial interpolating any given continuous function at the domain points $\{\frac{iv_1+jv_2+kv_3}{d}, i+j+k = d\}$. There are $n = \binom{d+2}{2}$ such points and therefore

$$\Pi_b^d : C^0(\partial\Omega) \to \mathbf{R}^M \tag{5.9}$$

with $M = n \times nbf$. On the other hand, it can be seen that there's a matrix B such that given the B-net $\mathbf{c}$ of an element u of $S_d^r(\Omega)$

$$B\mathbf{c} = \Pi_b^d(u_{|\partial\Omega}). \tag{5.10}$$

Indeed for a tetrahedron $\langle v_1, v_2, v_3, v_4 \rangle$ with a single boundary face $\langle v_1, v_2, v_3 \rangle$, $\Pi_b^d(u_{|\partial\Omega})$ is a vector formed with the c_{ijk0}, $i+j+k = d$.

We now introduce a finite dimensional subspace S of V.

$$S = \{p \in S_d^r(\Omega), p_{|\partial\Omega} = \Pi_b^d(g)\}.$$

Let's fix the tetrahedral partition $\mathcal{T}$, and let $n = \dim P_d$. If nth is the number of tetrahedra in $\mathcal{T}$, the B-net of an element p of $S_d^r(\Omega)$ has length $N = n * nth$. We denote by p, an approximant of u and $\mathbf{c}$ the B-net of p. We therefore have

$$p_{|t} = \sum_{s=1}^{n} c_s^t B_s^d. \tag{5.11}$$

We seek to minimize $J(v) = \frac{1}{2}\int_\Omega |\nabla v|^2 - \int_\Omega fv$ over $S \subseteq S_d^r(\Omega)$. Let $\mathbf{G}$ encode the B-net of the interpolant of g on the boundary. In terms of B-nets, elements of S, are described by

$$\{\mathbf{c} \in \mathbf{R}^N, H\mathbf{c} = 0 \text{ and } R\mathbf{c} = \mathbf{G}\}. \tag{5.12}$$

We write $\Pi^d(f)_{|t} = \sum_{\alpha=1}^n f_\alpha^t B_\alpha^d$ and represent the B-net of $\Pi^d(f)$ by $\mathbf{F}$. Let $u = \sum_t \sum_{\beta=1}^n c_\gamma^t B_\gamma^d$ be an element of $S_d^r(\Omega)$. We introduce the local mass matrix $M^t = (\int_t B_\alpha^d B_\beta^d)_{\alpha,\beta=1,\ldots,n}$ and denote by M the corresponding global mass matrix. $K^t = (\int_t \nabla B_\alpha^d \nabla B_\beta^d)_{\alpha,\beta=1,\ldots,n}$ is the local stiffness matrix and we denote by K the global stiffness matrix. Arguing as in the continuous case, one shows that J has a unique minimizer in S. In terms of B-nets J is written as

$$\begin{aligned}
J(\mathbf{c}) &= \frac{1}{2}\int_\Omega |\nabla u|^2 - \int_\Omega fu \\
&= \frac{1}{2}\sum_t \int_t |\nabla u|^2 - \sum_t \int_t fu \\
&= \frac{1}{2}\sum_t \sum_{\alpha=1,\beta=1}^n c_\alpha^t c_\beta^t \int_t \nabla B_\alpha^d \nabla B_\beta^d - \sum_t \sum_{\alpha=1,\beta=1}^n f_\alpha^t c_\beta^t \int_t B_\alpha^d B_\beta^d \\
&= \frac{1}{2}\sum_t (c^t)^T K^t c^t - \sum_t (F^t)^T M^t \\
&= \frac{1}{2}(\mathbf{c})^T K\mathbf{c} - (\mathbf{F})^T M.
\end{aligned}$$

So the idea is to minimize J under the constraints in (5.12). This leads to the existence of Lagrange multipliers λ_1 and λ_2 such that

$$\begin{aligned}
K\mathbf{c} + H^T\lambda_1 + R^T\lambda_2 &= M\mathbf{F}, \\
H\mathbf{c} &= 0, \\
R\mathbf{c} &= \mathbf{G}.
\end{aligned}$$

We assemble this into an equation

$$\begin{pmatrix} H^T & R^T & K \\ 0 & 0 & H \\ 0 & 0 & R \end{pmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \mathbf{c} \end{bmatrix} = \begin{bmatrix} M\mathbf{F} \\ 0 \\ \mathbf{G}, \end{bmatrix}$$

where we used the symmetry of K and M. We computed a least squares solution of $Ax - b = 0$ where $A = \begin{pmatrix} H^T & R^T & K \\ 0 & 0 & H \\ 0 & 0 & R \end{pmatrix}$ and $b = \begin{bmatrix} M\mathbf{F} \\ 0 \\ \mathbf{G} \end{bmatrix}$. The way we have arranged the system turns out to be more efficient on MATLAB than the natural way which would follow from the system of equations.

5.1.3 Spline approximations of the Neumann problem

The treatment of the Neumann boundary conditions is similar in ideas to the treatment of the previous section. Equivalently, the problem is to minimize

$$K(u) = \frac{1}{2}\int_\Omega |\nabla u|^2 - \int_\Omega fu - \int_{\partial\Omega} gu$$

over

$$W = \{w \in H^1(\Omega), \int_\Omega w = 0\}.$$

Notice that the Neumann boundary condition is implicitly contained in the variational formulation. We indicate how to approximate the terms $\int_\Omega w = 0$ and $\int_{\partial\Omega} gu$.

On a tetrahedron t, for $p = \sum_{i+j+k+l=d} c_{ijkl} B^d_{ijkl}$, $\int_t p = \sum_{i+j+k+l=d} c_{ijkl}$. One then sums over all tetrahedra to get $\int_\Omega p$ for an approximant p of w. We show explicitly how to compute $\int_f gu$ for the boundary face face $= \langle v_1, v_2, v_3\rangle$ of $t = \langle v_1, v_2, v_3, v_4\rangle$. Although g and $u_{|\partial\Omega}$ are functions of 3 variables, they can be represented as bivariate splines on a face as described in [Lai and Schumaker]. Let then

$$u_{|\partial\Omega} = \sum_{i+j+k=d} c_{ijk} B^d_{ijk}$$

and

$$g = \sum_{i+j+k=d} b_{ijk} B^d_{ijk},$$

with $m = \binom{d+2}{2}$, we can write

$$\begin{aligned} \int_{\text{face}} gu &= \sum_{\alpha=1,\beta=1}^{m} c_\alpha b_\beta \int_{\text{face}} B^d_\alpha B^d_\beta \\ &= B^T M_b C_0, \end{aligned}$$

where B encodes the B-net of g, C_0 the B-net of $u_{|\partial\Omega}$ and $M_b = (\int_{\text{face}} B^d_\alpha B^d_\beta)_{\alpha,\beta=1,\ldots,m}$ and the B^d_α's are bivariate Bernstein polynomials. On the other hand, if C encodes the B-net of u on t, we have $RC = C_0$ for some matrix R. Therefore

$$\int_{\text{face}} gu = B^T M_b R C.$$

5.1.4 Numerical results

We simply required the solution to be globally continuous for the Dirichlet and Neumann problem. We used two different kind of domains and give the errors in the L^∞ norm of the exact solution against the computed solution.

Domain 1: It is formed by the union of two tetrahedra which share a common face.

Domain 2: This is a cube of volume one which has been subdivided into six tetrahedra.

Dirichlet problem

We used three different functions referred here as Case 1, Case 2 and Case 3.

Case 1

$$g = \frac{1}{(1+x+y+z)}$$

Case 2

$$g = \exp(x+y+z)$$

Case 3

$$g = x(1-x)y(1-y)z(1-z)$$

Case 1 on Domain 1

Tetrahedra	d=1	d=2	d=3	d=4
2	1.7843e-01	4.5769e-02	1.3600e-02	4.1600e-03
2*8=16	8.5714e-02	1.3333e-02	2.4341e-03	4.8623e-04
16*8=128	3.3333e-02	2.9091e-03	2.4242e-04	1.2929e-05

Tetrahedra	d=5	d=6
2	1.3840e-03	4.7872e-04
2*8=16	8.8778e-05	1.4836e-05
16*8=128	8.7533e-07	1.0004e-07

Case 1 on Domain 2

Tetrahedra	d=1	d=2	d=3	d=4
6	2.4970e-01	8.8714e-02	3.0960e-02	9.9400e-03
6*8=48	1.2840e-01	2.8725e-02	6.6883e-03	1.4325e-03
48*8=384	5.6960e-02	7.2349e-03	7.1840e-04	7.3832e-05

Tetrahedra	d=5	d=6
6	3.5671e-03	1.1280e-03
6*8=48	2.1880e-04	4.8372e-05

Case 2 on Domain 1

Tetrahedra	d=1	d=2	d=3	d=4
2	1.5133e+00	2.1341e-01	2.6450e-02	3.0118e-03
2*8=16	5.7198e-01	3.9187e-02	2.4344e-03	1.4138e-04
16*8=128	1.7646e-01	5.9955e-03	1.5414e-04	2.1317e-06

Tetrahedra	d=5	d=6
2	2.9148e-04	2.7069e-05
2*8=16	6.2745e-06	2.1893e-07
16*8=128	.0073e-08	9.4241e-10

Case 2 on Domain 2

Tetrahedra	d=1	d=2	d=3	d=4
6	6.4017e+00	1.3922e+00	2.3880e-01	3.2070e-02
6*8=48	2.7623e+00	2.9100e-01	2.5136e-02	1.7554e-03
48*8=384	9.5226e-01	4.9066e-02	1.5654e-03	5.0882e-05

Tetrahedra	d=5	d=6
6	4.0221e-03	3.9298e-04
6*8=48	7.9726e-05	4.6256e-06

Case 3 on Domain 1

Tetrahedra	d=1	d=2	d=3	d=4
2	1.5625e-02	5.9040e-03	4.2236e-03	1.4040e-03
2*8=16	8.7891e-03	2.2748e-03	5.9727e-04	1.8647e-04
16*8=128	3.2971e-03	4.7640e-04	6.0321e-05	3.5633e-06

Tetrahedra	d=5	d=6
2	1.0287e-03	2.6021e-17
2*8=16	1.4766e-05	7.4593e-17
16*8=128	1.3051e-07	2.2465e-16

Case 3 on Domain 2

Tetrahedra	d=1	d=2	d=3	d=4
6	1.5625e-02	7.1057e-03	5.3931e-03	1.2742e-03
6*8=48	9.8141e-03	2.0937e-03	5.6773e-04	1.2085e-04
48*8=384	4.4806e-03	4.7148e-04	4.6600e-05	4.6553e-06

Tetrahedra Size	d=5	d=6
6	6.4223e-04	1.5451e-16
6*8=48	9.5319e-06	3.5388e-16

Neumann problem

We used three different functions for the Neumann problem referred here as Case 1, Case 2 and Case 3.

Case 1

$$g = x(1-x)y(1-y)z(1-z)$$

Case 2

$$g = \frac{1}{(1+x+y+z)}$$

Case 3

$$g = \exp(x+y+z)$$

Case 1 on Domain 1

Tetrahedra	2	16	128
d=2	6.0156e-03	2.9306e-03	6.7992e-04
d=3	8.9427e-03	7.9836e-04	6.8210e-05
d=4	1.7623e-03	1.9041e-04	8.0530e-06
d=5	8.5063e-04	1.3493e-05	2.1078e-07

Case 1 on Domain 2

Tetrahedra	6	48
d=2	6.9085e-03	4.3158e-03
d=3	1.0547e-02	1.0358e-03
d=4	3.5039e-03	3.1497e-04
d= 5	1.4400e-03	2.3231e-05

Case 2 on Domain 1

Tetrahedra	2	16	128
d=2	8.5733e-02	2.4247e-02	5.2087e-03
d=3	2.5626e-02	4.6303e-03	5.9077e-04
d=4	7.6899e-03	8.5796e-04	5.7704e-05
d=5	2.0860e-03	1.5101e-04	5.5329e-06

Case 2 on Domain 2

Tetrahedra	6	48
d=2	2.9245e-01	5.9019e-02
d=3	6.6962e-02	1.4796e-02
d=4	3.6261e-02	3.6754e-03
d= 5	8.6099e-03	8.7173e-04

Case 3 on Domain 1

Tetrahedra	2	16	128
d=2	6.7937e-01	8.6980e-02	1.1454e-02
d=3	4.1252e-02	2.9073e-03	2.3591e-04
d=4	8.5352e-03	2.6622e-04	8.5578e-06
d=5	3.9107e-04	5.5686e-06	1.1284e-07

Case 3 on Domain 2

Tetrahedra	6	48
d=2	3.6439e+00	6.1377e-01
d=3	3.6204e-01	3.6476e-02
d=4	1.0515e-01	4.1420e-03
d= 5	7.1862e-03	1.7644e-04

We have presented numerical evidence that our scheme is convergent. We now continue with the biharmonic equation.

5.2 The Case of the biharmonic Equation

Let Ω be an open bounded subset of $\mathbf{R}^3$ with a Lipschitz continuous boundary $\partial\Omega$. We will study the boundary value problem

$$\begin{cases} \Delta^2 u &= f \text{ in } \Omega \\ u &= g \text{ on } \partial\Omega \\ \frac{\partial u}{\partial \mathbf{n}} &= h \text{ in } \partial\Omega, \end{cases}$$

where $u : \overline{\Omega} \to \mathbf{R}$ is the unknown and $f : \Omega \to \mathbf{R}$, are given with h and g defined on $\partial\Omega$ with values in $\mathbf{R}$ satisfying the following compatibility condition: There exists $u_0 : \overline{\Omega} \to \mathbf{R}$ such that

$$\begin{aligned} u_0 &= g \quad \text{on } \partial\Omega, \\ \frac{\partial u_0}{\partial \mathbf{n}} &= h. \end{aligned}$$

We have

$$\Delta^2 u = \frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4} + \frac{\partial^4 u}{\partial z^4} + 2\frac{\partial^4 u}{\partial x^2 \partial y^2} + 2\frac{\partial^4 u}{\partial x^2 \partial z^2} + 2\frac{\partial^4 u}{\partial y^2 \partial z^2}.$$

Let's first assume that the solution u is smooth and multiply the PDE by a test function $v \in \mathcal{D}(\Omega)$. We get after integrating over Ω,

$$\int_\Omega \Delta^2 u \; v = \int_\Omega f \; v.$$

By Green's formula, we have

$$\begin{aligned} \int_\Omega \Delta^2 u \; v &= \int_\Omega \Delta(\Delta u) \; v \\ &= -\int_\Omega \nabla(\Delta u) \cdot \nabla v + \int_{\partial\Omega} \frac{\partial \Delta u}{\partial \mathbf{n}} \; v \\ &= -\int_\Omega \nabla(\Delta u) \cdot \nabla v \text{ as } v = 0 \text{ on } \partial\Omega \\ &= \int_\Omega \Delta u \; \Delta v \mathrm{d}x + \int_{\partial\Omega} \frac{\partial v}{\partial \mathbf{n}} \; \Delta u. \end{aligned}$$

since v has compact support in Ω, $\frac{\partial v}{\partial \mathbf{n}} = 0$ in a neighborhood of $\partial\Omega$. So

$$\int_\Omega \Delta^2 u = \int_\Omega \Delta u \ \Delta v.$$

We therefore have

$$\int_\Omega \Delta u \ \Delta v \mathrm{d}x = \int_\Omega f \ v, \ \forall v \in \mathcal{D}(\Omega). \tag{5.13}$$

Recall that $H_0^2(\Omega)$ is the closure of $\mathcal{D}(\Omega)$ in $H^2(\Omega)$ and

$$H_0^2(\Omega) = \{u \in H^2(\Omega), u = 0 \text{ on } \partial\Omega, \frac{\partial u}{\partial \mathbf{n}} = 0 \text{ on } \partial\Omega\},$$

where $u = 0$ on $\partial\Omega$ and $\frac{\partial u}{\partial \mathbf{n}} = 0$ on $\partial\Omega$ are to be understood in the trace sense.

By approximation, (5.13) holds for all v in $H_0^2(\Omega)$. We notice that $u - u_0$ satisfies homogeneous boundary values, i.e.

$$u - u_0 = 0 \text{ on } \partial\Omega \text{ and}$$
$$\frac{\partial(u - u_0)}{\partial \mathbf{n}} = 0 \text{ on } \partial\Omega$$

On the other hand,

$$\begin{aligned}\int_\Omega \Delta(u - u_0)\Delta v &= \int_\Omega \Delta u \Delta v - \int_\Omega \Delta u_0 \Delta v \\ &= \int_\Omega f v - \Delta u_0 v.\end{aligned}$$

We introduce

$$a(u, v) = \int_\Omega \Delta u \ \Delta v, \ \forall u, v \in H_0^2(\Omega),$$

and assume $f \in H^{-2}(\Omega)$, u_0 in $H^2(\Omega)$. We denote

$$\langle f, v \rangle = \int_\Omega f \ v. \tag{5.14}$$

We are lead to consider the problem:

$$(P) \qquad \text{Find } w \in H_0^2(\Omega)$$
$$a(w, v) = \langle f, v \rangle - a(u_0, v), \ \forall v \in H_0^2(\Omega).$$

Then $u = w + u_0$ solves the problem. Notice that we have reduced the problem to finding the solution in $H_0^2(\Omega)$. In view of the Lax-Milgram lemma, we need to show that a is continuous and elliptic on $H_0^2(\Omega)$. We have

$$|a(u,v)| = |\int_\Omega \Delta u \ \Delta v| \leq ||\Delta u||_{L^2(\Omega)} ||\Delta v||_{L^2(\Omega)}$$

and

$$\begin{aligned} ||\Delta u||_{L^2(\Omega)} &\leq ||\frac{\partial^2 u}{\partial x^2}||_{L^2(\Omega)} + ||\frac{\partial^2 u}{\partial y^2}||_{L^2(\Omega)} + ||\frac{\partial^2 u}{\partial z^2}||_{L^2(\Omega)} \\ &\leq ||u||_{H_0^2(\Omega)}, \end{aligned}$$

so

$$|a(u,v)| \leq ||u||_{H_0^2(\Omega)} ||v||_{H_0^2(\Omega)}.$$

On the other hand, for $v \in \mathcal{D}(\Omega)$,

$$\begin{aligned} ||\Delta v||^2_{L^2(\Omega)} &= \int_\Omega |\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}|^2 \\ &= \int_\Omega (\frac{\partial^2 u}{\partial x^2})^2 + (\frac{\partial^2 u}{\partial y^2})^2 + (\frac{\partial^2 u}{\partial z^2})^2 \\ &\quad + 2\frac{\partial^2 u}{\partial x^2}\frac{\partial^2 u}{\partial y^2} + 2\frac{\partial^2 u}{\partial y^2}\frac{\partial^2 u}{\partial z^2} + 2\frac{\partial^2 u}{\partial x^2}\frac{\partial^2 u}{\partial z^2}. \end{aligned} \tag{5.15}$$

Using Green's formula and interchanging derivatives, we have

$$\begin{aligned} \int_\Omega \frac{\partial^2 u}{\partial x^2}\frac{\partial^2 u}{\partial y^2} &= -\int_\Omega \frac{\partial u}{\partial x}\frac{\partial^3 u}{\partial x \partial y^2} \\ &= \int_\Omega \frac{\partial^2 u}{\partial x \partial y}\frac{\partial^2 u}{\partial x \partial y} \end{aligned}$$

and similar treatments for the terms with mixed derivatives in (5.15). This gives

$$||\Delta v||^2_{L^2(\Omega)} = |v|^2_{2,\Omega}, \ \forall v \in \mathcal{D}(\Omega).$$

By density, this is true for all $v \in H_0^2(\Omega)$. Since $||v||_{2,\Omega}$ and $|v|_{2,\Omega}$ are two equivalent norms on $H_0^2(\Omega)$, there's a constant $M > 0$ such that

$$a(v,v) = ||\Delta v||^2_{L^2(\Omega)} \geq M||v||^2_{2,\Omega}.$$

We conclude that the problem (P) has a unique solution $w_0 = u - u_0$ in $H^2(\Omega)$. By the corollary (4.2.2) of the Lax-Milgram lemma, w_0 minimizes the functional

$$J(w) = \frac{1}{2}a(w,w) - \langle f, w\rangle + a(u_0, w)$$

over $H_0^2(\Omega)$. We show that u minimizes the functional

$$L(v) = \frac{1}{2}a(v,v) - \langle f, v\rangle$$

over $V = \{v \in H^2(\Omega), v - u_0 \in H_0^2(\Omega)\}$. This type of argument has been fully explained when we dealt with the Dirichlet problem for the Poisson equation.

For $v \in V$, $v = u_0 + w$ for some $w \in H_0^2(\Omega)$. We have

$$L(v) = L(u_0 + w) = J(w) + L(u_0),$$

which shows that the minimum of L is reached for $u_0 + w_0 = u$.

5.2.1 Spline Approximations of the biharmonic Equation

We seek to replace the problem of minimizing $L(v)$ over V by a minimization problem over a subset of V. Recall that

$$S_d^r(\Omega) = \{p \in C^r(\Omega), p_{|t} \in P_d,\ \forall t \in \Delta\},$$

where Δ is a tetrahedral decomposition of Ω.

For $r \geq 1$, we have

$$S_d^r(\Omega) \subseteq S_d^1(\Omega) \subset H^2(\Omega).$$

Next, we want to construct a subset of $S_d^r(\Omega)$ whose elements satisfy the boundary conditions $u = g$ and $\frac{\partial u}{\partial \mathbf{n}} = h$ on $\partial\Omega$. If a tetrahedral partition is fixed, let nth be the number of tetrahedra and nbf the number of boundary faces. Recall that the boundary interpolation operator Π_b^d is defined as follows:

$$\Pi_b^d : C^0(\partial\Omega) \to \mathbf{R}^{nbf*dbf},$$

with $dbf = \binom{d+2}{2}$.

On the other hand, there's a matrix B such that given the B-net $\mathbf{c}$ of an element u of $S_d^r(\Omega)$

$$B\mathbf{c} = \Pi_b^d(u_{|\partial\Omega}).$$

To satisfy the boundary condition $\frac{\partial u}{\partial \mathbf{n}} = h$, we look at the $c_{ijk1}, i+j+k = d-1$. If $\mathbf{n}$ is the unit normal to the boundary face $face = \langle v_1, v_2, v_3\rangle$ with T-coordinates (a_1, a_2, a_3, a_4) with respect to $\langle v_1, v_2, v_3, v_4\rangle$ and if

$$p = \sum_{i+j+k+l=d} c_{ijkl} B_{ijkl}^d$$

on $\langle v_1, v_2, v_3, v_4\rangle$ then

$$\frac{\partial p}{\partial \mathbf{n}} = d \sum_{i+j+k+l=d-1} c_{ijkl}^{(1)} B_{ijkl}^{d-1},$$

with $c_{ijkl}^{(1)}(a_1, a_2, a_3, a_4) = a_1 c_{i+1,j,k,l} + a_2 c_{i,j+1,k,l} + a_3 c_{i,j,k+1,l} + a_4 c_{i,j,k,l+1}$.

To simplify the discussion, let's consider only one boundary face $face = \langle v_1, v_2, v_3\rangle$. If we write $\Pi_b^{d-1}(\frac{\partial u_0}{\partial \mathbf{n}}) = \{d_{ijk}, i+j+k = d-1\}$ then

$$d_{ijk} = d c_{ijk0}^{(1)},$$

so

$$\frac{d_{ijk}}{d} = a_1 c_{i+1,j,k,0} + a_2 c_{i,j+1,k,0} + a_3 c_{i,j,k+1,0} + a_4 c_{i,j,k,1}.$$

Now $a_4 \neq 0$ otherwise $\mathbf{n}$ is in the face (v_1, v_2, v_3) and the c_{ijk0} can be computed from $\Pi_b^d(u_0)$ so

$$c_{ijk1} = \frac{d_{ijk}}{d} - a_1 c_{i+1,j,k,0} - a_2 c_{i,j+1,k,0} - a_3 c_{i,j,k+1,0}.$$

The $c_{ijk1}, i+j+k = d-1$ can be associated to a layer of coefficients along the boundary face $face = \langle v_1, v_2, v_3\rangle$. This process can be repeated along each boundary

face defining a layer of coefficients along the boundary $L(u_0, \frac{\partial u_0}{\partial \mathbf{n}})$. Therefore there's a matrix N such that

$$N\mathbf{c} = L(u_0, \frac{\partial u_0}{\partial \mathbf{n}}).$$

Identifying a spline with its B-net, with n=dimP_d, we approximate V by

$$S = \{\mathbf{c} \in \mathbf{R}^{n*nth},\ H\mathbf{c} = 0,\ B\mathbf{c} = \Pi_b^d(u_{|\partial\Omega}), N\mathbf{c} = L(u_0, \frac{\partial u_0}{\partial \mathbf{n}})\}.$$

For simplicity, we write

$$\Pi_b^d(u_{|\partial\Omega}) = G_1,$$
$$L(u_0, \frac{\partial u_0}{\partial \mathbf{n}}) = G_2.$$

We now proceed to give the expression of the functional L on elements of S. Identifying again a spline with its B-net, let $u = \sum_t \sum_{\gamma=1}^n c_\gamma^t B_\gamma^d$ be an element of $S_d^r(\Omega)$ and $\Pi(f) = \sum_t \sum_{\alpha=1}^n f_\alpha^t B_\alpha^d$ where Π is the global interpolation operator. We represent the B-net of $\Pi(f)$ by $\mathbf{F}$. We also introduce the local mass matrix $M^t = (\int_t B_\alpha^d B_\beta^d)_{\alpha,\beta=1,\ldots,n}$ and denote by M the corresponding global mass matrix. $K^t = (\int_t \Delta B_\alpha^d \Delta B_\beta^d)_{\alpha,\beta=1,\ldots,n}$ is the local bending matrix and we denote by K the global bending matrix. L is written as

$$\begin{aligned}
L(\mathbf{c}) &= \frac{1}{2}\int_\Omega |\Delta u|^2 - \int_\Omega fu \\
&= \frac{1}{2}\sum_t \int_t |\Delta u|^2 - \sum_t \int_t fu \\
&= \frac{1}{2}\sum_t \sum_{\alpha=1,\beta=1}^n c_\alpha^t c_\beta^t \int_t \Delta B_\alpha^d \Delta B_\beta^d - \sum_t \sum_{\alpha=1,\beta=1}^n f_\alpha^t c_\beta^t \int_t B_\alpha^d B_\beta^d \\
&= \frac{1}{2}\sum_t (c^t)^T K^t c^t - \sum_t (F^t)^T M^t \\
&= \frac{1}{2}(\mathbf{c})^T K\mathbf{c} - (\mathbf{F})^T M.
\end{aligned}$$

We seek to minimize L on $\mathbf{R}^{n*nth}$ under the constraints $H\mathbf{c} = 0$, $B\mathbf{c} = G_1$ and $N\mathbf{c} = G_2$. This leads to the existence of Lagrange multipliers λ_1, λ_2, and λ_3 such

that

$$K\mathbf{c} + H^T\lambda_1 + B^T\lambda_2 + N^T\lambda_3 = M\mathbf{F},$$
$$H\mathbf{c} = 0,$$
$$B\mathbf{c} = \mathbf{G_1},$$
$$N\mathbf{c} = \mathbf{G_2}.$$

This can be assembled into an equation

$$\begin{pmatrix} H^T & B^T & N^T & K \\ 0 & 0 & 0 & H \\ 0 & 0 & 0 & B \\ 0 & 0 & 0 & N \end{pmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \mathbf{c} \end{bmatrix} = \begin{bmatrix} M\mathbf{F} \\ 0 \\ \mathbf{G_1} \\ \mathbf{G_2} \end{bmatrix},$$

where we used the symmetry of K and M. The matrix $A = \begin{pmatrix} H^T & B^T & N^T & K \\ 0 & 0 & 0 & H \\ 0 & 0 & 0 & B \\ 0 & 0 & 0 & N \end{pmatrix}$ is in general singular so the solution may not be unique. We compute a least squares solution of $Ax - b = 0$, where $b = \begin{bmatrix} M\mathbf{F} \\ 0 \\ \mathbf{G_1} \\ \mathbf{G_2} \end{bmatrix}$. When $||Ax - b||$ is sufficiently close to 0, we conclude that the least squares solution is sufficiently close to the unique solution of the problem.

5.2.2 Numerical results

These are numerical solutions for the 3D biharmonic equation which are continuously differentiable across tetrahedral elements which do not share a face with the

boundary. The domains we considered are

Domain 1: It is formed by the union of two tetrahedra which share a common face.

Domain 2: This is a cube of volume one which has been subdivided into six tetrahedra.

Domain 3: The domain here is a cube of volume one subdivided into twelve tetrahedra.

We used three different functions referred here as Case 1, Case 2 and Case 3.

Case 1

$$g = \exp(-x^2 - y^2 - z^2)$$

Case 2

$$g = \frac{1}{(1 + x + y + z)}$$

Case 3

$$g = x(1 - x)y(1 - y)z(1 - z)$$

Case 1 on Domain 1

Tetrahedra	2	16	128
d=2	1.0670e-01	2.0087e-02	5.3872e-003
d=3	4.7370e-02	4.3863e-03	4.6657e-004
d=4	7.2332e-03	9.5736e-04	1.9150e-004
d=5	2.5899e-03	2.3387e-04	Out of memory

Case 1 on Domain 2

Tetrahedra	6	48
d=2	1.5571e-01	5.9921e-02
d=3	6.4337e-02	9.1870e-03
d=4	1.0803e-02	5.9974e-03
d=5	1.2830e-02	Out of memory

Case 1 on Domain 3

Tetrahedra	12	96
d=2	9.4943e-02	1.5165e-02
d=3	1.3830e-02	3.9260e-03

Case 2 on Domain 1

Tetrahedra	2	16	128
d=2	6.2603e-02	2.5915e-02	5.4551e-003
d=3	3.0805e-02	4.2892e-03	5.4194e-004
d=4	7.2832e-03	8.4560e-04	2.5875e-004
d=5	2.5425e-03	1.6797e-04	Out of Memory

Case 2 on Domain 2

Tetrahedra	6	48
d=2	1.2833e-01	8.2672e-02
d=3	1.0845e-01	1.5910e-02
d=4	2.3205e-02	6.2875e-03
d=5	2.1838e-02	Out of memory

Case 2 on Domain 3

Tetrahedra	12	96
d=2	1.1648e-01	3.4038e-02
d=3	2.0580e-02	3.4276e-03

Case 3 on Domain 1

Tetrahedra	2	16	128
d=2	9.0998e-03	5.1188e-03	1.2035e-003
d=3	9.4522e-03	1.3539e-03	2.0061e-004
d=4	3.1107e-03	3.1121e-04	1.6199e-004
d=5	1.8457e-03	1.4199e-04	Out of memory

Case 3 on Domain 2

Tetrahedra	6	48
d=2	1.5625e-02	7.3145e-03
d=3	1.3468e-02	1.9666e-03
d=4	5.1264e-03	1.6385e-03
d=5	8.4341e-03	Out of memory

Case 3 on Domain 3

Tetrahedra	12	96
d=2	1.5625e-02	5.0647e-03
d=3	5.1140e-03	1.5260e-03

Chapter 6

The Navier-Stokes Equations

In this chapter we derive the Navier-Stokes equations from physical considerations, then we consider approximations of the Stokes equations and the Navier-Stokes equations.

6.1 Derivation of The Equations

We give a heuristic derivation of the Navier-Stokes equations describing the motion of an incompressible viscous Newtonian fluid in $\mathbf{R}^3$. The derivation is based on [Doering and Gibbon'95]. We'll use three considerations: material properties, Newton's second law and conservation of mass.

The viscosity of the fluid describes its tendency to resist shearing motions. In mechanics, shearing forces are described by the stress tensor $\mathbf{S}$. The component S_{ij} of the stress tensor is the force per unit area in the jth direction acting across an area element whose normal is in the ith direction. Forces in the direction of the normal to an area element, (S_{ii}), are associated to the pressure and those that act in the plane of the element are associated with shear stresses. We have the following decomposition of the stress tensor into portions due to the pressure P and the shear stress tensor T_{ij},

$$S_{ij} = -\delta_{ij} P + T_{ij}. \tag{6.1}$$

Since the fluid is Newtonian, the shear stress tensor $\mathbf{T}$ is a linear function of the rate of strain tensor $U_{ij} = \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}$, i.e.

$$\mathbf{T} = \alpha \mathbf{U}, \tag{6.2}$$

where α is a material parameter, the viscosity of the fluid. We are using here $\frac{\partial}{\partial x_i}, i = 1, 2, 3$ to represent respectively $\frac{\partial}{\partial x}, \frac{\partial}{\partial y}$ and $\frac{\partial}{\partial z}$. The rate of the strain tensor characterizes the deviation of the fluid motion from a rigid body motion.

The dependent variables in the description of the motion of the fluid are the density of the fluid $\rho(\mathbf{x}, t)$, the velocity vector field $\mathbf{u}(\mathbf{x}, t)$, and the pressure $P(\mathbf{x}, t)$ for $\mathbf{x} = (x, y, z)$ in $\mathbf{R}^3$.

We consider an infinitesimal element of the fluid of volume δV and mass δm located at position $\mathbf{x}$ at time t moving with the velocity $\mathbf{u}(\mathbf{x}, t)$. The forces that act on the fluid depend on the direction of the force and the orientation of the area across which the forces act. So if we consider a rectangular shaped portion of fluid centered at $\mathbf{x}$ with side lengths $(\delta x, \delta y, \delta z)$, the net force on the fluid in the jth direction is

$$\begin{aligned}
\delta F_j = &(S_{1j}(x + \frac{\delta x}{2}, y, z) - S_{1j}(x - \frac{\delta x}{2}, y, z))\delta y \delta z \\
&+ (S_{2j}(x, y + \frac{\delta y}{2}, z) - S_{2j}(x, y - \frac{\delta y}{2}, z))\delta x \delta z \\
&+ (S_{3j}(x, y, z + \frac{\delta z}{2}) - S_{3j}(x, y, z - \frac{\delta z}{2}))\delta x \delta z.
\end{aligned}$$

Using a Taylor expansion around $\mathbf{x}$ and keeping only the first order terms, we get

$$\delta F_j = (\frac{\partial S_{1j}(x, y, z)}{\partial x} + \frac{\partial S_{2j}(x, y, z)}{\partial y} + \frac{\partial S_{3j}(x, y, z)}{\partial z})\ \delta x \delta y \delta z.$$

Hence,

$$\delta \mathbf{F} = \nabla \cdot \mathbf{S}\ \delta V, \tag{6.3}$$

where for a tensor S_{ij}, $\nabla \cdot S$ is the vector of components $(\frac{\partial S_{1j}}{\partial x} + \frac{\partial S_{2j}}{\partial y} + \frac{\partial S_{3j}}{\partial z})$. Using (6.1) and (6.2), we have

$$\nabla \cdot \mathbf{S} = -\nabla P + \alpha \nabla \cdot \mathbf{U}. \tag{6.4}$$

Letting $\mathbf{f}$ encode the external body forces per unit mass e.g. gravity and $\delta\mathbf{F}$ encode the internal forces, Newton's second law for the element of fluid mass δm at position δx is

$$\frac{\mathrm{d}}{\mathrm{d}t}(\delta m\ \mathbf{u}(\mathbf{x}, t)) = \delta\mathbf{F} + \delta m\mathbf{f}. \tag{6.5}$$

Here $\frac{\mathrm{d}}{\mathrm{d}t}$ refers to the convective derivative defined as the rate of change with respect to an observer moving with the fluid. To be precise, for a function $f(\mathbf{x}, \mathbf{t})$,

$$\left(\frac{\mathrm{d}f(\mathbf{x}, t)}{\mathrm{d}t}\right)_{\text{fixed position}} = \lim_{\delta \to 0} \frac{f(\mathbf{x}, t + \delta t) - f(\mathbf{x}, t)}{\delta t} = \frac{\partial f(\mathbf{x}, t)}{\partial t}$$

and

$$\begin{aligned}\left(\frac{\mathrm{d}f(\mathbf{x}, t)}{\mathrm{d}t}\right)_{\text{moving}} &= \lim_{\delta \to 0} \frac{f(\mathbf{x} + \mathbf{u}\delta t, t + \delta t) - f(\mathbf{x}, t)}{\delta t} \\ &= \frac{\partial f(\mathbf{x}, t)}{\partial t} + \mathbf{u} \cdot \nabla f(\mathbf{x}, t).\end{aligned}$$

(6.5) gives

$$\frac{\mathrm{d}\delta m}{\mathrm{d}t}\mathbf{u}(\mathbf{x}, t) + \delta m \frac{\mathrm{d}\mathbf{u}(\mathbf{x}, t)}{\mathrm{d}t} = \delta\mathbf{F} + \delta m\mathbf{f}. \tag{6.6}$$

We now use the consideration that mass is conserved:

$$\frac{\mathrm{d}\delta m}{\mathrm{d}t} = 0.$$

Therefore (6.6) implies

$$\delta m(\frac{\partial \mathbf{u}(\mathbf{x}, t)}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u}(\mathbf{x}, t)) = \delta\mathbf{F} + \delta m\mathbf{f} = \nabla \cdot \mathbf{S}\ \delta V + \delta m\mathbf{f},$$

so

$$\begin{aligned}\frac{\partial \mathbf{u}(\mathbf{x},t)}{\partial t} + \mathbf{u}\cdot\nabla\mathbf{u}(\mathbf{x},t) &= \frac{1}{\rho(\mathbf{x},t)}\nabla\cdot\mathbf{S} + \mathbf{f} \\ &= \frac{1}{\rho(\mathbf{x},t)}(-\nabla P + \alpha\nabla\cdot\mathbf{U}) + \mathbf{f},\end{aligned} \tag{6.7}$$

where $\rho(\mathbf{x},t) = \frac{\delta m}{\delta V}$ is the density of the fluid. As the fluid is incompressible its volume does not change under motion. Therefore since the mass is conserved so is the density. Therefore $\frac{\mathrm{d}\rho}{\mathrm{d}t} = 0$ and $\rho(\mathbf{x},t)$ =constant. We write $\rho(\mathbf{x},t) = \rho$. On the other hand, as mass is conserved

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\frac{\delta m}{\delta V} = -\frac{\delta m}{(\delta V)^2}\frac{\mathrm{d}\delta V}{\mathrm{d}t} \tag{6.8}$$

and

$$\delta V = \delta x\delta y\delta z,$$

so

$$\frac{\mathrm{d}\delta V}{\mathrm{d}t} = \frac{\mathrm{d}\delta x}{\mathrm{d}t}\delta y\delta z + \delta x\frac{\mathrm{d}\delta y}{\mathrm{d}t}\delta z + \delta x\delta y\frac{\mathrm{d}\delta z}{\mathrm{d}t}.$$

For an observer moving with the fluid, the length elements $\delta x, \delta y$ and δz increase or decrease according to the relative velocity of their endpoints. So

$$\frac{\mathrm{d}\delta x}{\mathrm{d}t} = u_1(x+\frac{\delta x}{2},y,z,t) - u_1(x-\frac{\delta x}{2},y,z,t) = \frac{\partial u_1}{\partial x}\delta x$$

and likewise for the other components. We then get

$$\frac{\mathrm{d}\delta V}{\mathrm{d}t} = (\nabla\cdot\mathbf{u})\delta V,$$

so using (6.8)

$$\frac{\mathrm{d}\rho}{\mathrm{d}t} = -\frac{\delta m}{\delta V}\nabla\cdot\mathbf{u}.$$

Since $\frac{\mathrm{d}\rho}{\mathrm{d}t} = 0$ we see that mathematically the condition of incompressibility is simply

$$\nabla\cdot\mathbf{u} = 0.$$

A consequence of the incompressibility condition is a simplification of the term $\nabla \cdot \mathbf{U}$, the divergence of the rate of strain tensor; $\nabla \cdot \mathbf{U}$ is the vector of components

$$\begin{aligned}\sum_{j=1}^{3} \frac{\partial}{\partial x_j}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) &= \sum_{j=1}^{3} \frac{\partial^2 u_i}{\partial x_j^2} + \sum_{j=1}^{3} \frac{\partial^2 u_i}{\partial x_j \partial x_i} \\ &= \Delta u_i + \frac{\partial}{\partial x_i}(\nabla . \mathbf{u}) \\ &= \Delta u_i .\end{aligned}$$

We now rewrite (6.7) as

$$\frac{\partial \mathbf{u}(\mathbf{x}, t)}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u}(\mathbf{x}, t) = -\frac{\nabla P}{\rho} + \frac{\alpha}{\rho} \Delta \mathbf{u} + \mathbf{f} \tag{6.9}$$

We set $p = \frac{P}{\rho}$ and $\nu = \frac{\alpha}{\rho}$. Here, p is the kinematic pressure and ν the kinematic viscosity which will be called in the sequel for simplicity pressure and viscosity respectively. The Navier-Stokes equations are

$$\begin{cases} \frac{\partial \mathbf{u}(\mathbf{x},t)}{\partial t} - \nu \Delta \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u}(\mathbf{x}, t) + \nabla p &= \mathbf{f} \\ \nabla \cdot \mathbf{u} &= 0. \end{cases}$$

We shall consider in this dissertation two simplifications of the equations. We'll deal with the stationary case

$$\frac{\partial \mathbf{u}}{\partial t} = 0$$

and we shall investigate the situation where the velocity is sufficiently small to ignore the nonlinear term $\mathbf{u} \cdot \nabla \mathbf{u}(\mathbf{x}, t)$, the so-called Stokes equations.

6.2 Spline Approximations of the Stokes Equations

We consider numerical approximations of the 3D Stokes equations in velocity-pressure formulation. The pressure is eliminated from the equations by using a set of velocity fields which are divergence free. The later is discretized by means of splines of arbitrary degree and arbitrary smoothness. We then minimized the energy

functional associated with the variational problem over the set of splines to get the velocity vector. The pressure term is computed by solving a Poisson problem with Neumann boundary conditions.

6.2.1 Existence and uniqueness of solution

The Stokes equations are a linearized version of the Navier-Stokes equations. For an incompressible viscous fluid in a bounded domain Ω of $\mathbf{R}^3$, the Stokes equations are

$$\begin{cases} -\nu\Delta\mathbf{u} + \nabla p &= \mathbf{f} \text{ in } \Omega \\ \text{div } \mathbf{u} &= 0 \text{ in } \Omega \\ \mathbf{u} &= \mathbf{g} \text{ on } \partial\Omega. \end{cases} \tag{6.10}$$

The unknowns here are the velocity $\mathbf{u} = (u_1, u_2, u_3)^T$ of the fluid and the pressure p; ν is the kinematic viscosity, $\mathbf{f} = (f_1, f_2, f_3)$ represents the externally applied forces (e.g. gravity) and $\mathbf{g} = (g_1, g_2, g_3)$ the velocity at the boundary. We immediately notice using the divergence theorem that

$$0 = \int_\Omega \text{div } \mathbf{u} = \int_{\partial\Omega} \mathbf{g} \cdot \mathbf{n}.$$

$\mathbf{g}$ must therefore satisfy the compatibility condition $\int_{\partial\Omega} \mathbf{g} \cdot \mathbf{n} = 0$.

Let V_0 be the closure in $H_0^1(\Omega)^3$ of

$$\{\mathbf{v} \in \mathcal{D}(\Omega)^3, \text{such that div } \mathbf{v} = 0\}.$$

Since Ω is assumed to have piecewise planar boundary,

$$V_0 = \{\mathbf{v} \in H_0^1(\Omega)^3 \text{ such that div } \mathbf{v} = 0\},$$

[cf. Galdi'94]. If we take the inner product of the first equation in (6.10) with $\mathbf{v} \in \mathcal{D}(\Omega)^3$ satisfying div $\mathbf{v} = 0$ and by a density argument, we get a weak form of the equations

$$\nu \int_\Omega \nabla\mathbf{u} \cdot \nabla\mathbf{v} = \int_\Omega \mathbf{f} \cdot \mathbf{v}, \qquad \forall \mathbf{v} \in V_0$$

since

$$\int_\Omega \nabla p \cdot \mathbf{v} = -\int_\Omega p \text{ div } \mathbf{v} = 0.$$

On the other hand, since the equations involve ∇p, the pressure will be unique up to an additive constant. To have uniqueness, one can prescribe the value of the pressure at a point or require it to have zero mean. We therefore introduce

$$L^2_0(\Omega) = \{p \in L^2(\Omega), \int_\Omega p = 0\}.$$

We have the following existence and uniqueness results, [Girault and Raviart'86]:

Theorem 6.2.1 *Let Ω be a bounded and connected open subset of $\mathbf{R}^3$ with a Lipschitz continuous boundary $\partial\Omega$. For $\mathbf{f} \in H^{-1}(\Omega)^3$ and $\mathbf{g} \in H^{\frac{1}{2}}(\partial\Omega)^3$ satisfying*

$$\int_{\partial\Omega} \mathbf{g} \cdot \mathbf{n} = 0,$$

the problem: Find $(\mathbf{u}, p) \in H^1(\Omega)^3 \times L^2_0(\Omega)$ such that

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

has a unique solution. (Here the first two equations should be interpreted in the sense of distributions and the last one in the trace sense.) Moreover letting

$$V = \{\mathbf{u} \in H^1(\Omega)^3, \mathbf{u} = \mathbf{g} \text{ on } \partial\Omega, \text{ div } \mathbf{u} = 0 \text{ in } \Omega\},$$

the velocity $\mathbf{u}$ is the unique minimizer in V of the functional

$$J(\mathbf{u}) = \frac{\nu}{2}\int_\Omega \nabla\mathbf{u} \cdot \nabla\mathbf{u} - \int_\Omega \mathbf{f} \cdot \mathbf{u}.$$

Proof: Recall that the weak form of the equations is:

$$(P) \quad \text{Find } \mathbf{u} \in V \text{ such that } \nu\int_\Omega \nabla\mathbf{u} \cdot \nabla\mathbf{v} = \int_\Omega \mathbf{f} \cdot \mathbf{v}, \qquad \forall \mathbf{v} \in V_0.$$

Let $a(\mathbf{u},\mathbf{v}) = \nu \int_\Omega \nabla\mathbf{u}\cdot\nabla\mathbf{v}$. The bilinear form a is continuous on V. Since

$$a(\mathbf{v},\mathbf{v}) = \nu\int_\Omega |\nabla\mathbf{v}|^2 = \nu|\mathbf{v}|^2_{(H_0^1(\Omega))^3},$$

where $|\mathbf{v}|^2_{(H_0^1(\Omega))^3} = |v_1|^2 + |v_2|^2 + |v_3|^2$ defines a norm on $H_0^1(\Omega)^3$. By Poincare's inequality, a is V_0 elliptic.

Now, there's $\mathbf{u}_0 \in (H^1(\Omega))^3$ such that div $\mathbf{u}_0 = 0$ in Ω and $\mathbf{u}_0 = \mathbf{g}$ on $\partial\Omega$, [Girault and Raviart'86]. Let $\mathbf{w} = \mathbf{u} - \mathbf{u}_0$. For $\mathbf{u} \in V$, $\mathbf{w} \in V_0$ satisfies

$$a(\mathbf{w},\mathbf{v}) = \int_\Omega \mathbf{f}\cdot\mathbf{v} + a(\mathbf{u}_0,\mathbf{v}), \qquad \forall\mathbf{v}\in V_0.$$

By the Lax-Milgram lemma, such a problem has a unique solution $\mathbf{w}$. As a consequence problem (P) has a unique solution $\mathbf{u} = \mathbf{w} + \mathbf{u}_0$.

Since a is symmetric, by the corollary (4.2.2) of the Lax-Milgram lemma, $\mathbf{w} = \mathbf{u} - \mathbf{u}_0$ is the unique minimizer over V_0 of

$$K(\mathbf{v}) = \frac{1}{2}a(\mathbf{v},\mathbf{v}) - \left(\int_\Omega \mathbf{f}\cdot\mathbf{v} + a(\mathbf{u}_0,\mathbf{v})\right).$$

We show that $\mathbf{u}$ is the unique minimizer in V of

$$J(\mathbf{v}) = \frac{1}{2}a(\mathbf{v},\mathbf{v}) - \int_\Omega \mathbf{f}\cdot\mathbf{v}.$$

Notice that $V = V_0 + \mathbf{u}_0$. For $\mathbf{v}\in V_0$,

$$J(\mathbf{v}+\mathbf{u}_0) = K(\mathbf{v}) + J(\mathbf{u}_0).$$

Since $\mathbf{w}$ minimizes K , $\mathbf{w} + \mathbf{u}_0 = \mathbf{u}$ minimizes J in V. Such a minimizer is unique since it is a solution of problem (P). This completes the proof.

Let $\mathbf{u}\in V$ be a solution of (P). We have

$$\nu\int_\Omega \nabla\mathbf{u}\cdot\nabla\mathbf{v} = \int_\Omega \mathbf{f}\cdot\mathbf{v}, \qquad \forall\mathbf{v}\in\mathcal{D}(\Omega)^3 \text{ such that div } \mathbf{v} = 0.$$

Using Green's formula, this gives

$$\int_\Omega (-\nu\Delta\mathbf{u} - \mathbf{f})\mathbf{v} = 0, \qquad \forall\mathbf{v}\in\mathcal{D}(\Omega)^3 \text{ such that div } \mathbf{v} = 0.$$

This implies the existence of $p \in L^2(\Omega)$ such that

$$-\nu\Delta\mathbf{u} - \mathbf{f} = -\nabla p,$$

as elements of $\mathcal{D}(\Omega)'$, [Girault and Raviart'86]. This shows the existence of $(\mathbf{u}, p)$ satisfying (6.10).

6.2.2 Spline approximations

We first consider approximations of the velocity vector field and then we approximate the pressure using a Neumann equation that couples the pressure and the velocity on the boundary.

Approximation of the velocity

We now construct a finite dimensional approximation of V. We denote by p_i, $1 \leq i \leq 3$ an approximant of u_i and $\mathbf{c_i}$ the B-net of p_i. We therefore have

$$p_{i|t} = \sum_{s=1}^{n} c_{i,s}^t B_s^d, \qquad i = 1, 2, 3. \tag{6.11}$$

c_i^t encodes the B-net of p_i on t and c_i its B-net on Δ. Let

$$S = \{(p_1, p_2, p_3) \ \in S_d^r(\Omega)^3, \ (p_{1|\partial\Omega}, p_{2|\partial\Omega}, p_{3|\partial\Omega}) = (\Pi_b^d(g_1), \Pi_b^d(g_2), \Pi_b^d(g_3)),$$
$$\frac{\partial p_1}{\partial x_1} + \frac{\partial p_2}{\partial x_2} + \frac{\partial p_3}{\partial x_3} = 0 \text{ in each } t \in \Delta\}$$

where Π_b^d is the boundary interpolation operator. Recall that there's a matrix R such that

$$R\mathbf{c_i} = \Pi_b^d(p_{i|\partial\Omega}), \ \forall i = 1, 2, 3.$$

We will denote $\Pi_b^d(g_i)$ by G_i. On the other hand, there are matrices D_i such that $D_i\mathbf{c_i}$ is the B-net of $\frac{\partial p_i}{\partial x_i}$. D_i has size $(m * nth, n * nth)$, with n =dimP_d and m =dimP_{d-1}. For example, let

$$p = \sum_{\alpha+\beta+\gamma+\delta=d} c_{\alpha\beta\gamma\delta} B_{\alpha\beta\gamma\delta}^d$$

be the B-form of a polynomial of degree d on $t = \langle v_1, v_2, v_3, v_4 \rangle$. If (a_1, a_2, a_3, a_4) are the T-coordinates of the unit vector of the x_i axis with respect to t, then

$$\frac{\partial p_i}{\partial x_i} = \sum_{\alpha+\beta+\gamma+\delta=d-1} c^{(1)}_{\alpha\beta\gamma\delta}(a_1, a_2, a_3, a_4) B^{d-1}_{\alpha\beta\gamma\delta}$$

with

$$c^{(1)}_{\alpha\beta\gamma\delta}(a_1, a_2, a_3, a_4) = d(a_1 c_{\alpha+1,\beta,\gamma,\delta} + a_2 c_{\alpha,\beta+1,\gamma,\delta} + a_3 c_{\alpha,\beta,\gamma+1,\delta} + a_4 c_{\alpha,\beta,\gamma,\delta+1}).$$

For $d = 2$, the B-net of p is

$$(c_{2000}, c_{1100}, c_{0200}, c_{1010}, c_{0110}, c_{0020}, c_{1001}, c_{0101}, c_{0011}, c_{0002})^T$$

and the one of $\frac{\partial p_i}{\partial x_i}$ is

$$(c^{(1)}_{1000}, c^{(1)}_{0100}, c^{(1)}_{0010}, c^{(1)}_{0001})^T$$

and D_i has form

$$\begin{pmatrix} a_1 & a_2 & 0 & a_3 & 0 & 0 & a_4 & 0 & 0 & 0 \\ 0 & a_1 & a_2 & 0 & a_3 & 0 & 0 & a_4 & 0 & 0 \\ 0 & 0 & 0 & a_1 & a_2 & a_3 & 0 & 0 & a_4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & a_1 & a_2 & a_3 & a_4 \end{pmatrix}.$$

With a fixed triangulation of nth tetrahedra, identifying a spline with its B-net, S can be described as

$$S = \{(c_1, c_2, c_3) \in (\mathbf{R}^{n*nth})^3, Hc_i = 0, i = 1, 2, 3, Rc_i = G_i, i = 1, 2, 3, D_1 c_1 + D_2 c_2 + D_3 c_3 = 0\}.$$

We now show that J still has a unique minimizer in S. Let

$$S_0 = \{(p_1, p_2, p_3) \in S^r_d(\Omega)^3, \ (p_{1|\partial\Omega}, p_{2|\partial\Omega}, p_{3|\partial\Omega}) = (0, 0, 0), \frac{\partial p_1}{\partial x_1} + \frac{\partial p_2}{\partial x_2} + \frac{\partial p_3}{\partial x_3} = 0 \text{ in each } t \in \Delta\}.$$

We have $S_0 \subset (H_0^1(\Omega))^3$, and a is also S_0- elliptic. Therefore K has a unique minimizer in S_0. Now

$$S_0 + ((\Pi_b^d(g_1), \Pi_b^d(g_2), \Pi_b^d(g_3))) = S$$

and arguing as in the continuous case we see that J has a unique minimizer in S which is solution of the variational problem:

$$\text{Find } \mathbf{u} \in S \text{ such that } a(\mathbf{u}, \mathbf{v}) = \langle \mathbf{f}, \mathbf{v} \rangle \ \forall \mathbf{v} \ \in S.$$

We now give an explicit expression of J in terms of the B-nets of elements of S. we have

$$\begin{aligned} J(\mathbf{u}) &= \frac{\nu}{2} \sum_t \int_t |\nabla \mathbf{u}|^2 - \sum_t \int_t \mathbf{f} \cdot \mathbf{u} \\ &= \frac{\nu}{2} \sum_{j=1}^{3} \sum_t \int_t |\nabla u_j|^2 - \sum_{j=1}^{3} \int_t f_j u_j. \end{aligned}$$

Let $\mathbf{u} = (p_1, p_2, p_3)$ with the p_i as defined in (6.11). We write $\Pi_b^d(f_{i|t}) = F_i^t$ which can also be written $\Pi_b^d(f_{i|t}) = \sum_{\alpha=1}^{n} f_{i,\alpha}^t B_\alpha^d$ and $\Pi_b^d(f_i) = F_i$. We also introduce the local mass matrix $M^t = (\int_t B_\alpha^d B_\beta^d)_{\alpha,\beta=1,\ldots,n}$ and denote by M the corresponding global mass matrix. $K^t = (\int_t \nabla B_\alpha^d \nabla B_\beta^d)_{\alpha,\beta=1,\ldots,n}$ is the local stiffness matrix and we denote by K the global stiffness matrix. We have

$$\begin{aligned} J(\mathbf{u}) = J(c_1, c_2, c_3) &= \frac{\nu}{2} \sum_{j=1}^{3} \sum_t \sum_{\alpha=1,\beta=1}^{n} c_{j,\alpha}^t c_{j,\beta}^t \int_t \nabla B_\alpha^d \nabla B_\beta^d \\ &\quad + \sum_{j=1}^{3} \sum_t \sum_{\alpha=1,\beta=1}^{n} f_{j,\alpha}^t c_{j,\beta}^t \int_t B_\alpha^d B_\beta^d \\ &= \frac{\nu}{2} \sum_{j=1}^{3} \sum_t \sum_{\alpha=1,\beta=1}^{n} (c_j^t)^T K^t c_j^t + \sum_{j=1}^{3} \sum_t (F_j^t) M^t c_j^t \\ &= \frac{\nu}{2} \sum_{j=1}^{3} (c_j)^T K c_j + \sum_{j=1}^{3} (F_j)^T M c_j. \end{aligned}$$

We now introduce a few more notations. Let

$$\mathbf{c} = (c_1, c_2, c_3)^T, \quad \mathbf{F} = (F_1, F_2, F_3)^T, \quad \mathbf{G} = (G_1, G_2, G_3)^T$$

and

$$\overline{M} = \begin{pmatrix} M & 0 & 0 \\ 0 & M & 0 \\ 0 & 0 & M \end{pmatrix}, \qquad \overline{K} = \begin{pmatrix} K & 0 & 0 \\ 0 & K & 0 \\ 0 & 0 & K \end{pmatrix},$$

$$\overline{H} = \begin{pmatrix} H & 0 & 0 \\ 0 & H & 0 \\ 0 & 0 & H \end{pmatrix}, \qquad \overline{R} = \begin{pmatrix} R & 0 & 0 \\ 0 & R & 0 \\ 0 & 0 & R \end{pmatrix}.$$

We finally introduce

$$D = \begin{bmatrix} D_1 & D_2 & D_3 \end{bmatrix},$$

where I is the identity matrix of $\mathbf{R}^{m*nth}$ with $m = \dim P_{d-1}$. With these notations

$$J(\mathbf{c}) = J(c_1, c_2, c_3) = \frac{\nu}{2}\mathbf{c}^T\overline{K}\mathbf{c} + \mathbf{F}^T\overline{M}\mathbf{c}$$

and

$$S = \{\mathbf{c} = (c_1, c_2, c_3) \in (\mathbf{R}^{n*nth})^3, \overline{H}\mathbf{c} = 0, \overline{R}\mathbf{c} = \mathbf{G}, D\mathbf{c} = 0\}.$$

The problem of minimizing J over S is equivalent to that of minimizing J over $(\mathbf{R}^{n*nth})^3$ under the constraints

$$\overline{H}\mathbf{c} = 0, \qquad \overline{R}\mathbf{c} = \mathbf{G}, \text{ and} \qquad D\mathbf{c} = 0.$$

We recall that this problem has a unique solution $\mathbf{c}$. On the other hand there are Lagrange multipliers λ_1, λ_2 and λ_3 such that

$$\begin{aligned} \nu\overline{K}\mathbf{c} + \overline{H}^T\lambda_1 + \overline{R}^T\lambda_2 + D^T\lambda_3 &= \overline{M}\mathbf{F}, \\ \overline{H}\mathbf{c} &= 0, \\ \overline{R}\mathbf{c} &= \mathbf{G}, \\ D\mathbf{c} &= 0. \end{aligned}$$

This can be written as an equation

$$Ax = b$$

with

$$A = \begin{pmatrix} \overline{H}^T & \overline{R}^T & D^T & \nu\overline{K} \\ 0 & 0 & 0 & \overline{H} \\ 0 & 0 & 0 & \overline{R} \\ 0 & 0 & 0 & D \end{pmatrix} \qquad x = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \mathbf{c} \end{bmatrix} \qquad b = \begin{bmatrix} \overline{M}\mathbf{F} \\ 0 \\ \mathbf{G} \\ 0 \end{bmatrix}.$$

The equation can be solved using the matrix iterative algorithm or directly by seeking for a least squares solution.

Approximation of the pressure

The pressure is computed by using an approximation of the velocity. Assuming that u is smooth and taking the divergence of the first equation in(6.10), we get

$$-\Delta p = -\text{div } \mathbf{f}$$

since div $\mathbf{u} = 0$. This equation is supplied with Neumann boundary conditions

$$\frac{\partial p}{\partial \mathbf{n}} = \nabla p \cdot \mathbf{n} = \mathbf{f} \cdot \mathbf{n} + \nu(\Delta \mathbf{u}) \cdot \mathbf{n}, \qquad \text{on } \partial\Omega.$$

We check the compatibility condition for this Neumann problem.

$$\begin{aligned} \int_\Omega -\text{div } \mathbf{f} + \int_{\partial\Omega} \mathbf{f} \cdot \mathbf{n} + \nu(\Delta \mathbf{u}) \cdot \mathbf{n} &= \int_{\partial\Omega} \nu(\Delta \mathbf{u}) \cdot \mathbf{n} \\ &= \int_\Omega \nu \text{ div} \Delta \mathbf{u} \\ &= \nu \int_\Omega \Delta \text{ div } \mathbf{u} \\ &= 0, \end{aligned}$$

using the divergence theorem.

6.2.3 Numerical results

For the following numerical results, the approximations of the velocity vector field have continuous components over the domain and the pressure is globally continuous with derivative continuously differentiable across tetrahedra which do not share a face with the boundary.

Domain 1: This domain is formed by the union of two tetrahedra which share a common face.

Domain 2: We consider a cube of volume one which has been subdivided into six tetrahedra.

We consider three different vector fields $\mathbf{g} = (g_1, g_2, g_3)$ with a corresponding pressure p.

Case 1:

$$\begin{aligned} g_1 &= -\exp(x + 2y + 3z) \\ g_2 &= 2\exp(x + 2y + 3z) \\ g_3 &= -\exp(x + 2y + 3z) \\ p &= \exp(x + y + z) \end{aligned}$$

Case 2:

$$\begin{aligned} g_1 &= 1/(1 + x + y + z) \\ g_2 &= 1/(1 + x + y + z) \\ g_3 &= -2/(1 + x + y + z) \\ p &= \exp(x + y + z) \end{aligned}$$

Case 3:

$$
\begin{aligned}
g_1 &= x(1-x)y(1-y)z(1-z) \\
g_2 &= x(1-x)y(1-y)z(1-z)) \\
g_3 &= \tfrac{1}{6}z^2(y+x-1)(-x+2xy-y)(2z-3) \\
p &= \exp(x+y+z)
\end{aligned}
$$

Case 1 on Domain 1 with $d = 3$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
2	1.1233e+00	2.7279e+00	1.6173e+00	3.9290e+02
16	2.4754e-01	3.4290e-01	2.7805e-01	1.7609e+02

Case 1 on Domain 1 with $d = 4$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
2	6.1316e-01	5.4653e-01	7.1343e-01	9.7319e+01
16	3.9387e-02	4.1184e-02	4.4533e-02	1.2710e+01

Case 1 on Domain 1 with $d = 5$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
2	1.0361e-01	1.0863e-01	9.8393e-02	1.3007e+01
16	3.6605e-03	3.7574e-03	4.3928e-03	1.5159e+00

Case 1 on Domain 1 with $d = 6$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
2	1.9113e-02	1.8891e-02	2.1345e-02	2.9372e+00
16	3.0237e-04	3.1482e-04	3.7262e-04	9.7706e-02

Case 1 on Domain 2 with $d = 3$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
6	3.3633e+01	5.9431e+01	4.0397e+01	1.3466e+03
48	1.5083e+01	1.8709e+01	1.5222e+01	4.4382e+02

Case 1 on Domain 2 with $d = 4$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
6	1.7010e+01	4.4374e+01	3.5368e+01	3.8562e+02
48	9.4142e-01	2.2094e+00	1.8373e+00	3.5278e+01

Case 1 on Domain 2 with $d = 5$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
6	2.3804e+00	7.3711e+00	5.9629e+00	9.8470e+01

Case 1 on Domain 2 with $d = 6$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
6	3.9620e-01	1.2238e+00	1.0311e+00	2.7404e+01

Case 1 on Domain 2 with $d = 7$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
6	6.7456e-02	1.9789e-01	1.6260e-01	6.8411e+00

Case 2 on Domain 1 with $d = 3$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
2	1.0605e-02	1.4174e-02	1.9953e-02	1.2741e+00
16	2.5841e-03	2.3599e-03	3.5273e-03	2.1682e-01

Case 2 on Domain 1 with $d = 4$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
2	4.1169e-03	3.8110e-03	6.6898e-03	6.4602e-01
16	5.0246e-04	3.9061e-04	8.0557e-04	6.4369e-02

Case 2 on Domain 1 with $d = 5$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
2	1.5548e-03	1.2819e-03	2.1008e-03	1.1806e-01
16	1.2103e-04	6.5984e-05	1.6800e-04	8.7476e-03

Case 2 on Domain 1 with $d = 6$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
2	5.8664e-04	3.8984e-04	7.4313e-04	7.1244e-02
16	2.9003e-05	1.1554e-05	3.6974e-05	3.2856e-03

Case 2 on Domain 2 with $d = 3$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
6	3.2488e-02	3.6178e-02	5.4013e-02	2.4606e+00
48	2.2367e-02	2.2244e-02	2.8100e-02	6.8163e-01

Case 2 on Domain 2 with $d = 4$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
6	2.9860e-02	2.9388e-02	5.5366e-02	6.8644e-01
48	2.8115e-03	2.7051e-03	4.4312e-03	9.8769e-02

Case 2 on Domain 2 with $d = 5$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
6	7.1485e-03	6.9051e-03	1.3065e-02	2.4031e-01

Case 2 on Domain 2 with $d = 6$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
6	2.0444e-03	1.9881e-03	3.6732e-03	1.0345e-01

Case 2 on Domain 2 with $d = 7$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
6	6.0114e-04	5.6334e-04	1.0620e-03	4.1101e-02

Case 3 on Domain 1 with $d = 3$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
2	5.4823e-03	5.1512e-03	4.8780e-03	2.9874e-01
16	1.2800e-03	1.1522e-03	1.5770e-03	1.6408e-01

Case 3 on Domain 1 with $d = 4$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
6	2.6458e-03	2.3774e-03	2.1849e-03	2.9295e-01
48	4.2214e-04	2.3788e-04	3.7809e-04	5.8393e-02

Case 3 on Domain 1 with $d = 5$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
2	2.1517e-03	1.0184e-03	1.9192e-03	3.1441e-01
16	4.0894e-05	3.0021e-05	3.6268e-05	9.4051e-03

Case 3 on Domain 1 with $d = 6$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
2	3.9761e-09	2.3460e-09	3.9761e-09	1.5114e-03
16	5.2753e-11	5.3180e-11	5.2753e-11	3.9759e-05

Case 3 on Domain 2 with $d = 3$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
6	1.4097e-02	1.6179e-02	1.9143e-02	1.9427e+00
48	1.7650e-03	1.8291e-03	3.4444e-03	5.5009e-01

Case 3 on Domain 2 with $d = 4$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
6	3.4925e-03	3.6196e-03	8.8974e-03	2.5283e-01
48	3.7584e-04	3.7287e-04	4.4346e-04	9.3024e-02

Case 3 on Domain 2 with $d = 5$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
6	2.5998e-03	2.5998e-03	2.8143e-03	6.5884e-02

Case 3 on Domain 2 with $d = 6$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
6	2.5384e-07	2.5384e-07	2.5384e-07	1.3852e-02

Case 3 on Domain 2 with $d = 7$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
6	2.3598e-08	2.3598e-08	2.3598e-08	2.6995e-03

6.3 Spline Approximations of the Navier-Stokes Equations

The Navier-Stokes equations which govern the motion of an incompressible viscous fluid in a bounded domain Ω of $\mathbf{R}^3$ are

$$\begin{cases} -\nu\ \Delta \mathbf{u} + \sum_{j=1}^{3} u_j \frac{\partial \mathbf{u}}{\partial x_j} + \nabla p = \mathbf{f} & \text{in } \Omega, \\ \text{div } \mathbf{u} = 0 & \text{in } \Omega. \end{cases} \tag{6.12}$$

The unknowns here are the velocity $\mathbf{u} = (u_1, u_2, u_3)^T$ of the fluid and the pressure p. The kinematic viscosity ν and $\mathbf{f} = (f_1, f_2, f_3)$ which represents the externally applied forces (e.g. gravity) are given. The stress on the fluid is encoded in the nonlinear term.

6.3.1 The continuous problem

We shall deal with the Dirichlet boundary condition. To prescribe the velocity on the boundary $\partial\Omega$ of Ω, we set $\mathbf{u} = \mathbf{g}$ on $\partial\Omega$. In view of the divergence theorem, $\mathbf{g}$ must satisfy the compatibility condition

$$\int_{\partial\Omega} \mathbf{g} \cdot \mathbf{n} = 0 \tag{6.13}$$

where $\mathbf{n}$ is the unit outer normal to $\partial\Omega$.

Formally, by taking the scalar product of equation (6.12) with $\mathbf{v} \in H_0^1(\Omega)$ satisfying div $\mathbf{v} = 0$, we get a weak form of the Navier-Stokes equations: Find $\mathbf{u} \in H^1(\Omega)^3$ such that

$$\nu \int_\Omega \nabla\mathbf{u} \cdot \nabla\mathbf{v} + \sum_{j=1}^{3} \int_\Omega u_j \frac{\partial \mathbf{u}}{\partial x_j} \cdot \mathbf{v} = \int_\Omega \mathbf{f} \cdot \mathbf{v} \qquad \forall \mathbf{v} \in V_0 \tag{6.14}$$

$$\text{div } \mathbf{u} = 0 \qquad \text{in } \Omega \tag{6.15}$$

$$\mathbf{u} = \mathbf{g} \qquad \text{on } \partial\Omega, \tag{6.16}$$

where

$$V_0 = \{\mathbf{v} \in H_0^1(\Omega)^3, \text{div } \mathbf{v} = 0\}.$$

Let

$$L_0^2(\Omega) = \{u \in L^2(\Omega), \int_\Omega u = 0\} \quad \text{and}$$

$$H^{\frac{1}{2}}(\partial\Omega) = \{\tau(u), u \in H^1(\Omega)\},$$

where by $\tau(u)$, we mean the trace of u on $\partial\Omega$. We have the following well-known existence and uniqueness results, (cf. [Girault and Raviart'86]).

Theorem 6.3.1 *Let Ω be a bounded connected open subset of $\mathbf{R}^3$ with a Lipschitz continuous boundary. For $\mathbf{f} \in H^{-1}(\Omega)^3$ and $\mathbf{g} \in H^{\frac{1}{2}}(\partial\Omega)^3$ satisfying*

$$\int_{\partial\Omega} \mathbf{g} \cdot \mathbf{n} = 0,$$

the problem: find $(\mathbf{u}, p) \in H^1(\Omega)^3 \times L^2_0(\Omega)$ *such that*

$$\begin{cases} -\nu\ \Delta\mathbf{u} + \sum_{j=1}^{3} u_j \frac{\partial}{\partial x_j}\mathbf{u} + \nabla p &= \mathbf{f} \qquad in\ \Omega \\ div\ \mathbf{u} &= 0 \qquad in\ \Omega \\ \mathbf{u} &= \mathbf{g} \quad on\ \partial\Omega, \end{cases}$$

has a solution which is unique provided that ν *is sufficiently large. (Here the first two equations should be interpreted in the sense of distributions and the last one in the trace sense.)*

Unlike the linear case, this problem cannot be cast directly as a minimization problem. In [Gunzburger'86] a procedure to reduce this problem to the solution of a sequence of Stokes problems is described. Here we shall derive an algorithm which withstands our tests.

6.3.2 Spline approximations

We first compute the velocity vector field and then the pressure term. The difference with the previous section is the presence of the nonlinear term.

Approximations of the velocity

Let us fix the tetrahedral partition $\mathcal{T}$, and let $n = \dim P_d$. If nth is the number of tetrahedra in $\mathcal{T}$, the B-net of an element p of $S^r_d(\Omega)$ has length $N = n * nth$. We denote by p_i, $1 \le i \le 3$ an approximant of u_i and c_i the B-net of p_i. We therefore have

$$p_{i|t} = \sum_{s=1}^{n} c^t_{i,s} B^d_s \qquad i = 1, 2, 3.$$

We will consider a subspace W of the space of test functions. We'll use

$$W = \{(v_1, v_2, v_3) \in S^r_d(\Omega)^3, \quad (\Pi^b_d(v_1), \Pi^d_b(v_2), \Pi^b_d(v_3)) = (0, 0, 0),$$
$$\frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} + \frac{\partial v_3}{\partial x_3} = 0\},$$

where Π_b^d is the boundary interpolation operator which when acting on splines of degree d is simply the restriction on the boundary.

Let d_i encode the B-net of v_i. Thus, if (v_1, v_2, v_3) is in W we have

$$D_1 d_1 + D_2 d_2 + D_3 d_3 = 0, \tag{6.17}$$

$$H d_1 = H d_2 = H d_3 = 0, \tag{6.18}$$

$$R d_1 = R d_2 = R d_3 = 0, \tag{6.19}$$

where H is the smoothness matrix, R is a matrix realization of Π_d^b, i.e. Rd_i is the B-net of v_i on the boundary for all $i = 1, 2, 3$, and D_i is a discrete derivative. If p has B-net c, $D_i c$ is the B-net of $\frac{\partial u}{\partial x_i}$. Let also $G_i, i = 1, 2, 3$ denote the B-net of the components g_i of $\mathbf{g}$. It is convenient to use $\mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix}$ and $\mathbf{d} = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix}$. We also write $\mathbf{G} = \begin{pmatrix} G_1 \\ G_2 \\ G_3 \end{pmatrix}$ and therefore have

$$\overline{H}\mathbf{d} \stackrel{\text{def}}{=} \begin{pmatrix} H & 0 & 0 \\ 0 & H & 0 \\ 0 & 0 & H \end{pmatrix} \mathbf{d} = 0,$$

$$\overline{R}\mathbf{d} \stackrel{\text{def}}{=} \begin{pmatrix} R & 0 & 0 \\ 0 & R & 0 \\ 0 & 0 & R \end{pmatrix} \mathbf{d} = 0, \quad \text{and}$$

$$D\ \mathbf{d} = \begin{pmatrix} D_1 & D_2 & D_3 \end{pmatrix} \mathbf{d} = 0,$$

where I is the identity matrix of $\mathbf{R}^{m*nth}$ with m=dimP_{d-1}. With these notations, (6.17), (6.18) and (6.19) are written

$$\overline{D}\mathbf{c} = 0, \tag{6.20}$$

$$\overline{H}\mathbf{c} = 0, \tag{6.21}$$

$$\overline{R}\mathbf{c} = 0. \tag{6.22}$$

We now proceed to the discretization of the equations. Equation (6.15) and (6.16) are discretized as

$$\overline{D}\mathbf{c} = \mathbf{0}, \tag{6.23}$$

$$\overline{R}\mathbf{c} = \mathbf{G}, \tag{6.24}$$

with $(p_1, p_2, p_3) \in S_d^r(\Omega)^3$ giving

$$\overline{H}\mathbf{c} = 0. \tag{6.25}$$

Equation (6.14) can also be written

$$\nu \sum_t \int_t \nabla \mathbf{u} \nabla \mathbf{v} + \sum_{j=1}^{3} \sum_t \int_t u_j \frac{\partial \mathbf{u}}{\partial x_j} \cdot \mathbf{v} = \sum_t \int_t \mathbf{f} \cdot \mathbf{v} = \sum_t \sum_{j=1}^{3} \int_t f_j v_j.$$

Let's write $\Pi^d(f_j)_{|t} = \sum_{\alpha=1}^n f_{j,\alpha}^t B_\alpha^d$ for the interpolation of f_j and $v_{j|t} = \sum_{\beta=1}^n d_{j,\beta}^t B_\beta^d$ and let F_j^t, F_j, d_j^t and d_j encode the B-net of $f_{j|t}$, f_j, $v_{j|t}$ and v_j respectively.

Let $M^t = (\int_t B_\alpha^d B_\beta^d)_{\alpha,\beta=1,\ldots,n}$ and M the corresponding mass matrix. We have

$$\begin{aligned}
\sum_t \int_t \Pi_b^d(f_j) v_j &= \sum_t \sum_{\alpha=1,\beta=1}^{n} f_{j,\alpha}^t d_{j,\beta}^t B_\alpha^d B_\beta^d \\
&= \sum_t (F_j^t)^T M^t d_j^t \\
&= (F_j)^T M d_j.
\end{aligned}$$

Let $\overline{M} = \begin{pmatrix} M & 0 & 0 \\ 0 & M & 0 \\ 0 & 0 & M \end{pmatrix}$ and $\mathbf{F} = \begin{pmatrix} F_1 \\ F_2 \\ F_3 \end{pmatrix}$. We have

$$\int_\Omega \mathbf{f} \cdot \mathbf{v} = (F_1)^T M d_1 + (F_2)^T M d_2 + (F_3)^T M d_3 = \mathbf{F}^T \overline{M} \mathbf{d}$$

Similarly we introduce the local and global stiffness matrix

$$K^t = (\int_t \nabla B^d_\alpha \cdot \nabla B^d_\beta)_{\alpha,\beta=1,\ldots,n}$$

and K. We write $\overline{K} = \begin{pmatrix} K & 0 & 0 \\ 0 & K & 0 \\ 0 & 0 & K \end{pmatrix}$ and we have

$$\begin{aligned} \int_\Omega \nabla \mathbf{u} \cdot \nabla \mathbf{v} &= \sum_{j=1}^{3} \int_\Omega \nabla u_j \cdot \nabla v_j \\ &= c_1^T K d_1 + c_2^T K d_2 + c_3^T K d_3 \\ &= \mathbf{c}^T \overline{K} \mathbf{d}. \end{aligned}$$

Thus if we define the form a_0 by

$$a_0(\mathbf{u}, \mathbf{v}) = \int_\Omega \nabla \mathbf{u} \cdot \nabla \mathbf{v}, \tag{6.26}$$

the discrete version of a_0 would be

$$a_0(\mathbf{u}, \mathbf{v}) = \mathbf{c}^T \overline{K} \mathbf{d}$$

in the sense that $\mathbf{c}$ and $\mathbf{d}$ encode the B-net of $\mathbf{u}$ and $\mathbf{v}$ respectively.

Let's also introduce

$$a_1(\mathbf{w}; \mathbf{u}, \mathbf{v}) = \sum_{j=1}^{3} \int_\Omega w_j \frac{\partial \mathbf{u}}{\partial x_j} \cdot \mathbf{v}. \tag{6.27}$$

We have

$$\int_\Omega w_j \frac{\partial \mathbf{u}}{\partial x_j} \cdot \mathbf{v} = \sum_{i=1}^{3} \int_\Omega w_j \frac{\partial u_i}{\partial x_j} v_i = \sum_{i=1}^{3} \sum_t \int_\Omega w_j \frac{\partial u_i}{\partial x_j} v_i.$$

Let's write

$$u_{i|t} = \sum_{\alpha=1}^{n} c_{i,\alpha}^t B_\alpha^d, \qquad v_{i|t} = \sum_{\beta=1}^{n} d_{i,\beta}^t B_\beta^d, \qquad w_{j|t} = \sum_{\gamma=1}^{n} e_{j,\gamma}^t B_\gamma^d$$

so

$$\int_\Omega w_j \frac{\partial u_i}{\partial x_j} v_i = \sum_t \sum_{\alpha,\beta,\gamma=1}^{n} e_{j,\gamma}^t c_{i,\alpha}^t d_{i,\beta}^t \int_t B_\gamma^d \frac{\partial B_\alpha^d}{\partial x_j} B_\beta^d.$$

Let $\mathrm{Mat}_{j\beta}^t = \left(\int_t B_\gamma^d \frac{\partial B_\alpha^d}{\partial x_j} B_\beta^d \right)_{\alpha,\gamma=1,\ldots,n}$ so that

$$\int_t w_j \frac{\partial u_i}{\partial x_j} v_i = \sum_{\beta=1}^{n} d_{i,\beta}^t (e_j^t)^T \mathrm{Mat}_{j\beta}^t c_i^t = B_{ji}^t(\mathbf{e}, \mathbf{c}) d_i^t,$$

where $B_{ji}^t(\mathbf{e}, \mathbf{c}) = \left[(e_j^t)^T \mathrm{Mat}_{j1}^t c_i^t, \ldots, (e_j^t)^T \mathrm{Mat}_{jnth}^t c_i^t \right]$ and if we let B_{ji} assembled with the B_{ji}^t, we can write

$$\int_\Omega u_j \frac{\partial u_i}{\partial x_j} v_i = B_{ji}(\mathbf{e}, \mathbf{c}) d_i$$

with $B_{ji}(\mathbf{e}, \mathbf{c}) = [B_{ji}^1(\mathbf{e}, \mathbf{c}), \ldots, B_{ji}^{nth}(\mathbf{e}, \mathbf{c})]$.

Finally, we have

$$\begin{aligned} \sum_{j=1}^{3} \int_\Omega w_j \frac{\partial \mathbf{u}}{\partial x_j} \cdot \mathbf{v} = \sum_{i,j=1}^{3} \int_\Omega w_j \frac{\partial u_i}{\partial x_j} v_i &= \sum_{i,j=1}^{3} B_{ji}(\mathbf{e}, \mathbf{c}) d_i \\ &= \sum_{i=1}^{3} (\sum_{j=1}^{3} B_{ji}(\mathbf{e}, \mathbf{c})) d_i. \end{aligned}$$

Therefore

$$a_1(\mathbf{w}; \mathbf{u}, \mathbf{v}) = \left[\sum_{j=1}^{3} B_{j1}(\mathbf{e}, \mathbf{c}) \quad \sum_{j=1}^{3} B_{j2}(\mathbf{e}, \mathbf{c}) \quad \sum_{j=1}^{3} B_{j3}(\mathbf{e}, \mathbf{c}) \right] \mathbf{d}.$$

Now since

$$\sum_{j=1}^{3} \int_\Omega u_j \frac{\partial \mathbf{u}}{\partial x_j} \cdot \mathbf{v} = a_1(\mathbf{u}; \mathbf{u}, \mathbf{v}),$$

equation (6.14) is discretized as

$$\nu \mathbf{c}^T \overline{K} \mathbf{d} + \left[\sum_{j=1}^{3} B_{j1}(\mathbf{c}, \mathbf{c}) \quad \sum_{j=1}^{3} B_{j2}(\mathbf{c}, \mathbf{c}) \quad \sum_{j=1}^{3} B_{j3}(\mathbf{c}, \mathbf{c}) \right] \mathbf{d} = \mathbf{F}^T \overline{M} \mathbf{d}.$$

If one considers the following linear functional in $\mathbf{d}$,

$$J(\mathbf{d}) = \left(\nu \mathbf{c}^T \overline{K} + \left[\sum_{j=1}^{3} B_{j1}(\mathbf{c}, \mathbf{c}) \quad \sum_{j=1}^{3} B_{j2}(\mathbf{c}, \mathbf{c}) \quad \sum_{j=1}^{3} B_{j3}(\mathbf{c}, \mathbf{c}) \right] - \mathbf{F}^T \overline{M} \right) \mathbf{d},$$

$J(\mathbf{d}) = 0$ for all $\mathbf{d}$ satisfying (6.20), (6.21) and (6.22). A fortiori, taking the derivative with respect to $\mathbf{d}$ we must have:

$$\nu \mathbf{c}^T \overline{K} + \left[\sum_{j=1}^{3} B_{j1}(\mathbf{c}, \mathbf{c}) \quad \sum_{j=1}^{3} B_{j2}(\mathbf{c}, \mathbf{c}) \quad \sum_{j=1}^{3} B_{j3}(\mathbf{c}, \mathbf{c}) \right]$$
$$+ \lambda_1^T \overline{H} + \lambda_2^T \overline{R} + \lambda_3^T D = \mathbf{F}^T \overline{M}$$

for some Lagrange multipliers $\lambda_1, \lambda_2, \lambda_3$.

This added with (6.23), (6.24) and (6.25) provided the non-linear equations which were solved. We consider two methods for solving the non-linear system of equations, using a simple iteration algorithm and using Newton's method . The equations are

$$\nu \overline{K} \mathbf{c} + \left[\sum_{j=1}^{3} B_{j1}(\mathbf{c}, \mathbf{c}) \quad \sum_{j=1}^{3} B_{j2}(\mathbf{c}, \mathbf{c}) \quad \sum_{j=1}^{3} B_{j3}(\mathbf{c}, \mathbf{c}) \right]^T$$
$$+ \overline{H}^T \lambda_1 + \overline{R}^T \lambda_2 + D^T \lambda_3 = \overline{M} \mathbf{F},$$
$$\overline{H} \mathbf{c} = 0,$$
$$\overline{R} \mathbf{c} = \mathbf{G},$$
$$D \mathbf{c} = 0.$$

Notice that

$$B_{ji}^t(\mathbf{e},\mathbf{c})^T = \begin{pmatrix} (e_j^t)^T \mathrm{Mat}_{j1}^t c_i^t \\ \vdots \\ (e_j^t)^T \mathrm{Mat}_{j,nth}^t c_i^t \end{pmatrix} \tag{6.28}$$

$$= \begin{pmatrix} (e_j^t)^T \mathrm{Mat}_{j1}^t \\ \vdots \\ (e_j^t)^T \mathrm{Mat}_{j,nth}^t \end{pmatrix} c_i^t \tag{6.29}$$

$$\stackrel{\text{def}}{=} B_j^t(\mathbf{e}) c_i^t. \tag{6.30}$$

And after assembling, we may write

$$B_{ji}(\mathbf{e},\mathbf{c})^T = B_j(\mathbf{e}) c_i, \tag{6.31}$$

with $B_j(\mathbf{e})$ having the size of K and

$$\left(\sum_{j=1}^3 B_{j1}(\mathbf{e},\mathbf{c}) \quad \sum_{j=1}^3 B_{j2}(\mathbf{e},\mathbf{c}) \quad \sum_{j=1}^3 B_{j3}(\mathbf{e},\mathbf{c}) \right)^T$$

$$= \begin{pmatrix} \sum_{j=1}^3 B_{j1}(\mathbf{e},\mathbf{c})^T \\ \sum_{j=1}^3 B_{j2}(\mathbf{e},\mathbf{c})^T \\ \sum_{j=1}^3 B_{j3}(\mathbf{e},\mathbf{c})^T \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{j=1}^3 B_j(\mathbf{e}) c_1 \\ \sum_{j=1}^3 B_j(\mathbf{e}) c_2 \\ \sum_{j=1}^3 B_j(\mathbf{e}) c_3 \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{j=1}^3 B_j(\mathbf{e}) & 0 & 0 \\ 0 & \sum_{j=1}^3 B_j(\mathbf{e}) & 0 \\ 0 & 0 & \sum_{j=1}^3 B_j(\mathbf{e}) \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix}$$

$$\stackrel{\text{def}}{=} \overline{B}(\mathbf{e})\mathbf{c},$$

where

$$B(\mathbf{e}) = \sum_{j=1}^{3} B_j(\mathbf{e})$$

and $\overline{B}(\mathbf{e})$ is defined analogous to the definition of $\overline{H}$. With these notations,

$$a_1(\mathbf{w}; \mathbf{u}, \mathbf{v}) = (\overline{B}(\mathbf{e})\mathbf{c})^T \mathbf{d} \tag{6.32}$$

and the nonlinear equations to be solved are

$$\begin{aligned} \nu\overline{K}\mathbf{c} + \overline{B}(\mathbf{c})\mathbf{c} + \overline{H}^T\lambda_1 + \overline{R}^T\lambda_2 + D^T\lambda_3 &= \overline{M}\mathbf{F} \\ \overline{H}\mathbf{c} &= 0 \\ \overline{R}\mathbf{c} &= \mathbf{G} \\ D\mathbf{c} &= 0. \end{aligned} \tag{6.33}$$

THE ALGORITHMS

A simple iteration algorithm:

Let $(\mathbf{c}^0, \lambda_1^0, \lambda_2^0, \lambda_3^0)$ be the solution of the linear problem (i.e. the associated Stokes equations) and define $(\mathbf{c}^{n+1}, \lambda_1^{n+1}, \lambda_2^{n+1}, \lambda_3^{n+1})$ as the solution of

$$\begin{aligned} \nu\overline{K}\mathbf{c}^{n+1} + \overline{B}(\mathbf{c}^n)\mathbf{c}^{n+1} + \overline{H}^T\lambda_1^{n+1} + \overline{R}^T\lambda_2^{n+1} + D^T\lambda_3^{n+1} &= \overline{M}\mathbf{F} \\ \overline{H}\mathbf{c}^{n+1} &= 0 \\ \overline{R}\mathbf{c}^{n+1} &= \mathbf{G} \\ D\mathbf{c}^{n+1} &= 0. \end{aligned}$$

We would like to point out that Theorem 4.4.1 can be applied to solve the previous systems. This follows from the properties of the forms a_0 and a_1.

Newton's method

It is convenient to consider a mapping

$$\Gamma : (\mathbf{c}, \lambda_1, \lambda_2, \lambda_3) \mapsto (\nu\overline{K}\mathbf{c} + \overline{B}(\mathbf{c})\mathbf{c} + \overline{H}^T\lambda_1 + \overline{R}^T\lambda_2 + D^T\lambda_3 - \overline{M}\mathbf{F}, \overline{H}\mathbf{c}, \overline{R}\mathbf{c} - \mathbf{G}, D\mathbf{c}).$$

We seek to solve $\Gamma(\mathbf{c}, \lambda_1, \lambda_2, \lambda_3) = 0$. We write $\mathbf{X}^n = (\mathbf{c}^n, \lambda_1^n, \lambda_2^n, \lambda_3^n)$. Let $\mathbf{X}^0$ be the solution of the linear problem and define $\mathbf{X}^{n+1}$ such that

$$\Gamma^{'}(\mathbf{X}^n)(\mathbf{X}^{n+1} - \mathbf{X}^n) = -\Gamma(\mathbf{X}^n). \tag{6.34}$$

The main difficulty in evaluating $\Gamma^{'}(\mathbf{X}^n)$ is to establish formulas for the Frechet derivative of the mapping

$$\mathbf{c} \mapsto \overline{B}(\mathbf{c})\mathbf{c} = \begin{pmatrix} \sum_{j=1}^3 B_j(\mathbf{c})c_1 \\ \sum_{j=1}^3 B_j(\mathbf{c})c_2 \\ \sum_{j=1}^3 B_j(\mathbf{c})c_3 \end{pmatrix}.$$

It is enough to consider the mapping

$$\mathbf{c} \mapsto \sum_{j=1}^3 B_j(\mathbf{c})c_i$$

which is assembled from the mappings

$$\mathbf{c} \mapsto B_{ji}^t(\mathbf{c}, \mathbf{c})^T = \begin{pmatrix} (c_j^t)^T \mathrm{Mat}_{j1}^t c_i^t \\ \vdots \\ (c_j^t)^T \mathrm{Mat}_{j,nth}^t c_i^t \end{pmatrix}.$$

We therefore introduce the bilinear mapping, for i, j fixed

$$(\mathbf{u}, \mathbf{v}) \mapsto \gamma_{ji}(\mathbf{u}, \mathbf{v}) = \begin{pmatrix} (u_j)^T \mathrm{Mat}_{j1} v_i \\ \vdots \\ (v_j)^T \mathrm{Mat}_{j,nth} v_i \end{pmatrix}$$

for a series of matrices $\mathrm{Mat}_{k=1,\ldots,nth}$. The derivative of

$$\beta_{ji} : \mathbf{c} \mapsto \gamma_{ji}(\mathbf{c}, \mathbf{c})$$

is

$$\beta_{ji}'(\mathbf{c})(\mathbf{h}) = \gamma_{ji}(\mathbf{c}, \mathbf{h}) + \gamma_{ji}(\mathbf{h}, \mathbf{c})$$

for $\mathbf{h} = (h_1, h_2, h_3)$. As a consequence, if γ denotes the mapping $\mathbf{c} \mapsto \overline{B}(\mathbf{c})\mathbf{c}$,

$$\gamma'(\mathbf{c})(\mathbf{h}) = \begin{pmatrix} \sum_{j=1}^3 B_j(\mathbf{c})h_1 \\ \sum_{j=1}^3 B_j(\mathbf{c})h_2 \\ \sum_{j=1}^3 B_j(\mathbf{c})h_3 \end{pmatrix} + \begin{pmatrix} \sum_{j=1}^3 B_j(\mathbf{h})c_1 \\ \sum_{j=1}^3 B_j(\mathbf{h})c_2 \\ \sum_{j=1}^3 B_j(\mathbf{h})c_3 \end{pmatrix}$$

with $B_j(\mathbf{c})h_i$ defined as in (6.31). In a more compact form,

$$\gamma'(\mathbf{c})(\mathbf{h}) = \overline{B}(\mathbf{c})\mathbf{h} + \overline{B}(\mathbf{h})\mathbf{c}.$$

We therefore have explicitly

$$\begin{aligned} \Gamma'(\mathbf{X}^n)(\mathbf{X}^{n+1} - \mathbf{X}^n) = (\nu\overline{K}(\mathbf{c}^{n+1} - \mathbf{c}^n) + \overline{B}(\mathbf{c}^n)(\mathbf{c}^{n+1} - \mathbf{c}^n) + \overline{B}(\mathbf{c}^{n+1} - \mathbf{c}^n)\mathbf{c}^n \\ + \overline{H}^T(\lambda_1^{n+1} - \lambda_1^n) + \overline{R}^T(\lambda_2^{n+1} - \lambda_2^n) + D^T(\lambda_2^{n+1} - \lambda_2^n) \\ \overline{H}^T(\mathbf{c}^{n+1} - \mathbf{c}^n), \overline{R}(\mathbf{c}^{n+1} - \mathbf{c}^n), D(\mathbf{c}^{n+1} - \mathbf{c}^n)). \end{aligned}$$

The equation (6.34) therefore implies

$$\begin{aligned} \nu\overline{K}\mathbf{c}^{n+1} + \overline{B}(\mathbf{c}^n)\mathbf{c}^{n+1} + \overline{B}(\mathbf{c}^{n+1} - \mathbf{c}^n)\mathbf{c}^n \\ + \overline{H}^T\lambda_1^{n+1} + \overline{R}^T\lambda_2^{n+1} + D^T\lambda_3^{n+1} = \overline{M}\mathbf{F} \end{aligned} \tag{6.35}$$

along with

$$\overline{H}\mathbf{c}^{n+1} = 0, \qquad \overline{R}\mathbf{c}^{n+1} - \mathbf{G} = 0, \text{ and} \qquad D\mathbf{c}^{n+1} = 0 \tag{6.36}$$

We now show that $\overline{B}(\mathbf{c}^{n+1} - \mathbf{c}^n)\mathbf{c}^n$ can be written $\widetilde{B}(\mathbf{c}^n)(\mathbf{c}^{n+1} - \mathbf{c}^n)$ with $\widetilde{B}(\mathbf{c}^n)\mathbf{c}^n = \overline{B}(\mathbf{c}^n)\mathbf{c}^n$. We turn back to (6.31) and (6.30).

$$\begin{aligned} B_{ji}^t(\mathbf{c}, \mathbf{c})^T &= \begin{pmatrix} (c_j^t)^T \mathrm{Mat}_{j1}^t c_i^t \\ \vdots \\ (c_j^t)^T \mathrm{Mat}_{j,nth}^t c_i^t \end{pmatrix} \\ &\stackrel{\text{def}}{=} B_j^t(\mathbf{c})c_i^t \end{aligned}$$

can also be written

$$B^t_{ji}(\mathbf{c})^T = \begin{pmatrix} (c^t_i)^T(\mathrm{Mat}^t_{j1})^T c^t_j \\ \vdots \\ (c^t_i)^T(\mathrm{Mat}^t_{j,nth})^T c^t_j \end{pmatrix} \stackrel{\text{def}}{=} \widetilde{B}^t_i(\mathbf{c})c^t_j,$$

and after assembling

$$B_{ji}(\mathbf{c},\mathbf{c})^T = B_j(\mathbf{c})c_i = \widetilde{B}_i(\mathbf{c})c_j,$$

which gives the announced relations.

(6.35) can then be written

$$\begin{aligned} \nu\overline{K}\mathbf{c}^{n+1} + \overline{B}(\mathbf{c}^n)\mathbf{c}^{n+1} + \widetilde{B}(\mathbf{c}^n)\mathbf{c}^{n+1} + \overline{H}^T\lambda_1^{n+1} + \overline{R}^T\lambda_2^{n+1} + \\ D^T\lambda_3^{n+1} = \overline{M}\mathbf{F} + \overline{B}(\mathbf{c}^n)\mathbf{c}^n. \end{aligned} \tag{6.37}$$

This equation along with (6.36) provide the equations solved for the Newton's method iteration. Theorem 4.4.1 can also be applied to solve the above systems because $\overline{B}(\mathbf{c}^{(n)})$ is anti-symmetric and $\overline{K} + \widetilde{B}(\mathbf{c}^{(n)})$ (and consequently its symmetric part) is positive definite with respect to $L = (\overline{H}^T, \overline{R}^T, D^T)^T$ for ν sufficiently large. Indeed for a spline with B-coefficients encoded in $\mathbf{x}$ which satisfy $L\mathbf{x} = 0$,

$$\mathbf{x}^T K\mathbf{x} + \mathbf{x}^T\widetilde{B}(\mathbf{c}^{(n)})\mathbf{x} \geq \nu C||\mathbf{x}||^2_{H^1(\Omega)^3} - C_1||\mathbf{c}^{(n)}||_{H^1(\Omega)^3}||\mathbf{x}||^2_{H^1(\Omega)^3}.$$

But since the sequence $(\mathbf{c}^{(n)})$ is bounded (we prove its convergence below), there's $C_2 > 0$ for which $||\mathbf{c}^{(n)}||_{H^1(\Omega)^3} \leq C_2$, so

$$\mathbf{x}^T K\mathbf{x} + \mathbf{x}^T\widetilde{B}(\mathbf{c}^{(n)})\mathbf{x} \geq (\nu C - C_1C_2)||\mathbf{x}||^2_{H^1(\Omega)^3},$$

which proves what was claimed.

Convergence of the algorithms

In this section, we study only the homogeneous problem, i.e. $\mathbf{g} = 0$ or $\overline{\mathbf{G}} = 0$. The arguments presented here follow classical arguments (cf, [Girault and Raviart'86]) and [Karakashian'82]. (In the later, Lagrange multipliers are used to enforce only the divergence free condition.) First let's give some properties of the form a_0 defined in (6.26) and a_1 defined in (6.27) with their discrete counterparts.

a_0 is $H_0^1(\Omega)$ elliptic, i.e.

$$a_0(\mathbf{v}, \mathbf{v}) \geq C||\mathbf{v}||_{H_0^1(\Omega)^3}$$

for some constant $C > 0$.

For $\mathbf{v}$ in $S_d^r(\Omega)^3$ with B-net $\mathbf{d}$ satisfying $\overline{R}\mathbf{d} = 0$,

$$\begin{aligned} ||\mathbf{v}||^2_{H_0^1(\Omega)^3} &= \int_\Omega |\mathbf{v}|^2 + \int_\Omega |\nabla \mathbf{v}|^2 \\ &= \mathbf{d}^T \overline{M} \mathbf{d} + \mathbf{d}^T \overline{K} \mathbf{d}, \end{aligned}$$

so the ellipticity condition says that if $\mathbf{d}$ satisfies $\overline{H}\mathbf{d} = 0$ and $\overline{R}\mathbf{d} = 0$ then

$$\mathbf{d}^T \overline{K} \mathbf{d} \geq C||\mathbf{d}||^2_{H_0^1(\Omega)^3} = C(\mathbf{d}^T \overline{M} \mathbf{d} + \mathbf{d}^T \overline{K} \mathbf{d}).$$

The form a_1 satisfies the following important property

$$a_1(\mathbf{w}; \mathbf{v}, \mathbf{v}) = 0.$$

when $\mathbf{v}, \mathbf{w}$ are in $H^1(\Omega)^3$ with div $\mathbf{w} = 0$ and $\mathbf{w} \cdot \mathbf{n}_{|\partial\Omega} = 0$. In particular if $\mathbf{e}$ satisfies

$$D\mathbf{e} = 0, \qquad \overline{H}\mathbf{e} = 0, \qquad \text{and } \overline{R}\mathbf{e} = 0 \tag{6.38}$$

which can also be written

$$L\mathbf{e} = 0,$$

with $\mathbf{L} = \begin{pmatrix} \overline{H} \\ \overline{R} \\ D \end{pmatrix}$, and $\mathbf{d}$ satisfies $\overline{H}\mathbf{d} = 0$, we have

$$(\overline{B}(\mathbf{e})\mathbf{d})^T\mathbf{d} = 0 \tag{6.39}$$

using the expression of a_1 in (6.32). We also point out that the trilinear form a_1 is continuous on $H^1(\Omega)^3 \times H^1(\Omega)^3 \times H^1(\Omega)^3$, that is there is $C_1 > 0$ such that

$$|a_1(\mathbf{w};\mathbf{u},\mathbf{v}) = |(\overline{B}(\mathbf{e})\mathbf{c})^T\mathbf{d}| \leq C_1||\mathbf{e}||_{H^1(\Omega)^3}||\mathbf{c}||_{H^1(\Omega)^3}||\mathbf{d}||_{H^1(\Omega)^3}.$$

These results can be found in [Girault and Raviart'86].

We derive an a priori error estimate of a solution of (6.33). We rewrite it as

$$\begin{aligned} \nu\overline{K}\mathbf{c} + \overline{B}(\mathbf{c})\mathbf{c} + L^T\lambda &= \overline{M}\mathbf{F} \\ L\mathbf{c} &= 0, \end{aligned} \tag{6.40}$$

with $\lambda = \begin{pmatrix} \lambda_1, & \lambda_2, & \lambda_3 \end{pmatrix}^T$. We multiply the first of (6.40) on the left by $\mathbf{c}^T$ and get

$$\nu\mathbf{c}^T\overline{K}\mathbf{c} = \mathbf{c}^T\overline{M}\mathbf{F}.$$

Using the ellipticity of a_0, we have

$$\nu C||\mathbf{c}||^2_{H^1(\Omega)^3} \leq ||\mathbf{c}||_{L^2(\Omega)^3}||\mathbf{F}||_{L^2(\Omega)^3},$$

so

$$||\mathbf{c}||_{H^1(\Omega)^3} \leq \frac{1}{\nu C}||\mathbf{F}||_{L^2(\Omega)^3}. \tag{6.41}$$

Uniqueness of the discrete velocity

We shall prove that there's a unique vector $\mathbf{c}$ solution of (6.33). The Lagrange multipliers λ_1, λ_2 and λ_3 may not be unique.

Let $(\mathbf{u}_1, \lambda_1, \lambda_2, \lambda_3)$ and $(\mathbf{u}_2, \beta_1, \beta_2, \beta_3)$ be two solutions of (6.33). Put $\epsilon = \mathbf{u}_1 - \mathbf{u}_2$. We have

$$\nu\overline{K}\epsilon + \overline{B}(\mathbf{u}_1)\mathbf{u}_1 - \overline{B}(\mathbf{u}_2)\mathbf{u}_2 + \overline{H}^T(\lambda_1 - \beta_1) + \overline{R}^T(\lambda_2 - \beta_2) + D^T(\lambda_3 - \beta_3) = 0$$

$$\overline{H}\mathbf{c} = 0$$

$$\overline{R}\mathbf{c} = 0$$

$$D\mathbf{c} = 0.$$

Multiplying the first of these equations on the left by ϵ^T, we get

$$\nu\epsilon^T\overline{K}\epsilon + \epsilon^T\overline{B}(\mathbf{u}_1)\epsilon + \epsilon^T\overline{B}(\epsilon)\mathbf{u}_2 = 0,$$

or

$$\nu\epsilon^T\overline{K}\epsilon + \epsilon^T\overline{B}(\epsilon)\mathbf{u}_2 = 0.$$

Using (6.41), the ellipticity of a_0 and the continuity of a_1 we get

$$\begin{aligned}\nu C||\epsilon||^2_{H^1(\Omega)^3} &\leq C_1||\epsilon||^2_{H^1(\Omega)^3}||\mathbf{u}_2||_{H^1(\Omega)^3} \\ &\leq \frac{C_1}{\nu C}||\mathbf{F}||_{L^2(\Omega)^3}||\epsilon||^2_{H^1(\Omega)^3}.\end{aligned}$$

So

$$(\nu C - \frac{C_1}{\nu C}||\mathbf{F}||_{L^2(\Omega)^3})||\epsilon||^2_{H^1(\Omega)^3} \leq 0.$$

We conclude that if $||\mathbf{F}||_{L^2(\Omega)^3}$ is sufficiently small, i.e.

$$||\mathbf{F}||_{L^2(\Omega)^3} \leq \frac{(\nu C)^2}{C_1},$$

the discrete equations have a unique solution $\mathbf{c}$.

To prove existence of a solution to (6.40), we shall need one additional lemma which can be found in [Temam'84].

Lemma 6.3.2 *(Brouwer fixed point theorem) Let H be a finite-dimensional Hilbert space with inner product $(.,.)$ and norm $||.||$. Let the map $g : H \to H$ be continuous. Suppose there exists $\eta > 0$ such that $(g(x), x) > 0$ for each x such that $||x|| = \eta$. Then there exists $x^* \in H$, with $||x^*|| \leq \eta$ such that $g(x^*) = 0$.*

With $\mathbf{c}$ fixed let us consider the mapping $\overline{J}$ defined on

$$S_0 = \{\mathbf{d} \in (\mathbf{R})^{3N}, L\mathbf{d} = 0\}$$

by

$$\overline{J}(\mathbf{d}) = \mathbf{d}^T \left(\nu \overline{K}\mathbf{c} + \overline{B}(\mathbf{c})\mathbf{c} - \overline{M}\mathbf{F}\right).$$

Owing to the continuity property of the trilinear form b, $\overline{J}$ is continuous on S_0 equipped with the norm of $H_0^1(\Omega)^3$. Let $(.,.)$ denote the associated inner product. By the Riesz representation theorem, there is a function $\mu : S_0 \to S_0$ such that

$$(\mathbf{d}, \mu(\mathbf{c})) = \overline{J}(\mathbf{d}).$$

We have by (6.39),

$$\begin{aligned}(\mathbf{c}, \mu(\mathbf{c})) = \overline{J}(\mathbf{c}) &= \nu \mathbf{c}^T \overline{K}\mathbf{c} - \mathbf{c}^T \overline{M}\mathbf{F} \\ &\geq \nu C ||\mathbf{c}||^2_{H_0^1(\Omega)^3} - ||\mathbf{c}||_{H_0^1(\Omega)^3} C_M ||\mathbf{F}||_{L^2(\Omega)^3} \\ &= ||\mathbf{c}||_{H_0^1(\Omega)^3} \left(\nu C ||\mathbf{c}||_{H_0^1(\Omega)^3} - C_M ||\mathbf{F}||_{L^2(\Omega)^3}\right),\end{aligned}$$

where C_M is the L_2 norm of the mass matrice $\overline{M}$. Therefore for

$$||\mathbf{c}||_{H_0^1(\Omega)^3} = 2\frac{C_M ||\mathbf{F}||_{L^2(\Omega)^3}}{\nu C} =: r,$$

we have $(\mathbf{c}, \mu(\mathbf{c})) > 0$. It follows from the Brouwer fixed point theorem that, there is $\mathbf{e}$ with $||\mathbf{e}||_{H_0^1(\Omega)^3} \leq r$ such that $\mu(\mathbf{e}) = 0$. This implies that $\overline{J}(\mathbf{d}) = 0$ on S_0. By the Lagrange multiplier method, there is λ for which $\overline{J}(\mathbf{d}) + \mathbf{d}^T L^T \lambda = 0$. This gives after simplification (6.40).

We have proved the following theorem:

Theorem 6.3.3 *(6.33) has a unique solution* $\mathbf{c}$ *provided the (vector) spline encoded in* $\mathbf{F}$ *has a sufficiently small* L^2 *norm or* ν *is small enough.*

Convergence of the simple iteration algorithm

The problem is to show convergence of the solution $\mathbf{c}^{n+1}$ of

$$\begin{aligned} \nu\overline{K}\mathbf{c}^{n+1} + \overline{B}(\mathbf{c}^n)\mathbf{c}^{n+1} + L^T\lambda^{\mathbf{n+1}} &= \overline{M}\mathbf{F} \\ L\mathbf{c}^{n+1} &= 0, \end{aligned} \tag{6.42}$$

to the solution $\mathbf{c}$ of (6.40) where we put $\lambda^{\mathbf{n+1}} = \left(\lambda_1^{n+1}, \ \lambda_2^{n+1}, \ \lambda_3^{n+1} \right)^T$. We have the following theorem

Theorem 6.3.4 *(6.42) has a unique solution* $\mathbf{c}^{n+1}$ *and the unique solution* $\mathbf{c}$ *of (6.33) is such that*

$$||\mathbf{c}^{n+1} - \mathbf{c}||_{H^1(\Omega)^3} \leq \gamma_1 ||\mathbf{c}^n - \mathbf{c}||_{H^1(\Omega)^3},$$

for a constant $\gamma_1 < 1$. *As a consequence* $\mathbf{c}^{n+1}$ *converges to* $\mathbf{c}$.

First let's show that (6.42) has a unique solution $\mathbf{c}^{n+1}$. Let $\mathbf{d}^{n+1}$ be another solution and β^{n+1} the associated Lagrange multiplier. Put $\epsilon^{n+1} = \mathbf{c}^{n+1} - \mathbf{d}^{n+1}$. Also put $\tau^{n+1} = \lambda^{n+1} - \beta^{n+1}$. We have

$$L\epsilon^{n+1} = 0 \tag{6.43}$$

and

$$\nu\overline{K}\epsilon^{n+1} + \overline{B}(\mathbf{c}^n)\epsilon^{n+1} + L^T\tau^{n+1} = 0.$$

Multiplying this last relation by $(\epsilon^{n+1})^T$ on the right and using (6.43) and (6.39), we get

$$(\epsilon^{n+1})^T\overline{K}\epsilon^{n+1} = 0$$

and ϵ^{n+1} satisfies $\overline{H}\epsilon^{n+1} = 0$ and $\overline{R}\epsilon^{n+1} = 0$. Therefore $\epsilon^{n+1} = 0$ since if $\mathbf{v}$ is smooth, zero on the boundary and $\int_\Omega |\nabla \mathbf{v}|^2 = 0$, $\mathbf{v} = 0$. This proves uniqueness. Existence can be proven as in the proof of Theorem (6.3.3).

Now, put $\epsilon^{n+1} = \mathbf{c}^{n+1} - \mathbf{c}$ and $\tau^{n+1} = \lambda^{n+1} - \lambda$. We have

$$\begin{aligned} \nu\overline{K}\epsilon^{n+1} + \overline{B}(\mathbf{c}^n)\mathbf{c}^{n+1} - \overline{B}(\mathbf{c})\mathbf{c} + L^T\tau^{n+1} &= 0, \\ L\epsilon^{n+1} &= 0, \end{aligned}$$

so

$$\begin{aligned} \nu\overline{K}\epsilon^{n+1} + \overline{B}(\mathbf{c}^n)\epsilon^{n+1} + (\overline{B}(\mathbf{c}^n) - \overline{B}(\mathbf{c}))\mathbf{c} + L^T\tau^{n+1} &= 0, \\ L\epsilon^{n+1} &= 0. \end{aligned}$$

Multiplying on the left by $(\epsilon^{n+1})^T$, after simplifications, we get

$$\nu(\epsilon^{n+1})^T\overline{K}\epsilon^{n+1} + (\epsilon^{n+1})^T(\overline{B}(\mathbf{c}^n) - \overline{B}(\mathbf{c})\mathbf{c}) = 0.$$

We notice that

$$\overline{B}(\mathbf{c}^n) - \overline{B}(\mathbf{c}) = \overline{B}(\mathbf{c}^n - \mathbf{c}) = \overline{B}(\epsilon^n).$$

It is actually convenient when necessary to use the forms a_0 and a_1 up to an identification of the splines with their B-forms. The previous relations then reads

$$\nu a_0(\epsilon^{n+1}, \epsilon^{n+1}) = -a_1(\epsilon^n; \mathbf{c}, \epsilon^{n+1})$$

which joined with the continuity of the form a_1 and the ellipticity of a_0 yields

$$C\nu||\epsilon^{n+1}||^2_{H^1(\Omega)^3} \leq C_1||\epsilon^n||_{H^1(\Omega)^3}||\epsilon^{n+1}||_{H^1(\Omega)^3}||\mathbf{c}||_{H^1(\Omega)^3}. \tag{6.44}$$

So

$$\begin{aligned} ||\epsilon^{n+1}||^2_{H^1(\Omega)^3} &\leq \frac{C_1}{C\nu}||\epsilon^n||_{H^1(\Omega)^3}||\mathbf{c}||_{H^1(\Omega)^3} \\ &\leq \gamma_1||\epsilon^n||_{H^1(\Omega)^3}, \end{aligned}$$

where $\gamma_1 = \frac{C_1||\mathbf{F}||_{L^2(\Omega)^3}}{(\nu C)^2}$. We have $\gamma < 1$ under the assumption that $||\mathbf{F}||_{L^2(\Omega)^3}$ is sufficiently small.

Convergence of Newton's iterations

We are interested in the sequence $\mathbf{c}^{n+1}$ defined by

$$\begin{aligned} \nu\overline{K}\mathbf{c}^{n+1} + \overline{B}(\mathbf{c}^n)\mathbf{c}^{n+1} + \widetilde{B}(\mathbf{c}^n)\mathbf{c}^{n+1} + L^T\lambda^{n+1} &= \overline{M}\mathbf{F} + \overline{B}(\mathbf{c}^n)\mathbf{c}^n \\ L\mathbf{c}^{n+1} &= \overline{\mathbf{G}}. \end{aligned} \tag{6.45}$$

We have the following convergence result

Theorem 6.3.5 *There exists $r > 0$ such that if $||\mathbf{c}-\mathbf{c}^0||_{H^1(\Omega)^3} < r$, there is a unique $\mathbf{c}^{n+1}$ solution of (6.45) and $||\mathbf{c} - \mathbf{c}^n||_{H^1(\Omega)^3} < r$ for all n with $||\mathbf{c} - \mathbf{c}^{n+1}||_{H^1(\Omega)^3} \leq \frac{1}{r}||\mathbf{c} - \mathbf{c}^n||^2_{H^1(\Omega)^3}$. Moreover, if there's $\eta < 1$ such that $||\mathbf{c} - \mathbf{c}^0||_{H^1(\Omega)^3} = r\eta$, then $\mathbf{c}^n$ converges to $\mathbf{c}$.*

Proof The existence of $\mathbf{c}^{n+1}$ in (6.45) can be proven as in the proof of Theorem (6.3.3). We prove first uniqueness. Given $\mathbf{c}^n$, let $\mathbf{d}^{n+1}$ be another solution of (6.45) and β^{n+1} the associated Lagrange multiplier. Put $\epsilon^{n+1} = \mathbf{c}^{n+1} - \mathbf{d}^{n+1}$. We have:

$$\begin{aligned} \nu\overline{K}\epsilon^{n+1} + \overline{B}(\mathbf{c}^n)\epsilon^{n+1} + \widetilde{B}(\mathbf{c}^n)\epsilon^{n+1} + L^T\lambda^{n+1} &= 0 \\ L\epsilon^{n+1} &= 0. \end{aligned}$$

Therefore

$$\begin{aligned} 0 &= \nu(\epsilon^{n+1})^T\overline{K}\epsilon^{n+1} + (\epsilon^{n+1})^T\overline{B}(\epsilon^{n+1})\mathbf{c}^n \\ &\geq ||\epsilon^{n+1}||^2_{H^1(\Omega)^3}(\nu C - C_1||\mathbf{c}^n||_{H^1(\Omega)^3}) \\ &\geq ||\epsilon^{n+1}||^2_{H^1(\Omega)^3}\big(\nu C - C_1||\mathbf{c} - \mathbf{c}^n||_{H^1(\Omega)^3} - C_1||\mathbf{c}||^2_{H^1(\Omega)^3}\big). \end{aligned}$$

Using the a priori error estimate (6.41), this gives

$$0 \geq ||\epsilon^{n+1}||^2_{H^1(\Omega)^3}\big(\nu C - C_1||\mathbf{c} - \mathbf{c}^n||_{H^1(\Omega)^3} - \frac{C_1}{\nu C}||\mathbf{F}||_{L^2(\Omega)^3}\big).$$

We want to show that

$$\nu C - C_1||\mathbf{c}-\mathbf{c}^n||_{H^1(\Omega)^3} - \frac{C_1}{\nu C}||\mathbf{F}||_{L^2(\Omega)^3} \geq 0,$$

which is equivalent to

$$||\mathbf{c}-\mathbf{d}^n||_{H^1(\Omega)^3} \leq \frac{-C_1||\mathbf{F}||_{L^2(\Omega)^3} + \nu^2 C^2}{\nu C C_1}.$$

We therefore let $r = \frac{1}{2}\frac{-C_1||\mathbf{F}||_{L^2(\Omega)^3})+\nu^2C^2}{\nu C C_1}$ and prove by induction that if $||\mathbf{c}-\mathbf{c}^0||_{H^1(\Omega)^3} < r$, then $||\mathbf{c}-\mathbf{c}^n||_{H^1(\Omega)^3} < r$ for all n. This will show that $\epsilon^{n+1} = 0$ and that the solution of (6.45) is unique. Notice that $r > 0$ under the assumption that $||\mathbf{F}||_{L^2(\Omega)^3}$ is sufficiently small. That assumption is actually needed to guarantee the uniqueness of the discrete solution.

Let $\mathbf{e}^{n+1} = \mathbf{c}^{n+1} - \mathbf{c}$ and $\tau^{n+1} = \lambda^{n+1} - \lambda$. We have

$$\begin{aligned}\nu\overline{K}\mathbf{e}^{n+1} + \overline{B}(\mathbf{c}^n)\mathbf{c}^{n+1} + \overline{B}(\mathbf{c}^{n+1})\mathbf{c}^n - \overline{B}(\mathbf{c})\mathbf{c} + L^T\tau^{n+1} &= \overline{B}(\mathbf{c}^n)\mathbf{c}^n \\ L\mathbf{e}^{n+1} &= 0,\end{aligned}$$

which gives

$$\nu\overline{K}\mathbf{e}^{n+1} + \overline{B}(\mathbf{c}^n)\mathbf{e}^{n+1} + \overline{B}(\mathbf{e}^n)\mathbf{c} + \overline{B}(\mathbf{c}^{n+1})\mathbf{c}^n = \overline{B}(\mathbf{c}^n)\mathbf{c}^n,$$

or

$$\nu\overline{K}\mathbf{e}^{n+1} + \overline{B}(\mathbf{c}^n)\mathbf{e}^{n+1} + \overline{B}(\mathbf{e}^n)\mathbf{c} + \overline{B}(\mathbf{e}^{n+1})\mathbf{c}^n - \overline{B}(\mathbf{e}^n)\mathbf{c}^n = 0.$$

We multiply this equation on the left by $(\mathbf{e}^{n+1})^T$, and get

$$\nu(\mathbf{e}^{n+1})^T\overline{K}\mathbf{e}^{n+1} = (\mathbf{e}^{n+1})^T\overline{B}(\mathbf{e}^n)\mathbf{e}^n - (\mathbf{e}^{n+1})^T\overline{B}(\mathbf{e}^{n+1})\mathbf{c}^n.$$

We therefore have

$$\begin{aligned}\nu C||\mathbf{e}^{n+1}||_{H^1(\Omega)^3} &\leq C_1\big(||\mathbf{e}^n||^2_{H^1(\Omega)^3} + ||\mathbf{e}^{n+1}||_{H^1(\Omega)^3}||\mathbf{c}^n||_{H^1(\Omega)^3}\big) \\ &\leq C_1\big(||\mathbf{e}^n||^2_{H^1(\Omega)^3} + ||\mathbf{e}^{n+1}||_{H^1(\Omega)^3}(||\mathbf{e}^n||_{H^1(\Omega)^3} + ||\mathbf{c}||_{H^1(\Omega)^3})\big).\end{aligned}$$

Using the a priori estimate (6.41) and the induction hypothesis, we have

$$\big(\nu C - C_1 r - C_1 \frac{||\mathbf{F}||_{L^2(\Omega)^3}}{\nu C}\big)||\mathbf{e}^{n+1}||_{H^1(\Omega)^3} \leq C_1 ||\mathbf{e}^n||^2_{H^1(\Omega)^3}.$$

A tedious computation yields

$$\frac{C_1}{\nu C - C_1 r - C_1 \frac{||\mathbf{F}||_{L^2(\Omega)^3}}{\nu C}} = \frac{1}{r}.$$

So

$$||\mathbf{e}^{n+1}||_{H^1(\Omega)^3} \leq \frac{1}{r}||\mathbf{e}^n||^2_{H^1(\Omega)^3} \leq r$$

completing the induction argument.

We have proved that $||\mathbf{c} - \mathbf{c}^{n+1}||_{H^1(\Omega)^3} \leq \frac{1}{r}||\mathbf{c} - \mathbf{c}^n||^2_{H^1(\Omega)^3}$. It follows that

$$\begin{aligned} ||\mathbf{c} - \mathbf{c}^n||_{H^1(\Omega)^3} &\leq \prod_{k=1}^{n} \frac{1}{r^{2^{k-1}}}||\mathbf{c} - \mathbf{c}^0||^{2^n}_{H^1(\Omega)^3} \\ &\leq \frac{1}{r^{\sum_{k=1}^{n} 2^{k-1}}}||\mathbf{c} - \mathbf{c}^0||^{2^n}_{H^1(\Omega)^3} \\ &\leq \frac{1}{r^{2^n - 1}}||\mathbf{c} - \mathbf{c}^0||^{2^n}_{H^1(\Omega)^3}, \end{aligned}$$

for $n = 1, 2, \ldots$ This shows that if there's $\eta < 1$ such that $||\mathbf{c} - \mathbf{c}^0||_{H^1(\Omega)^3} = r\eta$, then $\mathbf{c}^n$ converges to $\mathbf{c}$. Indeed in that case, we have

$$||\mathbf{c} - \mathbf{c}^n||_{H^1(\Omega)^3} \leq r\eta^{2^n}, \quad n = 1, 2, \ldots$$

Approximations of the pressure

It is known that the pressure satisfies a Poisson equation which is obtained by taking the divergence of the first equation in (6.12). We have

$$-\Delta p = -\text{div } \mathbf{f} \ + \ \text{div } (\mathbf{u} \cdot \nabla)\mathbf{u}$$

since div $\mathbf{u} = 0$. This equation is supplied with Neumann boundary conditions

$$\frac{\partial p}{\partial \mathbf{n}} = \nabla p \cdot \mathbf{n} = \mathbf{f} \cdot \mathbf{n} + \nu(\Delta \mathbf{u}) \cdot \mathbf{n} - ((\mathbf{u} \cdot \nabla)\mathbf{u}) \cdot \mathbf{n}$$

and is solved numerically for p using the techniques described in the section on the Poisson equation.

6.3.3 Numerical results

We have experimented the code on different tetrahedral partitions with different smoothness and different degrees. We choose known vector fields which are divergence free and compute the errors in the infinite norm. We first list numerical results corresponding to continuous vector fields and pressure of class C^1 across tetrahedral elements which do not share a face with the boundary. This demonstrates the choice we have in selecting which amount of smoothness can be required of the approximant. Typically one would construct piecewise continuous approximations of the pressure. Here the viscosity is set to 1.

Domain 1: This domain is formed by the union of two tetrahedra which share a common face.

Domain 2: We consider a cube of volume one which has been subdivided into six tetrahedra.

We consider three different vector fields $\mathbf{g} = (g_1, g_2, g_3)$ with a corresponding pressure p on the previous domains.

Case 1:

$$
\begin{aligned}
g_1 &= 2\,(y-z)\exp(x^2+y^2+z^2) \\
g_2 &= -2\,(x-z)\exp(x^2+y^2+z^2) \\
g_3 &= 2\,(x-y)\exp(x^2+y^2+z^2 \\
p &= \exp(x+y+z)
\end{aligned}
$$

Case 2:

$$\begin{aligned} g_1 &= x(1-x)y(1-y)z(1-z) \\ g_2 &= x(1-x)y(1-y)z(1-z)) \\ g_3 &= \tfrac{1}{6}z^2(y+x-1)(-x+2xy-y)(2z-3) \\ p &= \exp(x+y+z) \end{aligned}$$

Case 3:

$$\begin{aligned} g_1 &= 1/(1+x+y+z) \\ g_2 &= 1/(1+x+y+2*z) \\ g_3 &= -1/(1+x+y+z) - \tfrac{1}{2}1/(1+x+y+2z) \\ p &= \exp(x+y+z) \end{aligned}$$

These functions have been chosen so that the nonlinear term in the Navier-Stokes equations do not vanish. In the following tables, Error 1, Error 2 and Error 3 are the errors made in computing f_1, f_2 and f_3 respectively. We also indicate the error on the pressure term. The first time a domain appears, the size of the matrix which was solved is listed.

Case 1 on Domain 1 with $d = 3$

Tetrahedra	Size	Error 1	Error 2	Error 3	Pressure
2	290 × 290	7.9188e-01	1.8725e-01	8.8710e-01	3.4535e+01
16	2680 × 2680	3.9973e-01	1.0744e-01	3.9027e-01	1.5749e+01

Case 1 on Domain 1 with $d = 4$

Tetrahedra	Size	Error 1	Error 2	Error 3	Pressure
2	505 × 505	2.9228e-01	8.8375e-02	3.4786e-01	1.4645e+01
16	4580 × 4580	6.3349e-02	2.6866e-02	6.2141e-02	2.2879e+00

Case 1 on Domain 1 with $d = 5$

Tetrahedra	Size	Error 1	Error 2	Error 3	Pressure
2	805 × 805	2.8805e-01	1.8885e-01	2.8456e-01	5.9541e+00
16	7196 × 7196	8.4794e-03	3.9548e-03	8.5251e-03	6.7550e-01

Case 1 on Domain 1 with $d = 6$

Tetrahedra	Size	Error 1	Error 2	Error 3	Pressure
2	1204 × 1204	5.4170e-02	5.0070e-02	5.2498e-02	2.0899e+00
16	10640 × 10640	1.3485e-03	6.6982e-04	1.1929e-03	1.5991e-01

Case 1 on Domain 2 with $d = 3$

Tetrahedra	Size	Error 1	Error 2	Error 3	Pressure
6	960 × 960	5.3508e-01	6.8599e-01	6.8599e-01	3.9479e+01
48	8400 × 8400	3.6084e-01	3.6084e-01	3.6084e-01	9.2745e+00

Case 1 on Domain 2 with $d = 4$

Tetrahedra	Size	Error 1	Error 2	Error 3	Pressure
6	1650 × 1650	9.9685e-01	9.9685e-01	9.9685e-01	1.1079e+01
48	14280 × 14280	7.1163e-02	7.1163e-02	7.1163e-02	1.8966e+00

Case 1 on Domain 2 with $d = 5$

Tetrahedra	Size	Error 1	Error 2	Error 3	Pressure
6	2604 × 2604	1.9913e-01	1.9913e-01	1.9913e-01	4.2036e+00

Case 1 on Domain 2 with $d = 6$

Tetrahedra	Size	Error 1	Error 2	Error 3	Pressure
6	3864 × 3864	4.6539e-02	4.6539e-02	4.6539e-02	1.2909e+00

Case 1 on Domain 2 with $d = 7$

Tetrahedra	Size	Error 1	Error 2	Error 3	Pressure
6	5472 × 5472	1.0646e-02	1.0646e-02	1.0646e-02	6.8582e-01

For the next test vector fields the approximations of the velocity vector field are better and hence we get more accurate approximations of the pressure.

Case 2 on Domain 1 with $d = 3$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
2	5.4823e-03	5.1512e-03	4.8780e-03	2.9863e-01
16	1.2800e-03	1.1522e-03	1.5770e-03	1.6411e-01

Case 2 on Domain 1 with $d = 4$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
2	2.6458e-03	2.3774e-03	2.1849e-03	2.9286e-01
16	4.2213e-04	2.3787e-04	3.7811e-04	5.8395e-02

Case 2 on Domain 1 with $d = 5$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
2	2.1518e-03	1.0184e-03	1.9193e-03	3.1441e-01
16	4.0893e-05	3.0022e-05	3.6272e-05	9.4046e-03

Case 2 on Domain 1 with $d = 6$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
2	1.4059e-08	1.3591e-08	1.4851e-08	1.5114e-03
16	3.2372e-10	3.1183e-10	6.7459e-10	3.9750e-05

Case 2 on Domain 2 with $d = 3$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
6	1.4097e-02	1.6179e-02	1.9143e-02	1.9430e+00
48	1.7648e-03	1.8289e-03	3.4440e-03	5.5006e-01

Case 2 on Domain 2 with $d = 4$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
6	3.4929e-03	3.6200e-03	8.8970e-03	2.5272e-01
48	3.7583e-04	3.7285e-04	4.4346e-04	9.3031e-02

Case 2 on Domain 2 with $d = 5$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
6	2.6001e-03	2.6001e-03	2.8142e-03	6.5987e-02

Case 2 on Domain 2 with $d = 6$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
6	3.0414e-07	3.0414e-07	2.7326e-07	1.3856e-02

Case 2 on Domain 2 with $d = 7$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
6	3.4135e-08	3.4135e-08	4.2451e-08	2.6998e-03

Case 3 on Domain 1 with $d = 3$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
2	1.3220e-02	2.8758e-02	2.2337e-02	2.5579e+00
16	5.9963e-03	6.1040e-03	4.7915e-03	9.4864e-01

Case 3 on Domain 1 with $d = 4$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
2	7.2904e-03	9.0516e-03	8.8195e-03	1.3839e+00
16	7.7037e-04	1.4633e-03	1.2568e-03	2.4490e-01

Case 3 on Domain 1 with $d = 5$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
2	2.2836e-03	4.0212e-03	3.0422e-03	2.7893e-01
16	2.1804e-04	3.2009e-04	3.0179e-04	5.4180e-02

Case 3 on Domain 1 with $d = 6$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
2	1.0482e-03	1.5710e-03	1.3561e-03	1.8802e-01
16	6.2463e-05	6.3539e-05	6.7752e-05	2.1708e-02

Case 3 on Domain 2 with $d = 3$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
6	3.3312e-02	4.0606e-02	5.0388e-02	2.9399e+00
48	2.3622e-02	2.4789e-02	3.0950e-02	1.2750e+00

Case 3 on Domain 2 with $d = 4$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
6	3.0671e-02	3.7226e-02	6.6432e-02	9.4831e-01
48	3.0258e-03	4.1415e-03	6.4099e-03	1.9938e-01

Case 3 on Domain 2 with $d = 5$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
6	7.5366e-03	9.6083e-03	1.7657e-02	3.7409e-01

Case 3 on Domain 2 with $d = 6$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
6	2.1930e-03	3.0274e-03	5.7868e-03	1.9517e-01

Case 3 on Domain 2 with $d = 7$

Tetrahedra	Error 1	Error 2	Error 3	Pressure
6	7.5139e-04	1.1032e-03	1.8545e-03	7.0576e-02

Table 6.1: Accuracy of the velocity and pressure, $r = 0$

Tetrahedra	Size	Error 1	Error 2	Error 3	Pressure
6	961×960	1.4097e-002	1.6179e-002	1.9143e-002	3.4500e-001
12	2101×2100	1.4561e-003	1.4561e-003	1.7462e-003	4.1284e-001
24	4291×4290	1.4511e-003	1.4190e-003	1.5752e-003	2.3806e-001
48	8401×8400	1.0339e-003	1.0308e-003	1.5299e-003	1.6460e-001
96	16441×16440	9.0358e-004	9.0346e-004	6.6460e-004	1.6278e-001
192	31621×31620	1.5348e-004	1.5121e-004	2.3278e-004	5.1888e-002

We proceed to give additional numerical results for Case 2 on Domain 2. Now, on a cube initially subdivided into six sub-tetrahedra, we used the matrix iterative algorithm. The bisection method of refinement was used. The purpose of displaying these results is to show several level of refinements. We fixed the degree of the splines to 3 and vary the smoothness. We set $r = 0, 1$ and 2. However the pressure is only continuous. We also display here how accurate the divergence condition, the boundary conditions and smoothness conditions are. Although the reduced algorithm was used, we give here the size the matrices that are solved when using the other methods. The real size was simply 20 times the number of tetrahedra where 20 is the dimension of the space of polynomials of total degree 3. Error 1, Error 2 and Error 3 denote the errors on the first, second and third component of the velocity respectively. Where we judged necessary, we also display the maximum shape measure σ of the tetrahedral partition. The variations of σ is certainly the reason of oscillations in some of the approximations. Note that if a uniform refinement in eight tetrahedra is used, the approximations behave in a monotone way.

Table 6.2: Accuracy of the smoothness conditions, $r = 0$

Tetrahedra	Sigma	Error 1	Error 2	Error 3
6	4.1815	6.4152e-003	6.4152e-003	6.4152e-003
12	4.4142	2.0286e-003	2.0286e-003	4.5644e-004
24	3.8284	5.2176e-004	5.2176e-004	2.0871e-004
48	4.1815	4.4773e-004	4.4773e-004	2.9773e-004
96	4.4142	6.5638e-004	6.5638e-004	2.1166e-004
192	3.8284	4.2883e-005	4.2883e-005	2.0618e-005

Table 6.3: Accuracy of the boundary conditions, $r = 0$

Tetrahedra	Sigma	Error 1	Error 2	Error 3
6	4.1815	6.4152e-003	6.4152e-003	7.3923e-003
12	4.4142	8.8082e-004	8.8082e-004	1.2172e-003
24	3.8284	4.0930e-004	4.0930e-004	4.0930e-004
48	4.1815	4.4773e-004	4.4773e-004	4.4773e-004
96	4.4142	3.2819e-004	3.2819e-004	3.2819e-004
192	3.8284	3.6039e-005	3.6039e-005	3.6039e-005

6.4 Numerical simulation of fluid flows

Our final numerical experiment is the calculation of a flow in a cavity. The cavity domain Ω is the unit cube and the flow is caused by a tangential velocity applied to the side $y = 1$. We assume that all external forces vanish. Since they are independent of time, the flow limits to a steady state modelled by (6.12). For the boundary conditions, we take $\mathbf{g} = (g_1, g_2, g_3)$ with $g_2 = g_3 = 0$ and $g_1 = 0$ except on the side $y = 1$ where $g_1 = 1$. We have displayed the configuration of the flow for Reynolds number 400, in the center plane $z = \frac{1}{2}$ using the first and second component of the velocity, in the center plane $x = \frac{1}{2}$ using the second and third components and finally in the center plane $y = \frac{1}{2}$ using the first and third components. We used

Table 6.4: Accuracy of the divergence free condition, $r = 0$

Tetrahedra	Divergence
6	2.3303e-003
12	4.4041e-004
24	1.0435e-004
48	7.4622e-005
96	5.4698e-005
192	4.2883e-006

Table 6.5: Accuracy of the velocity and pressure, $r = 1$

Tetrahedra	Sigma	Error 1	Error 2	Error 3	Pressure
6	4.1815	6.0223e-003	6.2017e-003	1.2326e-002	3.0920e-001
12	4.4142	2.3063e-003	2.2623e-003	1.5898e-003	3.7065e-001
24	3.8284	1.7874e-003	1.8019e-003	1.4641e-003	2.8609e-001
48	4.1815	1.1168e-003	1.1432e-003	1.4807e-003	1.9697e-001
96	4.4142	1.6100e-003	1.6127e-003	7.5146e-004	1.2666e-001

Table 6.6: Accuracy of the smoothness conditions, $r = 1$

Tetrahedra	Sigma	Error 1	Error 2	Error 3
6	4.1815	8.5805e-003	7.0732e-003	1.2994e-002
12	4.4142	2.1879e-003	2.2069e-003	1.6728e-003
24	3.8284	8.7277e-004	8.7071e-004	9.0578e-004
48	4.1815	1.0591e-003	1.0608e-003	1.1549e-003
96	4.4142	6.7895e-004	6.9228e-004	6.1978e-004

Table 6.7: Accuracy of the boundary conditions, $r = 1$

Tetrahedra	Sigma	Error 1	Error 2	Error 3
6	4.1815	1.0924e-002	1.1170e-002	1.1592e-002
12	4.4142	2.6329e-003	2.3176e-003	2.8183e-003
24	3.8284	2.3098e-003	2.1577e-003	2.0573e-003
48	4.1815	1.0022e-003	9.3040e-004	1.2877e-003
96	4.4142	4.9163e-004	4.8677e-004	8.2081e-004

Table 6.8: Accuracy of the divergence free condition, $r = 1$

Tetrahedra	Divergence
6	3.9109e-003
12	7.3256e-004
24	3.3587e-004
48	1.8103e-004
96	8.3684e-005

Table 6.9: Accuracy of the velocity and pressure, $r = 2$

Tetrahedra	Sigma	Error 1	Error 2	Error 3	Pressure
6	4.1815	6.0223e-003	6.2017e-003	1.2326e-002	3.0920e-001
12	4.4142	2.2747e-003	2.2683e-003	1.5856e-003	3.7023e-001
24	3.8284	1.8173e-003	1.8082e-003	1.4962e-003	2.8492e-001
48	4.1815	1.0874e-003	1.1067e-003	1.4351e-003	1.8918e-001
96	4.4142	1.4268e-003	1.4084e-003	7.5803e-004	1.3249e-001

Table 6.10: Accuracy of the smoothness conditions, $r = 2$

Tetrahedra	Sigma	Error 1	Error 2	Error 3
6	4.1815	8.5805e-003	7.0732e-003	1.2994e-002
12	4.4142	2.1342e-003	2.1335e-003	1.7477e-003
24	3.8284	9.2243e-004	9.3067e-004	9.2330e-004
48	4.1815	1.0547e-003	1.0749e-003	1.2324e-003
96	4.4142	6.2688e-004	5.7562e-004	7.8092e-004

Table 6.11: Accuracy of the boundary conditions, $r = 2$

Tetrahedra	Error 1	Error 2	Error 3
6	1.0924e-002	1.1170e-002	1.1592e-002
12	2.7720e-003	2.6273e-003	2.8825e-003
24	2.3418e-003	2.1429e-003	2.0910e-003
48	9.3480e-004	9.2857e-004	1.3968e-003
96	5.0745e-004	4.9045e-004	8.4306e-004

Table 6.12: Accuracy of the divergence free condition, $r = 2$

Tetrahedra	Divergence
6	3.9109e-003
12	7.3900e-004
24	3.5634e-004
48	2.0256e-004
96	1.2620e-004

continuous splines of degree 7 over the cube subdivided in six tetrahedra. These results agree with the ones of [Wang and Sheu'97] who used quadratic polynomials on cubic elements.

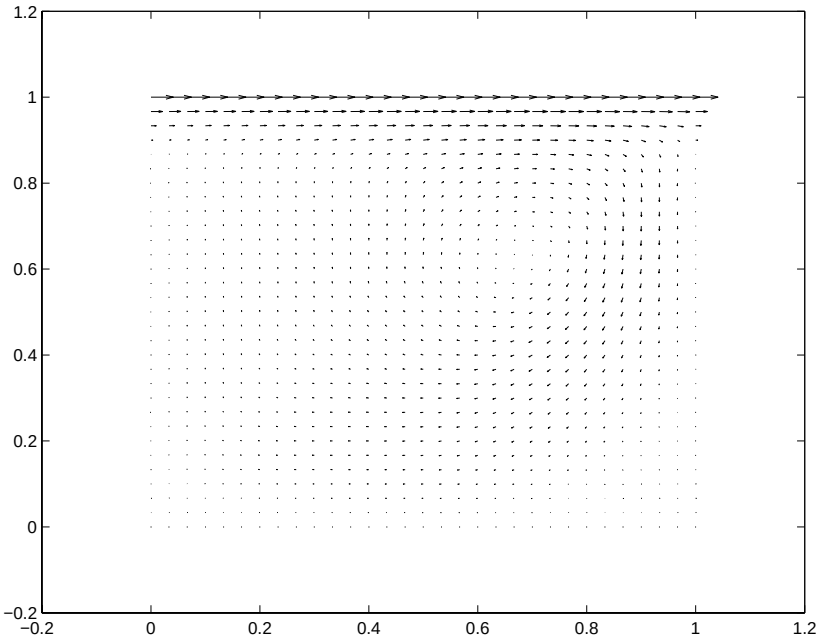

Figure 6.1: Fluid profile in the $x - y$ plane

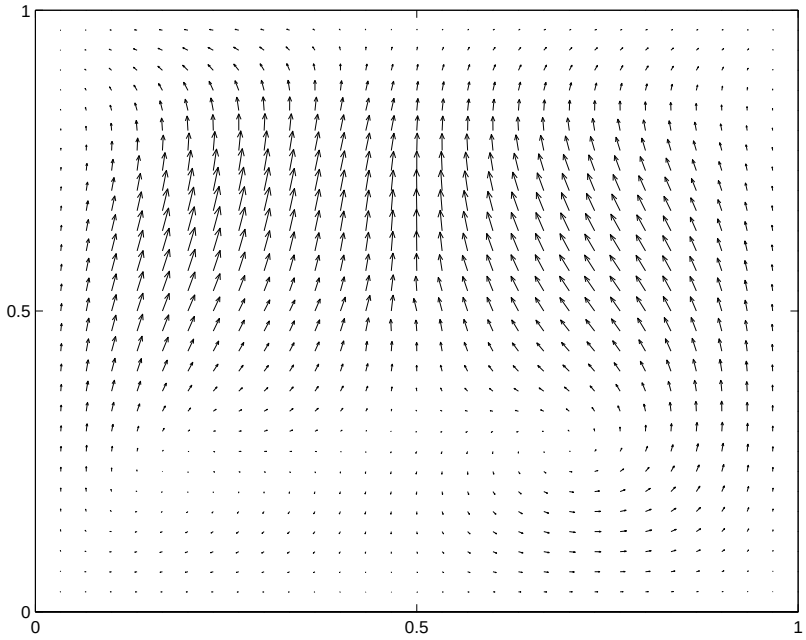

Figure 6.2: Fluid profile in the $y - z$ plane

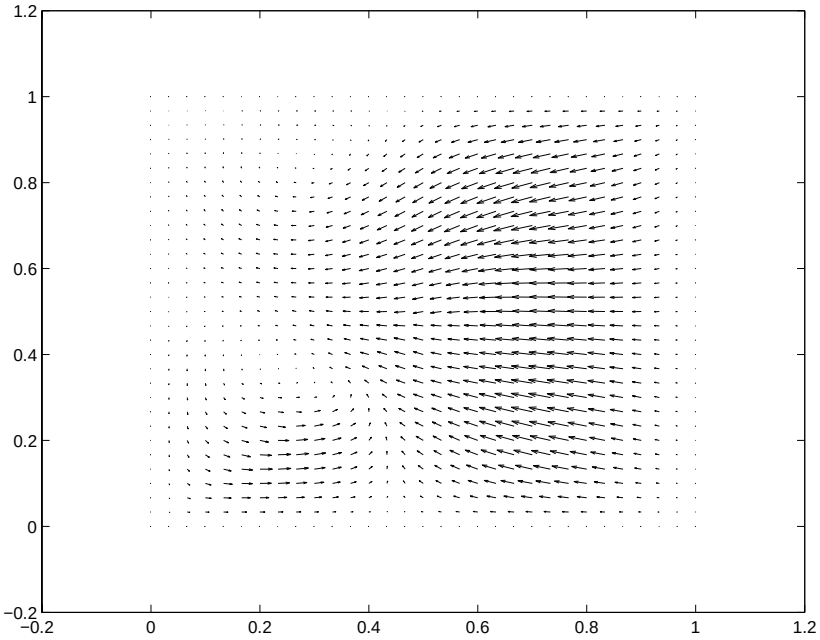

Figure 6.3: Fluid profile in the $x - z$ plane

Chapter 7

Concluding Remarks

We have in this dissertation numerically solve various PDE's using multivariate splines, B-nets and smoothness conditions. Three dimensional problems typically require a lot of memory for computation and this study does not make an exception to this rule. The approximations seem to be very sensitive to the maximum shape measure of the tetrahedral partition. We have identified a refinement strategy and indicate some of its properties suggested by numerical results. When triangulating the unit cube, this algorithm gives the same results as the one in [Ong'94]. The numerical part of this dissertation took a lot of time. When one mistake is made, the programs which already take time to run have to be launched again. Various work can follow this study. We plan to extend these ideas to time-dependent problems, consider the Navier-Stokes equations in stream function formulation and numerically solve the Navier-Stokes equations on an exterior domain.

Bibliography

[1] G.A. Awanou and M.J. Lai, On Convergence Rate of the Augmented Lagrangian Algorithm for Nonsymmetric Saddle Point Problems, 2003. (Preprint)

[2] A. Liu and B. Joe, On the Shape of Tetrahedra from Bisection, Math. of Computation Vol. 63, N° 207, pp 141-154, 1994.

[3] A. Liu and B. Joe, Quality Local Refinement of Tetrahedral Meshes Based on 8-Subtetrahedron Subdivision, Math. of Computation Vol. 65, N° 215, pp 1183-1200, 1996 1994.

[4] I. Babuska, The Finite Element Method with Lagrangian Multipliers, Numer. Math, N° 20, pp 179-192, 1972/73.

[5] I. Babuska and J.M. Melenk, The Partition of Unity Finite Element Method, Comp. Methods in Appl. Mech. Engrg., N° 139, pp 289-314, 1996.

[6] D. Braess, Finite Elements, Springer-Verlag, 1992.

[7] S.C. Brenner and L.R. Scott, The Mathematical Theory of Finite Element Methods, Springer-Verlag, 1994.

[8] P. Ciarlet, Introduction to Numerical Linear Algebra and Optimisation, Cambridge University Press, 1989.

[9] de Boor, C., B-form Basics, Geometric Modeling, edited by G. Farin, SIAM Publication, Philadelphia, pp 131-148 1987.

[10] B. Cockburn, G.E. Karniadakis and C.W. Shu, Discontinuous Galerkin methods, Springer-Verlag, 2000.

[11] C. Doering and J.D. Gibbon, Applied Analysis of the Navier-Stokes Equations, Cambridge University Press, 1995.

[12] M.R. Dorr, On the Discretization of Interdomain Coupling in Elliptic Boundary-value Problems, in Domain Decomposition Methods, T.F. Chan, R. Glowinski, J. Periaux and O.B. Widlund editors, SIAM Publication, Philadelphia, pp 17-46, 1988.

[13] M. Fortin and R. Glowinski, Augmented Lagrangian Methods, Elsevier, 1983.

[14] G.P. Galdi, An introduction to the mathematical theory of the Navier-Stokes equations, Vol I, Springer-Verlag, New York, 1994.

[15] M. Gunzburger, Finite Element Methods for Viscous Incompressible Flows, Academic Press, 1989.

[16] O. Karakashian, On a Galerkin-Lagrange multiplier method for the Stationary Navier-Stokes Equations, SIAM J. of Num. Anal., Vol. 19, N° 5, pp 909-923,1982.

[17] A. Quarteroni and A. Valli, Numerical Approximation of Partial Differential Equations, Springer, 1997.

[18] V. Girault and P.A. Raviart, Finite Element Method for Navier-Stokes Equations, Springer-Verlag, 1986.

[19] M.J. Lai and L. Schumaker, Splines on Triangulations, in Preparation.

[20] P.Lax, N. Milgram, Parabolic Equations, Contributions to the Theory of Partial Differential Equations, Princeton, 1954.

[21] M.E.G. Ong, Uniform Refinement of a Tetrahedron, SIAM J. of Sci. Comput. Vol. 15, N° 5, pp 1134-1144, 1994.

[22] R. Temam, Navier-Stokes Equations. Theory and Numerical Analysis , North-Holland Publishing Co., Amsterdam, 1984.

[23] M.M.T. Wang and T.W.H. Sheu, An Element-by-element BICGSTAB Iterative Method for Three-dimensional Steady Navier-Stokes Equations, J. Comp. Applied Math., Vol 79, pp 147-165, 1997.

[24] S. Zhang, Multi-level Iterative Techniques, Ph.D. Thesis, Dept of Mathematics, Penn. State U., 1988.