

# FEMoctave, Finite Element Algorithms in Octave

Andreas Stahel, Bern University of Applied Sciences

Version 2.1.2, created on 21st August 2023



©Andreas Stahel, 2023

“FEMoctave” by Andreas Stahel, BFH, Biel, Switzerland is licensed under a Creative Commons Attribution-ShareAlike 3.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/3.0/> or send a letter to Creative Commons, 444 Castro Street, Suite 900, Mountain View, California, 94041, USA.

You are free: to copy, distribute, transmit the work, to adapt the work and to make commercial use of the work. Under the following conditions: You must attribute the work to the original author (but not in any way that suggests that the author endorses you or your use of the work). Attribute this work as follows:

Andreas Stahel: FEMoctave, FEM algorithms in Octave.

If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one.

<b>Table of Contents</b>	<b>1</b>
<b>1 Introduction</b>	<b>7</b>
<b>2 The Problems to be Examined</b>	<b>9</b>
2.1 The domain $\Omega \subset \mathbb{R}^2$ and its boundary $\Gamma = \partial\Omega = \Gamma_1 \cup \Gamma_2$	9
2.2 The general elliptic problem	9
2.3 The symmetric elliptic problem	10
2.4 The symmetric eigenvalue problem	10
2.5 The general parabolic problem	10
2.6 The symmetric parabolic problem	10
2.7 The hyperbolic problem	11
2.8 1D boundary value problems	11
2.9 1D initial boundary value problems of order 1	12
2.10 1D initial boundary value problems of order 2	12
2.11 1D eigenvalue value problems	12
2.12 Nonlinear 1D boundary value problems	12
2.13 Plane elasticity	12
2.13.1 Description of strain	13
2.13.2 Description of stress and Hooke's law	13
2.13.3 The plane stress problem	16
2.13.4 The plane strain problem	17
2.14 Elasticity problems for axisymmetric solids, using cylindrical coordinates	18
<b>3 Illustrative Examples</b>	<b>20</b>
3.1 Solving elliptic problems, static heat equations	20
3.1.1 A symmetric problem	20
3.1.2 Laplace equation in cylindrical coordinates	21
3.1.3 Diffusion on an L-shaped domain	22
3.1.4 A diffusion convection problem	23
3.2 Solving eigenvalue problems	24
3.3 Solving parabolic problems, dynamic heat equations	27
3.4 Solving hyperbolic problems, wave equations	27
3.5 Solving 1D steady state boundary value problems	29
3.6 Solving 1D dynamic initial boundary value problems of order 1, a heat equation	32
3.7 Solving 1D dynamic initial boundary value problems of order 2, a wave equation	33
3.8 Solving nonlinear 1D boundary value problems	35
3.8.1 A nonlinear 1D BVP solved by <code>BVP1DNL()</code>	35
3.8.2 A nonlinear 1D BVP solved by successive substitution	36
3.9 Plane elasticity	37
3.9.1 A plane stress example	37
3.9.2 A plane strain example	40
3.10 An axially symmetric elasticity example	43
<b>4 The Commands of FEMoctave</b>	<b>45</b>
4.1 Commands for 2D meshes: creation and modification	45
4.1.1 Structure of a mesh	45
4.1.2 Create a uniform mesh on a rectangle: <code>CreateMeshRect()</code>	46
4.1.3 Using triangle: <code>CreateMeshTriangle()</code> and <code>ReadMeshTriangle()</code>	48
4.1.4 Adapting meshes and creating holes by using options of <code>CreateMeshTriangle()</code>	50

4.1.5	Converting meshes: upgrading and downgrading	53
4.1.6	Use <code>delaunay()</code> to create a mesh: <code>Delaunay2Mesh()</code>	55
4.1.7	Deforming meshes by <code>MeshDeform()</code>	56
4.2	Evaluation and displaying results	56
4.2.1	Display results on meshes, <code>FEMtrimesh()</code> , <code>FEMtrisurf()</code> and <code>FEMtricontour()</code>	56
4.2.2	Evaluate the gradient of a function at the nodes: <code>FEMevaluateGradient()</code>	57
4.2.3	Evaluate a function and its gradient at the Gauss points: <code>FEMevaluateGP()</code>	58
4.2.4	Integrate a function over the domain: <code>FEMIntegrate()</code>	58
4.2.5	Evaluation at arbitrary points or along curves, integration along curves: <code>FEMgriddata()</code>	59
4.3	How to define functions	60
4.3.1	Functions for static problems	61
4.3.2	Functions for dynamic problems	62
4.4	Solving elliptic problems	62
4.4.1	Symmetric elliptic problems: <code>BVP2Dsym()</code>	62
4.4.2	General elliptic problems: <code>BVP2D()</code>	63
4.5	Solving 2D eigenvalue problems: <code>BVP2Deig()</code>	63
4.6	Solving parabolic problems: <code>IBVP2D()</code> and <code>IBVP2Dsym()</code>	65
4.7	Solving hyperbolic problems: <code>I2BVP2D()</code>	66
4.8	Solving 1D steady state problems, <code>BVP1D()</code>	66
4.9	Solving 1D dynamic problems of order 1, <code>IBVP1D()</code>	70
4.10	Solving 1D dynamic problems of order 2, <code>I2BVP1D()</code>	72
4.11	Solving 1D eigenvalue problems: <code>BVP1Deig()</code>	73
4.12	Solving nonlinear 1D boundary value problems: <code>BVP1DNL()</code>	74
4.13	Solving plane stress and plane strain problems: <code>PlaneStress()</code> , <code>PlaneStrain()</code>	76
4.13.1	Evaluating plane stress and plane strain solutions	78
4.13.2	Evaluation of basic strain and stress: <code>EvaluateStrain()</code> , <code>EvaluateStress()</code>	79
4.13.3	Evaluation of stress expressions: <code>EvaluateVonMises()</code> , <code>EvaluatePrincipalStress()</code> and <code>EvaluateTresca()</code>	80
4.14	Solving axisymmetric elasticity problems, <code>AxiStress()</code>	82
4.14.1	Evaluating axisymmetric solutions	82
4.14.2	Evaluation of strains and stress for axisymmetric problems	83
4.15	Internal commands in FEMoctave	85
4.15.1	Linear elements: <code>FEMEquation.cc</code> and <code>FEMEquation.m</code>	85
4.15.2	Quadratic elements: <code>FEMEquationQuad.cc</code> and <code>FEMEquationQuad.m</code>	86
4.15.3	Cubic elements: <code>FEMEquationCubic.cc</code> and <code>FEMEquationCubic.m</code>	86
4.15.4	Effect of right hand side for dynamic problems: <code>FEMInterpolWeight()</code>	86
4.15.5	Effect of the Dirichlet values: <code>FEMInterpolBoundaryWeight()</code>	87
4.15.6	Determine a few small eigenvalues: <code>eigSmall()</code>	87
4.15.7	Generating the equations for elasticity problems	88
4.16	External programs	89
<b>5</b>	<b>Tools for Didactical Purposes</b>	<b>90</b>
5.1	Observe the convergence of the error as $h \rightarrow 0$	90
5.2	Some Element Stiffness Matrices	92
5.2.1	Element contributions for equilateral triangles	92
5.2.2	From FEM to a finite difference approximation	95
5.2.3	Element stiffness matrices for 1D problems	98
5.2.4	Element stiffness matrices for elasticity problems	99
5.3	Behavior of a FEM solution within triangular elements	101

5.4	Estimate the number of nodes and triangles in a mesh and the effect on the sparse matrix . . . .	104
5.5	Compare linear, quadratic and cubic elements . . . . .	106
5.6	Are second order elements C1 conforming? . . . . .	107
5.7	Superconvergence for a 1D BVP . . . . .	110
5.8	Stability of time steppers, or lack thereof . . . . .	112
5.9	The shear locking effect by linear elements . . . . .	114
5.10	Bending of an Euler beam . . . . .	119
5.11	Missing boundary constraints and null spaces . . . . .	123
<b>6</b>	<b>The Mathematics of the Algorithms for 2D FEM</b>	<b>126</b>
6.1	Classical solutions and weak solutions . . . . .	126
6.2	A few triangular elements . . . . .	128
6.3	Transformation, interpolation and Gauss integration . . . . .	129
6.3.1	Transformation of coordinates and integration over a general triangle . . . . .	129
6.3.2	Gauss integration on the standard triangle with 3 Gauss points . . . . .	131
6.3.3	Gauss integration on the standard triangle with 7 Gauss points . . . . .	132
6.4	Construction of first order elements . . . . .	133
6.4.1	Linear interpolation on a triangle . . . . .	134
6.4.2	Integration of $f \phi$ . . . . .	134
6.4.3	Integration of $b_0 u \phi$ . . . . .	135
6.4.4	Integration of $a \nabla u \cdot \nabla \phi$ . . . . .	135
6.4.5	Integration of $u \vec{b} \cdot \nabla \phi$ . . . . .	136
6.4.6	Integration over boundary segments . . . . .	137
6.5	Construction of second order elements . . . . .	139
6.5.1	The basis functions for a second order element and quadratic interpolation . . . . .	139
6.5.2	Determine values at the Gauss points and apply Gauss integration . . . . .	140
6.5.3	Integration of $f \phi$ . . . . .	141
6.5.4	Integration of $b_0 u \phi$ . . . . .	142
6.5.5	Transformation of the gradient to the standard triangle . . . . .	142
6.5.6	Partial derivatives at the nodes . . . . .	145
6.5.7	Integration of $u \vec{b} \cdot \nabla \phi$ and $a \nabla u \cdot \nabla \phi$ . . . . .	146
6.5.8	Integration over boundary segments . . . . .	147
6.6	Construction of third order elements . . . . .	149
6.6.1	The basis functions for a third order element and cubic interpolation . . . . .	150
6.6.2	Determine values at the Gauss points and apply Gauss integration . . . . .	151
6.6.3	Integration of $f \phi$ and $b_0 u \phi$ . . . . .	153
6.6.4	Transformation of the gradient to the standard triangle . . . . .	154
6.6.5	Integration of $u \vec{b} \cdot \nabla \phi$ and $a \nabla u \cdot \nabla \phi$ . . . . .	156
6.6.6	Partial derivatives at the nodes . . . . .	157
6.6.7	Integration over boundary segments . . . . .	158
6.6.8	From a polynomial interpolation to the Gauss integration points . . . . .	160
6.7	Convergence of the approximate solutions $u_h$ to the exact solution $u$ . . . . .	161
6.8	Dynamic problems . . . . .	162
6.8.1	Dynamic problems of the heat equation type . . . . .	163
6.8.2	Using eigenvalues for dynamic problems of the heat equation type . . . . .	164
6.8.3	Dynamic problems of the wave equation type . . . . .	164
6.8.4	Using eigenvalues for dynamic problems of the wave equation type . . . . .	166
6.9	Inverse power iteration or <code>eigs()</code> to determine small eigenvalues of positive definite matrices . . . . .	167

<b>7</b>	<b>The Algorithms for 1D FEM</b>	<b>168</b>
7.1	The problems to be examined . . . . .	168
7.2	Interpolation, Gauss integration and the element stiffness matrices . . . . .	169
7.3	Taking boundary conditions into account . . . . .	172
7.4	Solving the BVP with a system of linear equations . . . . .	174
7.5	Evaluation of the solution between nodes and evaluation of the derivative . . . . .	175
7.6	The first order dynamic problem . . . . .	176
7.6.1	An explicit time step . . . . .	177
7.6.2	An implicit time step . . . . .	177
7.6.3	A Crank–Nicolson time step . . . . .	177
7.6.4	An L–stable Runge–Kutta solver, DIRK . . . . .	177
7.7	The second order dynamic problem . . . . .	178
7.8	Nonlinear boundary value problems, Newton’s method and partial substitution . . . . .	179
<b>8</b>	<b>Plane Elasticity and Axially Symmetric Elasticity</b>	<b>181</b>
8.1	The plane stress problem . . . . .	181
8.2	Construction of first order elements . . . . .	182
8.2.1	Integration of $f_1 \phi_1 + f_2 \phi_2$ . . . . .	182
8.2.2	Integration of the terms involving derivatives of $\phi_1$ and $\phi_2$ . . . . .	183
8.2.3	The boundary integral . . . . .	184
8.3	Construction of second order elements . . . . .	184
8.3.1	Integration of $f_1 \phi_1 + f_2 \phi_2$ . . . . .	184
8.3.2	Integration of the terms involving derivatives of $\phi_1$ and $\phi_2$ . . . . .	185
8.3.3	The boundary integral . . . . .	186
8.4	Construction of third order elements . . . . .	186
8.4.1	Integration of $f_1 \phi_1 + f_2 \phi_2$ . . . . .	186
8.4.2	Integration of the terms involving derivatives of $\phi_1$ and $\phi_2$ . . . . .	187
8.4.3	The boundary integral . . . . .	188
8.5	The plane strain problem . . . . .	189
8.6	Elasticity for axially symmetric setups . . . . .	190
8.7	Construction of first order elements . . . . .	191
8.7.1	Integration of $r (f_r \phi_r + f_z \phi_z)$ . . . . .	191
8.7.2	Integration of the terms involving derivatives of $\phi_r$ and $\phi_z$ . . . . .	192
8.7.3	The boundary integral . . . . .	193
8.8	Construction of second order elements . . . . .	193
8.8.1	Integration of $r (f_r \phi_r + f_z \phi_z)$ . . . . .	193
8.8.2	Integration of the terms involving derivatives of $\phi_r$ and $\phi_z$ . . . . .	194
8.8.3	The boundary integral . . . . .	195
8.9	Construction of third order elements . . . . .	196
8.9.1	Integration of $r (f_r \phi_r + f_z \phi_z)$ . . . . .	196
8.9.2	Integration of the terms involving derivatives of $\phi_r$ and $\phi_z$ . . . . .	196
8.9.3	The boundary integral . . . . .	196
<b>9</b>	<b>Examples, Examples, Examples</b>	<b>197</b>
9.1	An elliptic problem with variable coefficients . . . . .	197
9.2	An animated wave . . . . .	198
9.3	An elliptic problem with radial symmetry, superconvergence . . . . .	199
9.4	An example with limited regularity . . . . .	203
9.5	A potential flow problem . . . . .	205

9.6	A potential flow problem in a circular pipe . . . . .	207
9.7	A minimal surface problem . . . . .	210
9.8	Computing a capacitance . . . . .	211
9.8.1	State the problem . . . . .	211
9.8.2	Create the mesh and solve the BVP . . . . .	212
9.8.3	Compute the capacitance . . . . .	213
9.9	Torsion of beams, Prandtl stress function . . . . .	215
9.9.1	The setup with the warp function and the Prandtl stress function . . . . .	215
9.9.2	On a disk with radius $R$ . . . . .	217
9.9.3	On a square . . . . .	218
9.9.4	On a rectangle . . . . .	219
9.10	Dynamic heat conduction problems . . . . .	219
9.10.1	With a narrow section in the domain . . . . .	219
9.10.2	With a section with lower thermal conductivity . . . . .	221
9.10.3	Cooling of a cylinder . . . . .	224
9.10.4	Heat waves . . . . .	227
9.10.5	Static heat equation in a ball in $\mathbb{R}^3$ , solved as a 1D problem . . . . .	229
9.10.6	Dynamic heat equation in a cylinder, solved as a 1D problem . . . . .	232
9.11	Wave propagation, Kirchhoff diffraction . . . . .	234
9.11.1	A dynamic solution . . . . .	234
9.12	Sound waves in $\mathbb{R}^2$ and $\mathbb{R}^3$ . . . . .	236
9.12.1	A sound wave in $\mathbb{R}^3$ with cylindrical coordinates . . . . .	237
9.12.2	A sound wave in $\mathbb{R}^2$ . . . . .	238
9.12.3	Sound waves in $\mathbb{R}^3$ and $\mathbb{R}^2$ as 1D problems . . . . .	239
9.13	Reflection and transmission by a change of impedance . . . . .	241
9.14	Schrödinger's harmonic oscillator . . . . .	243
9.15	The EIT forward problem . . . . .	244
9.16	Stretching of a beam . . . . .	251
9.17	Keller's nonlinear boundary value problems . . . . .	252
9.17.1	Partial successive substitution . . . . .	252
9.17.2	Newton's method . . . . .	253
9.17.3	Using <code>BVP1DNL()</code> . . . . .	254
9.17.4	A very similar problem with multiple solutions . . . . .	254
9.18	A mathematical pendulum problem . . . . .	255
9.19	A BVP with multiple nonlinear contributions . . . . .	256
9.20	Fisher's equation . . . . .	258
9.21	From Salt Lake City to Zürich, the shortest connection on a sphere . . . . .	260
9.21.1	A solution based on successive substitution . . . . .	261
9.21.2	A solution using <code>BVP1DNL()</code> . . . . .	263
9.22	A 1D nonlinear bending beam problem . . . . .	264
9.22.1	Solving the BVP using Newton's algorithm . . . . .	264
9.22.2	Solving the BVP with the command <code>BVP1DNL()</code> . . . . .	266
9.23	Mass transfer in a porous catalyst . . . . .	267
9.24	Troesch's equation . . . . .	269
9.25	Motion of a string . . . . .	270
9.26	A pipe under pressure . . . . .	271
9.26.1	As a plane strain problem . . . . .	271
9.26.2	As an axisymmetric problem . . . . .	275
9.26.3	The analytical solution . . . . .	277

---

9.27	A crook with a weight attached . . . . .	278
9.28	A wrench . . . . .	283
9.29	A rotating rubber cylinder . . . . .	285
9.30	A washer fastener examined as spring . . . . .	287
9.30.1	The setup . . . . .	287
9.30.2	Evaluate the force by integrating the normal stress . . . . .	290
9.30.3	Evaluate the force by an energy argument . . . . .	292
9.30.4	Comparison of linear, quadratic and cubic elements . . . . .	292
9.30.5	Different boundary conditions . . . . .	293
	<b>Bibliography</b>	<b>294</b>
	<b>List of Figures</b>	<b>295</b>
	<b>List of Tables</b>	<b>298</b>
	<b>Index</b>	<b>299</b>

There is no such thing as “*the perfect notes*” and improvements are always possible. I welcome feedback and constructive criticism. Please let me know if you use/like/dislike the lecture notes. Please send your observations and remarks to [Andreas.Stahel@gmx.com](mailto:Andreas.Stahel@gmx.com) .

# 1 Introduction

- Goals of this project:
  - Provide support material for teaching FEM. The material provided might help other instructors to explain or illustrate the methods and effects of finite element algorithms.
  - Use *Octave* to implement first, second and third order triangular elements in 2D for scalar boundary value problems. For elasticity plane stress and plane strain problems are examined. For linear 1D boundary value problems and initial boundary value problems second order elements are used. A nonlinear solver for 1D problems is implemented. This leads to the *Octave* package **FEMoctave**.
  - Provide examples on how to solve steady state and dynamic heat equations, the wave equation and 2D elasticity equations. A few nonlinear 1D examples are provided too.
- Tools provided by this project:
  - Find this document on the internet at <https://andreasstahel.github.io/FEMoctave/FEMdoc.pdf> and the complete *Octave* package at <https://andreasstahel.github.io/FEMoctave/FEMoctave.tgz>.
  - Documentation and codes are also on GitHub at <https://github.com/AndreasStahel/FEMoctave> and with *Octave* you should be able to install it by calling

```
pkg install https://github.com/AndreasStahel/FEMoctave/archive/v2.1.2.tar.gz
```
  - I work exclusively with Unix systems, but it is possible to use the package on other systems by modifying the `Makefile`.
  - The only external program used in FEMoctave is `triangle`, an excellent mesh generator. The source code of `triangle` is not included. Find source code and documentation at [www.cs.cmu.edu/~quake/triangle.html](http://www.cs.cmu.edu/~quake/triangle.html).

This is **not**:

- an introduction to *Octave* (or `MATLAB`). Users are assumed to be familiar with the basics of using *Octave*. If this is not the case, may I use the occasion for a shameless add for my book *Octave and Matlab for Engineering Applications* by Springer, [Stah22].
- an introduction to FEM algorithms. For a basic (and affordable) introduction consider [TongRoss08]. The basic concept is not explained in these notes for FEMoctave, but many details are spelled out. I used some of the presentations for a class *Numerical Methods* for biomedical engineers at the University of Bern. There the main ideas of FEM are spelled out. Find the lecture notes for this class on my web site at <https://andreasstahel.github.io/Notes/NumMethods.pdf>.
- an introduction to partial differential equations (PDE). The user of FEMoctave is assumed to know which boundary and initial boundary value problems are well posed, i.e. will have a unique solution.

The structure of this document is as follows:

- 1 **Introduction**: a self reference.
- 2 **The Problems to be Examined**: for each type of problem one example is presented. This is a good starting point to find out what type of problems are examined in these notes.
- 3 **Illustrative Examples**: a few examples are worked out, code and results shown. Read this section if you want to start working with FEMoctave.

- 4 **The Commands of FEMoctave:** all commands of FEMoctave are briefly explained and some documentation is provided. This is comparable to a manual.
- 5 **Tools for Didactical Purposes:** some results and illustrations that might be useful when teaching a class on the mathematical basics of FEM.
- 6 **The Mathematics of the Algorithms:** the mathematics of the 2D FEM algorithms is spelled out. Linear, quadratic and cubic elements on triangles are constructed. A matrix formulation is used wherever possible.
- 7 **The Algorithms for 1D FEM:** the mathematics for second order element of one dimensional problems are spelled out. Examined are static problems and dynamic problems of order 1 and 2, i.e. heat and wave equations.
- 8 **Elasticity:** the mathematical aspects of an FEM algorithm to solve plane stress and plane strain problems are presented. The algorithms for axially symmetric elasticity problems are explained, all leading to a matrix formulation.
- 9 **Examples, Examples, Examples:** as the title says.

## 2 The Problems to be Examined

This section consists of a brief list all types of problems that can be solved with this software. A list of the necessary commands is given in Table 1 on page 11. The instruction on how to use the commands are given in Section 4. Some typical examples are worked out in Section 9.

### 2.1 The domain $\Omega \subset \mathbb{R}^2$ and its boundary $\Gamma = \partial\Omega = \Gamma_1 \cup \Gamma_2$

Throughout this presentation work with bounded domains  $\Omega \subset \mathbb{R}^2$  with two disjoint section  $\Gamma_1$  and  $\Gamma_2$  of the boundary  $\Gamma = \partial\Omega$ .

- On the section  $\Gamma_1$  a Dirichlet boundary condition is applied, i.e.  $u(x, y) = g_1(x, y)$  for a known function  $g_1$ .
- On the section  $\Gamma_2$  a Neumann or Robin boundary condition is applied, i.e. the outer normal derivative of  $u$  equals  $g_2 + g_3 u$  for a known functions  $g_2$  and  $g_3$ .

In the example shown in Figure 1 the solution satisfies  $u = +3$  on the circular part  $\Gamma_1$  and  $\frac{\partial}{\partial y}u = -1$  along the  $x$ -axis. The solution  $u(x, y)$  solves  $\Delta u = \nabla \cdot \nabla u = \text{div grad } u = 0$  and minimizes the functional

$$F(u) = \iint_{\Omega} \frac{1}{2} \|\nabla u\|^2 dA - \int_{\Gamma_2} u ds$$

amongst all functions  $u$  which satisfy  $u(x, y) = +3$  on  $\Gamma_1$ .

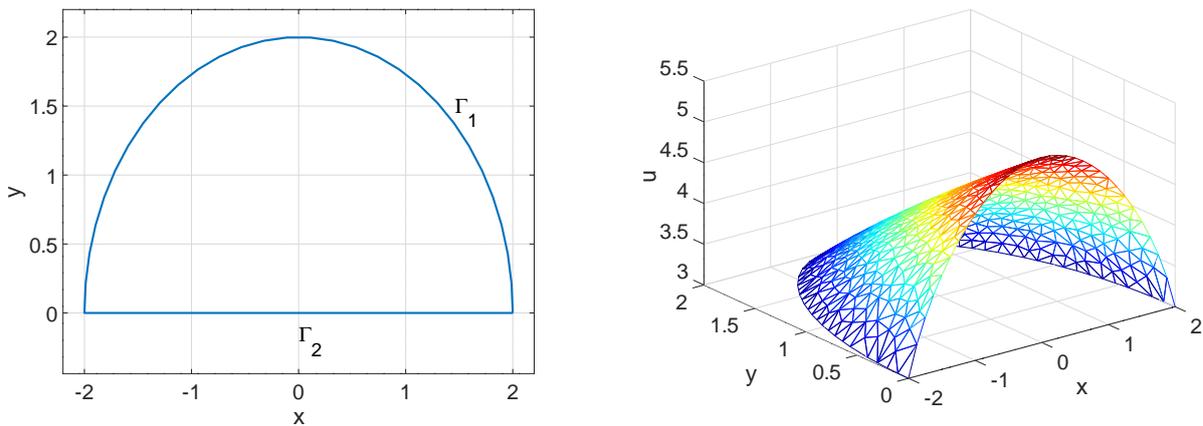


Figure 1: A semi-disk as domain in  $\mathbb{R}^2$  and a solution of a BVP

### 2.2 The general elliptic problem

Let  $\Omega \subset \mathbb{R}^2$  be a bounded domain with a nice boundary  $\Gamma$ , consisting of two disjoint sections  $\Gamma_1$  and  $\Gamma_2$ . For given functions  $a, b_0, \vec{b}, f$  and  $g_i$  we seek a solution of the second order **boundary value problem** (BVP)

$$\begin{aligned} -\nabla \cdot (a \nabla u - u \vec{b}) + b_0 u &= f & \text{for } (x, y) \in \Omega \\ u &= g_1 & \text{for } (x, y) \in \Gamma_1 \\ \vec{n} \cdot (a \nabla u - u \vec{b}) &= g_2 + g_3 u & \text{for } (x, y) \in \Gamma_2 \end{aligned} \quad (1)$$

It is assumed that there is a unique solution  $u$ . Consult your book on the theory of PDEs to determine whether the BVP has in fact a unique solution. Examples of this type of equation are given in Section 3.1.4.

### 2.3 The symmetric elliptic problem

If there is no convection contribution  $\vec{b}$  in (1) one ends up with a self-adjoint problem.

$$\begin{aligned} -\nabla \cdot (a \nabla u) + b_0 u &= f & \text{for } (x, y) \in \Omega \\ u &= g_1 & \text{for } (x, y) \in \Gamma_1 \\ a \frac{\partial u}{\partial n} &= g_2 + g_3 u & \text{for } (x, y) \in \Gamma_2 \end{aligned} \quad (2)$$

The resulting matrix  $\mathbf{A}$  will be symmetric and if  $a > 0$ ,  $b_0 \geq 0$  and  $\Gamma_1 \neq \emptyset$  or  $b_0 > 0$ , then the BVP has a unique solution and the resulting matrix is strictly positive definite.

Using Calculus of Variations one can show that solving (2) is equivalent to minimizing the functional  $F$  below among all functions  $u$  vanishing on  $\Gamma_1$ .

$$F(u) = \iint_{\Omega} \frac{1}{2} a \langle \nabla u, \nabla u \rangle + \frac{1}{2} b_0 u^2 - f u \, dA - \int_{\Gamma_2} g_2 u + \frac{1}{2} g_3 u^2 \, ds.$$

Examples of this type are given in Sections 3.1.1, 3.1.2, 3.1.3, 9.4 and 9.15.

### 2.4 The symmetric eigenvalue problem

For given functions  $a$ ,  $b_0$ ,  $f$  and  $g_3$  seek values of  $\lambda$  and nontrivial solutions  $u$  of the **eigenvalue problem** below.

$$\begin{aligned} -\nabla \cdot (a \nabla u) + b_0 u &= \lambda f u & \text{for } (x, y) \in \Omega \\ u &= 0 & \text{for } (x, y) \in \Gamma_1 \\ a \frac{\partial u}{\partial n} &= g_3 u & \text{for } (x, y) \in \Gamma_2 \end{aligned} \quad (3)$$

An example of this type is given in Section 3.2.

### 2.5 The general parabolic problem

If all functions depend on time  $t$  and the spacial variables  $x$  and  $y$  consider the general dynamic heat equation.

$$\begin{aligned} \rho \frac{\partial}{\partial t} u - \nabla \cdot (a \nabla u - u \vec{b}) + b_0 u &= f & \text{for } (x, y, t) \in \Omega \times (0, T] \\ u &= g_1 & \text{for } (x, y, t) \in \Gamma_1 \times (0, T] \\ \vec{n} \cdot (a \nabla u - u \vec{b}) &= g_2 + g_3 u & \text{for } (x, y, t) \in \Gamma_2 \times (0, T] \\ u &= u_0 & \text{on } \Omega \text{ at } t = 0 \end{aligned} \quad (4)$$

This is an Initial Boundary Value Problem (IBVP). Mathematicians call this a parabolic problem, engineers think of dynamic heat equations. Examples are shown in Sections 3.3 and 9.10.4.

### 2.6 The symmetric parabolic problem

Consider the symmetric situation of (4) to find the symmetric parabolic problem below.

$$\begin{aligned} \rho \frac{\partial}{\partial t} u - \nabla \cdot (a \nabla u) + b_0 u &= f & \text{for } (x, y, t) \in \Omega \times (0, T] \\ u &= g_1 & \text{for } (x, y, t) \in \Gamma_1 \times (0, T] \\ a \frac{\partial \nabla u}{\partial n} &= g_2 + g_3 u & \text{for } (x, y, t) \in \Gamma_2 \times (0, T] \\ u &= u_0 & \text{on } \Omega \text{ at } t = 0 \end{aligned} \quad (5)$$

If  $u(x, y)$  and  $\lambda$  are solutions of the eigenvalue problem (3) with  $f = g_1 = g_2 = 0$ , then the dynamic problem (5) is solved by  $e^{-\lambda t} u(x, y)$ . See also Section 6.8.2.

## 2.7 The hyperbolic problem

Examine an IBVP of hyperbolic type, with the wave equation  $\ddot{u} = \Delta u$  as the typical example.

$$\begin{aligned}
 \rho \frac{\partial^2}{\partial t^2} u + 2\alpha \frac{\partial}{\partial t} u - \nabla \cdot (a \nabla u - u \vec{b}) + b_0 u &= f & \text{for } (x, y, t) \in \Omega \times (0, T] \\
 u &= g_1 & \text{for } (x, y, t) \in \Gamma_1 \times (0, T] \\
 \vec{n} \cdot (a \nabla u + u \vec{b}) &= g_2 + g_3 u & \text{for } (x, y, t) \in \Gamma_2 \times (0, T] \\
 u &= u_0 & \text{on } \Omega \text{ at } t = 0 \\
 \frac{\partial}{\partial t} u &= v_0 & \text{on } \Omega \text{ at } t = 0
 \end{aligned} \tag{6}$$

Examples are shown in Sections 3.4, 9.2 and 9.11. The effect of eigenvalues is described in Section 6.8.4.

command	type of problem	section
BVP2Dsym()	solve a symmetric elliptic BVP	4.4.1
BVP2D()	solve a general elliptic BVP	4.4.2
BVP2Deig()	solve a symmetric elliptic eigenvalue problem	4.5
IBVP2D()	solve a parabolic IBVP	4.6
IBVP2Dsym()	solve a symmetric parabolic IBVP	4.6
I2BVP2D()	solve a hyperbolic IBVP	4.7
BVP1D()	solve a static 1D BVP	4.8
IBVP1D()	solve a first order 1D IBVP	4.9
I2BVP1D()	solve a second order 1D IBVP	4.10
IBVP1Deig()	solve a 1D eigenvalue BVP	4.11
IBVP1DNL()	solve nonlinear 1D BVP	4.12
PlaneStress()	solve a plane stress problem	4.13
PlaneStrain()	solve a plane strain problem	4.13

Table 1: Commands to solve boundary value problems, initial boundary value problems and elasticity problems

## 2.8 1D boundary value problems

A 1D BVP (boundary value problem) is an ordinary differential equation for the independent variable  $u(x)$  with  $x$  in an interval  $x_0 \leq x \leq x_n$  of the form

$$- (a(x) u'(x))' + b(x) u'(x) + c(x) u(x) = d(x) f(x) \tag{7}$$

with some boundary conditions at  $x = x_0$  and  $x = x_n$ .

$$\begin{aligned}
 u(x_i) &= g_D & \text{Dirichlet} \\
 a(x_i) u'(x_i) &= g_{N1} + g_{N2} u(x_i) & \text{Neumann}
 \end{aligned} \tag{8}$$

The coefficient functions  $a(x)$ ,  $b(x)$ ,  $c(x)$  and  $d(x)$  will be evaluated at the Gauss points. The function  $f(x)$  and the solution  $u(x)$  are evaluated (resp. determined) at the nodes.

Find examples in Sections 3.5, 5.6, 5.7 and 9.16. In Sections 9.17, 9.20, 9.21 and 9.22 nonlinear boundary value problems are solved.

## 2.9 1D initial boundary value problems of order 1

The dynamic problem is of the form

$$w(x) \frac{\partial}{\partial t} u(x, t) - (a(x) u'(x, t))' + b(x) u'(x, t) + c(x) u(x, t) = d(x) f(x, t) \quad (9)$$

with the initial condition  $u(x, 0) = u_0(x)$  and appropriate boundary conditions. These have to be independent on time  $t$ .

Find examples in Sections 3.6, 5.8 and 9.10.6.

## 2.10 1D initial boundary value problems of order 2

The dynamic problem is of the form

$$w_2(x) \frac{\partial^2}{\partial t^2} u(x, t) + w_1(x) \frac{\partial}{\partial t} u(x, t) - (a(x) u'(x, t))' + b(x) u'(x, t) + c(x) u(x, t) = d(x) f(x, t) \quad (10)$$

with the initial condition  $u(x, 0) = u_0(x)$  and appropriate boundary conditions. These have to be independent on time  $t$ .

Find examples in Sections 3.7, 9.12.3, 9.13 and 9.25.

## 2.11 1D eigenvalue value problems

For given functions  $a, b, c, w$  and  $g_{N2}$  seek values of  $\lambda$  and nontrivial solutions  $u$  of the **eigenvalue problem** below.

$$- (a(x) u'(x))' + b(x) u'(x) + c(x) u(x) = \lambda w(x) u(x) \quad (11)$$

with appropriate boundary conditions

$$\begin{aligned} u(x_i) &= 0 && \text{Dirichlet} \\ a(x_i) u'(x_i) &= g_{N2} u(x_i) && \text{Neumann} \end{aligned} \quad (12)$$

Find examples in Sections 3.2 and 9.14.

## 2.12 Nonlinear 1D boundary value problems

For given functions  $a, b, c$  and  $d$  and a function  $f(x, u)$  or  $f(x, u, u')$  search for solutions of

$$- (a(x, u(x), u'(x))) u'(x))' + b(x) u'(x) + c(x) u(x) = d(x) f(x, u(x), u'(x)) \quad (13)$$

with linear boundary conditions, Dirichlet or Neumann. The leading coefficient  $a$  may depend on  $u$  and  $u'$  too, i.e.  $a(x, u(x), u'(x))$ .

Find examples in Sections 3.8.1, 3.8.2, 9.17, 9.18, 9.19, 9.20, 9.21, 9.22, 9.23 and 9.24.

## 2.13 Plane elasticity

With FEMoctave plane elasticity problems can be examined, either plane stress or plane strain.

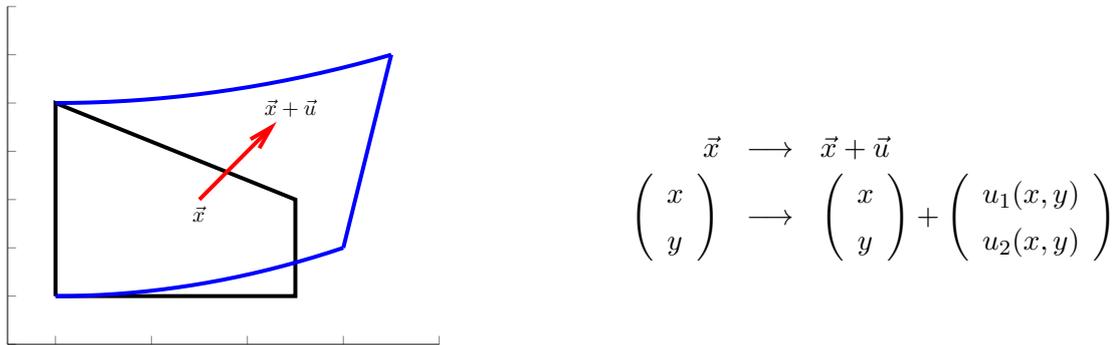


Figure 2: Deformation of an elastic solid

**2.13.1 Description of strain**

The first goal is to determine the displacement function  $\vec{u} = (u_1, u_2)$ . It describes the displacement of arbitrary points  $(x, y) \in \Omega \subset \mathbb{R}^2$ . Based on  $\vec{u}(x, y)$  the infinitesimal strain tensor is given by

$$\begin{bmatrix} \epsilon_{xx} & \epsilon_{xy} \\ \epsilon_{xy} & \epsilon_{yy} \end{bmatrix} = \begin{bmatrix} \frac{\partial u_1}{\partial x} & \frac{1}{2} \left( \frac{\partial u_1}{\partial y} + \frac{\partial u_2}{\partial x} \right) \\ \frac{1}{2} \left( \frac{\partial u_1}{\partial y} + \frac{\partial u_2}{\partial x} \right) & \frac{\partial u_2}{\partial y} \end{bmatrix}.$$

It contains the essential information of how a small section of the large solid is deformed, see Figure 3. Obviously

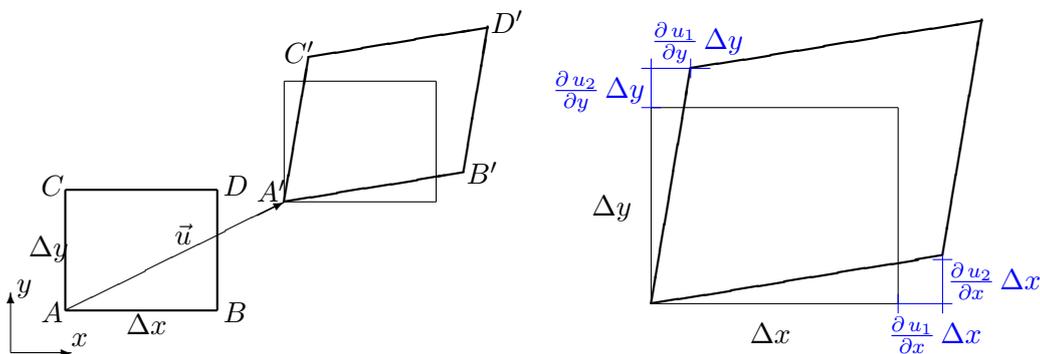


Figure 3: Definition of strain: rectangle before and after deformation

this can be used in the space  $\mathbb{R}^3$  too, leading to the  $3 \times 3$  strain matrix (or tensor of order 2)

$$\begin{bmatrix} \epsilon_{xx} & \epsilon_{xy} & \epsilon_{xz} \\ \epsilon_{xy} & \epsilon_{yy} & \epsilon_{yz} \\ \epsilon_{xz} & \epsilon_{yz} & \epsilon_{zz} \end{bmatrix} = \begin{bmatrix} \frac{\partial u_1}{\partial x} & \frac{1}{2} \left( \frac{\partial u_1}{\partial y} + \frac{\partial u_2}{\partial x} \right) & \frac{1}{2} \left( \frac{\partial u_1}{\partial z} + \frac{\partial u_3}{\partial x} \right) \\ \frac{1}{2} \left( \frac{\partial u_1}{\partial y} + \frac{\partial u_2}{\partial x} \right) & \frac{\partial u_2}{\partial y} & \frac{1}{2} \left( \frac{\partial u_2}{\partial z} + \frac{\partial u_3}{\partial y} \right) \\ \frac{1}{2} \left( \frac{\partial u_1}{\partial z} + \frac{\partial u_3}{\partial x} \right) & \frac{1}{2} \left( \frac{\partial u_2}{\partial z} + \frac{\partial u_3}{\partial y} \right) & \frac{\partial u_3}{\partial z} \end{bmatrix}$$

and the geometric interpretations in Table 2.

**2.13.2 Description of stress and Hooke’s law**

The deformation of the solid will lead to normal and shearing stress, with the units forces per area. Find a graphical interpretation of the 6 strains in space  $\mathbb{R}^3$  in Figure 4 and a description in Table 3.

symbol	formula	interpretation
$\epsilon_{xx}$	$\frac{\partial u_1}{\partial x}$	ratio of change of length divided by length in $x$ direction
$\epsilon_{yy}$	$\frac{\partial u_2}{\partial y}$	ratio of change of length divided by length in $y$ direction
$\epsilon_{zz}$	$\frac{\partial u_3}{\partial z}$	ratio of change of length divided by length in $z$ direction
$\epsilon_{xy} = \epsilon_{yx}$	$\frac{1}{2} \left( \frac{\partial u_1}{\partial y} + \frac{\partial u_2}{\partial x} \right)$	the angle between the $x$ and $y$ axis is diminished by $2\epsilon_{xy}$
$\epsilon_{xz} = \epsilon_{zx}$	$\frac{1}{2} \left( \frac{\partial u_1}{\partial z} + \frac{\partial u_3}{\partial x} \right)$	the angle between the $x$ and $z$ axis is diminished by $2\epsilon_{xz}$
$\epsilon_{yz} = \epsilon_{zy}$	$\frac{1}{2} \left( \frac{\partial u_2}{\partial z} + \frac{\partial u_3}{\partial y} \right)$	the angle between the $y$ and $z$ axis is diminished by $2\epsilon_{yz}$

Table 2: Normal and shear strains in space

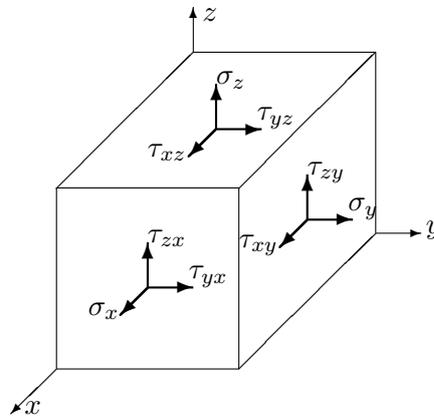


Figure 4: Components of stress in space

symbol	description
$\sigma_x$	normal stress at a surface orthogonal to $x = \text{const}$
$\sigma_y$	normal stress at a surface orthogonal to $y = \text{const}$
$\sigma_z$	normal stress at a surface orthogonal to $z = \text{const}$
$\tau_{xy} = \tau_{yx}$	tangential stress in $y$ direction at surface orthogonal to $x = \text{const}$ tangential stress in $x$ direction at surface orthogonal to $y = \text{const}$
$\tau_{xz} = \tau_{zx}$	tangential stress in $z$ direction at surface orthogonal to $x = \text{const}$ tangential stress in $x$ direction at surface orthogonal to $z = \text{const}$
$\tau_{yz} = \tau_{zy}$	tangential stress in $z$ direction at surface orthogonal to $y = \text{const}$ tangential stress in $y$ direction at surface orthogonal to $z = \text{const}$

Table 3: Description of normal and tangential stress in space

With FEMoctave there are three types of boundary conditions to be examined:

$$\begin{aligned} \vec{u} &= \vec{g}_D && \text{on Dirichlet boundary } \Gamma_1, \text{ i.e. prescribed displacement} \\ \text{force density} &= \vec{g}_N && \text{on Neumann boundary } \Gamma_2, \text{ i.e. prescribed force density} \\ \text{force density} &= \vec{0} && \text{on free boundary } \Gamma_3 \end{aligned} \quad (14)$$

The conditions can be set for each component, find the codes in Table 5, to be used when creating meshes by `CreateMeshRect()` or `CreateMeshTriangle()`.

For a linear material law the connection between stresses and strains is given by Hooke's law and uses two material parameters:

- $E$ : the Young's modulus of elasticity
- $\nu$ : the Poisson ratio, with  $0 \leq \nu \leq \frac{1}{2}$

FEMoctave is based on the general form of **Hooke's law** for isotropic (independent on direction) materials. It is a basic physical law<sup>1</sup>, confirmed by many measurements. The shown formulation is valid as long as all stress and strains are small. Hooke's law is the foundation of linear elasticity and any book on elasticity will show a formulation, e.g. [Prze68, §2.2]<sup>2</sup>, [Sout73, §2.7], or [Wein74, §10.1].

$$\begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{zz} \\ \varepsilon_{xy} \\ \varepsilon_{xz} \\ \varepsilon_{yz} \end{pmatrix} = \frac{1}{E} \begin{bmatrix} 1 & -\nu & -\nu \\ -\nu & 1 & -\nu \\ -\nu & -\nu & 1 \end{bmatrix} \cdot \begin{pmatrix} \sigma_x \\ \sigma_y \\ \sigma_z \end{pmatrix}, \quad (15)$$

$$\begin{pmatrix} \varepsilon_{xy} \\ \varepsilon_{xz} \\ \varepsilon_{yz} \end{pmatrix} = \frac{1+\nu}{E} \begin{pmatrix} \tau_{xy} \\ \tau_{xz} \\ \tau_{yz} \end{pmatrix}$$

or by inverting the matrix

$$\begin{pmatrix} \sigma_x \\ \sigma_y \\ \sigma_z \end{pmatrix} = \frac{E}{(1+\nu)(1-2\nu)} \begin{bmatrix} 1-\nu & \nu & \nu \\ \nu & 1-\nu & \nu \\ \nu & \nu & 1-\nu \end{bmatrix} \cdot \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{zz} \end{pmatrix}, \quad (16)$$

$$\begin{pmatrix} \tau_{xy} \\ \tau_{xz} \\ \tau_{yz} \end{pmatrix} = \frac{E}{1+\nu} \begin{pmatrix} \varepsilon_{xy} \\ \varepsilon_{xz} \\ \varepsilon_{yz} \end{pmatrix}.$$

This leads to an elastic energy density of

$$W = \frac{1}{2} \left\langle \begin{pmatrix} \sigma_x \\ \sigma_y \\ \sigma_z \end{pmatrix}, \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{zz} \end{pmatrix} \right\rangle + \left\langle \begin{pmatrix} \tau_{xy} \\ \tau_{xz} \\ \tau_{yz} \end{pmatrix}, \begin{pmatrix} \varepsilon_{xy} \\ \varepsilon_{xz} \\ \varepsilon_{yz} \end{pmatrix} \right\rangle \quad (17)$$

$$= \frac{1}{2} \frac{E}{(1+\nu)(1-2\nu)} \left\langle \begin{bmatrix} 1-\nu & \nu & \nu \\ \nu & 1-\nu & \nu \\ \nu & \nu & 1-\nu \end{bmatrix} \cdot \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{zz} \end{pmatrix}, \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{zz} \end{pmatrix} \right\rangle + \quad (18)$$

<sup>1</sup>One can verify that for homogeneous, isotropic materials a linear law must have this form, e.g. [Sege77]

<sup>2</sup>The missing factors 2 are due to the different definition of the shear strains.

$$+ \frac{E}{1+\nu} \left\langle \begin{pmatrix} \varepsilon_{xy} \\ \varepsilon_{xz} \\ \varepsilon_{yz} \end{pmatrix}, \begin{pmatrix} \varepsilon_{xy} \\ \varepsilon_{xz} \\ \varepsilon_{yz} \end{pmatrix} \right\rangle.$$

### 2.13.3 The plane stress problem

For a plane stress problem it is assumed that there are no stresses in  $z$ -direction, i.e.

$$\sigma_z = \tau_{xz} = \tau_{yz} = 0.$$

This leads to a simpler version of Hooke's law

$$\begin{pmatrix} \sigma_x \\ \sigma_y \\ \tau_{xy} \end{pmatrix} = \frac{E}{1-\nu^2} \begin{bmatrix} 1 & \nu & 0 \\ \nu & 1 & 0 \\ 0 & 0 & 1-\nu \end{bmatrix} \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix} \quad \text{and} \quad \begin{matrix} \varepsilon_{zz} = \frac{-\nu}{1-\nu} (\varepsilon_{xx} + \varepsilon_{yy}) \\ \varepsilon_{xz} = 0 \\ \varepsilon_{yz} = 0 \end{matrix}. \quad (19)$$

The energy density given by equation (17) simplifies to

$$\begin{aligned} W_{stress} &= \frac{1}{2} \left\langle \begin{pmatrix} \sigma_x \\ \sigma_y \\ 0 \end{pmatrix}, \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{zz} \end{pmatrix} \right\rangle + \left\langle \begin{pmatrix} \tau_{xy} \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \varepsilon_{xy} \\ \varepsilon_{xz} \\ \varepsilon_{yz} \end{pmatrix} \right\rangle = \frac{1}{2} \left\langle \begin{pmatrix} \sigma_x \\ \sigma_y \\ 2\tau_{xy} \end{pmatrix}, \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix} \right\rangle \\ &= \frac{E}{2(1-\nu^2)} \left\langle \begin{bmatrix} 1 & \nu & 0 \\ \nu & 1 & 0 \\ 0 & 0 & 2(1-\nu) \end{bmatrix} \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix}, \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix} \right\rangle \\ &= \frac{E}{2(1-\nu^2)} (\varepsilon_{xx}^2 + \varepsilon_{yy}^2 + 2\nu \varepsilon_{xx} \varepsilon_{yy} + 2(1-\nu) \varepsilon_{xy}^2). \quad (20) \end{aligned}$$

Since

$$\varepsilon_{xx}^2 + \varepsilon_{yy}^2 + 2\nu \varepsilon_{xx} \varepsilon_{yy} = \nu (\varepsilon_{xx} + \varepsilon_{yy})^2 + (1-\nu) (\varepsilon_{xx}^2 + \varepsilon_{yy}^2) \geq 0$$

the energy density  $W_{stress}$  is assured to be positive. With this the total energy of a plane stress problem can be written in the form<sup>3</sup>

$$\begin{aligned} U(\vec{u}) &= U_{elast} + U_{Vol} + U_{Surf} \quad (21) \\ &= \iint_{\Omega} \frac{1}{2} \frac{E}{(1-\nu^2)} \left\langle \begin{bmatrix} 1 & \nu & 0 \\ \nu & 1 & 0 \\ 0 & 0 & 2(1-\nu) \end{bmatrix} \cdot \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix}, \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix} \right\rangle dA - \\ &\quad - \iint_{\Omega} \vec{f} \cdot \vec{u} dA - \int_{\Gamma_2} \vec{g}_N \cdot \vec{u} ds. \end{aligned}$$

Using the Bernoulli principle this energy has to be minimized. It is this minimization problem that is solved, subject to the boundary conditions (14).

<sup>3</sup>We quietly dropped the constant thickness  $H$  from all expressions.

### 2.13.4 The plane strain problem

For a plane strain problem it is assumed that there are no strains in  $z$ -direction, i.e.

$$\varepsilon_{xz} = \varepsilon_{yz} = \varepsilon_{zz} = 0.$$

This leads to a simpler version of Hooke's law

$$\begin{pmatrix} \sigma_x \\ \sigma_y \\ \tau_{xy} \end{pmatrix} = \frac{E}{(1+\nu)(1-2\nu)} \begin{bmatrix} 1-\nu & \nu & 0 \\ \nu & 1-\nu & 0 \\ 0 & 0 & 1-2\nu \end{bmatrix} \cdot \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix}. \quad (22)$$

$$\sigma_z = \frac{E\nu(\varepsilon_{xx} + \varepsilon_{yy})}{(1+\nu)(1-2\nu)}, \quad \tau_{xz} = \tau_{yz} = 0$$

Observe that

$$\sigma_z = \frac{E\nu(\varepsilon_{xx} + \varepsilon_{yy})}{(1+\nu)(1-2\nu)} = \nu(\sigma_x + \sigma_y).$$

Modify the material parameters  $\nu$  and  $E$  to

$$\nu^* = \frac{\nu}{1-\nu} > \nu \quad \text{and} \quad E^* = \frac{E}{1-\nu^2} > E. \quad (23)$$

Then use elementary algebra to find

$$\begin{aligned} \nu &= \frac{\nu^*}{1+\nu^*}, \quad 1-\nu = 1 - \frac{\nu^*}{1+\nu^*} = \frac{1}{1+\nu^*} \\ \frac{1-2\nu}{1-\nu} &= \frac{1-2\frac{\nu^*}{1+\nu^*}}{1-\frac{\nu^*}{1+\nu^*}} = 1-\nu^* \quad \text{and} \quad \frac{\nu}{1-2\nu} = \frac{\frac{\nu^*}{1+\nu^*}}{1-2\frac{\nu^*}{1+\nu^*}} = \frac{\nu^*}{1-\nu^*} \\ E &= E^*(1-\nu^2) \end{aligned}$$

leading to a different notation for Hooke's law for the plane strain situation.

$$\begin{aligned} \begin{pmatrix} \sigma_x \\ \sigma_y \\ \tau_{xy} \end{pmatrix} &= \frac{E}{1+\nu} \begin{bmatrix} \frac{1-\nu}{1-2\nu} & \frac{\nu}{1-2\nu} & 0 \\ \frac{\nu}{1-2\nu} & \frac{1-\nu}{1-2\nu} & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix} = \frac{E^*(1-\nu^2)}{1+\nu} \begin{bmatrix} \frac{1}{1-\nu^*} & \frac{\nu^*}{1-\nu^*} & 0 \\ \frac{\nu^*}{1-\nu^*} & \frac{1}{1-\nu^*} & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix} \\ &= \frac{E^*}{(1-\nu^*)(1+\nu^*)} \begin{bmatrix} 1 & \nu^* & 0 \\ \nu^* & 1 & 0 \\ 0 & 0 & 1-\nu^* \end{bmatrix} \cdot \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix}. \end{aligned}$$

This is identical to Hooke's law (19) for the plane stress situation, but with  $E^*$  and  $\nu^*$  instead of  $E$  and  $\nu$ . The energy density is in this case given by

$$\begin{aligned} W_{strain} &= \frac{1}{2} \frac{E}{(1+\nu)(1-2\nu)} \left\langle \begin{bmatrix} 1-\nu & \nu & 0 \\ \nu & 1-\nu & 0 \\ 0 & 0 & 2(1-2\nu) \end{bmatrix} \cdot \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix}, \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix} \right\rangle \\ &= \frac{E(1-\nu)}{2(1+\nu)(1-2\nu)} \left( \varepsilon_{xx}^2 + \varepsilon_{yy}^2 + 2\frac{\nu}{1-\nu}\varepsilon_{xx}\varepsilon_{yy} + 2\frac{1-2\nu}{1-\nu}\varepsilon_{xy}^2 \right) \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2} \frac{E^*}{1 - (\nu^*)^2} \left\langle \begin{bmatrix} 1 & \nu^* & 0 \\ \nu^* & 1 & 0 \\ 0 & 0 & 2(1 - \nu^*) \end{bmatrix} \cdot \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix}, \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix} \right\rangle \\
&= \frac{E^*}{2(1 - (\nu^*)^2)} (\varepsilon_{xx}^2 + \varepsilon_{yy}^2 + 2\nu^* \varepsilon_{xx} \varepsilon_{yy} + 2(1 - \nu^*) \varepsilon_{xy}^2) \quad (24)
\end{aligned}$$

Now the plane strain energy density has the same form as the plane stress energy density, but with modified constants.

$$W_{stress} = \frac{E}{2(1 - \nu^2)} (\varepsilon_{xx}^2 + \varepsilon_{yy}^2 + 2\nu \varepsilon_{xx} \varepsilon_{yy} + 2(1 - \nu) \varepsilon_{xy}^2) .$$

For a plane strain problem Bernoulli's principle is used and the corresponding total energy minimized, similar to expression (21).

## 2.14 Elasticity problems for axisymmetric solids, using cylindrical coordinates

Examine a domain  $(x, r) = (r, z) \in \Omega \subset \mathbb{R}^2$  and revolve this domain about the  $z$ -axis to generate a volume in space  $\mathbb{R}^3$ . Assume that the displacements are axisymmetric, i.e.

$$\begin{pmatrix} u_1(x, y, z) \\ u_2(x, y, z) \\ u_3(x, y, z) \end{pmatrix} = \begin{pmatrix} u_r(r, z) \cos \varphi \\ u_r(r, z) \sin \varphi \\ u_z(r, z) \end{pmatrix} .$$

To determine the elastic energy in this deformed solid determine<sup>4</sup> the strains in the plane  $\varphi = 0$ .

$$\begin{aligned}
\varepsilon_{xx} &= \frac{\partial u_1}{\partial x} = \cos^2 \varphi \frac{\partial u_r}{\partial r} + \frac{\sin^2 \varphi}{r} u_r = \frac{\partial u_r}{\partial r} \\
\varepsilon_{yy} &= \frac{\partial u_2}{\partial y} = \sin^2 \varphi \frac{\partial u_r}{\partial r} + \frac{\cos^2 \varphi}{r} u_r = \frac{1}{r} u_r \\
\varepsilon_{zz} &= \frac{\partial u_3}{\partial z} \\
2\varepsilon_{xy} &= \frac{\partial u_1}{\partial y} + \frac{\partial u_2}{\partial x} = \cos \varphi \sin \varphi \frac{\partial u_r}{\partial r} - \frac{\cos \varphi \sin \varphi}{r} u_r + \cos \varphi \sin \varphi \frac{\partial u_r}{\partial r} - \frac{\sin \varphi \cos \varphi}{r} u_r = 0 \\
2\varepsilon_{xz} &= \frac{\partial u_1}{\partial z} + \frac{\partial u_3}{\partial x} = \cos \varphi \frac{\partial u_r}{\partial z} + \cos \varphi \frac{\partial u_z}{\partial r} - \frac{\sin \varphi}{r} \frac{\partial u_z}{\partial \varphi} = \frac{\partial u_r}{\partial z} + \frac{\partial u_z}{\partial r} \\
2\varepsilon_{yz} &= \frac{\partial u_2}{\partial z} + \frac{\partial u_3}{\partial y} = \sin \varphi \frac{\partial u_r}{\partial z} + \sin \varphi \frac{\partial u_z}{\partial r} + \frac{\cos \varphi}{r} \frac{\partial u_z}{\partial \varphi} = 0
\end{aligned}$$

This leads to

$$\begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{zz} \\ \varepsilon_{xy} \\ \varepsilon_{xz} \\ \varepsilon_{yz} \end{pmatrix} = \begin{pmatrix} \frac{\partial u_r}{\partial r} \\ \frac{1}{r} u_r \\ \frac{\partial u_z}{\partial z} \\ 0 \\ \frac{1}{2} \left( \frac{\partial u_r}{\partial z} + \frac{\partial u_z}{\partial r} \right) \\ 0 \end{pmatrix} = \begin{pmatrix} \varepsilon_{rr} \\ \varepsilon_{\varphi\varphi} \\ \varepsilon_{zz} \\ 0 \\ \varepsilon_{rz} \\ 0 \end{pmatrix} .$$

<sup>4</sup>For functions  $f(x, y, z) = F(r, \varphi, z)$  (i.e. the identical function written in Cartesian and polar coordinates) use the computational rule (chain rule) to conclude

$$\frac{\partial f}{\partial x} = \cos \varphi \frac{\partial F}{\partial r} - \frac{\sin \varphi}{r} \frac{\partial F}{\partial \varphi} \quad \text{and} \quad \frac{\partial f}{\partial y} = \sin \varphi \frac{\partial F}{\partial r} + \frac{\cos \varphi}{r} \frac{\partial F}{\partial \varphi} .$$

Observe that the angular strain  $\varepsilon_{\varphi\varphi}$  is given by the displacement  $\varepsilon_{\varphi\varphi} = \frac{1}{r} u_r$ . The energy density (17) in the  $rz$ -plane at angle  $\varphi = 0$  is given by

$$\begin{aligned} W(r, z) &= \frac{1}{2} \left\langle \begin{pmatrix} \sigma_x \\ \sigma_y \\ \sigma_z \end{pmatrix}, \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{zz} \end{pmatrix} \right\rangle + \left\langle \begin{pmatrix} \tau_{xy} \\ \tau_{xz} \\ \tau_{yz} \end{pmatrix}, \begin{pmatrix} 0 \\ \varepsilon_{xz} \\ 0 \end{pmatrix} \right\rangle \\ &= \frac{1}{2} \frac{E}{(1+\nu)(1-2\nu)} \left\langle \begin{bmatrix} 1-\nu & \nu & \nu \\ \nu & 1-\nu & \nu \\ \nu & \nu & 1-\nu \end{bmatrix} \begin{pmatrix} \varepsilon_{rr} \\ \frac{1}{r} u_r \\ \varepsilon_{zz} \end{pmatrix}, \begin{pmatrix} \varepsilon_{rr} \\ \frac{1}{r} u_r \\ \varepsilon_{zz} \end{pmatrix} \right\rangle + \frac{E}{1+\nu} \varepsilon_{rz}^2 \\ &= \frac{1}{2} \frac{E}{(1+\nu)(1-2\nu)} \left( (1-\nu) (\varepsilon_{rr}^2 + \varepsilon_{zz}^2 + \frac{1}{r^2} u_r^2) + 2\nu (\varepsilon_{rr}\varepsilon_{zz} + \frac{1}{r} u_r (\varepsilon_{rr} + \varepsilon_{zz})) \right) + \\ &\quad + \frac{E}{1+\nu} \varepsilon_{rz}^2 . \end{aligned}$$

To find the elastic energy in the deformed solid this expression can be integrated with respect to the angle  $\varphi$ , leading to an integral over the domain  $\Omega \subset \mathbb{R}^2$ . The contributions to the total energy by the volume and surface forces lead to similar expression, and finally to the total energy, similar to (21).

$$\begin{aligned} U(\vec{u}) &= U_{elast} + U_{Vol} + U_{Surf} \tag{25} \\ &= \iint_{\Omega} \frac{2\pi r E}{2(1+\nu)(1-2\nu)} \left( (1-\nu) (\varepsilon_{rr}^2 + \varepsilon_{zz}^2 + \frac{1}{r^2} u_r^2) + 2\nu (\varepsilon_{rr}\varepsilon_{zz} + \frac{1}{r} u_r (\varepsilon_{rr} + \varepsilon_{zz})) \right) dA + \\ &\quad + \iint_{\Omega} \frac{2\pi r E}{1+\nu} \varepsilon_{rz}^2 dA - \iint_{\Omega} 2\pi r \vec{f} \cdot \vec{u} dA - \int_{\Gamma_2} 2\pi r \vec{g}_N \cdot \vec{u} ds . \end{aligned}$$

Some of the contributions are similar to the elastic energy for plane stress problems (20) or (21), i.e.

$$W_{stress} = \frac{E}{2(1-\nu^2)} (\varepsilon_{xx}^2 + \varepsilon_{yy}^2 + 2\nu \varepsilon_{xx} \varepsilon_{yy} + 2(1-\nu) \varepsilon_{xy}^2) ,$$

but there are some more contributions. The factor  $r$  will change all expressions for the element stiffness matrices, i.e. new, slightly more complex Octave codes are required.

Using the Bernoulli principle this energy has to be minimized. It is this minimization problem that is solved, subject to the boundary conditions (14).

### 3 Illustrative Examples

Solving a BVP (Boundary Value Problem) or an IBVP (Initial Boundary Value Problem) with FEM usually involves three steps:

1. Generate the mesh to be used for the problem. With this step the type of element can be selected, i.e. linear, quadratic or cubic. The type of boundary conditions are specified.
2. Define the functions describing the problem and then apply the finite element algorithm to generate an approximate solution.
3. Visualize and analyze the obtained solution.

For all three steps FEMoctave provides tools and the following examples illustrate the procedures.

#### 3.1 Solving elliptic problems, static heat equations

##### 3.1.1 A symmetric problem

On a rectangle  $\Omega = [0, 1] \times [0, 2]$  with Dirichlet boundary  $\Gamma_1$  at  $x = 0$  and at  $y = 0$  and Neumann boundary  $\Gamma_2$  at  $x = 1$  and at  $y = 2$  seek a solution of

$$\begin{aligned} -\Delta u &= 0.25 & \text{for } (x, y) \in \Omega \\ u &= 0 & \text{for } (x, y) \in \Gamma_1 \\ \frac{\partial u}{\partial n} &= 0 & \text{for } (x, y) \in \Gamma_2 \end{aligned}$$

The solution is computed and displayed with the help of three commands.

- Divide the  $x$  and  $y$  axis in subintervalls of length 0.1 and generate the resulting rectangular mesh using `CreateMeshRect()`. Use the options `..., -1, -2, -1, -2`) to indicate the boundary conditions at the four edges in order lower, upper, left and right. In this example use the order Dirichlet, Neumann, Dirichlet, Neumann.
- Use `BVP2Dsym()` with constant coefficients to generate and solve the system of linear equation by the FEM.
- Use `FEMtrimesh()` to display the solution.

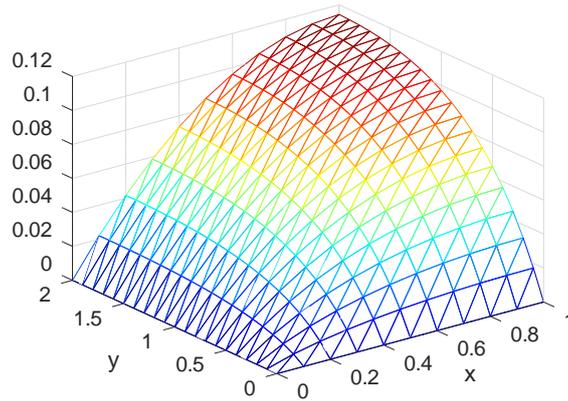
#### LaplaceRectangle.m

```
FEMmesh = CreateMeshRect([0:0.1:1], [0:0.1:2], -1, -2, -1, -2);
%%FEMmesh = MeshUpgrade(FEMmesh, 'quadratic'); %% uncomment to use quadratic elements
%%FEMmesh = MeshUpgrade(FEMmesh, 'cubic');      %% uncomment to use cubic elements

u = BVP2Dsym(FEMmesh, 1, 0, 0.25, 0, 0, 0);

figure(1); FEMtrimesh(FEMmesh, u);
           xlabel('x'); ylabel('y');
```

Find the result in Figure 5. The above code is using linear elements. To use quadratic or cubic elements uncomment one of the lines with `MeshUpgrade()`.

Figure 5: Solution of  $-\Delta u = 0.25$  on a rectangle

### 3.1.2 Laplace equation in cylindrical coordinates

The Laplace operator in cylindrical coordinates is given by

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = \frac{1}{\rho} \frac{\partial}{\partial \rho} \left( \rho \frac{\partial u}{\partial \rho} \right) + \frac{1}{\rho^2} \frac{\partial^2 u}{\partial \theta^2} + \frac{\partial^2 u}{\partial z^2}.$$

Assuming that the solution is independent on the angle  $\theta$ , then the Laplace equation  $-\Delta u(\rho, z) + b_0(\rho, z) = f(\rho, z)$  is given by

$$-\frac{\partial}{\partial \rho} \left( \rho \frac{\partial u}{\partial \rho} \right) - \frac{\partial}{\partial z} \left( \rho \frac{\partial u}{\partial z} \right) + \rho b_0(\rho, z) = \rho f(\rho, z).$$

Thus it is in the form of equation (2), with  $x = \rho$  and  $y = z$ . As an example consider  $b_0(\rho, z) = 10$  and  $f(\rho, z) = 2z$ . The domain  $\Omega$  to be examined is given by  $0 \leq \rho \leq 2$  and  $-1 \leq z \leq 2$  and the boundary conditions are

$$\begin{aligned} \frac{\partial u(0, z)}{\partial \rho} &= 0 \quad \text{symmetry for } -1 < z < 2 \\ \rho \frac{\partial u(2, z)}{\partial \rho} &= -1 \quad \text{flux out of domain for } -1 < z < 2 \\ u(\rho, -1) = u(\rho, 2) &= 0 \quad \text{given value for } 0 < \rho < 2 \end{aligned}$$

Since the coefficient functions in (2) are not constants define these functions in *Octave* and then use `BVP2Dsym()` to solve the problem. Observe that both Neumann boundary conditions are described by the same function  $g_2(\rho, z) = \frac{-\rho}{2}$ , since  $g_2(0, z) = 0$  and  $g_2(2, z) = -1$ . The code is shown below and find the result in Figure 6.

#### LaplaceCylindrical.m

```
FEMmesh = CreateMeshRect(linspace(0, 2, 20), linspace(-1, 2, 30), -1, -1, -2, -2);
%%FEMmesh = MeshUpgrade(FEMmesh, 'quadratic'); %% uncomment to use quadratic elements

function res = f(rz, dummy)   res = rz(:, 1) * 2. * rz(:, 2); endfunction
function res = b0(rz, dummy)  res = 10 * rz(:, 1);           endfunction
function res = a(rz, dummy)   res = rz(:, 1);             endfunction
function res = g2(rz)         res = -1 * rz(:, 1) / 2;       endfunction

u = BVP2Dsym(FEMmesh, 'a', 'b0', 'f', 0, 'g2', 0);
```

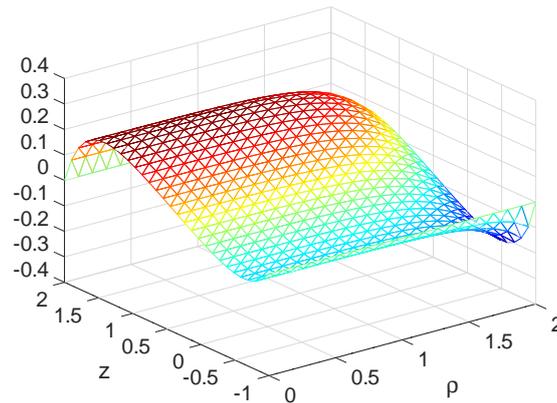


Figure 6: Solution of the Laplace equation in cylindrical coordinates

```
FEMtrimesh(FEMmesh,u);
xlabel('\rho'); ylabel('z');
```

### 3.1.3 Diffusion on an L-shaped domain

Examine a BVP on an L-shaped domain, as created in Section 4.1. The equation to be solved is

$$\begin{aligned} -\Delta u &= 1 & \text{for } (x,y) \in \Omega \\ \frac{\partial u}{\partial n} &= -2u & \text{for } (x,y) \in \Gamma \end{aligned}$$

For this problem there is no Dirichlet condition and it is solved in three steps.

- Generate the L-shaped domain with the help of `CreateMeshTriangle()`.
- Solve the equations with `BVP2Dsym()`.
- Display the result with `FEMtrimesh()` and `FEMtricontour()`.
- The code below uses linear elements. Uncommenting the line with `MeshUpgrade()` will solve the same problem using second or third order elements.

Find the code below and the result in Figure 7.

#### DiffusionLshape.m

```
nodes = [0,0,-2;1,0,-2;1,1,-2;-1,1,-2;-1,-1,-2;0,-1,-2];
FEMmesh = CreateMeshTriangle('Ldomain',nodes,0.02);
FEMmesh = MeshUpgrade(FEMmesh,'cubic'); %% uncomment to use cubic elements

u = BVP2Dsym(FEMmesh,1,0,1,0,0,-2);

figure(1); FEMtrimesh(FEMmesh,u);
          xlabel('x'); ylabel('y'); view(-30,30)

figure(2); FEMtricontour(FEMmesh,u);
          xlabel('x'); ylabel('y');
```

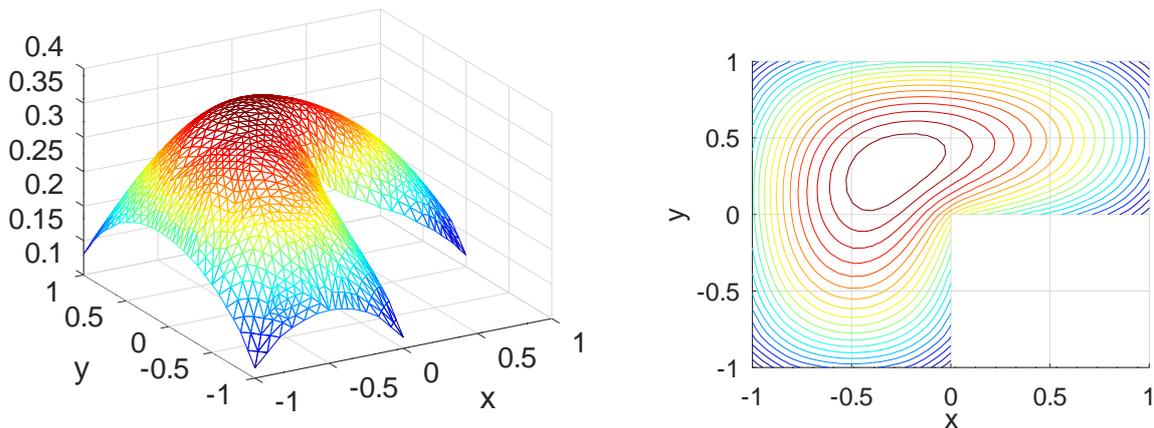


Figure 7: Solution of a diffusion problem on a L-shaped domain

### 3.1.4 A diffusion convection problem

Examine a steady state heat problem on the square  $\Omega = [0, 2] \times [0, 2]$  with constant heating ( $f(x, y) = +0.1$ ) and a strong convection in  $x$  direction ( $b_x(x, y) = 10$ ) and a weaker convection in  $y$  direction ( $b_y(x, y) = 5$ ). This leads to the PDE

$$-\Delta u + 10 \frac{\partial u}{\partial x} + 5 \frac{\partial u}{\partial y} = 0.1.$$

The temperature on all of the boundary vanishes. This is a problem of type (1). Solve the BVP with the code below and find the resulting level curves of the temperature in Figure 8.

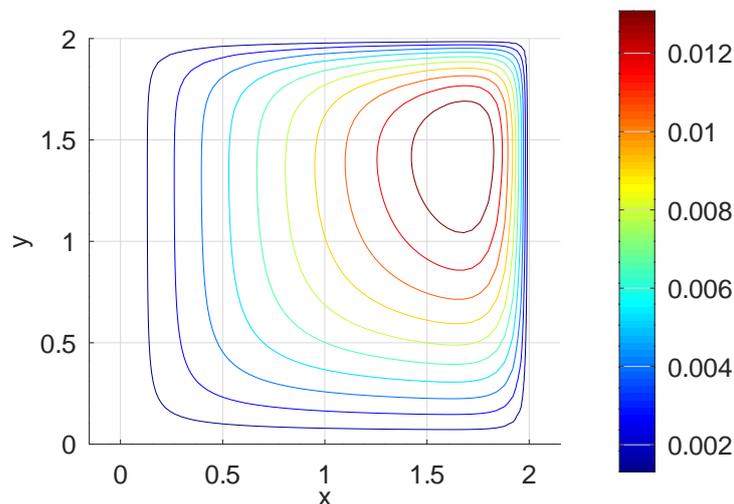


Figure 8: Solution of a diffusion convection problem

#### DiffusionConvection.m

```
FEMmesh = CreateMeshRect(linspace(0, 2, 51), linspace(0, 2, 51), -1, -1, -1, -1);
u = BVP2D(FEMmesh, 1, 0, 10, 5, 0.1, 0, 0, 0);
```

```
figure(1); FEMtricontour(FEMmesh,u,10);
colorbar(); xlabel('x'); ylabel('y'); grid on
```

The above code uses elements of order 1. To use elements of order 2 on a similar mesh one can first generate a mesh with linear elements and then use `MeshUpgrade()` to generate a mesh with elements of order 2. Convert the mesh back to linear elements, but with the identical nodes, i.e. use `MeshQuad2Linear()` and then display.

#### DiffusionConvection.m

```
FEMmesh = CreateMeshRect(linspace(0,2,26),linspace(0,2,26),-1,-1,-1,-1);
FEMmesh = MeshUpgrade(FEMmesh, 'quadratic'); %% make a mesh with elements of order 2
u = BVP2D(FEMmesh,1,0,10,5,0.1,0,0,0);
FEMmesh = MeshQuad2Linear(FEMmesh); %% convert to identical mesh with linear elements

figure(1); FEMtricontour(FEMmesh,u,10);
colorbar(); xlabel('x'); ylabel('y'); grid on
```

### 3.2 Solving eigenvalue problems

As a first eigenvalue problem compute the eigenvalues and eigenfunctions of the Laplace operator on the unit disc with Dirichlet boundary conditions, i.e. determine a scalar  $\lambda$  and nontrivial function  $u$  such that

$$-\Delta u = \lambda u \quad \text{on unit disc and} \quad u = 0 \quad \text{on the boundary.}$$

The goal is to compute the four smallest eigenvalues and display the fourth eigenfunction. Proceed in three steps.

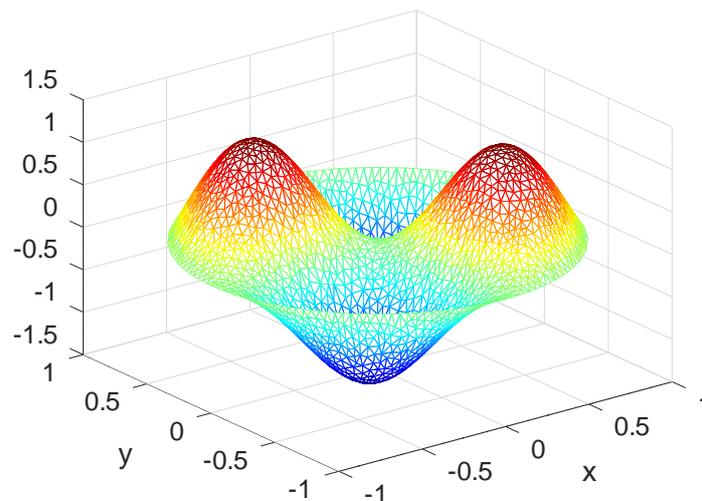
- Use `CreateTriangleMesh()` to generate the mesh on the unit disc.
- Use `BVP2Deig()` with constant coefficients to generate and solve the eigensystem.
- Use `FEMtrimesh()` to display the fourth eigenfunction. Find the result in Figure 9.
- To use second order element, use `MeshUpgrade()`.

The computed eigenvalues are  $\lambda_1 \approx 5.7857$ ,  $\lambda_2 = \lambda_3 \approx 14.6959$  and  $\lambda_4 \approx 26.4169$ . These values coincide nicely with the squares of the first zeros of the Bessel functions  $J_0(r)$ ,  $J_1(r)$  and  $J_2(r)$ , the values of the exact problem.

#### EigenvaluesDisc.m

```
xM = 0; yM = 0; R = 1; N = 160; alpha = linspace(0,N/(N+1)*2*pi,N)';
xy = [xM+R*cos(alpha),yM+R*sin(alpha),-ones(size(alpha))];

FEMmesh = CreateMeshTriangle('circle',xy,0.0005);
%%FEMmesh = MeshUpgrade(FEMmesh,'quadratic');
%%%%%% solve the eigenvalue problem, show the eigenvalues
%%[la,ve] = BVP2Deig(FEMmesh,1,0,1,0,4);
[la,ve,errorbound] = BVP2Deig(FEMmesh,1,0,1,0,4);
eigenvalues = la
errorbound
exact_values = [fsolve(@(x)besselj(0,x),2.3), fsolve(@(x)besselj(1,x),3.8),...
               fsolve(@(x)besselj(2,x),5)].^2
figure(1); FEMtrimesh(FEMmesh,ve(:,4));
xlabel('x'); ylabel('y');
```

Figure 9: The fourth eigenfunction of  $\Delta u = \lambda u$  on a disc

The result shows the first 4 eigenvalues and their corresponding error bounds. The error bounds of  $10^{-28}$  for the first eigenvalue is not to be taken too seriously, it just means *accurate up to machine precision* as eigenvalue of the global stiffness matrix. Observe that these are the eigenvalues of the FEM approximation to the boundary value problem. They are close to the eigenvalues of the continuous problem, i.e. the squares of the zeros of the Bessel functions.

#### Octave

```
eigenvalues =    5.7857
                14.6959
                14.6961
                26.4169

errorbound =    2.5479e-12  1.6604e-28
               2.9179e-12  7.0763e-16
               3.2020e-12  7.2782e-15
               3.5589e-12  2.3726e-28

exact_values =  5.7832  14.6820  26.3746
```

In cylindrical coordinates the Laplace operator is given by

$$\Delta u(r, \theta) = \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial u(r, \theta)}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 u(r, \theta)}{\partial \theta^2}.$$

For purely radial solutions the eigenvalue problem on a disk of radius  $R$  leads to

$$\begin{aligned} -(r u'(r))' &= \lambda r u(r) & \text{for } 0 < r < R \\ u'(0) &= 0 \\ u(R) &= 0 \end{aligned}.$$

This equation can be solved with the help of `BVP1Deig()`. Find the graphical result for the first four radial eigenfunctions in Figure 10(a). The exact values are the squares of the zeros<sup>5</sup> of the Bessel function  $J_0(r)$ .

<sup>5</sup>The function `fsolve()` is used to determine the zeros. To find approximate values use a plot of the function `besselj(0, x)`.

Obviously the first eigenvalue coincides with the above 2D approach. The other eigenvalues differ, since the above code takes angular dependence into account, while `BVP1Deig()` examines radial dependence exclusively.

```

                                EigenvaluesDisk1D.m
R = 1; N = 40; interval = linspace(0,R,N)';
f_r = @(r) r;
[r,eVal,eVec] = BVP1Deig(interval,f_r,0,0,f_r,[0,0],0,4);

figure(1); plot(r,eVec); xlabel('r'); ylabel('u');
            legend('1','2','3','4','location','southeast')
eVal_FEM    = eVal'
exact_values = [fsolve(@(x)besselj(0,x),2.3),fsolve(@(x)besselj(0,x),5.4),...
                fsolve(@(x)besselj(0,x),9),fsolve(@(x)besselj(0,x),12)].^2
-->
eVal_FEM =      5.7832    30.4713    74.8872    139.0416
exact_values =  5.7832    30.4713    74.8866    139.0403

```

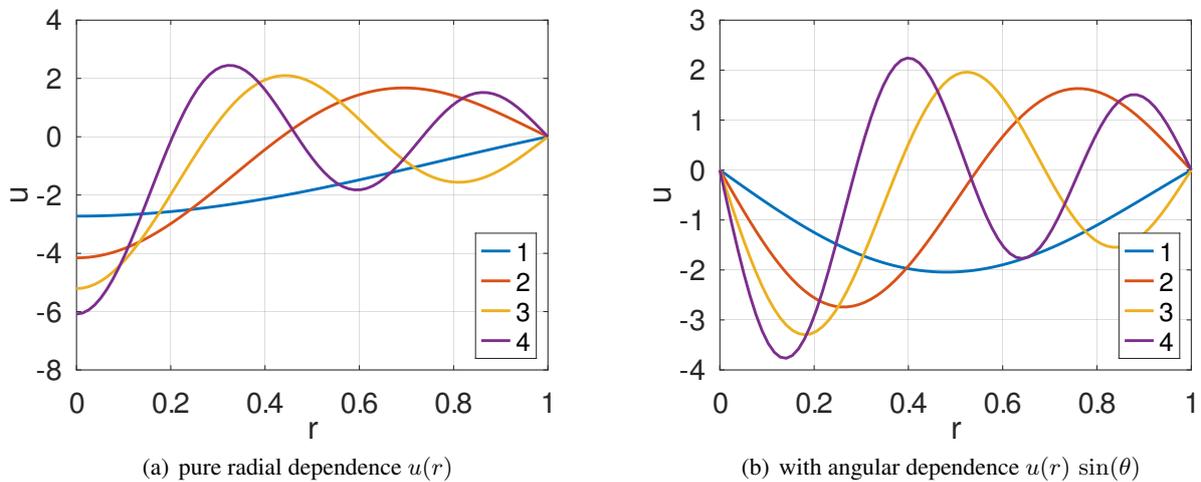


Figure 10: The first four radial eigenmodes of the Laplace operator on a disk of radius 1, for  $u(r)$  and  $u(r) \sin(\theta)$

To obtain the eigenfunctions with angular dependence use separation of variables. With function  $f(r, \theta) = u(r) \sin(n\theta + \delta)$  obtain

$$\begin{aligned} \Delta f(r, \theta) &= \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial f(r, \theta)}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 f(r, \theta)}{\partial \theta^2} \\ &= \left( \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial u(r)}{\partial r} \right) - n^2 \frac{1}{r^2} u(r) \right) \sin(n\theta + \delta). \end{aligned}$$

Thus the corresponding eigenvalue problem is

$$\begin{aligned} -(r u'(r))' + \frac{1}{r} n^2 u(r) &= \lambda r u(r) \quad \text{for } 0 < r < R \\ u'(0) &= 0 \\ u(R) &= 0 \end{aligned}$$

Find the result for  $n = 1$  of `BVP1Deig()` in Figure 10(b). The exact values are the squares of the zeros of the Bessel function  $J_n(r)$ .

**EigenvaluesDisk1D.m**

```

n = 1; n2_r = @(r)n^2./r;
[r,eVal2,eVec2] = BVP1Deig(interval,f_r,0,n2_r,f_r,[0,0],0,4);

figure(2); plot(r,eVec2); xlabel('r'); ylabel('u');
    legend('1','2','3','4','location','southeast')
eVal2_FEM = eVal2'
exact_values2 = [fsolve(@(x)besselj(n,x),4),fsolve(@(x)besselj(n,x),7),...
    fsolve(@(x)besselj(n,x),10),fsolve(@(x)besselj(n,x),13)].^2
-->
eVal2_FEM      = 14.682    49.219    103.500    177.523
exact_values2  = 14.682    49.218    103.499    177.521

```

**3.3 Solving parabolic problems, dynamic heat equations**

As an example solve the dynamic heat equation

$$\frac{\partial u}{\partial t} - \Delta u + 10 \frac{\partial u}{\partial x} + 5 \frac{\partial u}{\partial y} = 0.1 \quad \text{for } 0 < x, y < 2$$

with zero Dirichlet boundary conditions and the initial temperature

$$u(0, x, y) = u_0(x, y) = 0.005 x (2 - x)^2 y (2 - y).$$

The solution is computed at 7 equally spaced times  $t_i$  between 0 and 0.1. In-between 10 steps are taken, but the solution is not returned. Find the result of the code below in Figure 11. At time 0 the maximal value is attained at  $(x, y) = (\frac{2}{3}, 1)$ . The convection term  $+10 \frac{\partial u}{\partial x} + 5 \frac{\partial u}{\partial y}$  then moves the point of maximal temperature to the upper right section of the square. For large times  $t$  the solution will converge to the steady state solution shown in Figure 8 in Section 3.1.4.

**HeatDynamic.m**

```

% generate the mesh
FEMmesh = CreateMeshRect(linspace(0,2,31),linspace(0,2,31),-1,-1,-1,-1);
x = FEMmesh.nodes(:,1);y = FEMmesh.nodes(:,2);
% setup and solve the initial boundary value problem
m=1; a=1; b0=0; bx=10; by=5; f=0.1; gD=0; gN1=0; gN2=0;
t0=0; tend=0.1 ; steps = [6,10];
u0 = zeros(length(FEMmesh.nodes),1);
u0 = 0.005*(2-x).^2.*x.*y.*(2-y);
[u_dyn,t] = IBVP2D(FEMmesh,m,a,b0,bx,by,f,gD,gN1,gN2,u0,t0,tend,steps);
% show the animation on screen
u_max = max(u_dyn(:));
for t_ii = 1:length(t)
    figure(2); FEMtrimesh(FEMmesh,u_dyn(:,t_ii))
    xlabel('x'); ylabel('y'); caxis([0,u_max]); axis([0 2 0 2 0 u_max]); drawnow();
    figure(3); FEMtricontour(FEMmesh.elem,x,y,u_dyn(:,t_ii),linspace(0,0.99*u_max,11))
    xlabel('x'); ylabel('y'); caxis([0,u_max]); drawnow();
    pause(1)
endfor

```

**3.4 Solving hyperbolic problems, wave equations**

As an example solve the wave equation

$$\frac{\partial^2 u}{\partial t^2} - \Delta u = 0 \quad \text{for } x^2 + y^2 < 6$$

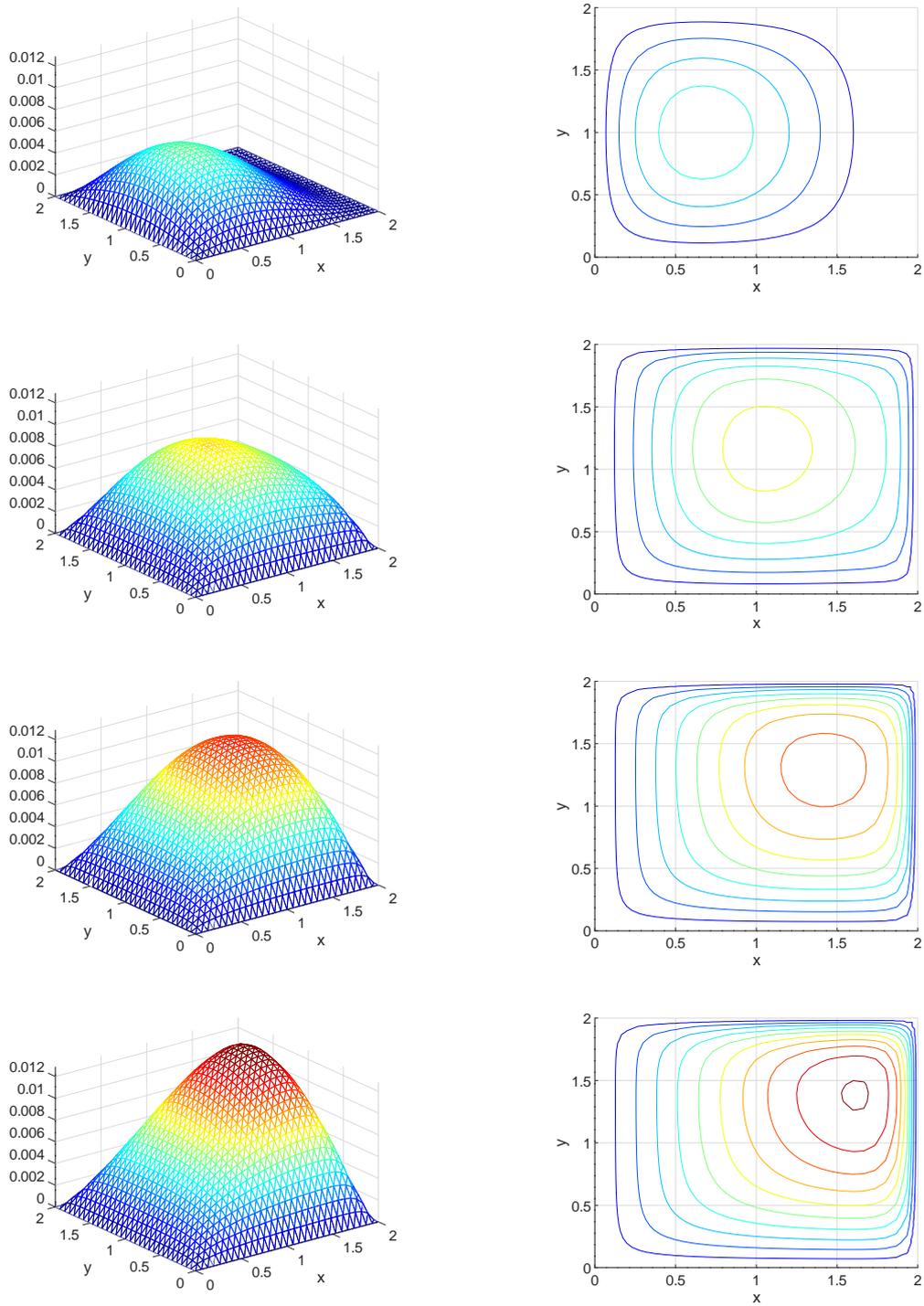


Figure 11: Solution of a dynamic heat equation

with zero Dirichlet boundary conditions, the initial displacement

$$u(0, x, y) = u_0(x, y) = 0.1 \exp(-(x-1)^2 - y^2) (R^2 - x^2 - y^2) / R^2$$

and zero initial velocity  $v_0 = 0$ . This assures compatible initial values, i.e. the boundary condition is satisfied at time  $t = 0$ . The solution is computed at 15 equally spaced times  $t_i$  between 0 and 7. In-between 30 steps are taken, but the solution is not returned. The solution is returned at 15 times, leading to Figure 12. This initial hump is traveling towards the boundary of the circle with speed 1, where it is reflected. More examples are shown in Sections 9.2 and 9.12.

#### WaveDynamic.m

```

%% generate a circle
alpha = linspace(0,2*pi,101)'; alpha = alpha(1:end-1); R = 6;
xy = [R*cos(alpha), R*sin(alpha), -ones(size(alpha))];
if 1 %% linear elements
    FEMmesh = CreateMeshTriangle('Circle',xy,0.03);
else %% quadratic elements
    FEMmesh = CreateMeshTriangle('Circle',xy,4*0.03);
    FEMmesh = MeshUpgrade(FEMmesh);
endif

x = FEMmesh.nodes(:,1); y = FEMmesh.nodes(:,2);
v0 = zeros(size(x));
u0 = 0.1*exp(-1*((x-1).^2+y.^2)); u0 = u0.*(R^2-x.^2-y.^2)/R^2;
%% setup and solve the initial boundary value problem
m=1; d=0; a=1; b0=0; bx=0; by=0; f=0; gD=0; gN1=0; gN2=0;
t0=0; tend=7 ; steps=[14,30];
tic();
[u_dyn,t] = I2BVP2D(FEMmesh,m,d,a,b0,bx,by,f,gD,gN1,gN2,u0,v0,t0,tend,steps);
toc()

figure(1) %% show the animation on screen
for t_ii = 1:length(t)
    FEMtrimesh(FEMmesh,u_dyn(:,t_ii))
    xlabel('x'); ylabel('y'); axis([-R R -R R -0.05 0.05])
    caxis([-0.05 0.05]); text(4,-2,0.04,sprintf('t=%2.1f',t(t_ii)))
    drawnow(); pause(0.3)
endfor
-->
Elapsed time is 0.93231 seconds.

```

### 3.5 Solving 1D steady state boundary value problems

To solve the BVP

$$\begin{aligned}
 -u''(x) &= (1-x)^2 & \text{for } 0 \leq x \leq 3 \\
 u(0) &= 2 \\
 u'(3) &= -2
 \end{aligned}$$

on an interval with 19 nodes use the code in ODE1.m, leading to Figure 13(a).

- The initial grid consists on  $N - 1 = 9$  subintervals. The algorithm adds the midpoints and thus there are  $2 \cdot 9 + 1 = 19$  nodes.
- The Dirichlet condition  $u(0) = 2$  allows to remove one equation and consequently  $\mathbf{A}$  will be a  $18 \times 18$  matrix.

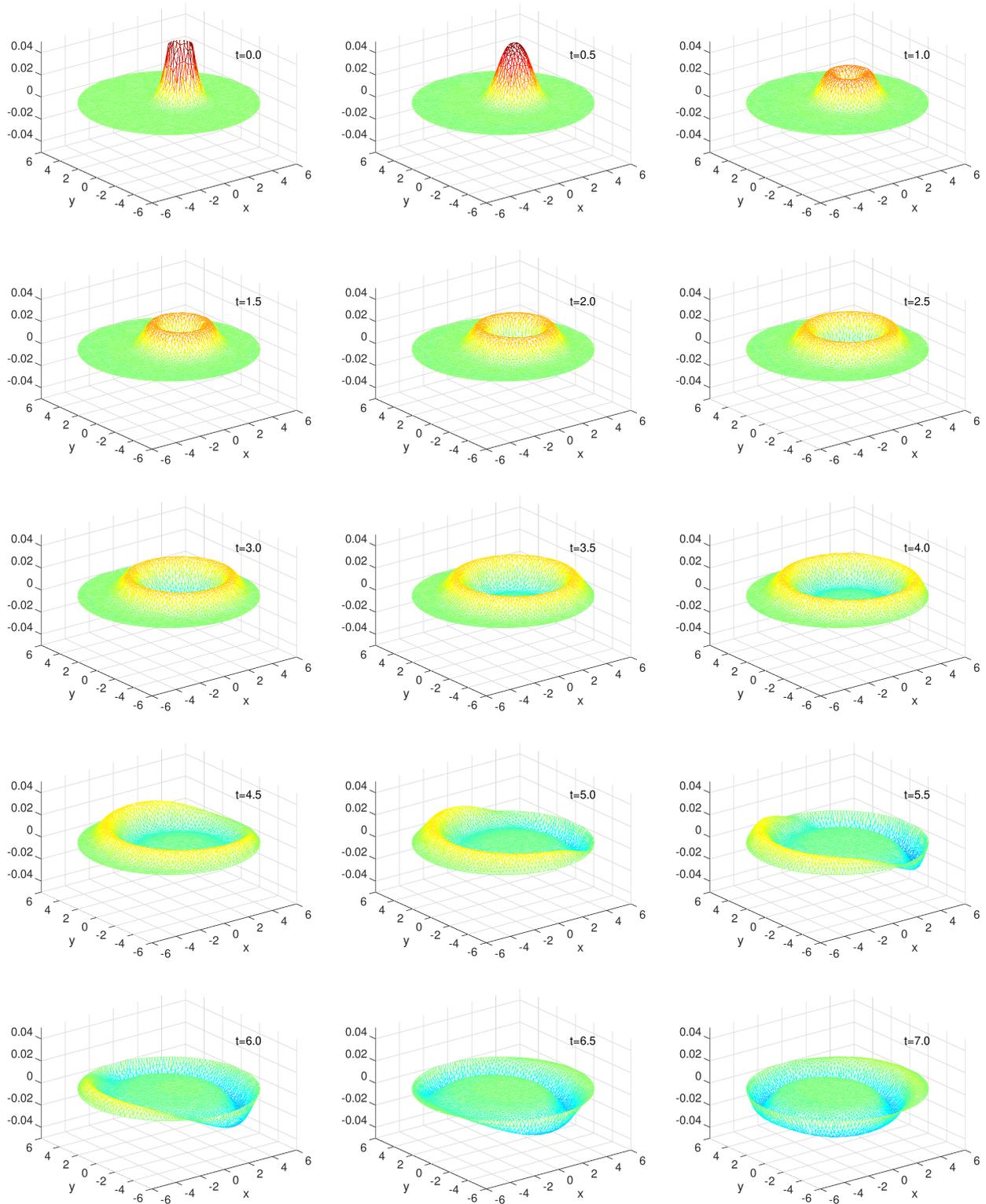


Figure 12: Solution of a wave equation

- Of the possible  $18 \cdot 18 = 324$  entries in the stiffness matrix  $\mathbf{A}$  only 68 are different from zero. The nonzero entries are in a band of width 5 around the diagonal of  $\mathbf{A}$ , visible in Figure 13(b). This figure was generated by (temporarily) adding a command `spy(A)` in the code `BVP1D.m`.

#### ODE1.m

```
N = 10; x = linspace(0,3,N);
[xn,u] = BVP1D(x,1,0,0,1,@(x)(1-x).^2,2,[-2,0]);
figure(1); plot(xn,u)
          xlabel('x'); ylabel('u')
```

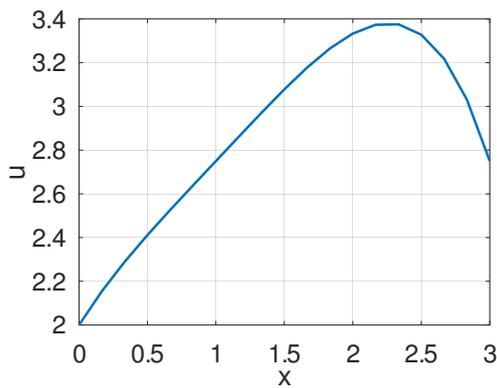
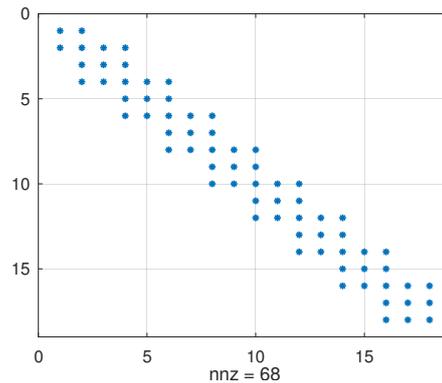
(a) the solution  $u(x)$ (b) the sparsity of  $\mathbf{A}$ 

Figure 13: The solution for a simple 1D boundary value problem and the sparsity of the matrix  $\mathbf{A}$

The constant and uniform heating of ball of radius  $R$  can be described by a boundary value problem. The constant  $f$  indicates how much thermal energy is added to the ball per volume and time.

$$\begin{aligned} -\frac{\partial}{\partial r} \left( r^2 \frac{\partial u(r)}{\partial r} \right) &= r^2 f & \text{for } 0 \leq r \leq R \\ \frac{\partial}{\partial r} u(0) &= 0 \\ u(R) &= 0 \end{aligned}$$

This BVP determines the steady state solution. The code below and the resulting Figure 14(a) confirms that the maximal temperature is attained at the center of the ball.

#### HeatBall1D.m

```
R = 5; f = 3;
N = 10; r = linspace(0,R,N);
[r,u] = BVP1D(r,@(r)r.^2,0,0,@(r)r.^2,f,[0,0],0);
figure(1); plot(r,u)
          xlabel('radius r'); ylabel('temperature u')
```

The thermal energy generated inside the radius  $r$  is given by  $f \frac{4\pi}{3} r^3$  and should be equal to the flux through the sphere with radius  $r$ , i.e.  $4\pi r^2 \frac{\partial u(r)}{\partial r}$ . The code below and the resulting Figure 14(b) confirms this observation.

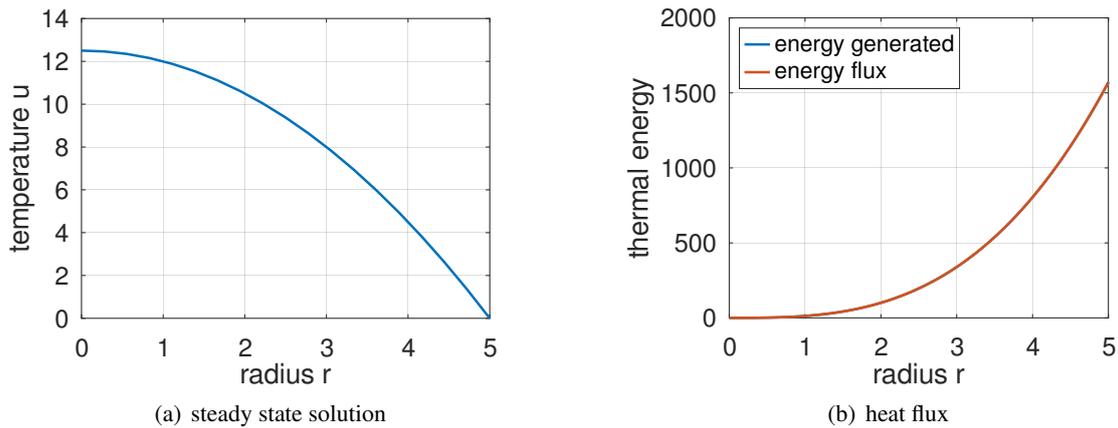


Figure 14: The solution of a steady state heat problem and the heat flux across spheres of radius  $r$

#### HeatBall1D.m

```
r_fine = linspace(0,R,1001);
[u_fine,du_fine] = pwquadinterp(r,u,r_fine);

figure(2); plot(r_fine,f*4/3*pi*r_fine.^3,r_fine,-4*pi*r_fine.^2.*du_fine)
            xlabel('radius r'); ylabel('thermal energy')
            legend('energy generated','energy flux','location','northwest')
```

### 3.6 Solving 1D dynamic initial boundary value problems of order 1, a heat equation

The initial boundary value problem describing the dynamics of heating a ball with radius  $R$  and initial temperature  $u(r, 0) = u_0(r) = 0$  is given by

$$\begin{aligned} r^2 \frac{\partial}{\partial t} u(x, t) - \frac{\partial}{\partial r} \left( r^2 \frac{\partial}{\partial r} u(r, t) \right) &= r^2 f(r, t) && \text{for } 0 < r < R \text{ and } t > 0 \\ \frac{\partial}{\partial r} u(0, t) &= 0 && \text{for } t > 0 \\ u(R, t) &= 0 && \text{for } t > 0 \\ u(r, 0) &= 0 && \text{for } 0 < r < R \end{aligned}$$

This assumes that the temperature  $u(r, t)$  depends on time  $t$  and radius  $r$  only and the temperature on the boundary  $r = R$  is kept at 0. Find the graphical output of the code below in Figures 15 and 16.

#### HeatingBallRadial.m

```
R = 3; BCleft = [0,0]; BCright = 0; f = 1;
u0 = 0; t0 = 0; t_end = 3;
steps = [10,10]; interval = linspace(0,R,11);
r_square = @(r) r.^2;
[r,u,t] = IBVP1D(interval,r_square,r_square,0,0,r_square,f,BCleft,BCright,...
                u0,t0,t_end,steps);

figure(1); plot(r,u(:,end))
            xlabel('radius r'); ylabel('temperature u at t=t_{end}')
figure(2); plot(t,u(1,:))
            xlabel('time t'); ylabel('temperature u at r=R')
figure(3); mesh(t,r,u)
```

```
xlabel('time t'); ylabel('radius r'); zlabel('temperature u')
figure(4); contour(t,r,u,[0.25:0.25:1.5])
xlabel('time t'); ylabel('radius r');
```

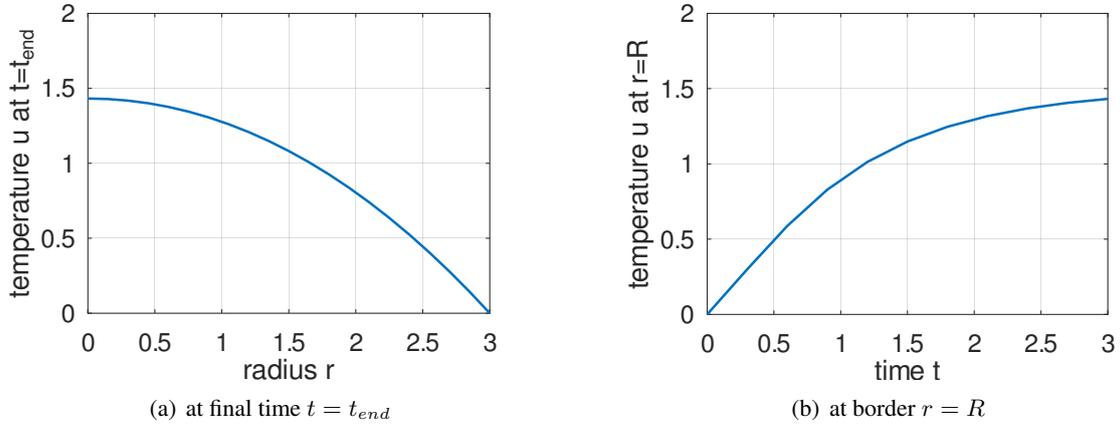


Figure 15: The solution of dynamic heating of a ball,  $u$  as function of  $r$  or  $t$

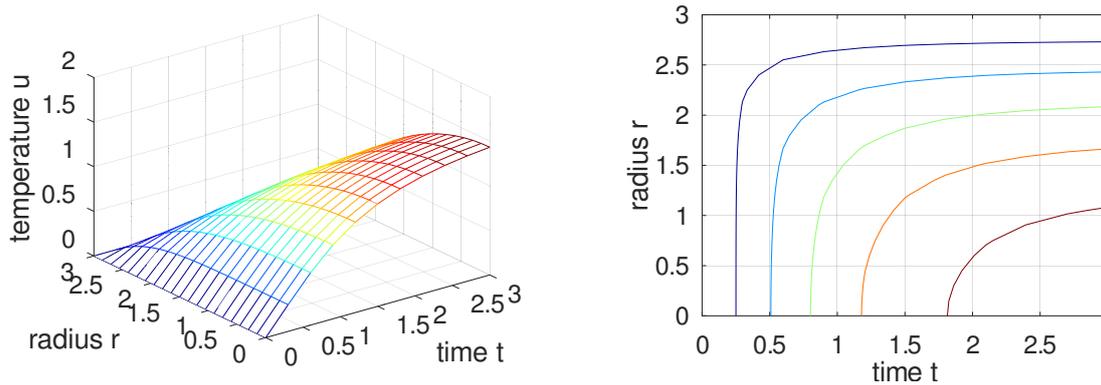


Figure 16: The solution of dynamic heating of a ball,  $u$  as function of  $r$  and  $t$ . The contours are at levels  $u = 0.25, 0.50, 0.75, 1.00$  and  $1.25$ .

### 3.7 Solving 1D dynamic initial boundary value problems of order 2, a wave equation

The initial boundary value problem describing the dynamics of a vibrating string with initial displacement

$$u(x, 0) = u_0(x) = \begin{cases} \sin(x) & \text{for } 0 \leq x \leq \pi \\ 0 & \text{for } \pi \leq x \leq 3\pi \end{cases}$$

and initial velocity

$$\frac{\partial}{\partial t} u(x, 0) = u_1(x) = \begin{cases} -\cos(x) & \text{for } 0 \leq x \leq \pi \\ 0 & \text{for } \pi \leq x \leq 3\pi \end{cases}$$

is given by

$$\begin{aligned} \frac{\partial^2}{\partial t^2} u(x, t) - \frac{\partial^2}{\partial x^2} u(x, t) &= 0 && \text{for } 0 < x < 3\pi \text{ and } t > 0 \\ u(0, t) &= 0 && \text{for } t > 0 \\ \frac{\partial}{\partial x} u(3\pi, t) &= 0 && \text{for } t > 0 \\ u(x, 0) &= u_0(x) && \text{for } 0 < x < 3\pi \\ \frac{\partial}{\partial t} u(x, 0) &= u_1(x) && \text{for } 0 < x < 3\pi \end{aligned}$$

Find the graphical output of the code below in Figures 17 and 18.

**Wave1D.m**

```
a = 1; b = 0; c = 0; d = 1; f = 0;
w2 = 1; w1 = 0; BCleft = 0; BCright = [0,0];
t0 = 0; tend = 25; steps = [100,20];
interval = linspace(0,3*pi,51)';
u0 = @(x)sin(x).*(x<=pi); u1 = @(x)-cos(x).*(x<=pi);

[x,u,t] = I2BVP1D(interval,w2,w1,a,b,c,d,f,BCleft,BCright,u0,u1,t0,tend,steps);

figure(1); mesh(t,x,u); xlabel('time t'); ylabel('position x'); zlabel('u')
    xlim([min(t),max(t)]); ylim([min(x),max(x)])
figure(2); contour(t,x,u,21); xlabel('time t'); ylabel('position x');
    colorbar
```

The initial hump of the form  $\sin(x)$  for  $0 \leq x \leq \pi$  is moving to the right with speed  $c = 1$ . At the border at  $x = 3\pi$  the pulse is reflected, caused by the Neumann boundary condition  $\frac{\partial}{\partial x} u(3\pi, t) = 0$ . Then it is moving back toward the border at  $x = 0$ . There it is reflected again, but with the negative of the amplitude, caused by the Dirichlet boundary condition  $u(0, t) = 0$ .

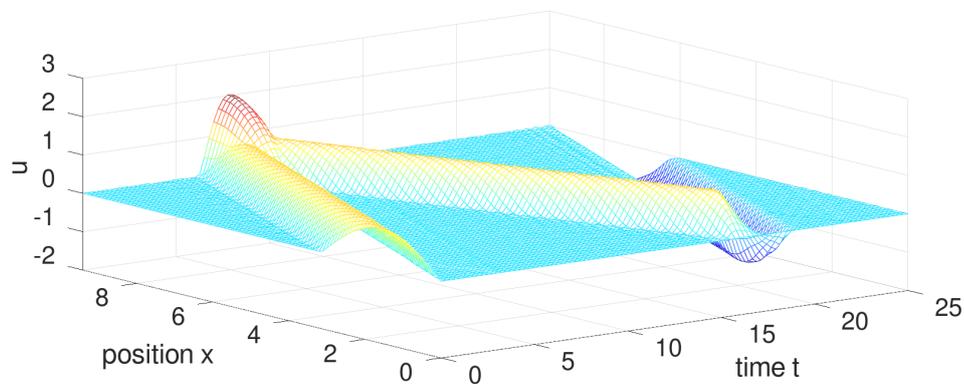


Figure 17: The amplitude of a vibrating string,  $u$  as function of  $x$  and  $t$ , the surface

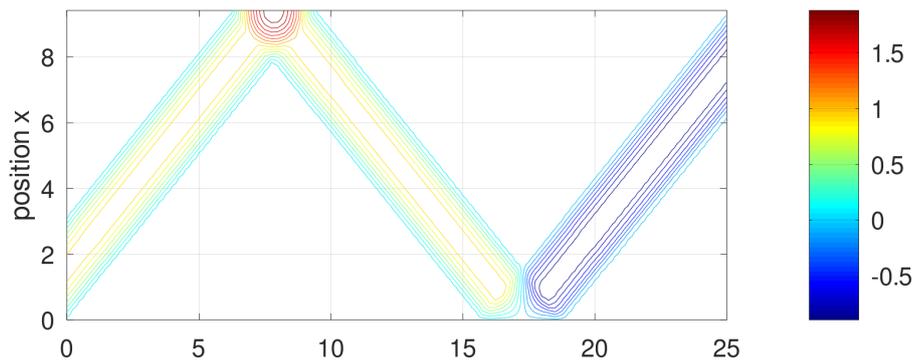


Figure 18: The amplitude of a vibrating string,  $u$  as function of  $x$  and  $t$ , the contour lines

### 3.8 Solving nonlinear 1D boundary value problems

#### 3.8.1 A nonlinear 1D BVP solved by BVP1DNL()

As example consider the boundary value problem

$$\begin{aligned} -\frac{d^2}{dx^2} u(x) &= \frac{1}{2} + \alpha(x u(x) + \sin(u(x))) & \text{for } -1 < x < 2 \\ u(-1) &= 1 \\ u(+2) &= 4 \end{aligned}$$

with the parameter  $\alpha = 0.1$ . The right hand side and its partial derivative are given by

$$\begin{aligned} f(x, u) &= \frac{1}{2} + \alpha(x u(x) + \sin(u(x))) \\ \frac{\partial}{\partial u} f(x, u) &= 0 + \alpha(x - \cos(u(x))) \end{aligned}$$

and as initial guess for the solution use  $u_0(x) = 1 + x$ . This nonlinear problem is then solved by the code below in `NL_Newton.m`. The iteration stopped after 3 steps. The progress of the iteration is displayed and the size of the last correction applied is  $\approx 5 \cdot 10^{-7}$ .

#### NL\_Newton.m

```

interval = linspace(-1,2,40)';
al = +0.1;
f = {@(x,u) 0.5+al*(x.*u+sin(u)), @(x,u)+al*(x-cos(u))};
BCleft = 1; BCright = 4;
u0 = @(x) 2+x;
[x,u,inform] = BVP1DNL(interval,1,0,0,1,f,BCleft,BCright,u0,'Display','iter');
inform
figure(1); plot(x,u); xlabel('x'); ylabel('u')
-->
iteration 1, RMS(correction) = 1.456674e-02, RMS(phi) = 1.695340e-02
iteration 2, RMS(correction) = 8.392369e-05, RMS(phi) = 9.773978e-05
iteration 3, RMS(correction) = 4.809044e-07, RMS(phi) = 5.599476e-07

inform = scalar structure containing the fields:
    info      = 1
    iter      = 3
AbsError = 4.8090e-07

```

### 3.8.2 A nonlinear 1D BVP solved by successive substitution

To find the shortest connection of the form  $y = u(x)$  between two points at  $x = a$  and  $x = b$  the length functional

$$L(u) = \int_a^b \sqrt{1 + (u'(x))^2} dx$$

has to be minimized. The corresponding Euler–Lagrange equation is

$$\frac{d}{dx} \left( \frac{1}{\sqrt{1 + (u'(x))^2}} \frac{du(x)}{dx} \right) = 0. \quad (26)$$

As example search the connection between  $(x, y) = (0, 1)$  and  $(1, 2)$ , i.e. with the two boundary conditions

$$u(0) = 1 \quad \text{and} \quad u(1) = 2.$$

The straight line is easily determined by `BVP1DNL()`.

#### ShortestConnection.m

```
interval = linspace(0,1,21)';
a = @(x,u,du) 1./sqrt(1+du.^2);
u0 = @(x) 2-cos(2*pi*x);
[x,u,inform] = BVP1DNL(interval,a,0,0,1,0,1,2,u0,'Tol',1e-4);
figure(1); plot(x,u); xlabel('x'); ylabel('u')
inform
-->
inform = scalar structure containing the fields:
    info = 1
    iter = 8
    AbsError = 1.1468e-04
```

Since the coefficient

$$a(x, u, u') = \frac{1}{\sqrt{1 + (u')^2}}$$

depends on  $u'$  a successive substitution is used and it takes many iteration to converge.

As second problem search for the shortest connection with  $u(0) = 1$  and  $a(u'(1)) u'(1) = \alpha$ , i.e.

$$\frac{1}{\sqrt{1 + (u'(1))^2}} u'(1) = \alpha.$$

For  $\alpha = 0.5$  the code below generates the result with 7 iterations. At the end of the code the values of  $u(1)$ ,  $u'(1)$  and the boundary condition are computed.

#### ShortestConnection.m

```
interval = linspace(0,1,21)';
a = @(x,u,du) 1./sqrt(1+du.^2);
alpha = 0.5;
u0 = @(x) 1+x-0.3*(1-cos(2*pi*x));
[x,u,inform] = BVP1DNL(interval,a,0,0,1,0,1,[alpha,0],u0,
    'tol',1e-6,'MaxIter',50,'tol',1e-4,'Display','iter');
figure(1); plot(x,u0(x),x,u); xlabel('x'); ylabel('u')
    legend('u_0','u','location','northwest')
inform
du = FEM1DEvaluateDu(x,u);
```

```

u_end_du_end = [u(end) du(end) du(end) *a(0,0,du(end))]
-->
iteration 1, RMS of correction = 2.424606e-01
iteration 2, RMS of correction = 3.402180e-02
iteration 3, RMS of correction = 1.014111e-02
iteration 4, RMS of correction = 2.689617e-03
iteration 5, RMS of correction = 6.834600e-04
iteration 6, RMS of correction = 1.715824e-04
iteration 7, RMS of correction = 4.294086e-05

inform = scalar structure containing the fields:
    info = 1
    iter = 7
    AbsError = 4.2941e-05

u_end_du_end = 1.5774    0.5774    0.5000

```

It is easy to see<sup>6</sup> that for  $\alpha > 1$  there is no solution. The above code for  $\alpha = 0.9$  will converge only after 28 iterations. For  $\alpha > 1$  it will not converge!

### 3.9 Plane elasticity

In this section a typical plane stress situation is examined and the related commands illustrated. This is followed by a similar plane strain situation.

#### 3.9.1 A plane stress example

On a trapezoidal domain visible in Figure 19(a) a plane stress problem is set up.

- The material parameters  $E$  and  $\nu$  describe copper.
- At the lower edge at  $y = 0$  the displacements are zero, i.e.  $u_1(x, 0) = u_2(x, 0) = 0$  for  $-0.05 \leq x \leq +0.05$ .
- The other edges are force free.
- On all of the domain a force density of  $\vec{f} = (0, \frac{100}{0.3 \cdot 0.1}) \approx (0, 3333)$  is given. Thus the solid is pushed in  $y$  direction.
- An initial mesh is generated with the help of `triangle` and then upgraded to a mesh with second order elements.

With a call of `PlaneStress()` the displacements  $\vec{u}_1$  and  $\vec{u}_2$  are computed and then displayed, leading to Figure 19.

#### PlaneStressExample.m

```

W = 0.1; H = 0.3; Load = 100; E = 110e9; nu = 0.35; %% copper
FEMmesh = CreateMeshTriangle('Example1', ...
    [-W/2 0, -11; +W/2 0 -22; W/4 H -22; -W/4 H -22], 0.0001);
figure(1); FEMtrimesh(FEMmesh)
    xlabel('x'); ylabel('y'); axis equal

```

<sup>6</sup>The Euler-Lagrange equation implies that  $u'(x) = \beta$  is a constant. Then the equation  $\frac{\beta}{\sqrt{1+\beta^2}} = a$  has to be solved, leading to  $\beta = \frac{\alpha}{\sqrt{1-\alpha^2}}$ . The monotonous increasing function  $f(z) = \frac{z}{\sqrt{1+z^2}}$  with the limits  $\lim_{z \rightarrow \pm\infty} f(z) = \pm 1$  shows that the BVP has a solution for  $-1 < \alpha < +1$ .

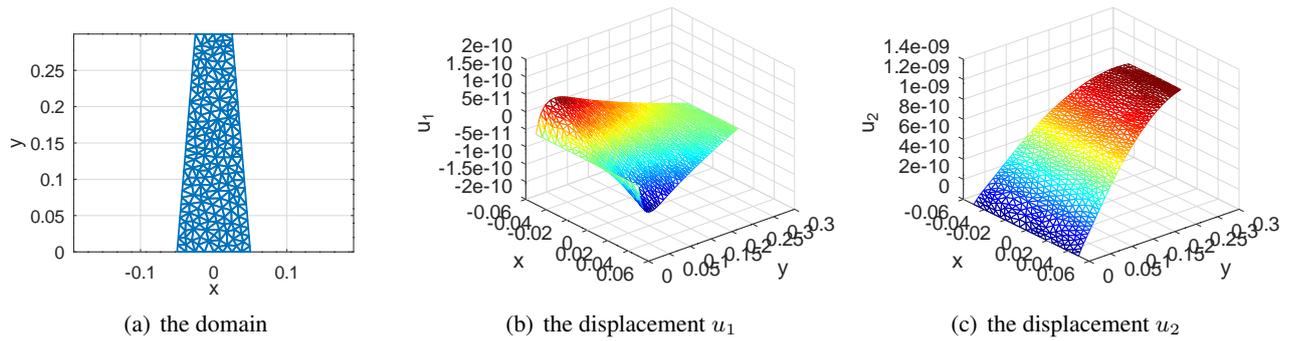


Figure 19: The computational domain and the two displacement functions  $u_1$  and  $u_2$

```
FEMmesh = MeshUpgrade(FEMmesh,'quadratic'); %% uncomment for second order elements
f = {0,Load/(H*W)}; gD = {0,0}; gN = {0,0};
[u1,u2] = PlaneStress(FEMmesh,E,nu,f,gD,gN);
figure(2); FEMtrimesh(FEMmesh,u1)
           xlabel('x'); ylabel('y'); zlabel('u_1'); view([50,30])
figure(3); FEMtrimesh(FEMmesh,u2)
           xlabel('x'); ylabel('y'); zlabel('u_2'); view([50,30])
```

With `EvaluateStrain()` the three strains  $\varepsilon_{xx}$ ,  $\varepsilon_{yy}$  and  $\varepsilon_{xy}$  are determined at the nodes and displayed, leading to Figure 20. The Saint–Venant’s principle at the lower edge  $y = 0$  is clearly visible.

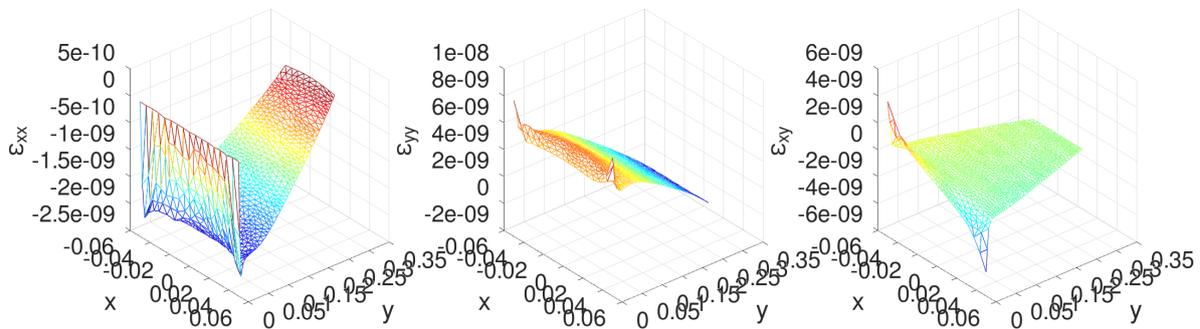
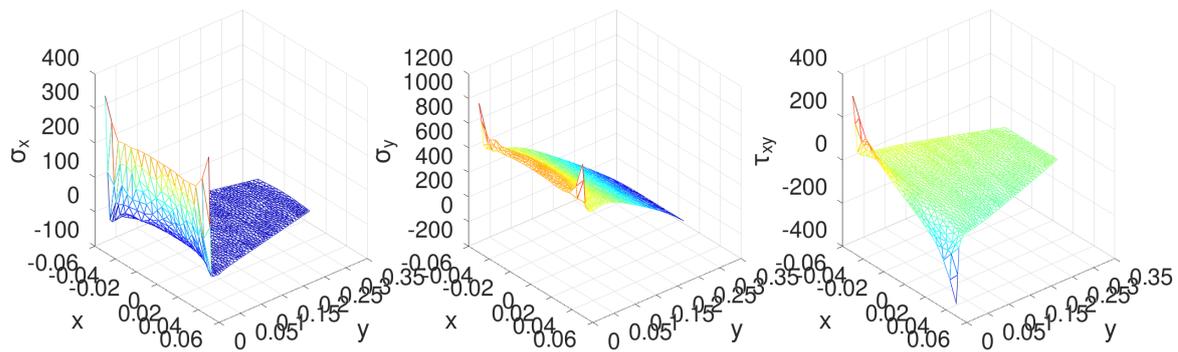


Figure 20: The normal strains  $\varepsilon_{xx}$ ,  $\varepsilon_{yy}$  and the shearing strain  $\varepsilon_{xy}$

#### PlaneStressExample.m

```
[eps_xx,eps_yy,eps_xy] = EvaluateStrain(FEMmesh,u1,u2);
figure(4);
subplot(1,3,1); FEMtrimesh(FEMmesh,eps_xx)
               xlabel('x'); ylabel('y'); zlabel('\epsilon_{xx}'); view([50,30])
subplot(1,3,2); FEMtrimesh(FEMmesh,eps_yy)
               xlabel('x'); ylabel('y'); zlabel('\epsilon_{yy}'); view([50,30])
subplot(1,3,3); FEMtrimesh(FEMmesh,eps_xy)
               xlabel('x'); ylabel('y'); zlabel('\epsilon_{xy}'); view([50,30])
```

With `EvaluateStress()` the three stresses  $\sigma_x$ ,  $\sigma_y$  and  $\tau_{xy}$  are determined at the nodes and displayed, leading to Figure 21. The Saint–Venant’s principle at the lower edge  $y = 0$  is again clearly visible.

Figure 21: The normal stresses  $\sigma_x$  and  $\sigma_y$  and the shearing stress  $\tau_{xy}$ 

#### PlaneStressExample.m

```
[sigma_x,sigma_y,tau_xy] = EvaluateStress(FEMmesh,u1,u2,E,nu);
figure(5);
subplot(1,3,1); FEMtrimesh(FEMmesh,sigma_x)
                xlabel('x'); ylabel('y'); zlabel('\sigma_x'); view([50,30])
subplot(1,3,2); FEMtrimesh(FEMmesh,sigma_y)
                xlabel('x'); ylabel('y'); zlabel('\sigma_y'); view([50,30])
subplot(1,3,3); FEMtrimesh(FEMmesh,tau_xy)
                xlabel('x'); ylabel('y'); zlabel('\tau_{xy}'); view([50,30])
```

With the two commands `EvaluateVonMises()` and `EvaluateTresca()` the von Mises stress and the Tresca stress are computed and displayed, leading to Figure 22. At the end of the code the two principal stresses  $\sigma_1$  and  $\sigma_2$  are computed, but not displayed.

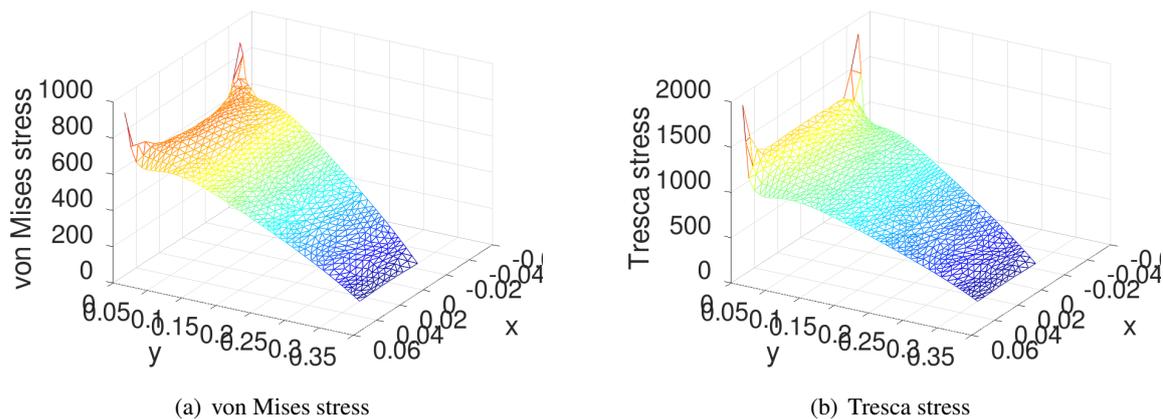


Figure 22: The von Mises and Tresca stress

#### PlaneStressExample.m

```
vonMises = EvaluateVonMises(sigma_x,sigma_y,tau_xy);
figure(6); FEMtrimesh(FEMmesh,vonMises)
                xlabel('x'); ylabel('y'); zlabel('von Mises stress'); view([120,30])
```

```
Tresca = EvaluateTresca(sigma_x, sigma_y, tau_xy);
figure(7); FEMtrimesh(FEMmesh, Tresca)
    xlabel('x'); ylabel('y'); zlabel('Tresca stress'); view([120, 30])
[s1, s2] = EvaluatePrincipalStress(sigma_x, sigma_y, tau_xy);
```

### 3.9.2 A plane strain example

On the trapezoidal domain visible in Figure 19(a) a plane strain problem is set up.

- The material parameters  $E$  and  $\nu$  describe copper.
- At the lower edge at  $y = 0$  the displacements are zero, i.e.  $u_1(x, 0) = u_2(x, 0) = 0$  for  $-0.05 \leq x \leq +0.05$ .
- At the upper edge at  $y = 0.3$  the horizontal displacements is set to  $+0.01$  and the vertical displacement is zero.
- The edges on the side are force free.
- There is no volume force applied to the domain, i.e.  $\vec{f} = \vec{0}$ .
- An initial mesh is generated by deforming a regular, rectangular mesh, and then upgraded to a mesh with second order elements.

With a call of `PlaneStrain()` the displacements  $\vec{u}_1$  and  $\vec{u}_2$  are computed and then displayed, leading to Figure 23. A coarser mesh on the same domain is generated and then used to display the original and deformed domain. Find the result in Figure 23(a) with the original domain in green and the deformed domain in red.

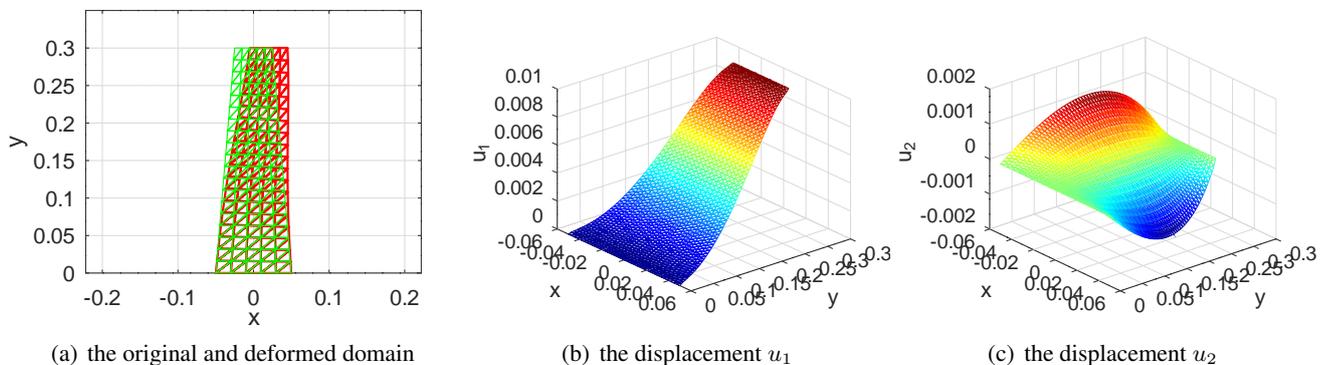


Figure 23: The computational domain and the two displacement functions  $u_1$  and  $u_2$

#### PlaneStrainExample.m

```
W = 0.1; H = 0.3; E = 110e9; nu = 0.35; %% copper

FEMmesh = CreateMeshRect(linspace(-W/2, W/2, 10), linspace(0, H, 30), -11, -11, -22, -22);
function xy_new = Deform(xy)
    xy_new = [xy(:, 1) .* (1 - 0.5/0.3 * xy(:, 2)) , xy(:, 2)];
endfunction
FEMmesh = MeshDeform(FEMmesh, 'Deform');
CMesh = CreateMeshRect(linspace(-W/2, W/2, 6), linspace(0, H, 20), -11, -11, -22, -22);
CMesh = MeshDeform(CMesh, 'Deform'); %% create a course mesh on the same domain
```

```

FEMmesh = MeshUpgrade(FEMmesh,'quadratic');

f = {0,0}; gN = {0,0};
function res = gD(xy)
    res = +(xy(:,2)>0.1)*0.01;
endfunction

[u1,u2] = PlaneStrain(FEMmesh,E,nu,f,{ 'gD',0},gN);

uli = FEMgriddata(FEMmesh,u1,CMesh.nodes(:,1),CMesh.nodes(:,2));
u2i = FEMgriddata(FEMmesh,u2,CMesh.nodes(:,1),CMesh.nodes(:,2));
factor = 2;
figure(11); trimesh(CMesh.elem,CMesh.nodes(:,1)+factor*uli,...
    CMesh.nodes(:,2)+factor*u2i,'color','red','linewidth',2);hold on ;
    trimesh(CMesh.elem,CMesh.nodes(:,1),...
    CMesh.nodes(:,2),'color','green','linewidth',1);
    hold off; axis equal; xlabel('x'); ylabel('y'); ylim([0,0.35])

figure(2); FEMtrimesh(FEMmesh,u1)
    xlabel('x'); ylabel('y'); zlabel('u_1'); view([50,30])
figure(3); FEMtrimesh(FEMmesh,u2)
    xlabel('x'); ylabel('y'); zlabel('u_2'); view([50,30])

```

With `EvaluateStrain()` the three strains  $\varepsilon_{xx}$ ,  $\varepsilon_{yy}$  and  $\varepsilon_{xy}$  are determined at the nodes and displayed, leading to Figure 24.

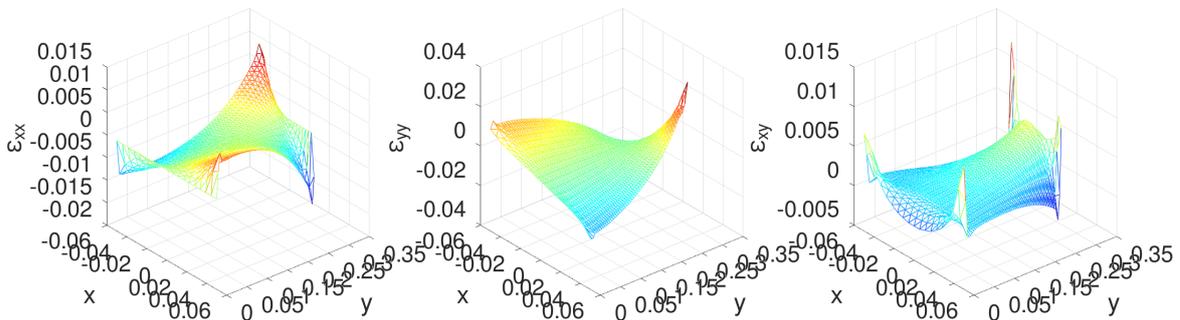


Figure 24: The normal strains  $\varepsilon_{xx}$ ,  $\varepsilon_{yy}$  and the shearing strain  $\varepsilon_{xy}$

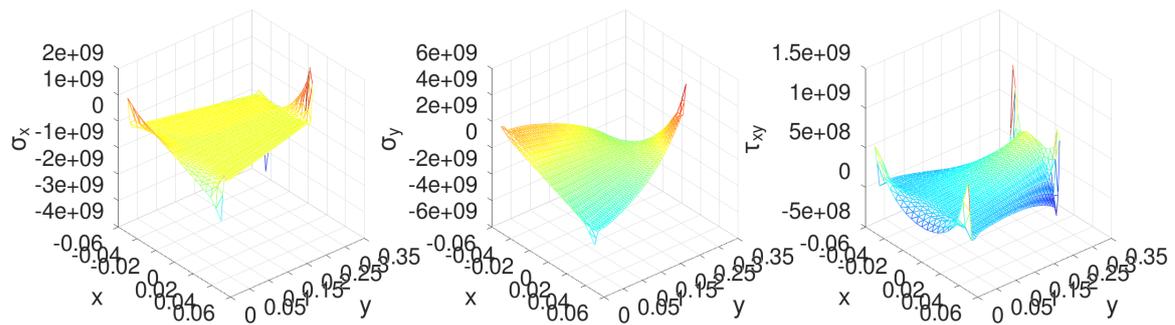
#### PlaneStrainExample.m

```

[eps_xx,eps_yy,eps_xy] = EvaluateStrain(FEMmesh,u1,u2);
figure(4);
subplot(1,3,1); FEMtrimesh(FEMmesh,eps_xx)
    xlabel('x'); ylabel('y'); zlabel('\epsilon_{xx}'); view([50,30])
subplot(1,3,2); FEMtrimesh(FEMmesh,eps_yy)
    xlabel('x'); ylabel('y'); zlabel('\epsilon_{yy}'); view([50,30])
subplot(1,3,3); FEMtrimesh(FEMmesh,eps_xy)
    xlabel('x'); ylabel('y'); zlabel('\epsilon_{xy}'); view([50,30])

```

With `EvaluateStress()` the three stresses  $\sigma_x$ ,  $\sigma_y$  and  $\tau_{xy}$  are determined at the nodes and displayed, leading to Figure 25. Observe that the function `EvaluateStress()` is called with for four return arguments, including  $\sigma_z$ . This assures that the plane strain expressions are used for the computations.

Figure 25: The normal stresses  $\sigma_x$  and  $\sigma_y$  and the shearing stress  $\tau_{xy}$ 

#### PlaneStrainExample.m

```
[sigma_x,sigma_y,tau_xy,sigma_z] = EvaluateStress(FEMmesh,u1,u2,E,nu);
figure(5); title('stress')
subplot(1,3,1); FEMtrimesh(FEMmesh,sigma_x)
    xlabel('x'); ylabel('y'); zlabel('\sigma_x'); view([50,30])
subplot(1,3,2); FEMtrimesh(FEMmesh,sigma_y)
    xlabel('x'); ylabel('y'); zlabel('\sigma_y'); view([50,30])
subplot(1,3,3); FEMtrimesh(FEMmesh,tau_xy)
    xlabel('x'); ylabel('y'); zlabel('\tau_{xy}'); view([50,30])
```

With the two commands `EvaluateVonMises()` and `EvaluateTresca()` the von Mises stress and the Tresca stress are computed and displayed, leading to Figure 22. Observe that four input arguments are given for the functions `EvaluateVonMises()` and `EvaluateTresca()`, including  $\sigma_z$ . This assures that the plane strain expressions are used for the computations. At the end of the code the two unknown principal stresses  $\sigma_1$  and  $\sigma_2$  are computed, but not displayed.

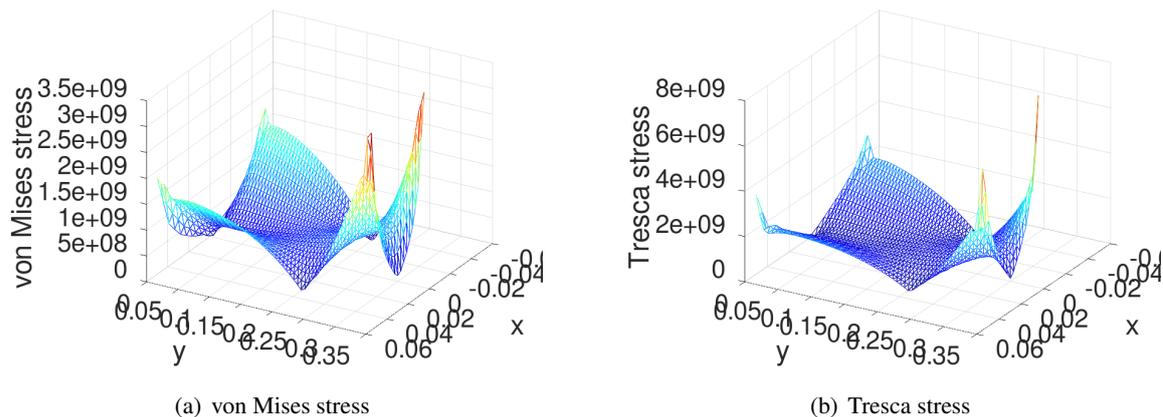


Figure 26: The von Mises and Tresca stress

#### PlaneStrainExample.m

```
vonMises = EvaluateVonMises(sigma_x,sigma_y,tau_xy,sigma_z);
figure(6); FEMtrimesh(FEMmesh,vonMises)
```

```

        xlabel('x'); ylabel('y'); zlabel("von Mises stress"); view([120,30])
Tresca = EvaluateTresca(sigma_x,sigma_y,tau_xy,sigma_z);
figure(7); FEMtrimesh(FEMmesh,Tresca)
        xlabel('x'); ylabel('y'); zlabel("Tresca stress"); view([120,30])

[s1,s2] = EvaluatePrincipalStress(sigma_x,sigma_y,tau_xy);

```

### 3.10 An axially symmetric elasticity example

A rectangular domain  $0 \leq r = x \leq R = 0.1$  and  $-2R \leq z \leq 2R$  is rotated about the  $z$ -axis and on the middle section  $-R \leq z \leq R$  of the surface an external pressure of the form

$$p(z) = \begin{cases} P(R^2 - z^2) & \text{for } |z| \leq R \\ 0 & \text{for } |z| > R \end{cases}$$

is applied. The aim is to determine the radial displacement  $u_r$  and the  $z$ -displacement  $u_z$ , as function of  $x = r$  and  $z$ .

- Due to the symmetry only the upper half of the cylinder has to be examined, with the boundary condition  $u_z = 0$  in the plane  $z = 0$ .
- Along the  $z$ -axis the boundary condition is  $u_r = 0$ .
- The upper edge is force free.
- Along the right edge at  $r = R$  the external pressure is applied.

As a first step create the mesh, define the pressure function and the material parameters. Then solve the problem using the function `AxiStress()`.

#### AxiSymmetricExample.m

```

R = 0.1;
if 0 %% nonuniform mesh
    Mesh = CreateMeshTriangle('AxiSymm',[0 0 -21; R 0 -32; R 2*R -22; 0 2*R -12],1e-4);
else
    Mesh = CreateMeshRect(linspace(0,R,10),linspace(0,2*R,20),-21,-22,-12,-32);
endif
Mesh = MeshUpgrade(Mesh,'quadratic');

function res = force(rz)
    R = 0.1; P = 1e5;    res = -P*max(R^2-rz(:,2).^2,0);
endfunction

E = 1e9; nu = 0.3; f = {0,0}; gD = {0,0}; gN = {'force',0};
[ur,uz] = AxiStress(Mesh,E,nu,f,gD,gN);

```

With this solution the original and deformed mesh can be displayed, leading to the left part of Figure 27.

#### AxiSymmetricExample.m

```

factor = 0.1*R/max(sqrt(ur.^2+uz.^2));
figure(1); trimesh(Mesh.elem,Mesh.nodes(:,1),Mesh.nodes(:,2),'color','green'); hold on
trimesh(Mesh.elem,Mesh.nodes(:,1)+factor*ur,Mesh.nodes(:,2)+factor*uz,...
        'color','red')
xlabel('r'); ylabel('z'); axis equal; hold off

```

With the displacements  $u_r$  and  $u_z$  evaluate stresses by using the functions `EvaluateStressAxi()` and `EvaluateVonMisesAxi()`.

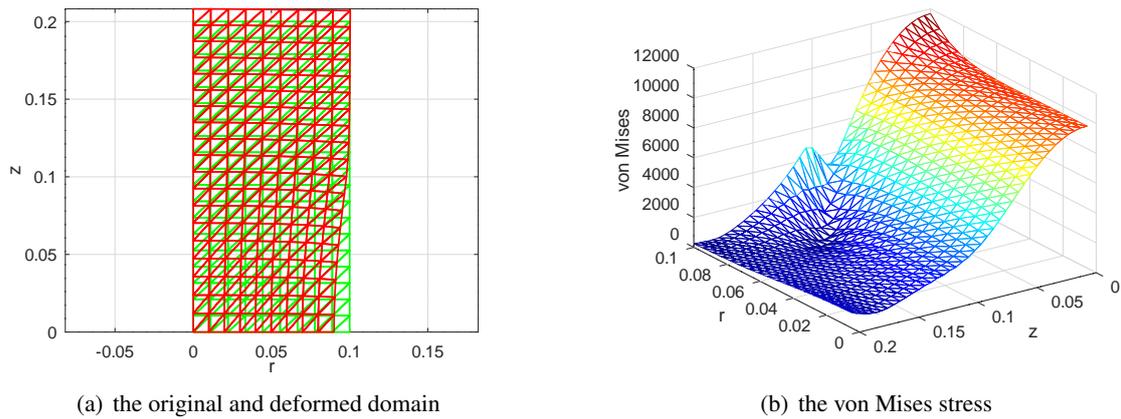


Figure 27: The original and deformed domain and the von Mises stress for an axially symmetric setup

#### AxisymmetricExample.m

```
[sigma_x,sigma_y,sigma_z,tau_xz] = EvaluateStressAxi(Mesh,ur,uz,E,nu);
figure(12); FEMtrimesh(Mesh,sigma_x)
    xlabel('r'); ylabel('z'); zlabel('\sigma_x')
figure(13); FEMtrimesh(Mesh,sigma_y)
    xlabel('r'); ylabel('z'); zlabel('\sigma_y')
figure(14); FEMtrimesh(Mesh,sigma_z)
    xlabel('r'); ylabel('z'); zlabel('\sigma_z')

vonMises = EvaluateVonMises(sigma_x,sigma_y,sigma_z,tau_xz);
figure(15); FEMtrimesh(Mesh,vonMises)
    xlabel('r'); ylabel('z'); zlabel('von Mises'); view(-125,30)
```

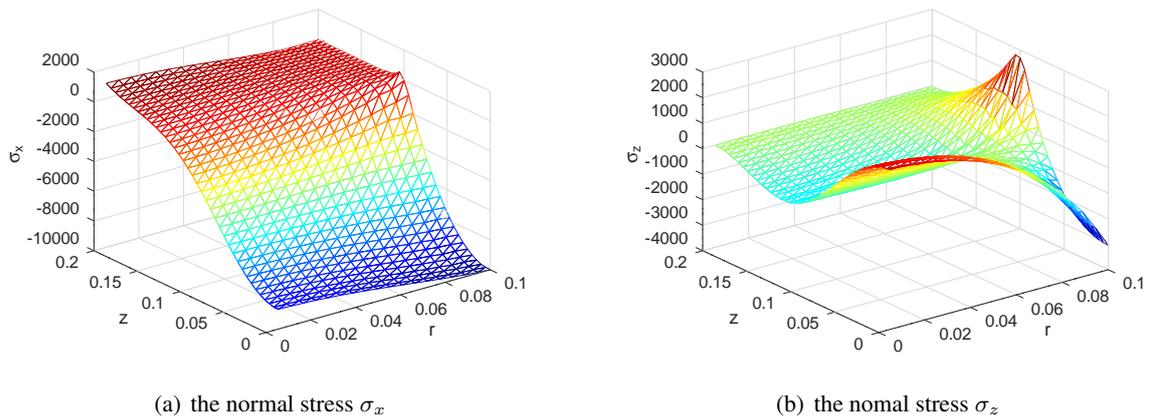


Figure 28: Stresses for an axially symmetric setup

## 4 The Commands of FEMoctave

In this section find the documentation for the commands provided by FEMoctave.

### 4.1 Commands for 2D meshes: creation and modification

#### 4.1.1 Structure of a mesh

The main information of a mesh, as shown in Section 6.1 is given by the position of the nodes (points), the corresponding triangles and the boundary edges. A mesh consists of

- $Nn$  nodes, with their  $(x, y)$  coordinates,
- $Ne$  elements, with 3 (or 6) nodes forming one triangle,
- $Nb$  boundary edges, with 2 (or 3) nodes forming one edge.

In FEMoctave this information is stored as a structure with an arbitrary name, but the elements of the structure require specific names, as shown in Table 4. The first 6 of these elements can be modified by the user and contain all the necessary information on the mesh to be used.

- `type`: a string indicating the order of the element, currently `linear`, `quadratic` or `cubic`.
- `nodes`: this  $Nn \times 2$  matrix contains the coordinates  $(x_i, y_i)$  of the nodes numbered by  $1 \leq i \leq Nn$ . The entries are real numbers.
- `nodesT`: this  $Nn$  vector of integers contains the information of the type of nodes. If the entry in row  $i$  equals 0 then node  $i$  is a DOF, i.e. the value of the solution is not prescribed. If the entry in row  $i$  equals 1 then node  $i$  is a Dirichlet node and the value of the solution is determined by the given function. For elasticity problems this is a  $Nn \times 2$  matrix with the information on both components.
- `elem`: for first order meshes this  $Ne \times 3$  matrix of integers contains in each row the numbers of three nodes forming one linear element (triangle). The triangles have a positive orientation. For second order elements it is a  $Ne \times 6$  matrix of integers. For third order elements it is a  $Ne \times 10$  matrix of integers.
- `elemT`: types of elements is not used yet.
- `edges`: this  $Nb \times 2$ ,  $Nb \times 3$  or  $Nb \times 4$  matrix of integers contains in each row the numbers of two, three or four nodes forming a boundary edge.
- `edgesT`: this  $Nb$  vector of integers contains the information of the type of edges. If the entry in row  $i$  equals  $-1$  then edge  $i$  is part of the Dirichlet boundary, i.e. the value of the solution is prescribed. If the entry in row  $i$  equals  $-2$  then edge  $i$  is part of the Neumann boundary, i.e. the value of the solution is not yet known. For elasticity problems this is a  $Nb \times 2$  matrix with the information on both components. See Table 5 for the codes.

All other elements of a mesh structure can be derived or computed from the above data.

- `elemArea`: this vector of real numbers contains the area of the individual triangles.
- `GP`: this matrix of reals contains the coordinates of all Gauss points for the numerical integration. There are 3 (or 7) Gauss points for each triangle.
- `GPT`: this vector of integer contains the type for each Gauss point. Currently not used.
- `nDOF`: this integer gives the total number of degrees of freedom (DOF) for the system to be solved.

Name	Size	Information
type	string	type of element, “linear”, “quadratic” or “cubic”
nodes	$Nn \times 2$	coordinates of nodes
nodesT	$Nn \times \{1, 2\}$	type of nodes, either 0 (free) or 1 (fixed)
elem	$Ne \times \{3, 6, 10\}$	list of nodes that make up the triangles
	$Ne \times 3$	for first order elements
	$Ne \times 6$	for second order elements
	$Ne \times 10$	for third order elements
elemT	$Ne \times 1$	type of elements
edges	$Nb \times \{2, 3, 4\}$	list of nodes that make up the boundary edges
edgesT	$Nb \times \{1, 2\}$	type of boundary edge, Dirichlet, Neumann or elasticity
elemArea	$Ne \times 1$	area of the triangles
GP		coordinates of the Gauss integration points
	$3 \cdot Ne \times 2$	for first order elements
	$7 \cdot Ne \times 2$	for second and third order elements
GPT	$(3 \text{ or } 7) \cdot Ne \times 1$	type of the Gauss integration points
nDOF	$1 \times \{1, 2\}$	total number of DOF of the system
node2DOF	$Nn \times \{1, 2\}$	renumbering from nodes to DOF

Table 4: Elements of a mesh structure

- `node2DOF`: This vector (or matrix for elasticity problems) gives for each node the number of the corresponding DOF. If the number equals 0 then it is a Dirichlet node.

The commands `CreateMeshRect()` and `CreateMeshTriangle()` create meshes with this structure.

The codes for the boundary conditions in Table 5 for elasticity problems might ask for a few examples of boundary conditions.

-11 : at this node the displacements are given by  $u_1(x, y) = gD1(x, y)$  and  $u_2(x, y) = gD2(x, y)$ .

-22 : at this node there are no surface forces, i.e. the node is on a free section of the boundary.

-12 : at this node the  $x$ -displacement  $u_1(x, y) = gD1(x, y)$  is given and there is no surface force in  $y$ -direction.

-31 : at this node the  $y$ -displacement  $u_2(x, y) = gD2(x, y)$  is given and surface force in  $x$ -direction is given by  $gN1(x, y)$ .

-23 : at this node there is no surface force in  $x$ -direction and the surface force in  $y$ -direction is given by  $gN2(x, y)$ .

#### 4.1.2 Create a uniform mesh on a rectangle: `CreateMeshRect()`

With the command `CreateMeshRect(x, y, Blow, Bup, Bleft, Bright)` you can create a mesh on a rectangle. The function takes 6 input arguments.

- The ordered vectors `x` and `y` contain the  $x$  and  $y$  coordinates of the mesh to be generated.

code	for scalar problems	
-1	Dirichlet condition , $u = g_1$ given	
-2	Neumann condition , $a \frac{\partial}{\partial n} u = g_2 + g_2 u$	
for elasticity problems		
code	in $x$ -direction	in $y$ -direction
-1*	displacement $u_1 = gD_1$ given	*
-*1	*	displacement $u_2 = gD_2$ given
-2*	force free section	*
-*2	*	force free section
-3*	force density $gN_1$ given	*
-*3	*	force density $gN_2$ given

Table 5: Codes for the boundary conditions

- For scalar problems the variables `Blow`, `Bup`, `Bleft` and `Bright` indicate the boundary condition on the corresponding edges. If the index is -1 then the edge is part of the Dirichlet boundary  $\Gamma_1$  and thus the value of the function is prescribed. If the index is -2 then the edge is part of the Neumann boundary  $\Gamma_2$  and thus information about the outer normal derivative is known, but not the value of the solution.
- For elasticity problems the variables `Blow`, `Bup`, `Bleft` and `Bright` indicate the boundary condition according to the codes in Table 5.

Examples of the usage are given in Sections 3.1.1 and 3.1.2.

#### CreateMeshRect()

```
Mesh = CreateMeshRect (X, Y, BLOW, BUP, BLEFT, BRIGHT)
```

Create a rectangular mesh with nodes at  $(x_i, y_j)$  with linear elements

parameters:

- \* `X, Y` are the vectors containing the coordinates of the nodes to be generated.
- \* `BLOW, BUP, BLEFT, BRIGHT` indicate the type of boundary condition at lower, upper, left and right edge of the rectangle
- \* for scalar problems
  - \* `B* = -1`: Dirichlet boundary condition
  - \* `B* = -2`: Neumann or Robin boundary condition
- \* for elasticity problems
  - \* `bi = -xy` : with two digits for  $x$  and  $y$  directions
  - \* `x/y = 1` : given displacement
  - \* `x/y = 2` : force free
  - \* `x/y = 3` : given force density

return values

- \* `MESH` is a structure with the information about the mesh. The mesh consists of `n_e` elements, `n_n` nodes and `n_ed` edges.
- \* `MESH.TYPE` a string with the type of triangle: linear
- \* `MESH.ELEM` `n_e` by 3 matrix with the numbers of the nodes forming triangular elements
- \* `MESH.ELEMAREA` `n_e` vector with the areas of the elements
- \* `MESH.ELEMT` `n_e` vector with the type of elements (not used)
- \* `MESH.NODES` `n_n` by 2 matrix with the coordinates of the nodes

```

* MESH.NODEST n_n vector with the type of nodes
* MESH.EDGES n_ed by 2 matrix with the numbers of the nodes forming edges
* MESH.EDGEST n_ed vector with the type of edge
* MESH.GP n_e*3 by 2 matrix with the coordinates of the Gauss points
* MESH.GPT n_e*3 vector of integers with the type of Gauss points
* MESH.NDOF number of DOF, degrees of freedom
* MESH.NODE2DOF n_n vector or n_n by 2 matrix of integers, mapping nodes to DOF

```

Sample call:

```

Mesh = CreateMeshRect(linspace(0,1,10),linspace(-1,2,20),-1,-1,-2,-2)
      will create a mesh with 200 nodes and 0<=x<=1, -1<=y<=+2

```

With `CreateMeshRect()` generate meshes with elements of order 1. With the help of `MeshUpgrade()` (Section 4.1.5) you can upgrade to the same mesh with elements of order 2 or 3.

#### 4.1.3 Using triangle: `CreateMeshTriangle()` and `ReadMeshTriangle()`

With the command `CreateMeshTriangle(name, xy, area)` you can create a mesh with the outer borders given in `xy`. The mesh will satisfy a minimal angle condition of  $30^\circ$  to avoid distorted triangles. The function takes 3 or 4 input arguments.

- The string 'name' is the file name to be used to store the information.
- The matrix `xy` contains the edge points of the domain and the information on the boundary conditions.
- `area` is the typical area of the triangles to be used.
- The optional argument `options` can specify more flags to the external call of the program `triangle`.

The mesh can then be read by calling `Mesh = ReadMeshTriangle('name.1')`. Examples of the usage are given in Sections 3.1.3 and 3.2 and in many of the examples in Section 9 starting on page 197.

#### CreateMeshTriangle()

```
MESH = CreateMeshTriangle(NAME,XY,AREA,OPTIONS)
```

Generate files with a mesh with linear elements using the external code `triangle`

parameters:

- \* NAME the base filename: the file `NAME.poly` will be generated then `triangle` will generate files `NAME.1.*` with the mesh
- \* XY vector containing the coordinates of the nodes forming the outer boundary. The last given node will be connected to the first given node to create a closed curve. Currently no holes can be generated.

The format for XY is `[x1,y1,b1;x2,y2,b2;...;xn,yn,bn]` where

- \* `xi` x-coordinate of node `i`
- \* `yi` y-coordinate of node `i`
- \* `bi` boundary marker for segment from node `i` to node `i+1`
  - \* for scalar problems
    - \* `bi = -1` Dirichlet boundary condition
    - \* `bi = -2` Neumann or Robin boundary condition
  - \* for elasticity problems
    - \* `B*` = `-xy` : with two digits for `x` and `y` directions
    - \* `x/y = 1` : given displacement
    - \* `x/y = 2` : force free

- \*  $x/y = 3$  : given force density
- \* AREA the typical area of the individual triangles to be used
- \* OPTIONS additional options to be used when calling triangle the options "pa" and the area will be added automatically. Default options are "q" resp. "qpa" to suppress the verbose information use "Q"

The information on the mesh generated is written to files and returned in the structure MESH, if the return argument is provided.

- \* The information can then be read and used by `Mesh = ReadMeshTriangle('NAME.1');`
- \* MESH is a structure with the information about the mesh.

The mesh consists of `n_e` elements, `n_n` nodes and `n_ed` edges.

- \* MESH.TYPE a string with the type of triangle: linear, quadratic or cubic
- \* MESH.ELEM `n_e` by 3 (or 6/10) matrix with the numbers of the nodes forming triangular elements
- \* MESH.ELEMAREA `n_e` vector with the areas of the elements
- \* MESH.ELEMT `n_e` vector with the type of elements (not used)
- \* MESH.NODES `n_n` by 2 matrix with the coordinates of the nodes
- \* MESH.NODEST `n_n` vector with the type of nodes (not used)
- \* MESH.EDGES `n_ed` by 2 (or 3/4) matrix with the numbers of the nodes forming edges
- \* MESH.EDGESET `n_ed` vector with the type of edge
- \* MESH.GP `n_e*(3/7)` by 2 matrix with the coordinates of the Gauss points
- \* MESH.GPT `n_e*(3/7)` vector of integers with the type of Gauss points
- \* MESH.NDOF number of DOF, degrees of freedom
- \* MESH.NODE2DOF `n_n` vector or `n_n` by 2 matrix of integers, mapping nodes to DOF

- With `CreateMeshTriangle()` generate meshes with elements of order 1. With the help of the command `MeshUpgrade()` (Section 4.1.5) you can upgrade to the same mesh with elements of order 2 or 3.
- If a return argument for `CreateMeshTriangle()` is provided, the mesh is returned.
- If no return argument is provided, the information is written to files. The generated mesh is then read by calling the function `ReadMeshTriangle()`.

This function can also be used to read meshes generated by direct call of the external program `triangle`. This allows to use all features of `triangle` and not only the very restricted setup used by `CreateMeshTriangle()`. In Section 4.1.4 find the options to generate meshes with holes and adapted mesh sizes. To find more about the features of `triangle` use the web page [www.cs.cmu.edu/~quake/triangle.html](http://www.cs.cmu.edu/~quake/triangle.html) or compile and install the code and then run `triangle -h` to examine the built-in help.

#### ReadMeshTriangle()

```
FEMMESH = ReadMeshTriangle(NAME.1)
read a mesh generated by CreateMeshTriangle(NAME)
parameter: NAME.1 the filename
return value: FEMMESH the mesh stored in NAME
```

Sample call:

```
CreateMeshTriangle('Test', [0,-1,-1;1,-1,-2;1,2,-1;0,2,-2],0.01)
Mesh = ReadMeshTriangle('Test.1');
will create a mesh with  $0 \leq x \leq 1$ ,  $-1 \leq y \leq +2$ 
and a typical area of 0.01 for each triangle
```

Find an example in Section 9.8.

With `CreateMeshTriangle()` and `ReadMeshTriangle()` one can only generate meshes with elements of order 1. With the help of `MeshUpgrade()` (Section 4.1.5) you can upgrade to the same mesh with elements of order 2 or 3.

#### 4.1.4 Adapting meshes and creating holes by using options of `CreateMeshTriangle()`

Give `CreateMeshTriangle()` more arguments to use some of the features of `triangle` to locally create finer meshes or generate domains with holes. If more than 4 arguments are provided, then line segments, point with mesh sizes or holes can be generated. Each of the additional arguments is a structure with a name as first entry. There are four types of options:

- **Segment:** to create additional line segments, used to modify mesh sizes. Besides the mandatory entry `name='Segment'` one additional entry `border` has to be provided. It lists the  $x$  and  $y$  coordinates of the points forming the segment, and the third entry 0 for each point.

```
Seg1.name = 'Segment';           %% mandatory name
Seg1.border = [0 0 0; 1 0 0; 1 2 0] %% the points on the segment
```

- **MeshSize:** to specify the mesh size in one part of the domain. Besides the entry `name='MeshSize'` two additional entries have to be provided. `where` with the  $x$  and  $y$  coordinates of the points where the maximal mesh size is given and `area` with the desired mesh size, i.e. the maximal area of the triangles.

```
Point1.name = 'MeshSize';      %% mandatory name
Point1.where = [1.5 0.2];      %% the point at which the mesh size is applied
Point1.area = 0.01;           %% maximal area in the selected area
```

- **Hole:** to create a hole in the domain. Besides the mandatory entry `name='Hole'` two additional entries `border` and `point` have to be provided. `border` lists the  $x$  and  $y$  coordinates of the points forming the hole, with the third entry indicating the type of boundary condition, according to Table 5. The entry `point` has to contain the coordinates of one point inside the hole.

```
Hole1.name = 'Hole'           %% mandatory name
Hole1.border = [1 1 -22; 3 2 -22; 3 4 -22; 1 2 -22] %% border of the hole
Hole1.point = [1.1 1.1];      %% one point in the hole
```

- **Option:** to give more options, not documented yet.

There are a few points to watch out for when using the above optional arguments to `CreateMeshTriangle()`:

- The lines created by `Segment` shall not interfere with the holes.
- If `Segment` is used to divide the domain into multiple sections, the endpoints of the segments have to be exactly on the borders of the domain, but you can (often) not use the points defining the borders of the domain. A possible way out is described in Example 4–2 below.

**4–1 Example :** As a first example for the additional options for the command `CreateMeshTriangle()` examine a domain with a hole in the middle section and a finer mesh at the lower edge. Find the result of the code below in Figure 29. In most parts of the domain the area of the triangles is approximately 0.001. Close to the hole the narrow section leads to smaller triangle and close to the lower edge the optional `Segment` leads to a finer mesh, visible in Figure 29(b).

```

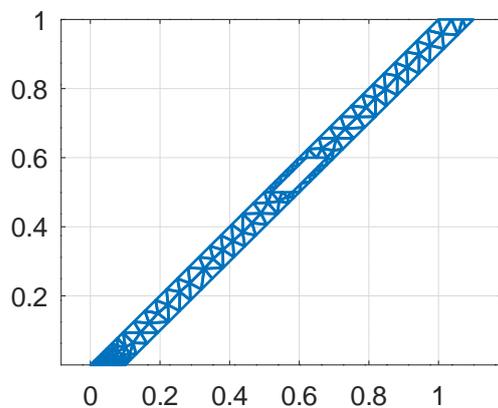
MeshBorder = [0 0 -11; 0.1 0 -22; 1.1 1 -23; 1 1 -22]; %% outer boundary of the domain

Hole.name   = 'Hole';    %% a hole in the middle section
Hole.border = [0.5+0.02 0.5 -22; 0.5+0.08 0.5 -22; 0.6+0.08 0.6 -22; 0.6+0.02 0.6 -22];
Hole.point  = [0.522 0.501];

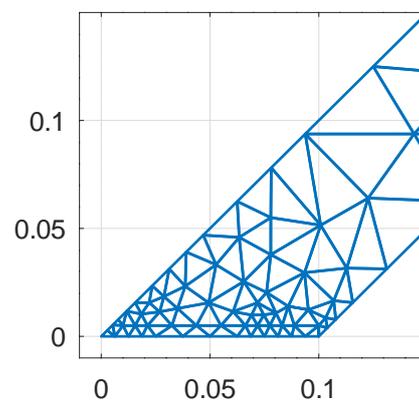
Segment.name = 'Segment'; %% close to the lower edge
Segment.border = [0.01 0.01 0; 0.09 0.01 0];

Mesh = CreateMeshTriangle('Mesh1',MeshBorder,0.01/9, Hole, Segment);
figure(1); FEMtrimesh(Mesh); axis equal

```



(a) the domain with a hole



(b) the finer mesh at the lower edge

Figure 29: A domain with a hole and a finer mesh at the lower edge

The above domain can be used for an elasticity computation. The lower edge is fixed and at the upper edge a vertical force is applied. At the lower edge Saint–Venant’s principle applies, i.e. shearing is expected and thus a finer mesh should be used. Find the graphical result in Figure 30. In addition the transversal deflection  $(u_2 - u_1)/\sqrt{2}$  along the center line is displayed. The result shows that the lever is behaving like a bending beam.

```

Mesh = MeshUpgrade(Mesh,'quadratic'); E = 100e9; nu = 0.3; f = 1;
[u1,u2] = PlaneStress(Mesh,E,nu,{0,0},{0,0},{0,f});

figure(2);clf; scale = 0.1/max(u2);
trimesh(Mesh.elem,Mesh.nodes(:,1)+scale*u1,Mesh.nodes(:,2)+scale*u2,...
        'color','red','linewidth',1)

hold on
trimesh(Mesh.elem,Mesh.nodes(:,1),Mesh.nodes(:,2),'color','green','linewidth',1)
hold off; axis([0 1.2 0 1.2])

yi = linspace(0,1); xi = yi+0.05;
uli = FEMgriddata(Mesh,u1,xi,yi); u2i = FEMgriddata(Mesh,u2,xi,yi);
bend = (u2i-uli)/sqrt(2);
figure(3); plot(yi,bend); xlabel('x'); ylabel('(u_2-u_1)/sqrt(2)')

```



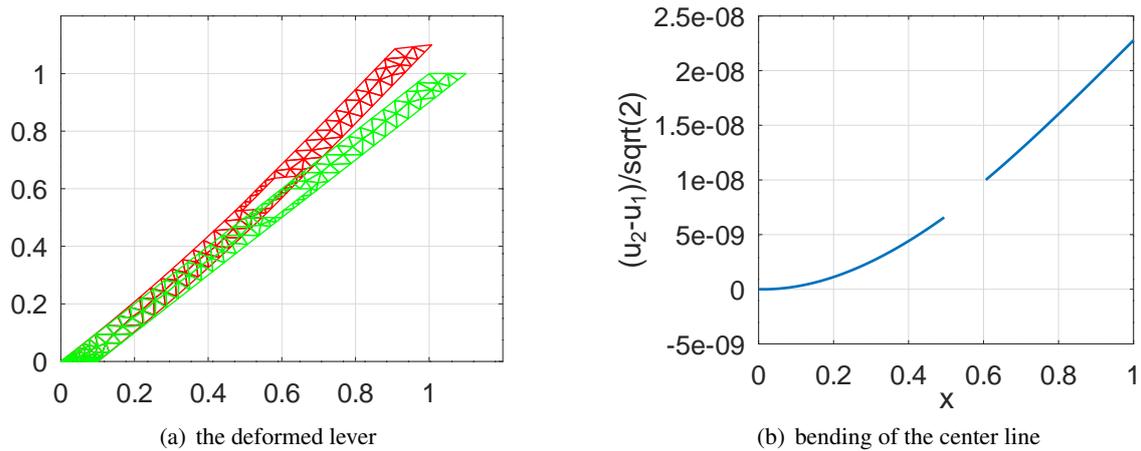


Figure 30: The deformed lever and the bending of the center line

**4-2 Example :** In this example a mesh with different size triangles is created on a unit square, see Figure 31. The

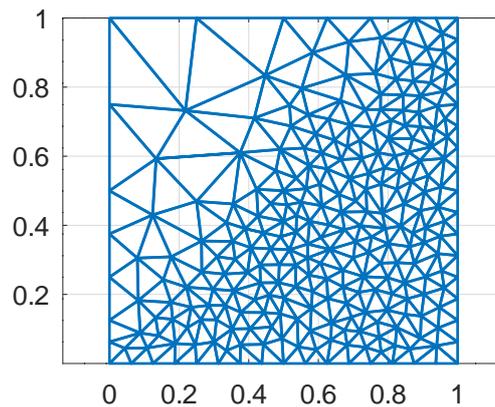


Figure 31: A domain with a two different mesh sizes

first section of the code below generates the file `Mesh2.poly`, in which the two corners  $(0,0)$  and  $(1,1)$  are listed twice. This causes FEMoctave to issue a warning `matrix singular to machine precision` when solving the resulting linear system. To avoid this problem edit the file `Mesh2.poly`, remove the two “extra” points and modify the connection of points 1 and 3.

#### Mesh2.poly

```
# nodes
6 2 0 1
1 0.000000000000 0.000000000000 -1
2 1.000000000000e+00 0.000000000000e+00 -1
3 1.000000000000e+00 1.000000000000e+00 -2
4 0.000000000000e+00 1.000000000000e+00 -1
5 0.000000000000e+00 0.000000000000e+00 0
6 1.000000000000e+00 1.000000000000e+00 0
# segments
5 1
```

```

1 1 2 -1
2 2 3 -2
3 3 4 -2
4 4 1 -1
5 5 6 0
# holes
0
# area markers
2
1 0.100000 0.900000 0 0.100000
2 0.900000 0.100000 0 0.002000
# generate mesh by : triangle -Qpq30a Mesh2.poly

```

Save the new file as `Mesh2_mod.poly` and run `triangle`, either by `triangle -pq30a Mesh2_mod.poly` from a command line or by `system('triangle -Qpq30a Mesh2_mod.poly')` within *Octave*.

#### Mesh2\_mod.poly

```

# nodes
4 2 0 1
1 0.0000000000000000 0.0000000000000000 -1
2 1.0000000000000000e+00 0.0000000000000000e+00 -1
3 1.0000000000000000e+00 1.0000000000000000e+00 -2
4 0.0000000000000000e+00 1.0000000000000000e+00 -1
# segments
5 1
1 1 2 -1
2 2 3 -2
3 3 4 -2
4 4 1 -1
5 1 3 0
# holes
0
# area markers
2
1 0.100000 0.900000 0 0.100000
2 0.900000 0.100000 0 0.002000
# generate mesh by : triangle -Qpq30a Mesh2_mod.poly

```

Then read the new mesh by `Mesh = ReadMeshTriangle('Mesh2_mod.1')`. The resulting mesh avoids the *Octave* warning. ◇

### 4.1.5 Converting meshes: upgrading and downgrading

Given a mesh `MeshLin` with first order elements one can generate the same mesh with elements of order 2 by calling the command `MeshUpgrade(MeshLin, 'quadratic')`. The numbering of the nodes of the linear elements is preserved in the mesh with the quadratic elements. The new nodes are placed at the midpoints of the edges of the triangles. With `MeshUpgrade(MeshLin, 'cubic')` a mesh with 10 node cubic elements is generated. Examine Figure 64 on page 129 on how the nodes are placed within the triangles.

#### MeshUpgrade()

```

MESHNEW = MeshUpgrade(MESHLIN,TYPE)
convert a mesh MESHLIN of order 1 to a mesh MESHNEW of order 2 or 3
parameters:
* MESHLIN the input mesh of order 1
* TYPE is a string, either 'quadratic' or 'cubic'

```

```

the default is 'quadratic'
return value: MESHNEW the output mesh of order 2 or 3

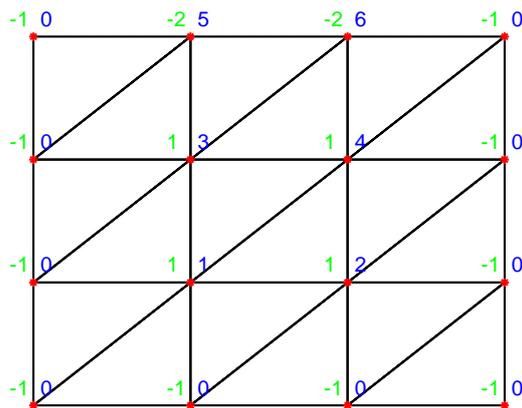
```

As example generate a mesh with elements of order 1 on the rectangle  $0 \leq x, y \leq 2$  with Dirichlet conditions on three edges and a Neumann condition on the upper edge at  $y = 2$ . In Figure 32 find the mesh with the types of nodes indicated and the numbering of the resulting degrees of freedom.

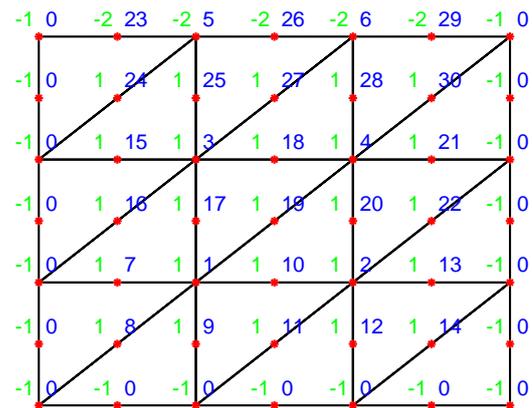
```

N = 3;
FEMmesh1 = CreateMeshRect(linspace(0,2,N+1),linspace(0,2,N+1),-1,-2,-1,-1);
FEMmeshQ = MeshUpgrade(FEMmesh1,'quadratic');

```



(a) linear elements



(b) quadratic elements

Figure 32: The same mesh with linear or quadratic elements. The types of the nodes are marked in green. Dirichlet nodes are marked by  $-1$ , Neumann nodes by  $-2$  and interior nodes by  $+1$ . The numbering of the resulting degrees of freedom is shown in blue. For Dirichlet nodes a DOF of 0 is used.

Using `MeshQuad2Linear()` one can convert a mesh of order 2 to a mesh of order 1. The nodes will remain unchanged, but there will be a factor of 4 more elements. With this function one can compare results based on first or second order elements, using exactly the same degrees of freedom.

```

MeshQuad2Linear()
MESH LIN = MeshQuad2Linear(MESH QUAD)
convert a mesh MESH QUAD of order 2 to a mesh MESH LIN of order 1
parameter: MESH QUAD the input mesh of order 2
return value: MESH LIN the output mesh of order 1

```

An example is shown in Section 3.1.4.

Using `MeshCubic2Linear()` one can convert a mesh of order 3 to a mesh of order 1. The nodes will remain unchanged, but there will be a factor of 9 more elements. With this function one can compare results based on first or third order elements, using exactly the same degrees of freedom.

```

MeshCubic2Linear()
MESH LIN = MeshCubic2Linear(MESH CUBIC)
convert a mesh MESH CUBIC of order 3 to a mesh MESH LIN of order 1
parameter: MESH CUBIC the input mesh of order 3
return value: MESH LIN the output mesh of order 1

```

#### 4.1.6 Use `delaunay()` to create a mesh: `Delaunay2Mesh()`

It is possible to use the *Octave* command `delaunay()` to generate a triangulation of a convex domain and then `Delaunay2Mesh()` to generate a mesh to be used by FEMoctave.

- The generated mesh consists of elements of order one. Use `MeshUpgrade()` to work with elements of order two or three.
- At first all boundary points are marked as Dirichlet points. Change the type description in the mesh if you want Neumann points.

#### Delaunay2Mesh()

```
FEMMESH = Delaunay2Mesh(TRI,X,Y)
generate a mesh with elements of order 1, using a Delaunay triangulation
parameters:
* TRI the Delaunay triangulation
* X,Y the coordinates of the points
return value
* FEMMESH is the mesh to be used by FEMoctave
```

Observe that the quality of the mesh might be very poor, e.g. triangles with very small angles. As example have a look at the upper edge on the right of the mesh in Figure 33. For almost all cases `triangle` will generate meshes of better quality. To generate the domain and the solution in Figure 33 use the code below.

#### TestDelaunay.m

```
[x,y] = meshgrid(linspace(-1,1,20)); x = x(:); y = y(:);
ind = find(y<1-0.5*x+0.001); x = x(ind); y = y(ind);
ind = find(x+y>-0.001); x = x(ind); y = y(ind);

tri = delaunay(x,y);
figure(1); triplot(tri,x,y);
hold on; plot(x,y,'*'); hold off
xlabel('x'); ylabel('y');
FEMmesh = Delaunay2Mesh(tri,x,y); FEMmesh = MeshUpgrade(FEMmesh,'quadratic');

u = BVP2Dsym(FEMmesh,1,0,4,0,0,0);
figure(2); FEMtrimesh(FEMmesh,u)
label('x'); ylabel('y'); view([100,45])
figure(3); FEMtricontour(FEMmesh,u);
xlabel('x'); ylabel('y');
```

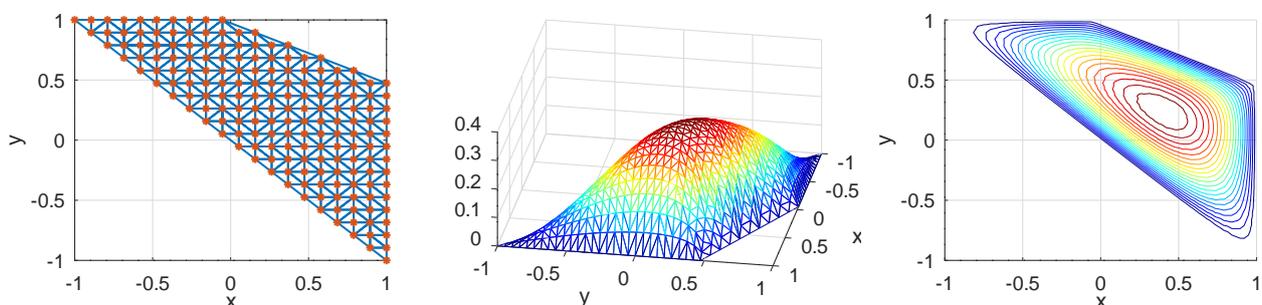


Figure 33: A mesh generated by a Delaunay triangulation and the solution of a BVP

### 4.1.7 Deforming meshes by MeshDeform()

With the function MeshDeform() the nodes of a linear mesh can be deformed.

#### MeshDeform()

```
MeshDeformed = MeshDeform(MESH,DEFORM)
    Deform the nodes of MESH by the transformation DEFORM
    parameters:
    * MESH the initial mesh with linear elements
      this has to be a mesh with linear elements
    * DEFORM the transformation formula
      the function DEFORM takes one argument XY, a n by 2 matrix with the
      x and y components in columns and returns the result in a n by 2 matrix.
    return value
    * DEFORMEDMESH the deformed mesh consists of linear elements
      use MESHUPGRADE to generate quadratic or cubic elements
```

One should pay attention to not deform the triangles in the mesh too badly by MeshDeform(), as this might decrease the accuracy of the solutions. The mesh generated by MeshCreateTriangle() will respect the condition of minimal 30° angles. After calling MeshDeform() this condition could be violated. Another option is to deform the borders of the mesh first, and then call CreateMeshTriangle(). In this case the minimal angle condition is respected.

To generate the quarter of a ring in Figure 48 on page 106 use polar coordinates

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} r \cdot \cos \varphi \\ r \cdot \sin \varphi \end{pmatrix} \quad \text{with } 1 \leq r \leq 2 \quad \text{and} \quad 0 \leq \varphi \leq \frac{\pi}{2}.$$

```
FEMmesh = CreateMeshTriangle('Test',[1,0,-1;2,0,-1;2,pi/2,-2;1,pi/2,-1],0.1^2);

function xy_new = Deform(xy) %% use polar coordinates
    xy_new = [xy(:,1).*cos(xy(:,2)), xy(:,1).*sin(xy(:,2))];
endfunction

FEMmesh = MeshDeform(FEMmesh,'Deform');
FEMtrimesh(FEMmesh)
```

Find an example in Section 9.1.

## 4.2 Evaluation and displaying results

### 4.2.1 Display results on meshes, FEMtrimesh(), FEMtrisurf() and FEMtricontour()

To display the results of the computations very elementary wrappers around trimesh(), trisurf() and tricontour() are provided.<sup>7</sup>

- With FEMtrimesh() display a function  $u$  as a 3D mesh. If no values for  $u$  are provided, the 2D mesh is displayed.
- With FEMtrisurf() display a function  $u$  as a 3D surface. The syntax is identical to FEMtrimesh().
- With FEMtricontour() display level curves of a function  $u$ . The syntax similar to the above.

<sup>7</sup>It is obviously possible to improve the wrappers, as non of the advanced features of trimesh() or trisurf() is passed through. If you want to use those, have a look at the elementary code in the FEMtri\* functions and copy the necessary lines in to your code.

All functions accept meshes with linear, quadratic or cubic elements.

- For quadratic elements the 6 nodes in each element are connected by straight lines, i.e. as if one second order triangle would be composed of 4 first order triangles.
- For cubic elements the 10 nodes in each element are connected by straight lines, i.e. as if one third order triangle would be composed of 9 first order triangles.

#### FEMtrimesh()

```
FEMtrimesh (MESH, U)
display a solution U on a triangular mesh
parameters:
  * MESH is the mesh
  * U values of the function to be displayed
  if U is not given, then the mesh is displayed in 2D
```

#### FEMtrisurf()

```
FEMtrisurf (MESH, U)
display a solution U as surface on a triangular mesh
parameters:
  * MESH is the mesh
  * U values of the function to be displayed
```

#### FEMtricontour()

```
FEMtricontour (MESH, U, V)
display contours of a solution U on a triangular mesh
parameters:
  * MESH is the mesh
  * U values of the function to be displayed
  * V contours to be used, default value is 21
  if V is scalar, it is the number of contours
  if V is a vector, it is the levels of the contours
```

### 4.2.2 Evaluate the gradient of a function at the nodes: FEMEvaluateGradient()

Given the values  $u$  of a function at the nodes, the two components of the gradient can be computed with the function `FEMEvaluateGradient()`.

#### FEMEvaluateGradient()

```
[UX,UY] = FEMEvaluateGradient (MESH,U)
evaluate the gradient of the function u at the nodes
parameters:
  * MESH is the mesh describing the domain and the boundary types
  * U vector with the values of the function at the node
return value
  * UX x component of the gradient of u
  * UY y component of the gradient of u
```

```
the values of the gradient are determined on each element
at the nodes the average of the gradient of the elements is used
```

The gradient is determined on each of the elements, using either linear, quadratic or cubic interpolation. Then at each node the average of the values of the gradient of the neighboring triangles is returned. This is different from the results generated by `FEMgriddata()`. Examples are given in Sections 5.1, 9.3, 9.4, 9.5 and 9.9. Due to using broadcasting in the *Octave* code (`bsxfun()`) the code is fast! This function could be used (or is that abused?) to evaluate derivatives of functions given on an irregular grid!

### 4.2.3 Evaluate a function and its gradient at the Gauss points: `FEMEvaluateGP()`

Given the values  $u$  of a function at the nodes, the values of  $u$  and its gradient can be computed at the Gauss points by calling `FEMEvaluateGP()`. For first order elements a piecewise linear interpolation is used, thus the gradients will be constant on each triangular element. For second order elements a quadratic interpolation is used. For third order elements a cubic interpolation is used.

#### `FEMEvaluateGP()`

```
[UGP, GRADUGP] = FEMEvaluateGP(MESH, U)
    evaluate the function and gradient at the Gauss points
    parameters:
        * MESH is the mesh describing the domain and the boundary types
        * U vector with the values of at the nodes
    return values
        * UGP values of u at the Gauss points
        * GRADUGP matrix with the values of the gradients in the columns
```

Examples are given in Sections 9.7 and 9.9.

### 4.2.4 Integrate a function over the domain: `FEMIntegrate()`

Given a function name, the values of a function at the nodes or at the Gauss points one can integrate this function over the domain given by the mesh. There are different methods used, all based on the Gauss integration presented in Section 6.3.2.

- If a function name is specified, then this function will be evaluated at the Gauss points and then integrated.
- If a scalar value is given, then the function is assumed to be constant.
- If a column vector is given with as many components as nodes in the mesh, then an element wise interpolation is used to obtain the values at the Gauss points. The function `FEMEvaluateGP()` is used to find the values at the Gauss points.
- If a column vector is given with as many components as Gauss points in the mesh, then these are used as values at the Gauss points.

#### `FEMIntegrate()`

```
NUMINTEGRAL = FEMIntegrate(MESH, U)
    integrate a function u over the domain given in Mesh
    parameters:
        * MESH is the mesh describing the domain
        * U the function to be integrated
            can be given as function name to be evaluated or as scalar
            value, or as a vector with the values at the nodes or the Gauss points.
    return value
        * NUMINTGERAL the numerical approximation of the integral
```

As a simple example integrate the function  $u(x, y) = xy^3$  over the unit square  $0 \leq x, y \leq 1$ . The exact integral equals  $\frac{1}{8}$ , but you have to subtract the exact value to see the difference to the numerical evaluation with the Gauss points. This is not unusual, since the Gauss integration leads to very accurate approximations, if the function is smooth. Linear elements use 3 integration points in each triangle, quadratic and cubic meshes use 7 integration points in each triangle. Thus integrations using a linear mesh might not be as accurate.

```

N = 40; Mesh = CreateMeshRect(linspace(0,1,N),linspace(0,1,N),-2,-2,-2,-2);
function res = f_int(xy) res = xy(:,1).*xy(:,2).^3; endfunction

integrall1 = FEMIntegrate(Mesh,'f_int') % using the function name
uGP = feval('f_int',Mesh.GP);
integral2 = FEMIntegrate(Mesh,uGP) % using the values at the Gauss points
-->
integrall1 = 0.12500
integral2 = 0.12500

```

To determine the area of a domain  $\Omega \subset \mathbb{R}^2$  one can integrate the constant 1 over the domain. More examples are given in Sections 5.1, 9.1, 9.7 and 9.9.

#### 4.2.5 Evaluation at arbitrary points or along curves, integration along curves: FEMgriddata()

Given a function by the values at the nodes of a mesh use the command FEMgriddata() to evaluate the function at arbitrary points.

- The value of the function and the partial derivatives can be evaluated.
- Depending on the mesh provided either a piecewise linear, quadratic or cubic interpolation is used.
- If a point  $(x_i, y_i)$  is on the edge of a triangle it is a matter of rounding which of the neighboring triangles is used for the interpolation. Since all elements used by FEMoctave are  $C^0$  conforming, this has no influence on the value of the function. The elements are not  $C^1$  conforming and thus the partial derivatives will jump across element boundaries. See also Section 5.3 starting on page 101.
- If a point  $(x_i, y_i)$  is not in a triangle, then NaN is returned.
- The evaluation is very fast, even for large numbers of elements and interpolation points.
- Evaluation along arbitrary curves is possible, and fast. Then use trapz() to integrate along curves. Find examples in Sections 9.5 and 9.15.

#### FEMgriddata()

```

[UI,UXI,UYI] = FEMgriddata(MESH,U,XI,YI)
evaluate the function (and gradient) at given points by interpolation
parameters:
* MESH is the mesh describing the domain
  If MESH consists of linear elements, piecewise linear interpolation is used.
  If MESH consists of quadratic elements, piecewise quadratic interpolation is used.
  If MESH consists of cubic elements, piecewise cubic interpolation is used.
* U vector with the values of the function at the nodes
* XI, YI coordinates of the points where the function is evaluated
return values:
* UI values of the interpolated function u
* UXI x component of the gradient of u
* UYI y component of the gradient of u

```

The values of the function and the gradient are determined on each element by a piecewise linear, quadratic or cubic interpolation.

If a point is not inside the mesh NaN is returned.

This function is similar to `FEMEvaluateGradient()`, but allows to evaluate at arbitrary points. At the nodes the value of the gradient in **one of the triangles** is returned. As a consequence the results generated by `FEMEvaluateGradient()` look smoother on occasion.

The code below evaluates a function on an L-shaped domain on a rectangular grid. Find the result in Figure 34.

```
nodes = [0,0,-2;1,0,-2;1,1,-2;-1,1,-2;-1,-1,-2;0,-1,-2];
Mesh = CreateMeshTriangle('Ldomain',nodes,0.002);
x = Mesh.nodes(:,1); y = Mesh.nodes(:,2);

function res = f_int2(xy) res = sin(pi*xy(:,1)).^2.*xy(:,2)+1; endfunction

u = feval('f_int2',Mesh.nodes);
N = 51; [xi,yi] = meshgrid(linspace(-1,1,N)); %% generate the uniform grid
tic(); ui3 = FEMgriddata(Mesh,u,xi,yi); toc()

figure(1); mesh(xi,yi,ui3)
        xlabel('x'); ylabel('y'); zlabel('u')
-->
Elapsed time is 0.0075829 seconds.
```

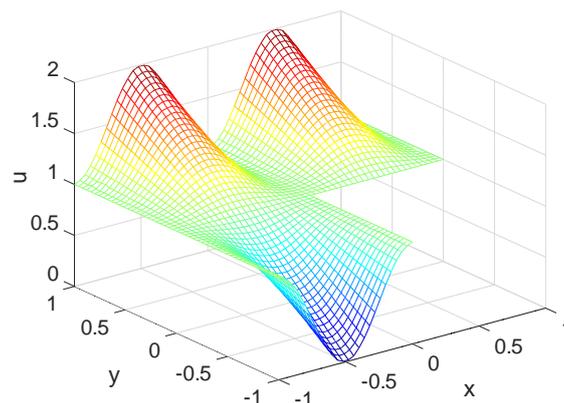


Figure 34: A function evaluated on a uniform grid

Examples are given in Sections 5.3, 9.3, 9.5, 9.8, 9.10 and 9.15.

### 4.3 How to define functions

There are three basic techniques to define functions in *Octave* to be used with *FEMoctave*.

- If the function is a constant you can simply use this scalar as input argument.
- You may provide the function name of the function to be called to compute the values of the function. Observe that the function **has to be vectorized**<sup>8</sup>. Due to a recent change in *Octave* the script versions

<sup>8</sup>Function on the boundary are actually called for one point at a time, but this might change. Thus it is advisable to write all functions vectorized.

should use a dummy second argument<sup>9</sup>. The function can be implemented as a name.m Octave function or as dynamically linked function name.oct, written in C++.

- You can provide a vector of the correct size with all the values of the function at the Gauss integration points of the mesh.

Section 9 contains many examples or you may examine the examples below.

#### 4.3.1 Functions for static problems

The functions `BVP2D()`, `BVP2Dsym()` and `BVP2Deig()` accept the coefficient functions as input parameters. These functions accept (currently) one parameter, a matrix with two columns. The first (resp. second) column contains the  $x$  (resp.  $y$ ) coordinates of the points at which the function is to be evaluated.

As a first example consider the function  $f(x, y) = 7$ . There are three options:

1. Pass the constant 7 as scalar to the FEMoctave function. This is the preferred approach.
2. Define a function

```

Octave
function res = ff(xy,dummy)
    res = 7*ones(size(yz)(1),1);
endfunction
```

and then pass the string 'ff' to the FEMoctave function.

3. Determine the vector of the correct size by

```

Octave
ffVec = 7 * ones(size(mesh.GP)(1),1);
```

and then pass the vector `ffVec` to the FEMoctave function.

For the second example function

$$f(x, y) = 7 + 2x$$

the option *constant* is not applicable. There are two equally valid methods.

1. Define a function

```

Octave
function res = ff(xy,dummy)
    res = 7 + 2*xy(:,1);
endfunction
```

and then pass the string 'ff' to the FEMoctave function.

2. Determine the vector of the correct size by

```

Octave
```

<sup>9</sup>In the script files (`FEMEquation.m` and similar) the function is called with the nodes types as second argument, to be used for different sections in the domain. If you only use the compiled versions (`FEMEquation.oct` and similar) the dummy argument is not required. I might remove this "feature" in a next release.

```
ffVec = 7 + 2*xy(:,1);
```

and then pass the vector `ffVec` to the `FEMoctave` function.

To implement the function

$$f(x, y) = J_0(r) = J_0(\sqrt{x^2 + y^2})$$

to be passed to the `FEMoctave` command use

```
function y = f(xy)
    y = besselj(0, sqrt(xy(:,1).^2+xy(:,2).^2));
endfunction
```

With this definition pass the string `'f'` to the `FEMoctave` function. Alternatively you can first compute the column vector `fVec` of this function at the Gauss points of the mesh by

```
fVec = f(mesh.GP);
```

and then pass the vector `fVec` to the `FEMoctave` function.

### 4.3.2 Functions for dynamic problems

The only change is the additional time  $t$ , to be passed as a second argument, i.e.  $f(x, y, t) = \dots$

## 4.4 Solving elliptic problems

The first few commands shown in Table 1 can be used to solve elliptic problem on a bounded domain  $\Omega \subset \mathbb{R}^2$ . In the next two sections the commands to solve a symmetric and a non-symmetric elliptic BVP are shown.

### 4.4.1 Symmetric elliptic problems: `BVP2Dsym()`

Equations given in the form of (2)

$$\begin{aligned} -\nabla \cdot (a \nabla u) + b_0 u &= f & \text{for } (x, y) \in \Omega \\ u &= g_1 & \text{for } (x, y) \in \Gamma_1 \\ a \frac{\partial u}{\partial n} &= g_2 + g_3 u & \text{for } (x, y) \in \Gamma_2 \end{aligned}$$

may be solved by

```
u = BVP2Dsym(mesh, a, b0, f, g1, g2, g3)
```

where the coefficient functions can be given as described in Section 4.3.1, as constants, strings or vectors. The return value `u` is a vector with the values of the solution at the nodes.

```
U = BVP2Dsym(MESH, A, B0, F, GD, GN1, GN2)
Solve a symmetric, elliptic boundary value problem
```

$$\begin{aligned} -\text{div}(a * \text{grad } u) + b0 * u &= f & \text{in domain} \\ u &= gD & \text{on Dirichlet boundary} \\ n * (a * \text{grad } u) &= gN1 + gN2 * u & \text{on Neumann boundary} \end{aligned}$$

parameters:

- \* MESH is the mesh describing the domain and the boundary types
- \* A, B0, F, GD, GN1, GN2 are the coefficients and functions describing the PDE. Any constant function can be given by its scalar value. The functions A, B0 and F may also be given as vectors with the values of the function at the Gauss points.

return value

- \* U is the vector with the values of the solution at the nodes

Find examples in Sections 3.1.1, 3.1.2, 3.1.3, 9.4, 9.5, 9.7 and 9.15.

#### 4.4.2 General elliptic problems: BVP2D ()

Equations given in the form of (1)

$$\begin{aligned} -\nabla \cdot (a \nabla u - u \vec{b}) + b_0 u &= f && \text{for } (x, y) \in \Omega \\ u &= g_1 && \text{for } (x, y) \in \Gamma_1 \\ \vec{n} \cdot (a \nabla u - u \vec{b}) &= g_2 + g_3 u && \text{for } (x, y) \in \Gamma_2 \end{aligned}$$

may be solved by

**Octave**

```
u = BVP2D(mesh, a, b0, bx, by, f, g1, g2, g3)
```

where the coefficient functions can be given as described in Section 4.3.1, as constants, strings or vectors. The expressions `bx` and `by` denote the two components of the convection vector  $\vec{b}$ . The return value `u` is a vector with the values of the solution  $u$  at the nodes. Find an example in Section 3.1.4.

**BVP2D()**

```
U = BVP2D(MESH, A, B0, BX, BY, F, GD, GN1, GN2)
```

Solve an elliptic boundary value problem

$$\begin{aligned} -\text{div}(a * \text{grad } u - u * (bx, by)) + b_0 * u &= f && \text{in domain} \\ u &= gD && \text{on Dirichlet boundary} \\ n * (a * \text{grad } u - u * (bx, by)) &= gN1 + gN2 * u && \text{on Neumann boundary} \end{aligned}$$

parameters:

- \* MESH is the mesh describing the domain and the boundary types
- \* A, B0, BX, BY, F, GD, GN1, GN2 are the coefficients and functions describing the PDE. Any constant function can be given by its scalar value. The functions A, B0, BX, BY and F may also be given as vectors with the values of the function at the Gauss points.

return value

- \* U is the vector with the values of the solution at the nodes

#### 4.5 Solving 2D eigenvalue problems: BVP2Deig ()

To solve an eigenvalue problem of the form (3)

$$\begin{aligned} -\nabla \cdot (a \nabla u) + b_0 u &= \lambda f u && \text{for } (x, y) \in \Omega \\ u &= 0 && \text{for } (x, y) \in \Gamma_1 \\ a \frac{\partial u}{\partial n} &= g_3 u && \text{for } (x, y) \in \Gamma_2 \end{aligned}$$

use

**Octave**

```
[Eval,Evec,errorbound] = BVP2Deig(mesh,a,b0,f,gN2,nVec,tol);
```

where the coefficient functions can be given as described in Section 4.3.1, as constants, strings or vectors.

- The function can be called with one (`Eval`) or two (`[Eval,Evec]`) return arguments. A possible third return argument (`[Eval,Evec,errorbound]`) is of limited use, since with newer versions of FEMoctave `eigs()` is used, instead of an inverse power iteration.
  - The first return value `Eval` is a column vector containing the estimated values of the eigenvalues  $\lambda_i$ .
  - If the second return value `Evec` is asked for, then a matrix will be returned. Each column contains the values of a normalized eigenfunction at the nodes.
  - The third return argument `errorbound` will return a matrix with two columns, containing information on the error bound of the eigenvalues. Observe the the error of the eigenvalue computation is given, not the error of the overall FEM problem. The error of the FEM discretization has to be estimated by other tools. Some mathematical details are given in Section 6.9.
    - \* The first column contains a conservative error estimate. The actual error of the eigenvalue is guaranteed to be smaller.
    - \* The second column contains a more aggressive error estimate. Under most circumstances the estimate is valid. For highly clustered eigenvalues the error is overestimated.. There are circumstances when the error of the largest eigenvalues is underestimated. If the error is extremely small, the estimate might indicate an even smaller error. Keep in mind the the error is always larger than machine accuracy permits.
- The integer parameter `nVec` indicate the number of smallest eigenvalues to be computed.
- The parameter `tol` will lead to the iteration stopping if the relative change from one step to the next is smaller than `tol`. If the parameter is not given, then a default value of  $10^{-5}$  is used.

An example of an eigenvalue problem is given in Section 3.2.

**BVP2Deig()**

```
[EVAL,EVEC,ERRORBOUND] = BVP2Deig(MESH,A,B0,W,GN2,NVEC,TOL)
```

determine the smallest eigenvalues `EVAL` and eigenfunctions `EVEC` for the BVP

$$\begin{aligned} -\operatorname{div}(a*\operatorname{grad} u) + b_0*u &= \operatorname{Eval}*w*u && \text{in domain} \\ u &= 0 && \text{on Dirichlet boundary} \\ n*(a*\operatorname{grad} u) &= gN2*u && \text{on Neumann boundary} \end{aligned}$$

parameters:

- \* `MESH` is the mesh describing the domain and the boundary types
- \* `A,B0,W,GN2` are the coefficients and functions describing the PDE. Any constant function can be given by its scalar value. The functions `A,B0` and `W` may also be given as vectors with the values of the function at the Gauss points.
- \* `NVEC` is the number of smallest eigenvalues to be computed
- \* `TOL` optional tolerance for the eigenvalue iteration, default  $1e-5$

return values:

- \* `EVAL` is the vector with the eigenvalues
- \* `EVEC` is the matrix with the eigenvectors as columns
- \* `ERORBOUND` is a matrix with error bounds of the eigenvalues

In Sections 6.8.2 and 6.8.4 find the consequences of the eigenvalues to solutions of dynamic heat and wave equations.

#### 4.6 Solving parabolic problems: IBVP2D () and IBVP2Dsym ()

To solve an initial boundary value problem (IBVP) of the form (4)

$$\begin{aligned} \rho \frac{\partial}{\partial t} u - \nabla \cdot (a \nabla u - u \vec{b}) + b_0 u &= f & \text{for } (x, y, t) \in \Omega \times (0, T] \\ u &= g_1 & \text{for } (x, y, t) \in \Gamma_1 \times (0, T] \\ \vec{n} \cdot (a \nabla u - u \vec{b}) &= g_2 + g_3 u & \text{for } (x, y, t) \in \Gamma_2 \times (0, T] \\ u &= u_0 & \text{on } \Omega \text{ at } t = 0 \end{aligned}$$

use the command IBVP2D (). Find an example in Section 3.3 and a description of the algorithm in Section 6.8.1.

##### IBVP2D()

```
[U, T] = IBVP2D(MESH, M, A, B0, BX, BY, F, GD, GN1, GN2, U0, T0, TEND, STEPS)
```

Solve an initial boundary value problem

$$\begin{aligned} m \cdot d/dt \ u - \operatorname{div}(a \cdot \operatorname{grad} \ u - u \cdot (bx, by)) + b_0 \cdot u &= f & \text{in domain} \\ u &= gD & \text{on Dirichlet boundary} \\ n \cdot (a \cdot \operatorname{grad} \ u - u \cdot (bx, by)) &= gN1 + gN2 \cdot u & \text{on Neumann boundary} \\ u(t_0) &= u_0 & \text{initial value} \end{aligned}$$

parameters:

- \* MESH is the mesh describing the domain and the boundary types
- \* M, A, B0, BX, BY, F, GD, GN1, GN2 are the coefficients and functions describing the PDE. Any constant function can be given by its scalar value. The functions M, A, B0, BX, BY and F may also be given as vectors with the values of the function at the Gauss points.
- \* F may be given as a string for a function depending on (x,y) and time t or a a vector with the values at nodes or as scalar. If F is given by a scalar or vector it is independent on time.
- \* U0 is the initial value, can be given as a constant, function name or as vector with the values at the nodes
- \* T0, TEND are the initial and final times
- \* STEPS is a vector with one or two positive integers. If STEPS = n, then n Crank Nicolson steps are taken and the results returned. If STEPS = [n, nint], then n\*nint steps are taken and (n+1) results returned.

return values

- \* U is a matrix with n+1 columns with the values of the solution at the nodes at different times T
- \* T is the vector with the values of the times at which the solutions are returned.

If there is no convection term  $\vec{b} = \vec{0}$ , then the resulting matrix  $\mathbf{A}$  is symmetric and (most often) positive definite. Thus one can use a Cholesky factorization for the time stepper. This is (or should be) faster. The structure of IBVP2Dsym () is almost identical to IBVP2D ().

##### IBVP2Dsym()

```
IBVP2Dsym(MESH, M, A, B0, F, GD, GN1, GN2, U0, T0, TEND, STEPS)
```

Solve a symmetric initial boundary value problem

$$\begin{aligned} m \cdot d/dt \ u - \operatorname{div}(a \cdot \operatorname{grad} \ u) + b_0 \cdot u &= f & \text{in domain} \\ u &= gD & \text{on Dirichlet boundary} \\ n \cdot (a \cdot \operatorname{grad} \ u) &= gN1 + gN2 \cdot u & \text{on Neumann boundary} \\ u(t_0) &= u_0 & \text{initial value} \end{aligned}$$

...

#### 4.7 Solving hyperbolic problems: I2BVP2D ()

Examine an IBVP (6) of hyperbolic type.

$$\begin{aligned} \rho \frac{\partial^2}{\partial t^2} u + 2\alpha \frac{\partial}{\partial t} u - \nabla \cdot (a \nabla u - u \vec{b}) + b_0 u &= f && \text{for } (x, y, t) \in \Omega \times (0, T] \\ u &= g_1 && \text{for } (x, y, t) \in \Gamma_1 \times (0, T] \\ \vec{n} \cdot (a \nabla u - u \vec{b}) &= g_2 + g_3 u && \text{for } (x, y, t) \in \Gamma_2 \times (0, T] \\ u &= u_0 && \text{on } \Omega \text{ at } t = 0 \\ \frac{\partial}{\partial t} u &= v_0 && \text{on } \Omega \text{ at } t = 0 \end{aligned}$$

To solve this wave type equation use the command I2BVP2D (). Find examples in Sections 9.2 and 9.12 and a description of the algorithm in Section 6.8.3.

#### I2BVP2D()

```
[U, T] = I2BVP2D(MESH, M, D, A, B0, BX, BY, F, GD, GN1, GN2, U0, V0, T0, TEND, STEPS)
```

Solve an initial boundary value problem

$$\begin{aligned} m \cdot d^2/dt^2 u + 2 \cdot d \cdot d/dt u - \text{div}(a \cdot \text{grad } u - u \cdot (b_x, b_y)) + b_0 \cdot u &= f && \text{in domain} \\ u &= g_D && \text{on Dirichlet boundary} \\ n \cdot (a \cdot \text{grad } u - u \cdot (b_x, b_y)) &= g_{N1} + g_{N2} \cdot u && \text{on Neumann boundary} \\ u(t_0) &= u_0 && \text{initial value} \\ d/dt u(t_0) &= v_0 && \text{initial velocity} \end{aligned}$$

parameters:

- \* MESH is the mesh describing the domain and the boundary types
- \* M, D, A, B0, BX, BY, F, GD, GN1, GN2 are the coefficients and functions describing the PDE.  
Any constant function can be given by its scalar value.  
The functions M, D, A, B0, BX, BY and F may also be given as vectors with the values of the function at the Gauss points.
- \* F may be given as a string for a function depending on (x, y) and time t or a vector with the values at nodes or as scalar.  
If F is given by a scalar or vector it is independent on time.
- \* U0, V0 are the initial value and velocity, can be given as a constant, function name or as vector with the values at the nodes
- \* T0, TEND are the initial and final times
- \* STEPS is a vector with one or two positive integers.  
If STEPS = n, then n steps are taken and the results returned.  
If STEPS = [n, nint], then n\*nint steps are taken and (n+1) results returned.

return values

- \* U is a matrix with n+1 columns with the values of the solution at the nodes at different times T
- \* T is the vector with the values of the times at which the solutions are returned.

#### 4.8 Solving 1D steady state problems, BVP1D ()

To solve a steady state boundary value problem (7), i.e.

$$- (a(x) u'(x))' + b(x) u'(x) + c(x) u(x) = d(x) f(x)$$

with some boundary conditions (8) at  $x = x_0$  and  $x = x_n$

$$\begin{aligned} u(x_i) &= g_D && \text{Dirichlet} \\ a(x_i) u'(x_i) &= g_{N1} + g_{N2} u(x_i) && \text{Neumann} \end{aligned}$$

use the command `BVP1D()`.

command	purpose
<code>BVP1D()</code>	solve a static boundary value problem in 1D
<code>IBVP1D()</code>	solve a first order initial boundary value problem in 1D
<code>I2BVP1D()</code>	solve a second order initial boundary value problem in 1D
<code>GenerateFEM1D()</code>	generate the matrices for <code>BVP1D()</code>
<code>pwquadinterp()</code>	piece-wise quadratic interpolation
<code>FEM1DEvaluateDu()</code>	evaluate the first derivative at the nodes
<code>FEM1DGaussPoints()</code>	determine the coordinates of the Gauss points and interpolation
<code>GenerateWeight1D()</code>	weight matrices for the dynamic 1D problems

Table 6: Commands to solve and examine 1D boundary value problems

For a call of `BVP1D()`

$$[x, u] = \text{BVP1D}(\text{interval}, a, b, c, d, f, \text{BCleft}, \text{BCright})$$

the following parameters are required:

- `interval`: the interval on which the BVP will be solved. It has the form `interval = [x1, x1, x2, ..., xn]`. On each subinterval  $[x_i, x_{i+1}]$  the midpoint  $\frac{x_i + x_{i+1}}{2}$  will be added and then quadratic function on the subinterval will be used for the computations.
- `a`, `b`, `c` and `d`: these coefficient functions can be given either as constant scalar value, as vector or as function handle<sup>10</sup> to determine the values.
  - The scalar constant will be used as value at the Gauss points.
  - The vector has to contain the values at the Gauss points.
  - The function handle will be evaluated at the Gauss points.

The code has to be vectorized, e.g. to describe the coefficient  $a(x) = x^2$  use the vectorized function handle `a = @(x) x.^2`.

- `f`: this function can be given as constant scalar, as vector or as a vectorized function handle. `f` will determine the values of the function  $f(x)$  at the nodes, i.e. the end- and mid-points of the subintervals.
  - The scalar constant will be used as value at the nodes.
  - The vector has to contain the values at the nodes.
  - The function handle will be evaluated at the nodes.
- There are two ways to define the inhomogeneous contribution  $d(x) f(x)$  for a BVP. Except for quadratic functions the results will be slightly different. E.g. for the BVP  $-u''(x) = \sin(x)$  the expression  $\sin(x)$  can be given by  $d(x) = \sin(x)$  and  $f(x) = 1$  or  $d(x) = 1$  and  $f(x) = \sin(x)$ .

<sup>10</sup>This is different from the way to define functions for the 2D codes in FEMoctave.

1. For  $d(x) = \sin(x)$  and  $f(x) = 1$  use the arguments  $d=@(x)\sin(x)$  and  $f=1$ . With this description  $\sin(x)$  will be evaluated at the Gauss points and then used for the RHS of the linear system. This is the approach preferred by this author.
  2. For  $d(x) = 1$  and  $f(x) = \sin(x)$  use the arguments  $d=1$  and  $f=@(x)\sin(x)$ . With this description  $\sin(x)$  will be evaluated at the nodes (end- and mid-points of the subintervals) and with a quadratic interpolation the values at the Gauss points will be determined and then used for the RHS of the linear system. This approach can be useful for iterative procedures.
- **BCleft** and **BCright**: with these parameters the boundary conditions are specified. If it is a single scalar Dirichlet conditions are used. If it is a vector of two scalars, Neumann conditions are used. E.g.
    - **BCleft**=3 describes the boundary condition  $u(x_0) = 3$
    - **BCleft**= [3, -2] describes the boundary condition  $a(x_0)u'(x_0) = 3 - 2u(x_0)$

There are two return arguments:

- **x** is a vector with the coordinates of the nodes, i.e. all end- and mid-points of the subintervals.
- **u** is a vector with the values of the (approximate) solution  $u(x)$  at the nodes.

Internally **BVP1D()** will call the function **GenerateFEM1D()** to approximate the solution of the boundary value problem by a linear system  $\mathbf{A}\vec{u} = \mathbf{M}\vec{f}$ . The code in **BVP1D()** implements the different boundary conditions.

#### BVP1D()

```
[X,U] = BVP1D(INTERVAL,A,B,C,D,F,BCLEFT,BCRIGHT)
```

```
solve a 1D boundary value probme (BVP)
-(a(x)*u'(x))' + b(x)*u'(x) + c(x)*u(x) = d(x)*f(x)
with boundary conditions at the two endpoints
* Dirichlet: u(x) = g_D
* Neumann: a(x)*u'(x) = g_N1 + g_N2*u(x)
```

parameters:

- \* **INTERVAL** the discretized interval for the BVP
- \* **A** constant, vector or function handle to evaluate  $a(x)$
- \* **B** constant, vector or function handle to evaluate  $b(x)$
- \* **C** constant, vector or function handle to evaluate  $c(x)$
- \* **D** constant, vector or function handle to evaluate  $d(x)$
- \* **F** constant, vector or function handle to evaluate  $f(x)$
- \* **BCLEFT** and **BCRIGHT** the two boundary conditions
  - \* for a Dirichlet condition specify a single value **G\_D**
  - \* for a Neumann condition specify the values [**G\_N1**,**G\_N2**]

return values

- \* **X** the nodes in the given interval
- \* **U** the values of the solution at the nodes

The function **GenerateFEM1D()** is used by **BVP1D()**, **IBVP1D()** and **I2BVP1D()** to generate the matrices required to solve the BVP and IBVP.

**GenerateFEM1D()**

```
[A,M,XNEW] = GenerateFEM1D(X,A,B,C,D)
```

generate the matrices A and M to discretize the expression  
 $-d/dx (a(x) d/dx u(x)) + b(x)*d/dx u(x) + c(x)*u(x) - d(x)*f(x)$   
 by  $A*u - M*f$

parameters:

- \* A function handle to evaluate the coefficient  $a(x)$ , vectorized
- \* B function handle to evaluate the coefficient  $b(x)$ , vectorized
- \* C function handle to evaluate the coefficient  $c(x)$ , vectorized
- \* D function handle to evaluate the coefficient  $d(x)$ , vectorized

return values

- \* A matrix discretizing the expressions involving  $u(x)$
- \* M matrix discretizing the evaluation of  $d(x)*f(x)$
- \* XNEW vector with the grid points

for an elementary demo use "demo GenerateFEM1D"

`BVP1D()` will determine the values of the solution at the nodes. If more values are required use a piecewise quadratic interpolation, i.e. the command `pwquadinterp()`. The values of the function and the first and second derivatives can be evaluated.

**pwquadinterp()**

```
yi = pwquadinterp(xdata,ydata,xi) % evaluate the values of the function
[yi, yi_x, yi_xx] = pwquadinterp(xdata,ydata,xi) % evaluate first and second derivatives
```

Use the data (xdata,ydata) to determine a piecewise quadratic function and then evaluate this function at the points xi.

With multiple return arguments derivatives are evaluated.

The function requires that:

- there are an odd number of data points,
- xdata and xi are both in increasing order,
- the xi values lie between xdata(1) and xdata(end).

The above command will evaluate at arbitrary points in the interval. If the values of the derivative are only required at the nodes, use the command `FEM1DEvaluateDu()`.

**FEM1DEvaluateDu()**

```
DU = FEM1DEvaluateDu(X,U)
evaluate the first derivative at the nodes x
```

parameters:

- \* X coordinates of the nodes, generated by `BVP1D()`
- \* U values of the function at the nodes

return values

- \* DU the values of the derivative at the nodes

For some problems the locations of the Gauss points are required. Use the function `FEM1DGaussPoints()`. This function also returns matrices to interpolate the values of a function and its derivative at the Gauss points, using the values at the nodes. See Section 7.2 for the details, equation (66) interpolates the values of the function.

**FEM1DGaussPoints()**

```
[XGAUSS, NODES2GAUSSU, NODES2GAUSSDU] = FEM1DGaussPoints(X)
```

determine the coordinates of the Gauss points and interpolation matrices

parameters:

- \* X coordinates of the nodes, generated by BVP1D()

return values

- \* XGAUSS coordinates of the Gauss points
- \* NODES2GAUSSU matrix to evaluate  $u$  at the Gauss points
- \* NODES2GAUSSDU matrix to evaluate  $u'$  at the Gauss points

The commands IBVP1D() and I2BVP1D() use the internal command GenerateWeight1D() to solve dynamic 1D problems.

**GenerateWeight1D()**

```
[XNEW, W1MAT, W2MAT] = GenerateWeight1D(X, W1, W2)
```

generate the weight matrices W1MAT and W2MAT

parameters:

- \* W1 constant, vector or function handle to evaluate the coefficient  $w_1(x)$ , vectorized
- \* W2 (optional) constant, vector or function handle to evaluate the coefficient  $w_2(x)$ , vectorized

return values

- \* XNEW vector with the grid points
- \* W1MAT weight matrix discretizing W1
- \* W2MAT (optional) weight matrix discretizing W2

**4.9 Solving 1D dynamic problems of order 1, IBVP1D()**

To solve a dynamic boundary value problem (9), i.e.

$$w(x) \frac{\partial}{\partial t} u(x, t) - (a(x) u'(x, t))' + b(x) u'(x, t) + c(x) u(x, t) = d(x) f(x, t)$$

with boundary conditions as in (8)

$$\begin{aligned} u(x_i) &= g_D && \text{Dirichlet} \\ a(x_i) u'(x_i) &= g_{N1} + g_{N2} u(x_i) && \text{Neumann} \end{aligned}$$

and an initial condition  $u(x, t_0) = u_0(x)$  use the command IBVP1D().

Most of the parameters for a call of

```
[x, u, t] = IBVP1D(interval, w, a, b, c, d, f, BCleft, BCright, u0, t0, tend, steps, varargin)
```

are very similar to the above call of BVP1D().

- $w$ : is the coefficient function for the weight  $w(x)$  and can be given as scalar value, vector of values at the Gauss points or function handle.

- $f$ : this function can be given as constant scalar, as vector or as a vectorized function handle of the form  $f = @(x, t)$ .  $f$  will determine the values of the function  $f(x, t)$  at the nodes, i.e. the end- and mid-points of the subintervals.
- $u_0$ : the initial value  $u(x, t_0) = u_0(x)$ , either as single scalar, vector of values at the nodes or a function handle.
- $t_0$ :  $t_{end}$ : initial and end time.
- $steps$ : number of steps to be taken by the time stepping algorithm.
  - If  $steps = n$  is a single, positive integer, then  $n$  steps will be taken and thus  $n + 1$  results returned, including the initial value  $u_0(x)$ .
  - If  $steps = [n \ nint]$  is a vector of two positive integers, then  $n$  steps will be taken and thus  $n + 1$  results returned. In between the returned results  $nint$  additional steps will be performed, such that the time step  $\Delta t$  will be smaller.

The length of one time step is given by  $\Delta t = \frac{t_{end}-t_0}{n*nint}$ .

- With the optional argument the type of time stepping algorithm can be selected. Use the string "solver" as name for the option and a string for the name of the algorithm.
  - "CN": the standard Crank–Nicolson algorithm. This is the default algorithm. It is consistent of order 2 and unconditionally stable, but not L–stable.
  - "implicit": the standard implicit solver. It is consistent of order 1, unconditionally stable and L–stable.
  - "explicit": the standard explicit solver. It is consistent of order 1, only conditionally stable and certainly not L–stable. For 1D problems it does not make sense to use this solver on real problems. It can be used to demonstrated the effect of conditional stability.
  - "RK": an implicit Runge–Kutta algorithm. It is consistent of order 2, unconditionally stable and L–stable. The computational effort is larger than for the other algorithms, but still small for 1D problems.

There are three return arguments:

- $x$  is a vector with coordinates of the nodes, i.e. all end- and mid–points of the subintervals.
- $u$  is a matrix with the values of the (approximate) solutions  $u(x)$  at the nodes. The first column contains the initial values. Each column contains the values for one time, given in the return argument  $t$ .
- $t$  is a vector with the times at which the values are returned in  $u$ .

Internally `IBVP1D()` will call the function `GenerateFEM1D()` to approximate the solution of the initial boundary value problem by a linear system of ordinary differential equations  $\mathbf{W} \frac{d}{dt} \vec{u}(t) + \mathbf{A} \vec{u}(t) = \mathbf{M} \vec{f}(t)$ . The code in `IBVP1D()` implements the boundary conditions and the time steppers.

#### IBVP1D()

```
[x, u, t] = IBVP1D(INTERVAL, W, A, B, C, D, F, BCLEFT, BCRIGHT, U0, T0, TEND, STEPS, OPTIONS)
```

solve a 1D initial boundary value problem (IBVP)

$$w(x) * d/dt u(x, t) - (a(x) * u'(x, t))' + b(x) * u'(x, t) + c(x) * u(x, t) = d(x) * f(x, t)$$

with initial condition  $u(x, t_0) = u_0(x)$  and boundary conditions at the two endpoints

```
* Dirichlet:      u(x,t) = g_D
* Neumann:      a(x)*u'(x,t) = g_N1 + g_N2*u(x)
```

parameters:

```
* INTERVAL the discretized interval for the BVP
* W constant, vector or function handle to evaluate w(x)
* A constant, vector or function handle to evaluate a(x)
* B constant, vector or function handle to evaluate b(x)
* C constant, vector or function handle to evaluate c(x)
* D constant, vector or function handle to evaluate d(x)
* F constant, vector or function handle to evaluate the f(x)
* BCLEFT and BCRIGHT the two boundary conditions
  * for a Dirichlet condition specify a single value G_D
  * for a Neumann condition specify the values [G_N1,G_N2]
* U0 constant, vector with the initial values at the nodes or a
  function handle to evaluate u(t0)
* T0, TEND are the initial and final times
* STEPS is a vector with one or two positive integers.
  * If STEPS = n, then n steps are taken and the n+1 results returned.
  * If STEPS = [n,nint], then n*nint steps are taken and (n+1) results returned.
* OPTIONS additional options, given as pairs name/value.
  Currently only the type stepping algorithm can be selected as "SOLVER"
  and then the possible values
  * "CN" the standard Crank-Nicolson (default)
  * "IMPLICIT" the standard implicit solver
  * "EXPLICIT" the standard explicit solver
  * "RK" an L-stable, implicit Runge-Kutta solver
```

return values

```
* X the nodes in the given interval
* U is a matrix with n+1 columns with the values of the solution at
  the nodes at different times T
* T is the vector with the values of the times at which the solutions are returned.
```

#### 4.10 Solving 1D dynamic problems of order 2, I2BVP1D ()

To solve a dynamic boundary value problem (10), i.e.

$$w_2(x) \frac{\partial^2}{\partial t^2} u(x,t) + w_1(x) \frac{\partial}{\partial t} u(x,t) - (a(x)u'(x,t))' + b(x)u'(x,t) + c(x)u(x,t) = d(x)f(x,t)$$

again with Dirichlet or Neumann boundary conditions and an initial value  $u(x, t_0) = u_0(x)$  and initial velocity  $\frac{\partial}{\partial t} u(x, t_0) = u_1(x)$  use the command I2BVP1D (). The syntax is extremely similar to the above IBVP1D (). The only essential difference is the specification of the initial velocity  $u_1(x)$ . Currently only one time stepper implemented, an unconditionally stable implicit scheme, consistent of order 2.

##### I2BVP1D()

```
[X,U,T] = I2BVP1D (INTERVAL,W1,A,B,C,D,F,BCLEFT,BCRIGHT,U0,U1,T0,TEND,STEPS)
```

solve a second order 1D initial boundary value problem (IBVP)

$$w_2(x) * d^2/dt^2 u(x,t) + w_1(x) * d/dt u(x,t) - (a(x) * u'(x,t))' + b(x) * u'(x,t) + c(x) * u(x,t) = d(x) * f(x,t)$$

with initial condition  $u(x, t_0) = u_0(x)$  and  $d/dt u(x, t_0) = u_1$  and

boundary conditions at the two endpoints

- \* Dirichlet:  $u(x,t) = g\_D$
- \* Neumann:  $a(x)u'(x,t) = g\_N1 + g\_N2u(x)$

parameters:

- \* INTERVAL the discretized interval for the BVP
- \* W2 constant, vector or function handle to evaluate  $w2(x)$
- \* W1 constant, vector or function handle to evaluate  $w1(x)$
- \* A constant, vector or function handle to evaluate  $a(x)$
- \* B constant, vector or function handle to evaluate  $b(x)$
- \* C constant, vector or function handle to evaluate  $c(x)$
- \* D constant, vector or function handle to evaluate  $d(x)$
- \* F constant, vector or function handle to evaluate the  $f(x,t)$
- \* BCLEFT and BCRIGHT the two boundary conditions
  - \* for a Dirichlet condition specify a single value  $G\_D$
  - \* for a Neumann condition specify the values  $[G\_N1,G\_N2]$
- \* U0 constant, vector with the initial values at the nodes or a function handle to evaluate  $u(t0)$
- \* U1 constant, vector with the initial velocities at the nodes or a function handle to evaluate  $u(t0)$
- \* T0, TEND are the initial and final times
- \* STEPS is a vector with one or two positive integers.
  - \* If STEPS =  $n$ , then  $n$  steps are taken and the  $n+1$  results returned.
  - \* If STEPS =  $[n,nint]$ , then  $n*nint$  steps are taken and  $(n+1)$  results returned.

return values

- \* X the nodes in the given interval
- \* U is a matrix with  $n+1$  columns with the values of the solution at the nodes at different times T
- \* T is the vector with the values of the times at which the solutions are returned.

#### 4.11 Solving 1D eigenvalue problems: BVP1Deig()

To solve an 1D eigenvalue problem of the form (11)

$$\begin{aligned}
 -(a(x)u'(x))' + b(x)u'(x) + c(x)u(x) &= \lambda w(x)u(x) && \text{for } x_0 < x < x_n \\
 u(x_i) &= 0 && \text{Dirichlet BC} \\
 a(x_i)u'(x_i) &= g_{N2}u(x_i) && \text{Neumann BC}
 \end{aligned}$$

use

**Octave**

```
[x,eVal,eVec,errorbound] = BVP1Deig(interval,a,b,c,w,BCleft,BCright,nVec,tol)
```

where the coefficient functions can be given as described in Section 4.8, as constants, vectors or function handles. There are (possibly) two additional arguments

- The integer parameter `nVec` indicates the number of smallest eigenvalues to be computed.
- The optional parameter `tol` will lead to the iteration stopping if the relative change from one step to the next is smaller than `tol`. If the parameter is not given, then a default value of `tol = 10-5` is used.

The function can be called with two (`x, Eval`) or three (`[x, Eval, Evec]`) return arguments. A possible fourth return argument (`[x, Eval, Evec, errorbound]`) is of limited use, since with newer versions of FEMoctave `eigs()` is used, instead of an inverse power iteration.

- The first argument `x` contains the nodes at which the solutions are evaluated.
- The second return value `eVal` is a column vector containing the estimated values of the eigenvalues  $\lambda_i$ .
- If the third return value `eVec` is asked for, then a matrix will be returned. Each column contains the values of a normalized eigenfunction at the nodes.
- The fourth return argument `errorbound` is similar to the command `BVP2Deig()` (Section 4.5, page 63).

An example of an 1D eigenvalue problem is given in Section 3.2.

#### BVP1Deig()

```
[X,EVAL,EVEC,ERRORBOUND] = BVP1Deig(INTERVAL,A,B,C,W,BCLEFT,BCRIGHT,NVEC,TOL)
```

determine the smallest eigenvalues `EVAL` and eigenfunctions `EVEC` for the BVP

$$\begin{aligned} -(a(x) * u'(x))' + b(x) * u'(x) + c(x) * u(x) &= eVal * w(x) * u(x) \\ u &= 0 \text{ on Dirichlet boundary} \\ a * u' &= g\_N2 * u \text{ on Neumann boundary} \end{aligned}$$

parameters:

- \* `INTERVAL` the discretized interval for the BVP
- \* `A` constant, vector or function handle to evaluate  $a(x)$
- \* `B` constant, vector or function handle to evaluate  $b(x)$
- \* `C` constant, vector or function handle to evaluate  $c(x)$
- \* `W` constant, vector or function handle to evaluate  $d(x)$
- \* `BCLEFT` and `BCRIGHT` the two boundary conditions
  - \* for a Dirichlet condition specify a single value `G_D`
  - \* for a Neumann condition specify the values `[G_N1,G_N2]`
- \* `NVEC` the number of smallest eigenvalues to be computed
- \* `TOL` optional tolerance for the eigenvalue iteration, default  $1e-5$

return values

- \* `X` the nodes in the given interval
- \* `EVAL` the eigenvalues of the solution at the nodes
- \* `EVEC` the matrix of eigenvectors of the solutions at the nodes
- \* `ERRORBOUND` a matrix with error bounds of the eigenvalues

#### 4.12 Solving nonlinear 1D boundary value problems: `BVP1DNL()`

Search for solutions of nonlinear boundary value problems given in equation (13)

$$-(a(x, u(x), u'(x)) u'(x))' + b(x) u'(x) + c(x) u(x) = f(x, u(x), u'(x))$$

with linear Dirichlet or Neumann boundary conditions.

- For the dependence on the function  $f$  Newton's method is used, based on the linear Taylor approximation

$$f(x, u + \phi, u' + \phi') \approx f(x, u, u') + \frac{\partial f}{\partial u} \phi + \frac{\partial f}{\partial u'} \phi'$$

- If the coefficient  $a$  depends on  $u$  or  $u'$  a partial substitution method is used.
- If  $a$  and  $f$  are nonlinear, a combination of Newton and substitution is used. The basic idea is spelled out in Algorithm 1. Find more details in Section 7.8.

**Algorithm 1:** The algorithm of BVP1DNL()

evaluate  $a_n = a(x, u_0(x), u'_0(x))$  and  $f_0 = f(x, u_0(x), u'_0(x))$

solve  $-(a_n u'_n)' + b u'_n + c u_n = f_0$

**repeat**

  set  $u_{old} = u_n$

  evaluate  $f_n = f(x, u_n(x), u'_n(x))$ ,  $f_u = \frac{\partial}{\partial u} f(x, u_n(x), u'_n(x))$  and  $f_{u'} = \frac{\partial}{\partial u'} f(x, u_n(x), u'_n(x))$

  evaluate the coefficients  $b_n = b - f_u$  and  $c_n = c - f_{u'}$

  solve  $-(a_n \phi')' + b_n \phi' + c_n \phi = +(a_n u'_n)' - b_n u'_n - c_n u_n + f_n$

  set  $u_n = u_n + \phi$ , i.e. one Newton step

  evaluate  $a_n = a(x, u_n(x), u'_n(x))$  and  $f_n = f(x, u_n(x), u'_n(x))$

  solve  $(a_n u'_n)' + b u'_n + c u_n = f_n$ , i.e one substitution step

**until** RMS of  $u_{old} - u_n$  small enough or too many iterations;

**return** the results

**BVP1DNL()**

BVP1DNL (INTERVAL, A, B, C, D, F, BCLEFT, BCRIGHT, U0, OPTIONS)

solve a nonlinear 1D boundary value problem (BVP)

$$-(a(x, u, u') * u'(x))' + b(x) * u'(x) + c(x) * u(x) = d(x) * f(x, u, u')$$

with boundary conditions at the two endpoints

\* Dirichlet:  $u(x) = g\_D$

\* Neumann:  $a(x, u, u') * u'(x) = g\_N1 + g\_N2 * u(x)$

parameters:

- \* INTERVAL the discretized interval for the BVP
- \* A constant, vector or function handle to evaluate  $a(x)$ ,  $a(x, u)$  or  $a(x, u, u')$  at the Gauss points.
  - \* A a constant or vector of values of  $a(x)$ .
  - \* A = @(X) a function handle to evaluate  $f(x)$ .
  - \* A = {@(X,U), @(X,U)} assumes that the function  $a(x, u)$  depends on  $x$  and  $u$ . The two function handles evaluate  $a(x, u)$  and the partial derivative  $a\_u(x, u)$ .
  - \* A = {@(X,U,U'), @(X,U,U') , @(X,U,U')} assumes that  $a(x, u, u')$  depends on  $x$ ,  $u$  and  $u'$ . The three function handles evaluate  $a(x, u, u')$  and the partial derivatives  $a\_u(x, u, u')$  and  $a\_u'(x, u, u')$ .
- \* B constant, vector or function handle to evaluate  $b(x)$  at Gauss points
- \* C constant, vector or function handle to evaluate  $c(x)$  at Gauss points
- \* D constant, vector or function handle to evaluate  $d(x)$  at Gauss points
- \* F constant, vector or function handle to evaluate  $f(x)$ ,  $f(x, u)$  or  $f(x, u, u')$  and the partial derivatives at nodes
  - \* F a constant or vector of values of  $f(x)$  at the nodes.
  - \* F = @(X) a function handle to evaluate  $f(x)$  at the nodes.
  - \* F = {@(X,U), @(X,U)} assumes that the function  $f$  depends on  $x$  and  $u$ . The two function handles evaluate  $f(x, u)$  and the partial derivative  $f\_u(x, u)$ .
  - \* F = {@(X,U,U'), @(X,U,U') , @(X,U,U')} assumes that  $f$  depends on  $x$ ,  $u$  and  $u'$ . The three function handles evaluate  $f(x, u, u')$  and the partial derivatives  $f\_u(x, u, u')$  and  $f\_u'(x, u, u')$ .
- \* BCLEFT and BCRIGHT the two boundary conditions
  - \* for a Dirichlet condition specify a single value G\_D
  - \* for a Neumann condition specify the values [G\_N1,G\_N2]
- \* U0 constant, vector or function handle to evaluate  $u_0(x)$  at the nodes.

This is the starting value for the iteration.

- \* OPTIONS additional options, given as pairs name/value.
- \* "TOL" the tolerance for the iteration to stop, given as pair [TOLREL,TOLABS] for the relative and absolute tolerance. The iteration stops if the absolute or relative error is smaller than the specified tolerance. RMS (root means square) values are used. If only TOLREL is specified TOLABS=TOLREL is used. The default values are TOLREL = TOLABS = 1E-5.
- \* "MAXITER" the maximal number of iterations to be used. The default value is 10.
- \* "DISPLAY" should information be displayed for the iterations
  - \* "OFF" no display, default
  - \* "ITER" display the number of the iteration and the RMS size of the update

return values

- \* X the nodes in the given interval
- \* U the values of the solution at the nodes
- \* INFORM a structure with information on the performance of the algorithm
  - \* INFORM.INFO=1 if the algorithm converged with the desired tolerance, -1 if not.
  - \* INFORM.ITER the number of iterations used.
  - \* INFORM.ABSERROR the RMS value of the last correction applied.

#### 4.13 Solving plane stress and plane strain problems: PlaneStress(), PlaneStrain()

For a plane stress problem the total energy in expression (21)

$$\begin{aligned}
 U(\vec{u}) &= U_{elast} + U_{Vol} + U_{Surf} \\
 &= \iint_{\Omega} \frac{1}{2} \frac{E}{(1-\nu^2)} \left\langle \begin{bmatrix} 1 & \nu & 0 \\ \nu & 1 & 0 \\ 0 & 0 & 2(1-\nu) \end{bmatrix} \cdot \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix}, \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix} \right\rangle dA - \\
 &\quad - \iint_{\Omega} \vec{f} \cdot \vec{u} dA - \int_{\Gamma_2} \vec{g}_N \cdot \vec{u} ds.
 \end{aligned}$$

is minimized, respecting the boundary conditions (14), i.e.

$$\begin{aligned}
 \vec{u} &= \vec{g}_D && \text{on Dirichlet boundary } \Gamma_1, \text{ i.e. prescribed displacement} \\
 \text{force density} &= \vec{g}_N && \text{on Neumann boundary } \Gamma_2, \text{ i.e. prescribed force density} \\
 \text{force density} &= \vec{0} && \text{on free boundary } \Gamma_3
 \end{aligned}$$

#### PlaneStress()

[U1,U2] = PlaneStress(MESH,E,NU,F,GD,GN)  
solve an plane stress problem

```

plane stress equation    in domain
      u = gD              on Gamma_1
force density = gN       on Gamma_2
force density = 0        on Gamma_3

```

parameters:

- \* MESH is the mesh describing the domain and the boundary types
- \* E,NU Young's modulus and Poisson's ratio for the material
- \* F = {F1,F2} a cell array with the two components of the volume forces
- \* GD = {GD1,GD2} a cell array with the two components of the

```

prescribed displacements on the boundary section Gamma_1
* GN = {GN1,GN2} a cell array with the two components of the
  surface forces on the boundary section Gamma_2
* Any constant function can be given by its scalar value
* Any function can be given by a string with the function name
* The functions E, NU, F1 and F2 may also be given as vectors
  with the values of the function at the Gauss points
return values
* U1 vector with the values of the x-displacement at the nodes
* U2 vector with the values of the y-displacement at the nodes

```

For a plane strain problem the total energy in expression (76)

$$\begin{aligned}
U(\vec{u}) &= U_{elast} + U_{Vol} + U_{Surf} \\
&= \iint_{\Omega} \frac{1}{2} \frac{E}{(1+\nu)(1-2\nu)} \left\langle \begin{bmatrix} 1-\nu & \nu & 0 \\ \nu & 1-\nu & 0 \\ 0 & 0 & 2(1-2\nu) \end{bmatrix} \cdot \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix}, \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix} \right\rangle dA - \\
&\quad - \iint_{\Omega} \vec{f} \cdot \vec{u} dA - \int_{\Gamma_2} \vec{g}_N \cdot \vec{u} ds \\
&= \iint_{\Omega} \frac{1}{2} \frac{E}{(1-(\nu^*)^2)} \left\langle \begin{bmatrix} 1 & \nu^* & 0 \\ \nu^* & 1 & 0 \\ 0 & 0 & 2(1-\nu^*) \end{bmatrix} \cdot \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix}, \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix} \right\rangle dA - \\
&\quad - \iint_{\Omega} \vec{f} \cdot \vec{u} dA - \int_{\Gamma_2} \vec{g}_N \cdot \vec{u} ds .
\end{aligned}$$

is minimized, again respecting the boundary conditions (14). Thus the code for `PlaneStrain()` is almost identical to `PlaneStress()`.

#### PlaneStrain()

```

[U1,U2] = PlaneStrain(MESH,E,NU,F,GD,GN)
  solve an plane strain problem

```

```

plane strain equation      in domain
           u = gD          on Gamma_1
force density = gN        on Gamma_2
force density = 0         on Gamma_3

```

parameters:

- \* MESH is the mesh describing the domain and the boundary types
- \* E,NU Young's modulus and Poisson's ratio for the material
- \* F = {F1,F2} a cell array with the two components of the volume forces
- \* GD = {GD1,GD2} a cell array with the two components of the prescribed displacements on the boundary section Gamma\_1
- \* GN = {GN1,GN2} a cell array with the two components of the surface forces on the boundary section Gamma\_2
- \* Any constant function can be given by its scalar value
- \* Any function can be given by a string with the function name
- \* The functions E, NU, F1 and F2 may also be given as vectors with the values of the function at the Gauss points

return values

- \* U1 vector with the values of the x-displacement at the nodes
- \* U2 vector with the values of the y-displacement at the nodes

#### 4.13.1 Evaluating plane stress and plane strain solutions

In Table 7 find the commands related to solving plane elasticity problems and analyzing their solutions. The functions `EvaluateStress()`, `EvaluateStrain()`, `EvaluateVonMises()`, `EvaluateTresca()` and `EvaluatePrincipalStress()` determine the values at the nodes of the mesh. Thus for many applications these function have to be followed by a call of `FEMgriddata()` to evaluate at arbitrary points.

Observe that the two computational paths

1. Evaluate the partial derivative  $\frac{\partial u_1}{\partial x}$  by a piecewise interpolation of the values of  $u_1$  at the nodes.
2. Evaluate the normal strain  $\varepsilon_{xx}$  at the nodes, followed by a piecewise interpolation to determine the value at the arbitrary point  $(x, y)$ .

will **NOT** generate identical results. The difference should be small, but can be substantial, in particular for first order elements.

1. The value of `eps_xx_1` is evaluated using the values of  $u_1$  at the nodes and then a piecewise linear or quadratic interpolation leads to the value of the partial derivative  $\frac{\partial u_1}{\partial x}$  at the point  $(x, y)$ .
2. The second option `eps_xx_2` will first find values of the strain  $\varepsilon_{xx}$  at the nodes, by taking an average of the partial derivatives  $\frac{\partial u_1}{\partial x}$  at the node in the different triangles touching the node. Then a piecewise linear or quadratic interpolation of the values of  $\varepsilon_{xx}$  at the nodes is used to estimate  $\varepsilon_{xx} = \frac{\partial u_1}{\partial x}$  at the point  $(x, y)$ .

```
[~,eps_xx_1,~] = FEMgriddata(FEMmesh,u1,x,y)
[eps_xx,eps_yy,tau_xy] = EvaluateStrain(FEMmesh,u1,u2);
eps_xx_2 = FEMgriddata(FEMmesh,eps_xx,x,y)
```

command	purpose
<code>PlaneStress()</code>	solve a plane stress problem
<code>PlaneStrain()</code>	solve a plane strain problem
<code>PStressEquationM()</code>	script to generate plane stress equations, order 1
<code>PStressEquationQuadM()</code>	script to generate plane stress equations, order 2
<code>PStressEquationCubicM()</code>	script to generate plane stress equations, order 3
<code>EvaluateStrain()</code>	given the displacement evaluate the strains at the nodes
<code>EvaluateStress()</code>	given the displacement evaluate the stresses at the nodes
<code>EvaluateVonMises()</code>	evaluate the von Mises stress at the nodes
<code>EvaluatePrincipalStress()</code>	evaluate the three principal stresses at the nodes
<code>EvaluateTresca()</code>	evaluate the Tresca stress at the nodes

Table 7: Commands to solve and examine plane elasticity problems

**4.13.2 Evaluation of basic strain and stress:** EvaluateStrain(), EvaluateStress()

Given the displacements  $\vec{u}_1$  and  $\vec{u}_2$  with the corresponding mesh use the function EvaluateStrain() to determine the normal and shearing strains at the nodes of the mesh. The same function can be used for plane stress and plane strain problems. The missing normal strain  $\varepsilon_{zz}$  in  $z$ -direction can be determined independently.

- For a plane stress setup use  $\varepsilon_{zz} = \frac{-\nu}{1-\nu} (\varepsilon_{xx} + \varepsilon_{yy})$ .
- For a plane strain setup the assumption is  $\varepsilon_{zz} = 0$ .

**EvaluateStrain()**

```
[EPS_XX,EPS_YY,EPS_XY] = EvaluateStrain(MESH,U1,U2)
```

evaluate the normal and shearing strains at the nodes

parameters:

- \* MESH is the mesh describing the domain
- \* U1 vector with the values of the x-displacements at the nodes
- \* U2 vector with the values of the y-displacements at the nodes

return values:

- \* EPS\_XX values of normal strain in x direction at the nodes
- \* EPS\_YY values of normal strain in y direction at the nodes
- \* EPS\_XY values of shearing strain at the nodes

Given the displacements  $\vec{u}_1$  and  $\vec{u}_2$  with the corresponding mesh use the function EvaluateStress() to determine the normal and shearing stresses at the nodes. Since Hooke's law is used to determine the stresses the material parameters  $E$  and  $\nu$  have to be provided. Use the same function for plane stress and plane strain problems, but with different arguments.

- For a plane stress setup ask for three return arguments  $\sigma_x$ ,  $\sigma_y$  and  $\tau_{xy}$ . All other components of the stress tensor are zero, based on the plane stress assumption.
- For a plane strain setup ask for four return arguments  $\sigma_x$ ,  $\sigma_y$ ,  $\tau_{xy}$  and  $\sigma_z$ . Based on Hooke's law the other shearing stresses are given by  $\tau_{xz} = \tau_{yz} = 0$ .

**EvaluateStress()**

```
[SIGMA_X,SIGMA_Y,TAU_XY,SIGMA_Z] = EvaluateStress(MESH,U1,U2,E,NU)
```

evaluate the normal and shearing stresses at the nodes, using Hooke's law for plane stress or plane strain setups

- \* [SIGMA\_X,SIGMA\_Y,TAU\_XY] = EvaluateStress(MESH,U1,U2,E,NU)  
with three return arguments assumes a plane stress situation
- \* [SIGMA\_X,SIGMA\_Y,TAU\_XY,SIGMA\_Z] = EvaluateStress(MESH,U1,U2,E,NU)  
with four return arguments assumes a plane strain situation

parameters:

- \* MESH is the mesh describing the domain
- \* U1 vector with the values of the x-displacements at the nodes
- \* U2 vector with the values of the y-displacements at the nodes
- \* E Young's modulus of elasticity, either as constant or as string with the function name
- \* NU Young's modulus of elasticity, either as constant or as string with the function name

return values:

- \* SIGMA\_X values of normal stress in x direction at the nodes
- \* SIGMA\_Y values of normal stress in y direction at the nodes
- \* TAU\_XY values of shearing strain at the nodes
- \* SIGMA\_Z values of normal stress in z direction at the nodes,  
only for plane strain situations

#### 4.13.3 Evaluation of stress expressions: EvaluateVonMises(), EvaluatePrincipalStress() and EvaluateTresca()

There are many expressions used for post processing elasticity problems. The following commands allow to evaluate a few of them at the nodes of the given mesh.

The **von Mises stress**  $\sigma_M$  is useful as an indicator for material failure for ductile materials, e.g. most metals. It is one of the most common output expressions used for mechanical FEM simulations. It is a measure for the differences of the principals stresses, since

$$\sigma_M^2 = \frac{1}{2} ((\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_3 - \sigma_1)^2) .$$

- For a plane stress setup use  $\sigma_z = \tau_{xz} = \tau_{yz} = 0$  to simplify the expression for the von Mises stress.

$$\begin{aligned} \sigma_M^2 &= \frac{1}{2} ((\sigma_x - \sigma_y)^2 + (\sigma_y - \sigma_z)^2 + (\sigma_z - \sigma_x)^2) + 3 (\tau_{xy}^2 + \tau_{yz}^2 + \tau_{zx}^2) \\ &= \frac{1}{2} ((\sigma_x - \sigma_y)^2 + \sigma_y^2 + \sigma_x^2) + 3 \tau_{xy}^2 = \sigma_x^2 + \sigma_y^2 - \sigma_x \sigma_y + 3 \tau_{xy}^2 \end{aligned}$$

- For a plane strain setup use  $\tau_{xz} = \tau_{yz} = 0$  to simplify the expression for the von Mises stress slightly.

$$\sigma_M^2 = \frac{1}{2} ((\sigma_x - \sigma_y)^2 + (\sigma_y - \sigma_z)^2 + (\sigma_z - \sigma_x)^2) + 3 \tau_{xy}^2$$

Select the plane stress or plane strain setup by calling the function EvaluateVonMises() with three or four input arguments.

- If the three arguments  $\sigma_x$ ,  $\sigma_y$  and  $\tau_{xy}$  are given, then a plane stress situation is used.
- If the four arguments  $\sigma_x$ ,  $\sigma_y$ ,  $\tau_{xy}$  and  $\sigma_z$  are given, then a plane strain situation is used.

#### EvaluateVonMises()

```
VONMISES = EvaluateVonMises(SIGMA_X, SIGMA_Y, TAU_XY, SIGMA_Z)
```

evaluate the von Mises stress at the nodes

- \* VONMISES = EvaluateVonMises(SIGMA\_X, SIGMA\_Y, TAU\_XY)  
with three input arguments assumes a plane stress situation
  - \* VONMISES = EvaluateVonMises(SIGMA\_X, SIGMA\_Y, TAU\_XY, SIGMA\_Z)  
with four input arguments assumes a plane strain situation
- parameters:
- \* SIGMA\_X values of normal stress in x direction at the nodes
  - \* SIGMA\_Y values of normal stress in y direction at the nodes
  - \* TAU\_XY values of shearing strain at the nodes
  - \* SIGMA\_Z values of normal stress in z direction at the nodes,  
only for plane strain situations

return values:

- \* VONMISES values of the von Mises stress at the nodes

By selecting an appropriate (local) coordinate system the shearing stresses will vanish and only the three **principal stresses**  $\sigma_1$ ,  $\sigma_2$  and  $\sigma_3$  are used. They are the eigenvalues of the stress matrix

$$\begin{bmatrix} \sigma_x & \tau_{xy} & \tau_{xz} \\ \tau_{xy} & \sigma_y & \tau_{yz} \\ \tau_{xz} & \tau_{yz} & \sigma_z \end{bmatrix}.$$

- For a plane stress problem determine the principal stresses  $\sigma_1$  and  $\sigma_2$  by solving a quadratic equation.

$$0 = \det \begin{bmatrix} \sigma_x - \sigma & \tau_{xy} \\ \tau_{xy} & \sigma_y - \sigma \end{bmatrix} = \sigma^2 - \sigma(\sigma_x + \sigma_y) + \sigma_x \sigma_y - \tau_{xy}^2$$

$$\sigma_{1,2} = \frac{1}{2} \left( (\sigma_x + \sigma_y) \pm \sqrt{(\sigma_x + \sigma_y)^2 - 4\sigma_x \sigma_y + 4\tau_{xy}^2} \right)$$

$$= \frac{1}{2} \left( (\sigma_x + \sigma_y) \pm \sqrt{(\sigma_x - \sigma_y)^2 + 4\tau_{xy}^2} \right)$$

For a plane stress problem the third principal stress is given by  $\sigma_3 = 0$ .

- For a plane strain setup the first two of the above principal stresses remain unchanged. The values of  $\sigma_3$  are determined by

$$\sigma_3 = \sigma_z = \frac{E\nu(\varepsilon_{xx} + \varepsilon_{yy})}{(1+\nu)(1-2\nu)} = \nu(\sigma_1 + \sigma_2) = \nu(\sigma_x + \sigma_y).$$

and returned by the function `EvaluateStress()`.

Thus there is no need for code to compute the values of  $\sigma_3$ .

#### EvaluatePrincipalStress()

```
[SIGMA_1,SIGMA_2] = EvaluatePrincipalStress(SIGMA_X,SIGMA_Y,TAU_XY)
```

evaluate the first two principal stresses at the nodes

parameters:

- \* SIGMA\_X values of normal stress in x direction at the nodes
- \* SIGMA\_Y values of normal stress in y direction at the nodes
- \* TAU\_XY values of shearing strain at the nodes

return values:

- \* SIGMA\_1 first principal stress at the nodes
- \* SIGMA\_2 second principal stress at the nodes

The **Tresca stress** is another indicator for material failure for ductile materials. The Tresca stress measures the differences of the principal stresses and is given by

$$\sigma_T = \max\{|\sigma_1 - \sigma_2|, |\sigma_2 - \sigma_3|, |\sigma_3 - \sigma_1|\}.$$

Select the plane stress or plane strain setup by calling the function `EvaluateTresca()` with three or four input arguments.

- If the three input arguments  $\sigma_x$ ,  $\sigma_y$  and  $\tau_{xy}$  are given, then a plane stress situation is used.
- If the four input arguments  $\sigma_x$ ,  $\sigma_y$ ,  $\tau_{xy}$  and  $\sigma_z$  are given, then a plane strain situation is used.

**EvaluateTresca()**

```
TRESCA = EvaluateTresca(SIGMA_X, SIGMA_Y, TAU_XY, SIGMA_Z)
```

evaluate the Tresca stress at the nodes

- \* `TRESCA = EvaluateTresca(SIGMA_X, SIGMA_Y, TAU_XY)`  
with three input arguments assumes a plane stress situation
- \* `TRESCA = EvaluateTresca(SIGMA_X, SIGMA_Y, TAU_XY, SIGMA_Z)`  
with four input arguments assumes a plane strain situation

parameters:

- \* `SIGMA_X` values of normal stress in x direction at the nodes
- \* `SIGMA_Y` values of normal stress in y direction at the nodes
- \* `TAU_XY` values of shearing strain at the nodes
- \* `SIGMA_Z` values of normal stress in z direction at the nodes,  
only for plane strain situations

return values:

- \* `TRESCA` Tresca stress at the nodes

**4.14 Solving axisymmetric elasticity problems, AxiStress()**

By minimizing the energy given by equation (25) on page 19 an axisymmetric elasticity problem can be solved. The construction of the elements is shown in Section 8.6 starting on page 190. The commands to solve axially symmetric problems and analyze their solutions are shown in Table 8.

**4.14.1 Evaluating axisymmetric solutions**

To determine the radial displacement  $u_r$  and the  $z$ -displacement  $u_z$  use the command `AxiStress()`.

**AxiStress()**

```
[UR, UZ] = AxiStress(MESH, E, NU, F, GD, GN)
```

solve an axisymmetric elasticity problem

```
plane stress equation      in domain
          u = gD           on Gamma_1
force density = gN         on Gamma_2
force density = 0          on Gamma_3
```

parameters:

- \* `MESH` is the mesh describing the domain and the boundary types
- \* `E, NU` Young's modulus and Poisson's ratio for the material
- \* `F = {F1, F2}` a cell array with the two components of the volume forces
- \* `GD = {GD1, GD2}` a cell array with the two components of the prescribed displacements on the boundary section `Gamma_1`
- \* `GN = {GN1, GN2}` a cell array with the two components of the surface forces on the boundary section `Gamma_2`
- \* Any constant function can be given by its scalar value
- \* Any function can be given by a string with the function name
- \* The functions `E`, `NU`, `F1` and `F2` may also be given as vectors with the values of the function at the Gauss points

return values

- \* `UR` vector with the values of the  $r$ -displacement at the nodes

\* UZ vector with the values of the z-displacement at the nodes

command	purpose
<code>AxiStress()</code>	solve an axially symmetric problem
<code>AxiStressEquationM()</code>	script to generate the equations, order 1
<code>AxiStressEquationQuadM()</code>	script to generate the equations, order 2
<code>AxiStressEquationCubicM()</code>	script to generate the equations, order 3
<code>EvaluateStrainAxi()</code>	given the displacement evaluate the strains at the nodes
<code>EvaluateStressAxi()</code>	given the displacement evaluate the stresses at the nodes
<code>EvaluateVonMisesAxi()</code>	evaluate the von Mises stress at the nodes
<code>EvaluatePrincipalStressAxi()</code>	evaluate the three principal stresses at the nodes
<code>EvaluateTrescaAxi()</code>	evaluate the Tresca stress at the nodes

Table 8: Commands to solve and examine axially symmetric elasticity problems

#### 4.14.2 Evaluation of strains and stress for axisymmetric problems

Based on Section 2.14 the strains for an axisymmetric problem with displacements  $u_r(r, z)$  and  $u_z(r, z)$  in the plane  $y = 0$  are given by

$$\begin{bmatrix} \varepsilon_{xx} & \varepsilon_{xy} & \varepsilon_{xz} \\ \varepsilon_{xy} & \varepsilon_{yy} & \varepsilon_{yz} \\ \varepsilon_{xz} & \varepsilon_{yz} & \varepsilon_{zz} \end{bmatrix} = \begin{bmatrix} \frac{\partial u_r}{\partial r} & 0 & \frac{1}{2}(\frac{\partial u_r}{\partial z} + \frac{\partial u_z}{\partial r}) \\ 0 & \frac{1}{r} u_r & 0 \\ \frac{1}{2}(\frac{\partial u_r}{\partial z} + \frac{\partial u_z}{\partial r}) & 0 & \frac{\partial u_z}{\partial z} \end{bmatrix}.$$

Using Hooke's law this leads to the stresses

$$\begin{aligned} \begin{pmatrix} \sigma_x \\ \sigma_y \\ \sigma_z \end{pmatrix} &= \frac{E}{(1+\nu)(1-2\nu)} \begin{bmatrix} 1-\nu & \nu & \nu \\ \nu & 1-\nu & \nu \\ \nu & \nu & 1-\nu \end{bmatrix} \cdot \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{zz} \end{pmatrix} \\ &= \frac{E}{(1+\nu)(1-2\nu)} \begin{bmatrix} 1-\nu & \nu & \nu \\ \nu & 1-\nu & \nu \\ \nu & \nu & 1-\nu \end{bmatrix} \cdot \begin{pmatrix} \frac{\partial u_r}{\partial r} \\ \frac{1}{r} u_r \\ \frac{\partial u_z}{\partial z} \end{pmatrix} \\ \tau_{zx} &= \frac{E}{1+\nu} \frac{1}{2} \left( \frac{\partial u_r}{\partial z} + \frac{\partial u_z}{\partial r} \right) \end{aligned}$$

Based on this the stresses and strains for axisymmetric problems can be evaluated. The codes are similar to the corresponding codes for plane elasticity problems, see Section 4.13.1 starting on page 78.

• **EvaluateStrainAxi()**  
`[EPS_XX,EPS_YY,EPS_ZZ,EPS_XZ] = EvaluateStrainAxi(MESH,UR,UZ)`

evaluate the normal and shearing strains at the nodes

parameters:

- \* MESH is the mesh describing the domain
- \* UR vector with the values of the r-displacements at the nodes
- \* UZ vector with the values of the z-displacements at the nodes

return values:

- \* EPS\_XX values of normal strain in x direction at the nodes
- \* EPS\_YY values of normal strain in y direction at the nodes
- \* EPS\_ZZ values of normal strain in z direction at the nodes
- \* EPS\_XZ values of shearing strain at the nodes

#### EvaluateStressAxi()

```
[SIGMA_X, SIGMA_Y, SIGMA_Z, TAU_XZ] = EvaluateStressAxi (MESH, UR, UZ, E, NU)
```

evaluate the normal and shearing stresses at the nodes, using Hooke's law

parameters:

- \* MESH is the mesh describing the domain
- \* UR vector with the values of the r-displacements at the nodes
- \* UZ vector with the values of the z-displacements at the nodes
- \* E Young's modulus of elasticity, either as constant or as string with the function name
- \* NU Young's modulus of elasticity, either as constant or as string with the function name

return values:

- \* SIGMA\_X values of normal stress in x direction at the nodes
- \* SIGMA\_Y values of normal stress in y direction at the nodes
- \* SIGMA\_Z values of normal stress in z direction at the nodes
- \* TAU\_XZ values of shearing strain at the nodes

#### EvaluateVonMisesAxi()

```
VONMISES = EvaluateVonMisesAxi (SIGMA_X, SIGMA_Y, SIGMA_Z, TAU_XZ)
```

evaluate the von Mises stress at the nodes

parameters:

- \* SIGMA\_X values of normal stress in x direction at the nodes
- \* SIGMA\_Y values of normal stress in y direction at the nodes
- \* SIGMA\_Z values of normal stress in z direction at the nodes
- \* TAU\_XZ values of shearing strain at the nodes

return values:

- \* VONMISES values of the von Mises stress at the nodes

- Based on the stress matrix

$$\begin{bmatrix} \sigma_x & 0 & \tau_{xz} \\ 0 & \sigma_y & 0 \\ \tau_{xz} & 0 & \sigma_z \end{bmatrix}$$

two principal stresses are given by solving a quadratic equation.

$$0 = \det \begin{bmatrix} \sigma_x - \sigma & \tau_{xz} \\ \tau_{xz} & \sigma_z - \sigma \end{bmatrix} = \sigma^2 - \sigma(\sigma_x + \sigma_z) + \sigma_x \sigma_z - \tau_{xz}^2$$

$$\begin{aligned}\sigma_{1,2} &= \frac{1}{2} \left( (\sigma_x + \sigma_z) \pm \sqrt{(\sigma_x + \sigma_z)^2 - 4\sigma_x\sigma_z + 4\tau_{xz}^2} \right) \\ &= \frac{1}{2} \left( (\sigma_x + \sigma_z) \pm \sqrt{(\sigma_x - \sigma_z)^2 + 4\tau_{xz}^2} \right)\end{aligned}$$

The third principal stress is given by  $\sigma_3 = \sigma_y$ .

#### EvaluatePrincipalStressAxi()

```
[SIGMA_1, SIGMA_2] = EvaluatePrincipalStressAxi(SIGMA_X, SIGMA_Z, TAU_XZ)
```

evaluate two principal stresses at the nodes

parameters:

- \* SIGMA\_X values of normal stress in x direction at the nodes
- \* SIGMA\_Z values of normal stress in z direction at the nodes
- \* TAU\_XZ values of shearing strain at the nodes

return values:

- \* SIGMA\_1 first principal stress at the nodes
- \* SIGMA\_2 second principal stress at the nodes

#### EvaluateTrescaAxi()

```
TRESCA = EvaluateTrescaAxi(SIGMA_X, SIGMA_Y, SIGMA_Z, TAU_XZ)
```

evaluate the Tresca stress at the nodes

parameters:

- \* SIGMA\_X values of normal stress in x direction at the nodes
- \* SIGMA\_Y values of normal stress in y direction at the nodes
- \* SIGMA\_Z values of normal stress in z direction at the nodes
- \* TAU\_XZ values of shearing strain at the nodes

return values:

- \* TRESCA Tresca stress at the nodes

## 4.15 Internal commands in FEMoctave

In this section a few internal commands are documented. Usually these commands are not used when solving boundary value problems or elasticity problems. But they contain the essential codes to generate the matrices and vectors required to solve the problems. The coding is based on the algorithms shown in Section 6, starting on page 126. They can also be useful to illustrate the essential steps of finite element algorithms, e.g. individual element stiffness matrices.

### 4.15.1 Linear elements: FEMEquation.cc and FEMEquation.m

This is the fundamental function that transforms a BVP to a system of linear equations. First order triangular elements are used. To speed it up it is written in C++, leading to the file FEMEquation.oct.

#### FEMEquation()

```
[A, B] = FEMEquation(MESH, A, B0, BX, BY, F, GD, GN1, GN2)
```

sets up the system of linear equations for a numerical solution of a PDE using a triangular mesh with elements of order 1

$$-\text{div}(a \cdot \text{grad } u - u \cdot (b_x, b_y)) + b_0 \cdot u = f \quad \text{in domain}$$

$$u = gD \quad \text{on Dirichlet boundary}$$

$$n*(a*\text{grad } u - u*(bx,by)) = gN1+g2N*u \quad \text{on Neumann boundary}$$

parameters:

- \* MESH triangular mesh of order 1 describing the domain and the boundary types
- \* A, B0, BX, BY, F, GD, GN1, GN2 are the coefficients and functions describing the PDE.  
Any constant function can be given by its scalar value.  
The functions A, B0, BX, BY and F may also be given as vectors with the values of the function at the Gauss points.

return values:

- \* A, B: matrix and vector for the linear system to be solved,  $A*u-B=0$

The script function `FEMEquation.m` performs the same task and is easier to read and understand, but considerably slower than the compiled code.

#### 4.15.2 Quadratic elements: `FEMEquationQuad.cc` and `FEMEquationQuad.m`

This is the fundamental function that transforms a BVP to a system of linear equations. Second order triangular elements are used. To speed it up it is written in C++.

##### FEMEquationQuad()

`[A, B] = FEMEquationQuad(MESH, A, B0, BX, BY, F, GD, GN1, GN2)`  
sets up the system of linear equations for a numerical solution of a PDE using a triangular mesh with elements of order 2

$$-\text{div}(a*\text{grad } u - u*(bx,by)) + b0*u = f \quad \text{in domain}$$

$$u = gD \quad \text{on Dirichlet boundary}$$

$$n*(a*\text{grad } u - u*(bx,by)) = gN1+g2N*u \quad \text{on Neumann boundary}$$

parameters:

- \* MESH triangular mesh of order 2 describing the domain and the boundary types
- \* A, B0, BX, BY, F, GD, GN1, GN2 are the coefficients and functions describing the PDE.  
Any constant function can be given by its scalar value.  
The functions A, B0, BX, BY and F may also be given as vectors with the values of the function at the Gauss points.

return values:

- \* A, B: matrix and vector for the linear system to be solved,  $A*u-B=0$

The script function `FEMEquationQuad.m` performs the same task and is easier to read and understand, but considerably slower than the compiled code.

#### 4.15.3 Cubic elements: `FEMEquationCubic.cc` and `FEMEquationCubic.m`

These two commands are very similar to the above section, but use triangular elements of order 3.

#### 4.15.4 Effect of right hand side for dynamic problems: `FEMInterpolWeight()`

For the time stepping in parabolic and hyperbolic problems many systems of linear equations have to be solved using the RHS  $f(t, x, y)$  for different values of the time  $t$ . Thus a function to keep track of the influence of  $f$  is useful, `FEMInterpolWeight()`. This function returns a sparse matrix `wMat` such that the RHS of the

system to be solved is given by  $\mathbf{wMat} \vec{f}$ .

---

#### FEMInterpolWeight()

---

```
WMAT = FEMInterpolWeight(FEMMESH,WFUNC)
```

create the matrix to determine the contribution of  $w*f$  to a IBVP or BVP  
the contribution of  $w*f$  is the determined by  $\mathbf{wMat}*f$ , where  $f$  is the  
vector with the values at the "free" nodes

$$\begin{aligned} -\text{div}(a*\text{grad } u) + b_0*u &= w*f && \text{in domain} \\ u &= g_D && \text{on Dirichlet boundary} \\ n*(a*\text{grad } u) &= g_{N1}+g_{N2}*u && \text{on Neumann boundary} \end{aligned}$$

parameters:

- \* MESH is the mesh describing the domain and the boundary types
- \* WFUNC is the weight function  $w$   
It may be given as a function name, a vector with the values  
at the Gauss points or as a scalar value

return value

- \* WMAT is the sparse weight matrix
- 

This function is used in `IBVP2D()`, `I2BVP2D()` and `IBVP2Dsym()`.

#### 4.15.5 Effect of the Dirichlet values: FEMInterpolBoundaryWeight()

If the same system has to be solved for many different Dirichlet values  $g_D$  on the boundary, one can generate the equation once and the only recompute the changes for different  $g_D$ .

---

#### FEMInterpolBoundaryWeight()

---

```
WMAT = FEMInterpolBoundaryWeight(FEMMESH,A,B0)
```

create the matrix to determine the contribution of  $g_D$  to a IBVP or BVP  
the contribution of  $g_D$  is the determined by  $\mathbf{wMat}*g_D$ , where  $g_D$  is  
the vector with the values at the Dirichlet nodes

$$\begin{aligned} -\text{div}(a*\text{grad } u) + b_0*u &= f && \text{in domain} \\ u &= g_D && \text{on Dirichlet boundary} \\ n*(a*\text{grad } u) &= g_{N1}+g_{N2}*u && \text{on Neumann boundary} \end{aligned}$$

parameters:

- \* FEMMESH is the mesh describing the domain and the boundary types.
- \* A,B0 are the coefficients and functions describing the PDE.

return value:

- \* WMAT is the sparse weight matrix
- 

#### 4.15.6 Determine a few small eigenvalues: eigSmall()

In the function `BVP2Deig()` a few small eigenvalues are determined with the help of the wrapper `eigSmall()` for the Octave function `eigs()`. Usually generalized eigenvalues are used in `FEMoctave`.

---

#### eigSmall()

---

```
[Lambda, {Ev,err}] = eigSmall(A,V,tol)
    solve  $A*Ev = Ev*diag(Lambda)$  standard eigenvalue problem
```

```
[Lambda, {Ev, err}] = eigSmall(A, B, V, tol)
    solve  $A \cdot Ev = B \cdot Ev \cdot \text{diag}(Lambda)$  generalized eigenvalue problem

A is a (sparse) mxm matrix
B is a (sparse) mxm matrix
V is either n, the number of desired eigenvalues
  or a mxn matrix, the initial eigenvectors for the iteration
tol is the relative error, used as the stopping criterion

X is a column vector with the eigenvalues
EV is a matrix whose columns represent normalized eigenvectors
err is a vector with the a posteriori error estimates for the eigenvalues

this implementation is based on using eigs()
```

#### 4.15.7 Generating the equations for elasticity problems

The codes `PStressEquationM.m`, `PStressEquationQuadM.m` and `PStressEquationCubicM.m` generate the linear system of equations to be solved for plane stress and plane strain problems. They are used in `PlaneStress()` and `PlaneStrain()`. The *Octave* codes are based on the algorithms in Section 8 (starting on page 181) and easier to read and understand than C++ code, which is not written (yet).

##### **PStressEquationM.m**

```
[gMat, gVec] = PStressEquationM(Mesh, EFunc, nuFunc, fFunc, gDFunc, gNFunc)
setup the equation for a plane stress problem with linear elements
```

##### **PStressEquationQuadM.m**

```
[gMat, gVec] = PStressEquationQuadM(Mesh, EFunc, nuFunc, fFunc, gDFunc, gNFunc)
setup the equation for a plane stress problem with quadratic elements
```

##### **PStressEquationCubicM.m**

```
[gMat, gVec] = PStressEquationCubicM(Mesh, EFunc, nuFunc, fFunc, gDFunc, gNFunc)
setup the equation for a plane stress problem with cubic elements
```

For axially symmetric problems similar commands are used, i.e. the script files `AxiStressEquationM.m`, `AxiStressEquationQuadM.m` and `AxiStressEquationCubicM.m`.

##### **AxiStressEquationM.m**

```
[gMat, gVec] = AxiStressEquationM(Mesh, EFunc, nuFunc, fFunc, gDFunc, gNFunc)
%% [gMat, gVec] = AxiStressEquationM(Mesh, EFunc, nuFunc, fFunc, gDFunc, gNFunc)
%%
%% setup the equation for an axisymmetric problem with linear elements
```

##### **AxiStressEquationQuadM.m**

```
[gMat, gVec] = AxiStressEquationQuadM(Mesh, EFunc, nuFunc, fFunc, gDFunc, gNFunc)
%% [gMat, gVec] = AxiStressEquationQuadM(Mesh, EFunc, nuFunc, fFunc, gDFunc, gNFunc)
%%
%% setup the equation for an axisymmetric problem with quadratic elements
```

##### **AxiStressEquationCubicM.m**

```
[gMat, gVec] = AxiStressEquationCubicM(Mesh, EFunc, nuFunc, fFunc, gDFunc, gNFunc)
%% [gMat, gVec] = AxiStressEquationCubicM(Mesh, EFunc, nuFunc, fFunc, gDFunc, gNFunc)
%%
%% setup the equation for an axisymmetric problem with cubic elements
```

The *Octave* codes might be replaced by a compiled code for speed reasons.

#### 4.16 External programs

To construct nonuniform triangular meshes FEMoctave uses an external program.

- **Triangle** to generate a good mesh. The source code is given in FEMoctave. Find documentation on the web page [www.cs.cmu.edu/~quake/triangle.html](http://www.cs.cmu.edu/~quake/triangle.html).
- **CuthillMcKee** to obtain a good numbering. Not necessary any more, since the sparse factorizations do a better job.
- **tricountour.m** is a code by Duane Hanselman available at the Mathworks web site `matlabcentral`. It was used by previous versions of the function `FEMtricountour()`. The current version of FEMoctave contains a simple implementation of `tricountour.m`. Neither code is able to generate good labels for the contours.

## 5 Tools for Didactical Purposes

In this section a few effects of FEM are illustrated. This could be useful to teach a class on the FEM.

- 5.1 The convergence of the solutions as  $h \rightarrow 0$  is examined, using an example. Find the orders of convergence for linear, quadratic and cubic elements.
- 5.2 Some element stiffness matrices are examined. A path from FEM to a finite difference approximation is shown.
- 5.3 The behavior of FEM solutions within an element is examined. Find a visualization of the accuracy of linear, quadratics or cubic element.
- 5.4 The number of nodes, triangles and their effect on the sparsity of the global stiffness matrix is examined.
- 5.5 Elements of order 1, 2 or three are used to solve the same problem. The sizes of the resulting matrices and error are examined.
- 5.6 A few examples illustrate the second order elements are not  $C^1$ -conforming, i.e. the first derivative might jump across borders of elements.
- 5.7 The effect of superconvergence is illustrated using a 1D boundary value problem.
- 5.8 The stability of four different time steppers for dynamic problems is visualized.
- 5.9 The effect of shear-locking for elasticity problems, solved with the help of linear elements, is explained and visualized. It is shown why quadratic elements do not suffer from possible shear-locking.
- 5.10 The standard problem of a bending Euler beam is solved with elements of order 1, 2 and 3 and on different meshes. The results are compared.
- 5.11 The (bad) effect of missing boundary constraints for elasticity problems is explained with the help of an example.

### 5.1 Observe the convergence of the error as $h \rightarrow 0$

Consider the unit square  $\Omega = [0, 1] \times [0, 1]$ . One can verify that the function  $u_e(x, y) = \sin(x) \cdot \sin(y)$  is the exact solution of the boundary value problem

$$\begin{aligned} -\nabla \cdot \nabla u &= -2 \sin(x) \cdot \sin(y) && \text{for } 0 \leq x, y \leq 1 \\ \frac{\partial u(x,1)}{\partial y} &= -\sin(x) \cdot \cos(1) && \text{for } 0 \leq x \leq 1 \text{ and } y = 1 \\ u(x, y) &= u_e(x, y) && \text{on the other sections of the boundary} \end{aligned} .$$

Let  $h > 0$  be the typical length of a side of a triangle. For second order elements  $2h$  is used and for third order elements  $3h$ , such that the computational effort is comparable to first order elements. Nonuniform meshes are used, to avoid superconvergence. By choosing different values of  $h$  one should observe smaller errors for smaller values of  $h$ . The sizes of the matrices vary (approximately) from  $50 \times 50$  to  $58'000 \times 58'000$ . The error is measured by computing the  $L_2$  norms of the difference of the exact and approximate solutions, for the values of the functions and its partial derivative with respect to  $y$ . These are the expressions used in the theoretical convergence estimates stated in Section 6.7. A double logarithmic plot leads to Figure 35.

- For linear elements:

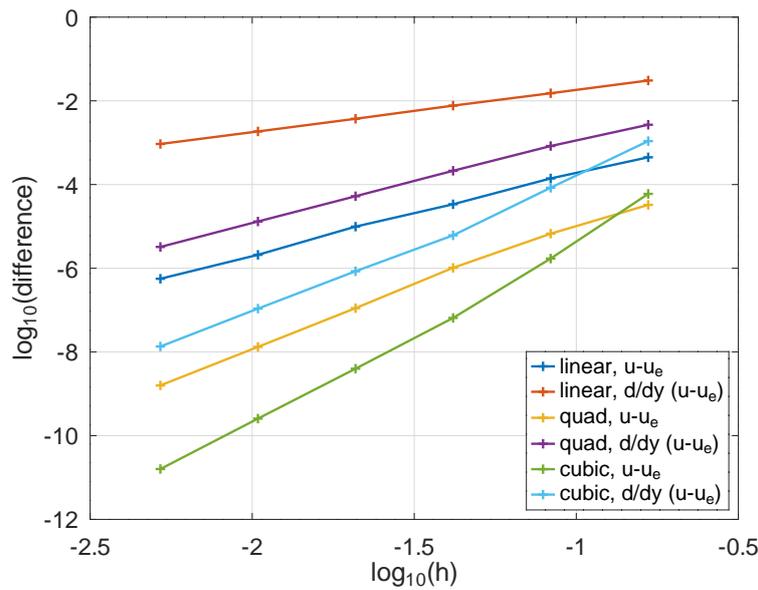


Figure 35: Convergence results for linear, quadratic and cubic elements

- The slope of the curve for the absolute values of  $u(x, y) - u_e(x, y)$  is approximately 2 and thus conclude that the error is proportional to  $h^2$ .
- The slope of the curve for the absolute values of  $\frac{\partial}{\partial y} (u(x, y) - u_e(x, y))$  is approximately 1 and thus conclude that the error of the gradient is proportional to  $h$ .
- For quadratic elements:
  - The slope of the curve for the absolute values of  $u(x, y) - u_e(x, y)$  is approximately 3 and thus conclude that the error is proportional to  $h^3$ .
  - The slope of the curve for the absolute values of  $\frac{\partial}{\partial y} (u(x, y) - u_e(x, y))$  is approximately 2 and thus conclude that the error of the gradient is proportional to  $h^2$ .
- For cubic elements:
  - The slope of the curve for the absolute values of  $u(x, y) - u_e(x, y)$  is approximately 4 and thus conclude that the error is proportional to  $h^4$ .
  - The slope of the curve for the absolute values of  $\frac{\partial}{\partial y} (u(x, y) - u_e(x, y))$  is approximately 3 and thus conclude that the error of the gradient is proportional to  $h^3$ .

These observations confirm the theoretical error estimates in Section 6.7 on page 161. It is rather obvious from Figure 35 that higher order elements generate more accurate solutions for a comparable computational effort.

#### TestConvergence.m

```

a = 1; b0 = 0; gN2 = 0; N = 6;
Npow = 6; % use Npow = 6 for final run

function res = u_exact(xy, dummy) res = sin(xy(:,1)).*sin(xy(:,2)); endfunction
function res = f(xy, dummy) res = 2*sin(xy(:,1)).*sin(xy(:,2)); endfunction
function res = u_y(xy) res = sin(xy(:,1)).*cos(xy(:,2)); endfunction

```

```

for ii = 1:Npaw
    Ni = N*2^(ii-1);    h(ii) = 1/(Ni); area = 0.5/(Ni)^2;
    FEMmesh1 = CreateMeshTriangle('TestConvergence',[0 0 -1;1 0 -1;1 1 -2;0 1 -1],area);
    FEMmesh2 = CreateMeshTriangle('TestConvergence',[0 0 -1;1 0 -1;1 1 -2;0 1 -1],4*area);
    FEMmesh2 = MeshUpgrade(FEMmesh2,'quadratic');
    FEMmesh3 = CreateMeshTriangle('TestConvergence',[0 0 -1;1 0 -1;1 1 -2;0 1 -1],9*area);
    FEMmesh3 = MeshUpgrade(FEMmesh3,'cubic');

    %% solve with first order elements
    u1 = BVP2Dsym(FEMmesh1,a,b0,'f','u_exact','u_y',gN2);
    Difference(ii) = sqrt(FEMIntegrate(FEMmesh1,(u1-u_exact(FEMmesh1.nodes)).^2));
    [ux,uy] = FEMEvaluateGradient(FEMmesh1,u1);
    DifferenceUy(ii) = sqrt(FEMIntegrate(FEMmesh1,(uy-u_y(FEMmesh1.nodes)).^2));

    %% now for second order elements
    u2 = BVP2Dsym(FEMmesh2,a,b0,'f','u_exact','u_y',gN2);
    DifferenceQ(ii) = sqrt(FEMIntegrate(FEMmesh2,(u2-u_exact(FEMmesh2.nodes)).^2));
    [ux,uy] = FEMEvaluateGradient(FEMmesh2,u2);
    DifferenceUyQ(ii) = sqrt(FEMIntegrate(FEMmesh2,(uy-u_y(FEMmesh2.nodes)).^2));

    %% now for third order elements
    u3 = BVP2Dsym(FEMmesh3,a,b0,'f','u_exact','u_y',gN2);
    DifferenceC(ii) = sqrt(FEMIntegrate(FEMmesh3,(u3-u_exact(FEMmesh3.nodes)).^2));
    [ux,uy] = FEMEvaluateGradient(FEMmesh3,u3);
    DifferenceUyC(ii) = sqrt(FEMIntegrate(FEMmesh3,(uy-u_y(FEMmesh3.nodes)).^2));
endfor
figure(1); plot(log10(h),log10(Difference),'+-',log10(h),log10(DifferenceUy),'+-',
    log10(h),log10(DifferenceQ),'+-',log10(h),log10(DifferenceUyQ),'+-',
    log10(h),log10(DifferenceC),'+-',log10(h),log10(DifferenceUyC),'+-')
    xlabel('log_{10}(h)'); ylabel('log_{10}(difference)')
    legend('linear, u-u_e','linear, d/dy (u-u_e)',
    'quad, u-u_e','quad, d/dy (u-u_e)','cubic, u-u_e','cubic, d/dy (u-u_e)',
    'location','southeast'); xlim([-2.5,-0.5])

```

## 5.2 Some Element Stiffness Matrices

### 5.2.1 Element contributions for equilateral triangles

Generate the trivial mesh consisting of a single equilateral triangle with the help of `CreateMeshTriangle()`. The code in `CreateTriangle.m` generates the mesh and Figure 36.

```

CreateTriangle.m
%% corners of an equilateral triangle
corners = 1*[0,0,-2;1,0,-2;0.5,sqrt(3)/2,-2];
mm = CreateMeshTriangle('one_triangle',corners,max(corners(:).^2))
plot([mm.nodes(:,1);mm.nodes(1,1)], [mm.nodes(:,2);mm.nodes(1,2)], 'o-r',
    mm.GP(:,1),mm.GP(:,2),'b*')
xlabel('x'); ylabel('y'); title('triangle, with Gauss points'); axis equal

```

For the PDE  $-\Delta u = 1$  generate the element stiffness matrix  $\mathbf{A}$  and the element vector  $\vec{f}$  by using the command `FEMEQuation()`.

$$\mathbf{A} = \frac{\sqrt{3}}{6} \begin{bmatrix} +2 & -1 & -1 \\ -1 & +2 & -1 \\ -1 & -1 & +2 \end{bmatrix}$$

$$\vec{b} = \frac{\sqrt{3}}{4 \cdot 3} \begin{pmatrix} -1 \\ -1 \\ -1 \end{pmatrix} = \frac{\text{area of triangle}}{3} \begin{pmatrix} -1 \\ -1 \\ -1 \end{pmatrix}$$

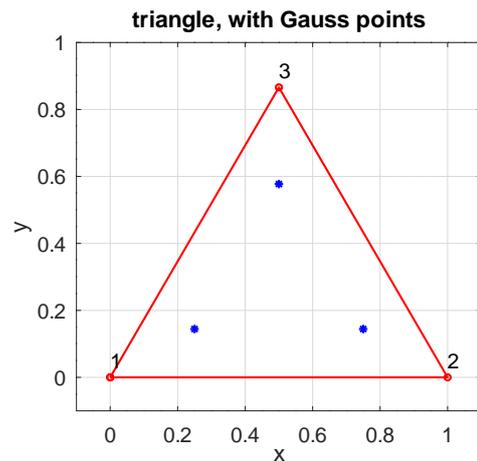


Figure 36: An linear, equilateral triangle, the Gauss integration points and the element stiffness matrix

```
[A,f] = FEMEquation (mm,1,0,0,0,1,0,0,0);
Element_Matrix = full(A)
Element_Vector = f
-->
Element_Matrix =    0.57735   -0.28868   -0.28868
                  -0.28868    0.57735   -0.28868
                  -0.28868   -0.28868    0.57735

Element_Vector =  -0.14434
                  -0.14434
                  -0.14434
```

This result corresponds to the exact result for the element stiffness matrix in Figure 36.

Using the same idea one can examine the contributions of the different terms to the element stiffness matrix. As example consider the term caused by  $b_0 u = 1 u$  in the PDE.

```
B = FEMEquation (mm,0,1,0,0,0,0,0,0);
B = full(B)
-->
B =    0.072169   0.036084   0.036084
      0.036084   0.072169   0.036084
      0.036084   0.036084   0.072169
```

The result confirms

$$\mathbf{B} = \frac{\text{area of triangle}}{12} \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}.$$

Examine a mesh consisting of equilateral triangles, as shown in Figure 37. Then examine the linear equation corresponding to an interior point at  $(x_i, y_i)$ .

- The node is corner of 6 triangles, thus the coefficient  $a_{i,i}$  of the global stiffness matrix consists of 6 contributions found on the diagonal in the element stiffness matrix  $\mathbf{A}$  in Figure 36, i.e.  $a_{i,i} = 6 \frac{+2}{2\sqrt{3}} = \frac{6}{\sqrt{3}}$ .

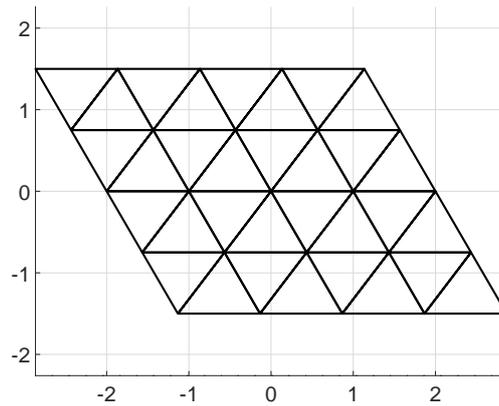


Figure 37: Uniform meshes consisting of equilateral triangles

- If a node at  $(x_j, y_j)$  shares two triangles with  $(x_i, y_i)$  then the entry  $a_{i,j}$  in the global stiffness matrix consists of 2 contributions found off the diagonal in the element stiffness matrix  $\mathbf{A}$  in Figure 36, i.e.  $a_{i,j} = 2 \frac{-1}{2\sqrt{3}} = \frac{-1}{\sqrt{3}}$ .
- If the function  $f$  in  $-\nabla^2 u = f$  is constant, then there will be 6 contributions from the six neighboring triangle. If the length of one side of a triangle equals  $h$ , then the area is  $\frac{\sqrt{3}}{4} h^2$ . Thus find  $b_i = 6 \frac{\text{area of triangle}}{3} (-f) = -\frac{\sqrt{3}}{2} h^2 f$ .

As a result find the equation for the node at  $(x_i, y_i)$ .

$$\frac{1}{h^2} \left( \frac{6}{\sqrt{3}} u(x_i, y_i) - \frac{1}{\sqrt{3}} \sum_{\text{neighbours}} u(x_j, y_j) \right) = +\frac{\sqrt{3}}{2} f$$

$$\frac{1}{h^2} \left( u(x_i, y_i) - \frac{1}{6} \sum_{\text{neighbours}} u(x_j, y_j) \right) = +\frac{1}{4} f$$

This is somewhat similar to a finite difference approximation. For each row of the global stiffness matrix the entry on the diagonal and 6 more will be different from 0.

One can examine second order elements and the resulting element stiffness matrix and vector for quadratic elements for the PDE  $-\Delta u = 1$ . The triangular, equilateral element and the matrix are shown in Figure 38. The vector is given by

$$\vec{b} = \frac{\sqrt{3}}{4 \cdot 3} \begin{pmatrix} 0 \\ 0 \\ 0 \\ -1 \\ -1 \\ -1 \end{pmatrix} = \frac{\text{area of triangle}}{3} \begin{pmatrix} 0 \\ 0 \\ 0 \\ -1 \\ -1 \\ -1 \end{pmatrix}$$

For the global stiffness matrix for the very regular mesh in Figure 37

$$\mathbf{A} = \frac{\sqrt{3}}{18} \begin{bmatrix} 6 & 1 & 1 & 0 & -4 & -4 \\ 1 & 6 & 1 & -4 & 0 & -4 \\ 1 & 1 & 6 & -4 & -4 & 0 \\ 0 & -4 & -4 & 24 & -8 & -8 \\ -4 & 0 & -4 & -8 & 24 & -8 \\ -4 & -4 & 0 & -8 & -8 & 24 \end{bmatrix}$$

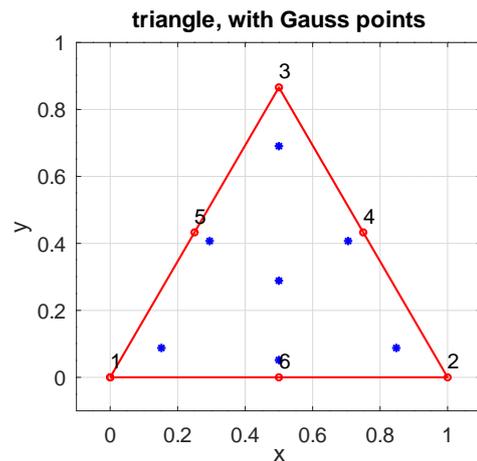


Figure 38: An equilateral, quadratic triangle, the Gauss integration points and the element stiffness matrix

- on each row of the matrix corresponding to a corner of the triangle the entry on the diagonal and 12 more will be different from 0. If the mesh is not as regular even 19 entries on each row might be different from zero.
- on each row of the matrix corresponding to a midpoint of a side of the triangle the entry on the diagonal and 6 more will be different from 0. If the mesh is not as regular even 9 entries on each row might be different from zero.

### 5.2.2 From FEM to a finite difference approximation

Generate the trivial mesh consisting of a single right triangle with the help of `CreateMeshTriangle`. The code in `CreateTriangle.m` generates the mesh and Figure 39. For the PDE  $-\Delta u = 1$  generate the element stiffness matrix  $\mathbf{A}$  and the element vector  $\vec{b}$  by using `FEMEquation()` or `FEMEquationM()`.

#### CreateTriangle.m

```
% corners of a right triangle
corners = 1*[0,0,-2;1,0,-2;0,1,-2];
CreateMeshTriangle('one_triangle',corners,max(corners(:).^2))
mm = ReadMeshTriangle('one_triangle.1');
[A,f] = FEMEquation(mm,1,0,0,0,1,0,0,0); %% using compiled code
Element_Matrix = full(A)
Element_Vector = f
-->
Element_Matrix =  1.00000  -0.50000  -0.50000
                 -0.50000   0.50000   0.00000
                 -0.50000   0.00000   0.50000

Element_Vector = -0.16667
                 -0.16667
                 -0.16667
```

Based on elements of the above type there is a connection of FEM to the finite difference method. Generate a rectangular grid, shown in Figure 40. Examine the PDE  $-\Delta u = \pi$  with Neumann boundary conditions. Use the command `FEMEquation` to generate the matrix  $\mathbf{A}$  and the vector  $\vec{b}$ , then the linear equation  $\mathbf{A} \vec{u} + \vec{b}$  has to

$$\mathbf{A} = \begin{bmatrix} +1 & -0.5 & -0.5 \\ -0.5 & +1 & 0 \\ -0.5 & 0 & +1 \end{bmatrix}, \quad \vec{b} = \frac{1}{6} \begin{pmatrix} -1 \\ -1 \\ -1 \end{pmatrix}$$

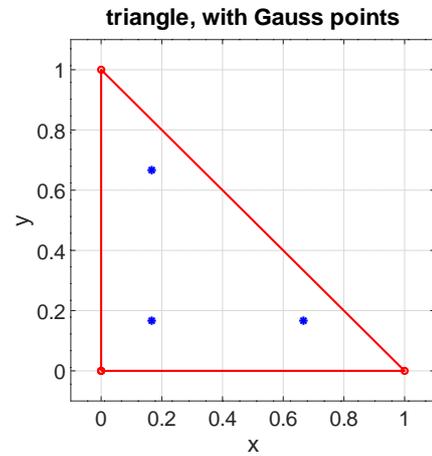
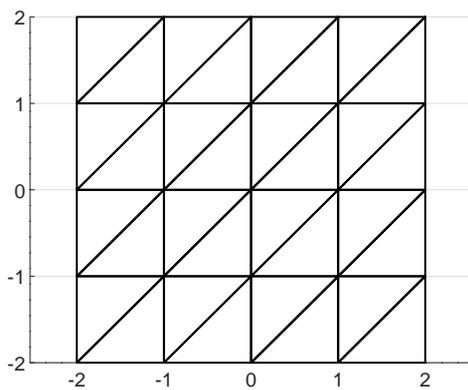
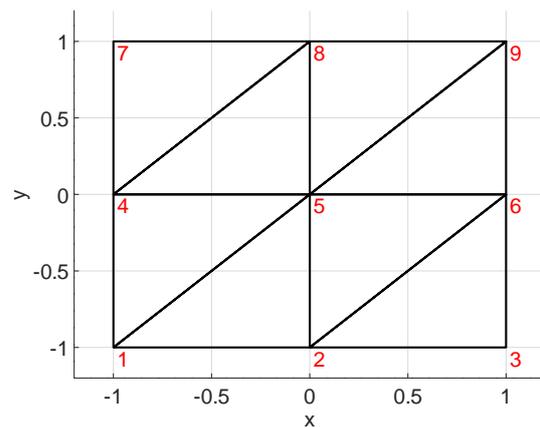


Figure 39: A right triangle, the Gauss integration points and the element stiffness matrix



(a) a section



(b) a small section, numbered

Figure 40: Uniform meshes consisting of rectangular triangles

be solved. The code displays the equation at node 5.

```
x = [-1,0,1];
FEMmesh = CreateMeshRect(x,x,-2,-2,-2,-2)
figure(1); clf
ShowMesh(FEMmesh.nodes,FEMmesh.elem)
xlabel('x'); ylabel('y')
axis(1.2*[-1,1,-1,1]*max(x))
hold on
for kk = 1:length(FEMmesh.nodes)
    text(FEMmesh.nodes(kk,1)+0.02,FEMmesh.nodes(kk,2)-0.07,num2str(kk),'color',[1 0 0])
endfor
hold off

a=1; b0=bx=by= 0; f=pi;
[A,b] = FEMEQuation(FEMmesh,a,b0,bx,by,f,0,0,0);
A5 = full(A(5,:))
b5 = b(5)
-->
A5 = 0 -1 0 -1 4 -1 0 -1 0
b5 = -3.1416
```

The results imply that the equation to be solved is

$$-u_2 - u_4 + 4u_5 - u_6 - u_8 = \pi.$$

Running the code again with  $x = [1, 0, 1] / 2$  will not change  $\mathbf{A}$ , but lead to  $b_5 = -\pi 4$ . Thus for a width  $h$  of the triangles the equation to be solved is

$$\frac{-u(x-h, y) - u(x, y-h) + 4u(x, y) - u(x+h, y) - u(x, y+h)}{h^2} = f(x, y).$$

This is the usual finite difference approximation of  $-\Delta u = f$ .

One can examine second order elements and the resulting element stiffness matrix and vector for quadratic elements for the PDE  $-\Delta u = 1$ . The element and the matrix are shown in Figure 41. The vector is given by

$$\mathbf{A} = \frac{1}{6} \begin{bmatrix} 6 & 1 & 1 & 0 & -4 & -4 \\ 1 & 3 & 0 & 0 & 0 & -4 \\ 1 & 0 & 3 & 0 & -4 & 0 \\ 0 & 0 & 0 & 16 & -8 & -8 \\ -4 & 0 & -4 & -8 & 16 & 0 \\ -4 & -4 & 0 & -8 & 0 & 16 \end{bmatrix}$$

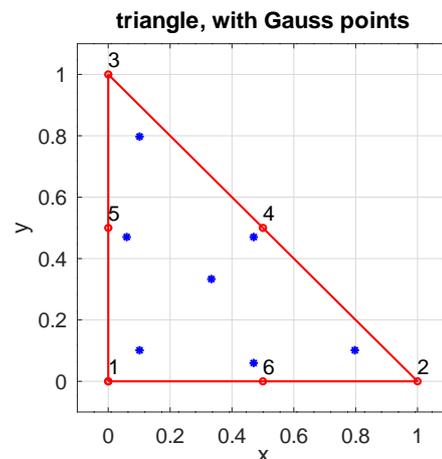


Figure 41: A right angle triangle, the Gauss integration points and the element stiffness matrix

$$\vec{b} = \frac{1}{2 \cdot 3} \begin{pmatrix} 0 \\ 0 \\ 0 \\ -1 \\ -1 \\ -1 \end{pmatrix} = \frac{\text{area of triangle}}{3} \begin{pmatrix} 0 \\ 0 \\ 0 \\ -1 \\ -1 \\ -1 \end{pmatrix}.$$

### 5.2.3 Element stiffness matrices for 1D problems

To examine an ODE of the form (7)

$$-(a(x)u'(x))' + b(x)u'(x) + c(x)u(x) + d(x)f(x) = 0$$

with the help of FEM the element stiffness matrix for each subinterval has to be determined. In the case of constant coefficients the formulas from Section 7.2 (page 169) can be simplified, mainly to examine the structure of the matrices. Four contributions have to be taken into account. Use code similar to

```
h = 1;
[A,M,x] = GenerateFEM1D([0 h],1,0,0,1);
A = full(A)
A_rat = rats(A*h)
M = full(M)
M_rat = rats(M/h)
```

to determine the matrices. With the notations  $u_- = u(-h/2)$ ,  $u_0 = u(0)$  and  $u_+ = u(+h/2)$  obtain

$$I_2 = \int_{-h/2}^{+h/2} u'(x) \phi'(x) dx \approx \left\langle \frac{1}{3h} \begin{bmatrix} +7 & -8 & +1 \\ -8 & +16 & -8 \\ +1 & -8 & +7 \end{bmatrix} \begin{pmatrix} u_- \\ u_0 \\ u_+ \end{pmatrix}, \begin{pmatrix} \phi_- \\ \phi_0 \\ \phi_+ \end{pmatrix} \right\rangle = \langle \mathbf{A}_2 \vec{u}, \vec{\phi} \rangle$$

$$I_1 = \int_{-h/2}^{+h/2} u'(x) \phi(x) dx \approx \left\langle \frac{1}{6} \begin{bmatrix} -3 & +4 & -1 \\ -4 & 0 & +4 \\ +1 & -4 & +3 \end{bmatrix} \begin{pmatrix} u_- \\ u_0 \\ u_+ \end{pmatrix}, \begin{pmatrix} \phi_- \\ \phi_0 \\ \phi_+ \end{pmatrix} \right\rangle = \langle \mathbf{A}_1 \vec{u}, \vec{\phi} \rangle$$

$$I_0 = \int_{-h/2}^{+h/2} u(x) \phi(x) dx \approx \left\langle \frac{h}{30} \begin{bmatrix} +4 & +2 & -1 \\ +2 & +16 & +2 \\ -1 & +2 & +4 \end{bmatrix} \begin{pmatrix} u_- \\ u_0 \\ u_+ \end{pmatrix}, \begin{pmatrix} \phi_- \\ \phi_0 \\ \phi_+ \end{pmatrix} \right\rangle = \langle \mathbf{A}_0 \vec{u}, \vec{\phi} \rangle$$

$$I_f = \int_{-h/2}^{+h/2} f(x) \phi(x) dx \approx \left\langle \frac{h}{30} \begin{bmatrix} +4 & +2 & -1 \\ +2 & +16 & +2 \\ -1 & +2 & +4 \end{bmatrix} \begin{pmatrix} f_- \\ f_0 \\ f_+ \end{pmatrix}, \begin{pmatrix} \phi_- \\ \phi_0 \\ \phi_+ \end{pmatrix} \right\rangle = \langle \mathbf{M}_e \vec{f}, \vec{\phi} \rangle$$

The contribution by one element to the linear system  $\mathbf{A} \vec{u} = \mathbf{M} \vec{f}$  is

$$(a \mathbf{A}_2 + b \mathbf{A}_1 + c \mathbf{A}_0) \vec{u} \quad \text{and} \quad \mathbf{M}_e \vec{f}.$$

As a very simple example examine the ODE  $-u''(x) = f$  on three intervals of length  $h$ . Use

$$\mathbf{M}_e \begin{pmatrix} f \\ f \\ f \end{pmatrix} = \frac{f}{6} \begin{pmatrix} 1 \\ 4 \\ 1 \end{pmatrix}$$

and the above element stiffness matrix  $\mathbf{A}_2$  to find the linear system

$$\frac{1}{3h} \begin{bmatrix} +7 & -8 & +1 & & & & \\ -8 & +16 & -8 & & & & \\ +1 & -8 & +14 & -8 & +1 & & \\ & & -8 & +16 & -8 & & \\ & & +1 & -8 & +14 & -8 & +1 \\ & & & -8 & +16 & -8 & \\ & & & +1 & -8 & +7 & \end{bmatrix} \begin{pmatrix} u_0 \\ u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \end{pmatrix} = \frac{f h}{6} \begin{pmatrix} 1 \\ 4 \\ 2 \\ 4 \\ 2 \\ 4 \\ 1 \end{pmatrix}$$

- The matrix is symmetric and shows a semi-bandwidth of 3, i.e. at most five entries in each row or column, around the diagonal.
- The above system of 7 equations does not have a unique solution. For a constant vector  $\vec{u} = \vec{1}$  the matrix multiplication leads to the zero vector. This is caused by the missing boundary conditions. With the additional constraints  $u_0 = u_6 = 0$  the system has a unique solution. In the above matrix the first and last rows and columns have to be removed.
- The second of the above equations is identical to the well known finite difference formula for the second derivative, i.e.

$$\frac{-u_0 + 2u_1 + u_2}{(h/2)^2} = f$$

and similar for the fourth and sixth equation. The third equation

$$\frac{u_0 - 8u_1 + 14u_2 - 8u_3 + u_4}{h^2} = f$$

shows a five point approximation of the second derivative and similar for the fifth equation

#### 5.2.4 Element stiffness matrices for elasticity problems

For the equilateral triangle in Figure 38 examine the symmetric element stiffness matrix for parameters  $E = 1$  and  $\nu = 0.3$  for linear elements.

$$\mathbf{A}_1 \approx \begin{bmatrix} 0.531 & -0.420 & -0.111 & 0.179 & -0.014 & -0.165 \\ -0.420 & 0.531 & -0.111 & 0.014 & -0.179 & 0.165 \\ -0.111 & -0.111 & 0.222 & -0.192 & 0.192 & 0 \\ 0.179 & 0.014 & -0.192 & 0.325 & -0.008 & -0.317 \\ -0.014 & -0.179 & 0.192 & -0.008 & 0.325 & -0.317 \\ -0.165 & 0.165 & 0 & -0.317 & -0.317 & 0.634 \end{bmatrix}$$

If the location of the corners of the triangle are slightly perturbed, then all entries are different from 0. On a

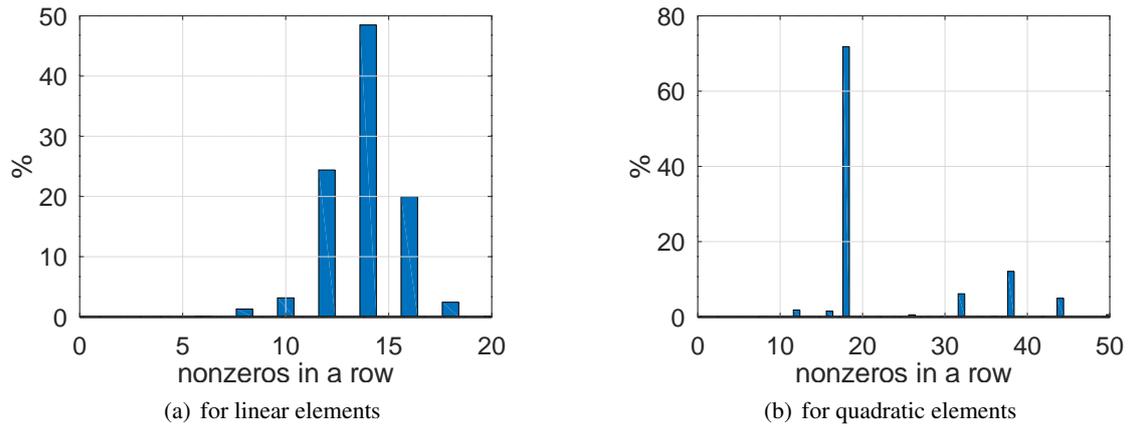


Figure 42: The number of nonzero entries in each row

mesh similar to Figure 37 (but not as uniform) with 14'040 degrees of freedom the number of nonzero entries in each row of the resulting matrix leads to the histogram in Figure 42(a). If each of the nodes would connect to 6 other nodes, then 14 nonzero entries per row are expected. The average observed on the examined mesh is 13.8 nonzeros in each row or column. Thus only  $\approx 1\%$  of the entries in the matrix are not zero, i.e. it is a very sparse matrix.

For quadratic elements the  $12 \times 12$  element stiffness matrix is given by

$$\mathbf{A}_2 \approx \begin{bmatrix}
 0.53 & 0.14 & 0.04 & 0 & -0.15 & -0.56 & 0.18 & 0 & 0.05 & 0 & -0.22 & -0.02 \\
 0.14 & 0.53 & 0.04 & -0.15 & 0 & -0.56 & 0 & -0.18 & -0.05 & 0.22 & 0 & 0.02 \\
 0.04 & 0.04 & 0.22 & -0.15 & -0.15 & 0 & 0.06 & -0.06 & 0 & 0.26 & -0.26 & 0 \\
 0 & -0.15 & -0.15 & 1.71 & -1.12 & -0.30 & 0 & 0.26 & 0.22 & 0 & 0 & -0.48 \\
 -0.15 & 0 & -0.15 & -1.12 & 1.71 & -0.30 & -0.26 & 0 & -0.22 & 0 & 0 & 0.48 \\
 -0.56 & -0.56 & 0 & -0.30 & -0.30 & 1.71 & 0.02 & -0.02 & 0 & -0.48 & 0.48 & 0 \\
 0.18 & 0 & 0.06 & 0 & -0.26 & 0.02 & 0.33 & 0 & 0.11 & 0 & -0.42 & -0.01 \\
 0 & -0.18 & -0.06 & 0.26 & 0 & -0.02 & 0 & 0.33 & 0.11 & -0.42 & 0 & -0.01 \\
 0.05 & -0.05 & 0 & 0.22 & -0.22 & 0 & 0.11 & 0.11 & 0.63 & -0.42 & -0.42 & 0 \\
 0 & 0.22 & 0.26 & 0 & 0 & -0.48 & 0 & -0.42 & -0.42 & 1.71 & -0.02 & -0.85 \\
 -0.22 & 0 & -0.26 & 0 & 0 & 0.48 & -0.42 & -0.00 & -0.42 & -0.02 & 1.71 & -0.85 \\
 -0.02 & 0.02 & 0.00 & -0.48 & 0.48 & -0.00 & -0.01 & -0.01 & -0.00 & -0.85 & -0.85 & 1.71
 \end{bmatrix}$$

and for a slight perturbation of the corners again all 144 entries are different from zero. On a mesh similar to Figure 37 (but not as uniform) with 56'700 degrees of freedom the number of nonzero entries in each row of the resulting matrix leads to the histogram in Figure 42(b) with an average of 22.7 nonzeros per row or column. For a corner of a triangle contacting 6 triangles expect  $6 \cdot 6 + 2 = 38$  nonzero entries. For a midpoint of a triangle expect  $2 \cdot 9 = 18$  nonzero entries. The midpoints outnumber the corners by a factor of three. Thus expect an average of  $\frac{3 \cdot 18 + 38}{4} = 23$  nonzero entries in each row of the matrix. Thus only  $\approx 0.4\%$  of the entries in the matrix are not zero, i.e. it is a very sparse matrix.

### 5.3 Behavior of a FEM solution within triangular elements

To examine the behavior of a solution within each of the triangular elements use the boundary value problem

$$\begin{aligned} -\Delta u &= -\exp(y) & \text{for } (x, y) \in \Omega \\ u(x, y) &= \exp(y) & \text{for } (x, y) \in \Gamma \end{aligned}$$

on the domain  $\Omega$  displayed in Figure 43(a). The exact solution is given by  $u(x, y) = \exp(y)$ , shown in Figure 43(b). The problem is solved twice:

1. using 32 triangular elements of order 1.
2. using 8 triangular elements of order 2.

The nodes used coincide for the two approaches, i.e four triangles in Figure 43(a) for the linear elements correspond to one of the eight triangles for the quadratic elements.

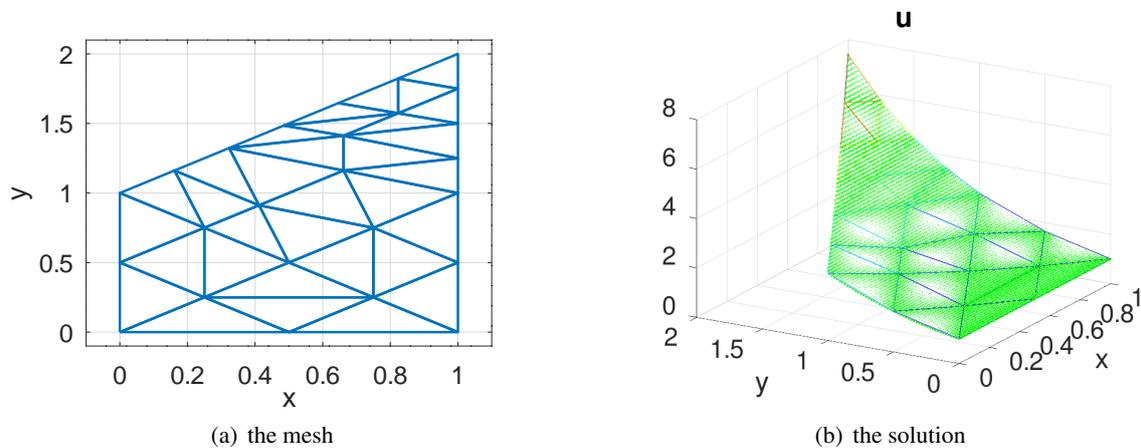
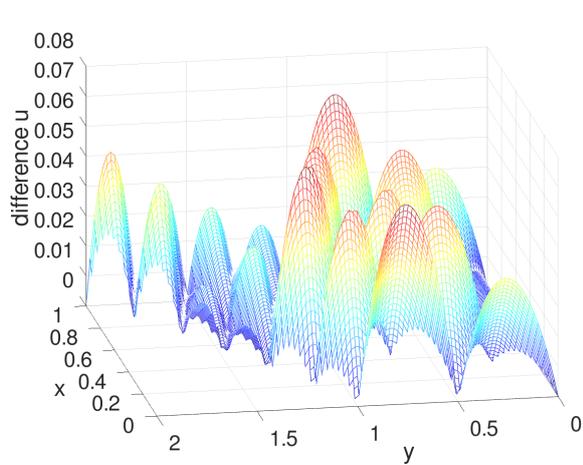


Figure 43: The mesh and the solution for a BVP

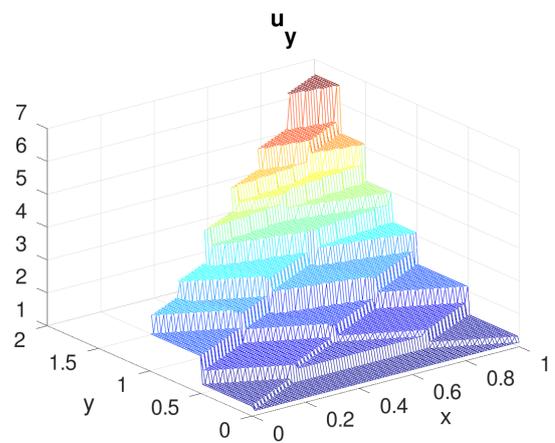
Figure 44(a) shows the difference of the computed solution with first order elements to the exact solution. Within each of the 32 elements the difference is not too far from a quadratic function. Figure 44(b) shows the values of the partial derivative  $\frac{\partial u}{\partial y}$ . It is clearly visible that the gradient is constant within each triangle, and not continuous across element borders.

Figure 45(a) shows the difference of the computed solution with second order elements to the exact solution. The error is considerably smaller than for linear elements, using identical degrees of freedom. Within each of the 8 elements the difference does not show a simple structure. Figure 45(b) shows the values of the partial derivative  $\frac{\partial u}{\partial y}$ . It is clearly visible that the gradient is not constant within the triangles. By a careful visual inspection one has to accept that the gradient is not continuous across element borders, but the jumps are considerably smaller than for linear elements. These elements are not  $C^1$ -conforming. Figure 46 shows the errors for the partial derivative  $\frac{\partial u}{\partial y}$  and confirms this observation.

In Figure 47 find the differences of the values of the solution and the partial derivative with respect to  $y$  for the same computation using cubic elements. Observe that the approximation errors are considerably smaller. The partial derivatives  $\frac{\partial u}{\partial x}$  and  $\frac{\partial u}{\partial y}$  is not continuous across the limits of the triangles, since these third order elements are not  $C^1$ -conforming.

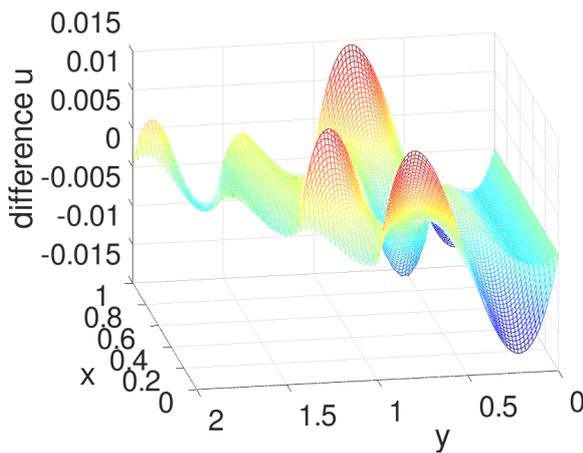


(a) the difference to the exact solution

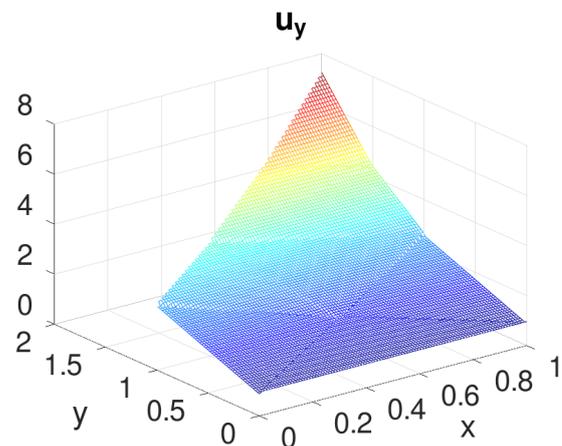


(b) the values of  $\frac{\partial u}{\partial y}$

Figure 44: Difference to the exact solution and values of  $\frac{\partial u}{\partial y}$ , using a first order mesh



(a) the difference to the exact solution



(b) the values of  $\frac{\partial u}{\partial y}$

Figure 45: Difference to the exact solution and values of  $\frac{\partial u}{\partial y}$ , using a second order mesh

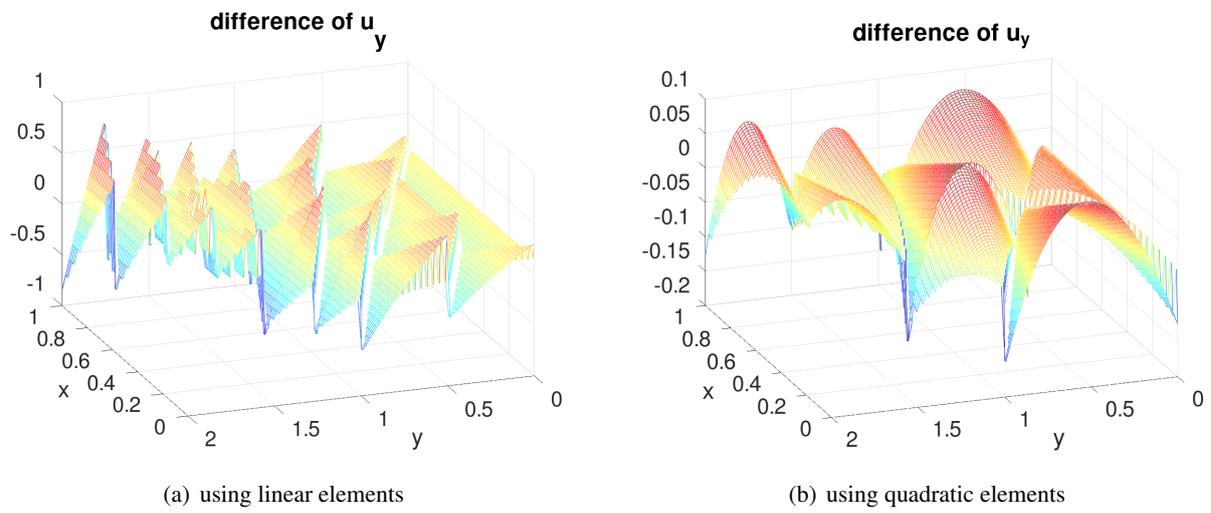


Figure 46: Difference of the approximate values of  $\frac{\partial u}{\partial y}$  to the exact values

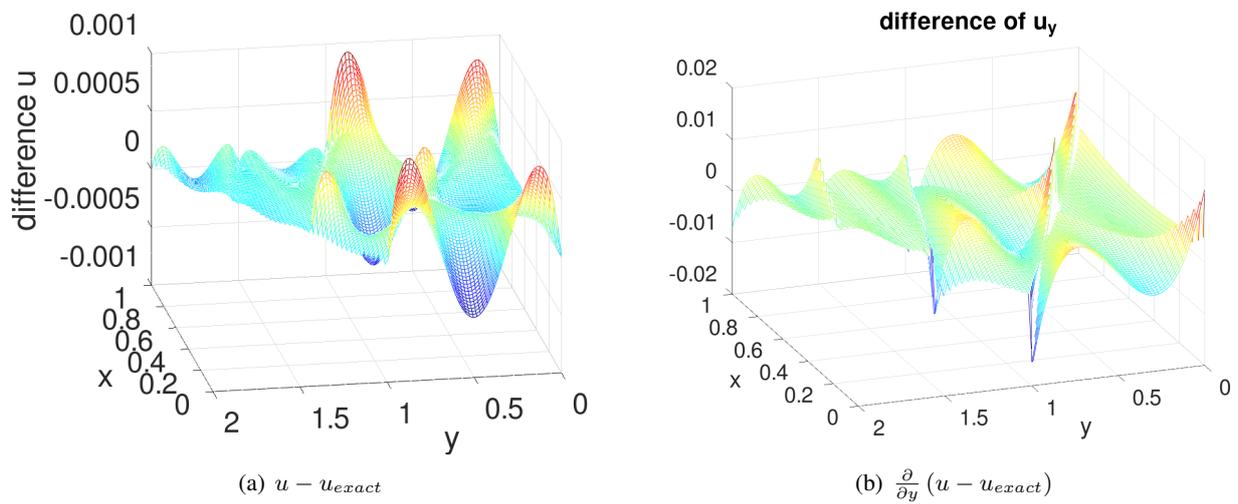


Figure 47: Difference of the approximate values of  $u$  and  $\frac{\partial u}{\partial y}$  to the exact values for cubic elements

**FEMInsideElement**

```

N = 2; MeshType = 'quadratic' %% use 'linear', 'quadratic' or 'cubic'
Mesh = CreateMeshTriangle('test',[0 0 -1;1 0 -1;1 2 -1; 0 1 -1],1/N^2);
switch MeshType
  case 'quadratic'
    Mesh = MeshUpgrade(Mesh,'quadratic');
  case 'cubic'
    Mesh = MeshUpgrade(Mesh,'cubic');
endswitch

xi = linspace(0.2,1.1,5); yi = xi*0.8+0.05;
Ngrid = 100; [xi,yi] = meshgrid(linspace(0,1,Ngrid),linspace(0,2,Ngrid));

figure(1); FEMtrimesh(Mesh)
          xlabel('x'); ylabel('y'); xlim([-0.1,1.1]); ylim([-0.1,2.1])

function res = u_exact(xy)    res = +exp(xy(:,2)) ; endfunction
function u    = f(xy)        u = -exp(xy(:,2)); endfunction

u_ex = reshape(u_exact([xi(:),yi(:)]),Ngrid,Ngrid);
u = BVP2Dsym(Mesh,1,0,'f','u_exact',0,0);
[ui,uxi,uyi] = FEMgriddata(Mesh,u,xi,yi);

figure(2); FEMtrimesh(Mesh,u);    hold on
          plot3(xi,yi,ui,'g. ');  hold off;
          xlabel('x'); ylabel('y'); title('u'); view([-60 25])
figure(3); mesh(xi,yi,uyi)
          xlabel('x'); ylabel('y'); title('u_y')
figure(4); mesh(xi,yi,uyi-u_ex)
          xlabel('x'); ylabel('y'); title('difference of u_y'); view([-110, 30])

```

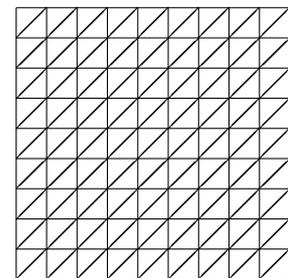
**5.4 Estimate the number of nodes and triangles in a mesh and the effect on the sparse matrix**

Let  $\Omega \subset \mathbb{R}^2$  be a domain with a triangular mesh with many triangles. There is a connection between

$$N = \text{number of nodes and } T = \text{number of triangles.}$$

Examine the typical mesh on the right and consider only triangles and nodes inside the mesh, as the number of contributions by the borders are considerably smaller for large meshes.

- each triangle has three corners
- each (internal) corner is touched by 6 triangles
- each triangle has 3 midpoints of edges and each of the midpoints is shared by 2 triangles
- For first order elements the nodes are the corners of the triangles.



$$N \approx \frac{1}{6} T \cdot 3 = \frac{1}{2} T$$

Thus the number  $N$  of nodes is approximately half the number  $T$  of triangles.

- For second order elements the nodes are the corners of the triangles and the midpoints of the edges. Each midpoint is shared by two triangles.

$$N \approx \frac{1}{2}T + \frac{3}{2}T = 2T$$

Thus the number  $N$  of nodes is approximately twice the number  $T$  of triangles.

- For third order elements the nodes are the corners of the triangles, two points each edge and the central point. Each point on an edge is shared by two triangles.

$$N \approx \frac{1}{2}T + \frac{2 \cdot 3}{2}T + T = \frac{9}{2}T$$

Thus the number  $N$  of nodes is approximately 4.5 times the number  $T$  of triangles.

The above implies that the number of degrees of freedom to solve a problem with second or third order elements with a typical diameter  $h$  of the triangles is approximately equal to using linear elements on triangles with diameter  $h/2$  (quadratic) or  $h/3$  (cubic).

The above estimates also allow to estimate how many entries in the sparse matrix resulting from an FEM algorithm will be different from zero.

- For linear elements each node typically touches 6 triangles and each of the involved corners is shared by two triangles. Thus there might be  $6 + 1 = 7$  nonzero entries in each row of the matrix.
- For second order triangles distinguish between corners and midpoints.
  - Each corner touches typically six triangles and thus expect up to  $6 \times 3 + 1 = 19$  nonzero entries in the corresponding row of the matrix.
  - Each midpoint touches two triangles and two of the corner points are shared. Thus expect up to  $2 + 2 \times 3 + 1 = 9$  nonzero entries in the corresponding row of the matrix.

The midpoints outnumber the corners by a factor of three. Thus expect an average of  $\frac{3 \cdot 9 + 19}{4} = 11.5$  nonzero entries in each row of the matrix.

- For third order triangles distinguish between corners, points on edges and center points.
  - Each corner touches typically six triangles and thus expect up to  $6 \times 6 + 1 = 37$  nonzero entries in the corresponding row of the matrix.
  - Each point on an edge touches two triangles and four points on the same edge are shared. Thus expect up to 16 nonzero entries in the corresponding row of the matrix.
  - Each center point leads to 10 nonzero entries.

There are approximately  $C$  corners points,  $2C$  midpoints and on the  $3C$  edges find  $6C$  points. Thus expect an average of  $\frac{1 \cdot 37 + 6 \cdot 16 + 2 \cdot 10}{2 + 6 + 1} = \frac{153}{9} = 17$  nonzero entries in each row of the matrix.

- The above estimates are not correct for equations with constant coefficients or horizontal or vertical edges. Then expect fewer nonzero entries in each row of the matrix.

This points to about a factor of  $\frac{11.5}{7} \approx 1.6$  more nonzero entries in the matrix for quadratic elements for the same number of degrees of freedom. For cubic elements expect a factor of  $\frac{17}{7} \approx 2.4$ . This implies that the computational effort is larger, the actual effect depends on the linear solver used.

### 5.5 Compare linear, quadratic and cubic elements

To examine the performance of the different order elements examine the BVP

$$\begin{aligned} -\nabla \left( (1+x^2) \nabla u(x,y) \right) &= -4(1+x^2) \exp(-2y) && \text{for } (x,y) \in \Omega \\ \frac{\partial u(y,0)}{\partial x} &= 0 && \text{for } 1 \leq y \leq 2 \\ u(x,y) &= \exp(-2y) && \text{on other sections of the boundary} \end{aligned}$$

on the domain shown in Figure 48. The exact solution is given by  $u_e(x,y) = \exp(-2y)$ . For different values of

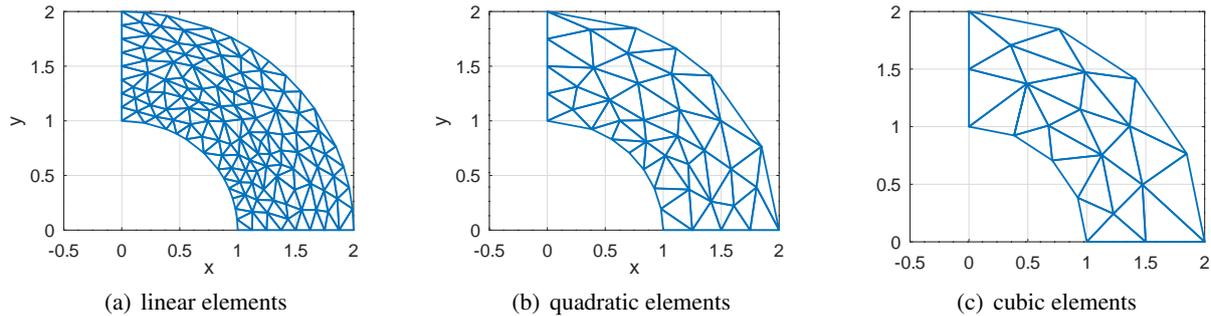


Figure 48: Meshes for linear, quadratic and cubic elements, leading to similar size linear systems to be solved.

the typical element size  $h$  for linear elements the three types of elements are used.

- For quadratic elements use  $h_{quad} = 2h$  to aim for the same number of degrees of freedom, i.e. the same size of linear system of equations to be examined. For cubic elements  $h_{cubic} = 3h$  is used. This leads to meshes shown in Figure 48. Observe that the mesh for cubic elements is not as good as the mesh for linear elements to approximate the deformed domain, caused by the larger elements.
- For each solution  $u$  determine the  $L_2$  error, i.e.

$$\text{error} = \left( \iint_{\Omega} |u(x,y) - u_e(x,y)|^2 dA \right)^{1/2}.$$

- For each setup determine the size  $n \times n$  of the matrix  $\mathbf{A}$  for the linear system to be solved.
- For each setup determine the number of nonzero entries in the sparse matrix  $\mathbf{A}$  and then the average number of nonzeros in each row of  $\mathbf{A}$ .
- When different values  $h_1$  and  $h_2$  are used the expression the errors are expected to be proportional to  $h^k$ , with the order of convergence  $k$ . Thus if  $h$  is replaced by  $h/2$  expect ratios of 2, 4, 8 or 16 for the  $L_2$  errors, according to the theoretical results shown in Section 6.7 on page 161.
- If  $h$  is replaced by  $h/2$  expect the number of elements and the size of the matrix  $\mathbf{A}$  to be multiplied by 4. The number of nonzero entries in each row should not change drastically.

The results in Table 9 confirm the theoretical estimates of the errors and the number of nonzero entries in the matrix  $\mathbf{A}$ .

Element	linear		quadratic		cubic	
width $h$ of elements	0.025	0.0125	0.050	0.0250	0.075	0.0375
number of elements	3944	15912	998	3944	432	1764
size $n$ of matrix	1920	7850	1920	7850	1896	7842
$L_2$ error	$2.2 \cdot 10^{-4}$	$6.4 \cdot 10^{-5}$	$1.8 \cdot 10^{-5}$	$1.4 \cdot 10^{-6}$	$8.4 \cdot 10^{-7}$	$5.6 \cdot 10^{-8}$
ratio of $L_2$ errors		$\approx 2.9$		$\approx 4.7$		$\approx 15$
nonzeros per row	6.8	6.9	11.0	11.2	16.1	16.6

Table 9: Results for elements of order 1, 2 and 3

## 5.6 Are second order elements $C^1$ conforming?

The command `BVP1D()` uses second order elements to solve two-point boundary value problems. Since the values of the piecewise quadratic functions coincide at the limit of two neighboring subintervals the elements are  $C^0$  conforming, i.e. the numerical solution is continuous. The “open” question is whether these elements are  $C^1$  conforming, i.e. are first derivatives continuous?

This is illustrated by the code below and the resulting Figure 49. The BVP solved is

$$-u''(x) = \text{sign}(x) \quad \text{for} \quad -1 < x < +1 \quad \text{and} \quad u(-1) = u(+1) = 0.$$

One might expect special behavior at  $x = 0$ .

- The approximate solution  $\vec{u}$  is determined by calling `BVP1D()`.
- The derivative is evaluated
  - by `pwquadinterp()` with a very high resolution.
  - by `FEM1DEvaluateDu()` at the nodes.

Then both are plotted and no difference between the two is visible, i.e. the first derivative  $u'(x)$  seems to be continuous. Zooming in around  $x \approx 0$  confirms the observation.

Observe: since  $-u''(x) = \pm 1$  is solved by  $\mp \frac{1}{2}x^2$  the exact solution  $u_{\text{exact}}(x)$  of the above problem consists of two quadratic functions, patched together at  $x = 0$ , in spite of the discontinuous right hand side  $\text{sign}(x)$ .

$$u_{\text{exact}}(x) = \begin{cases} \frac{1}{2}x(1+x) & \text{for } -1 \leq x \leq 0 \\ \frac{1}{2}x(1-x) & \text{for } 0 \leq x \leq +1 \end{cases}$$

The algorithm in `BVP1D()` is based on piecewise quadratic approximations and will thus generate the exact solution.

### Test\_C1conforming.m

```
n = 8; x = linspace(-1,1,n+1)';
[x,u] = BVP1D(x,1,0,0,@(x)sign(x),1,0,0);
figure(1); plot(x,u,'+-')
        xlabel('x'); ylabel('u')
x_fine = linspace(-1,1,10001)'; [u_fine,du_fine] = pwquadinterp(x,u,x_fine);
du = FEM1DEvaluateDu(x,u);
figure(2); plot(x_fine,du_fine,x,du,'+')
        xlabel('x'); ylabel('du/dx'); legend('interpolated','at nodes')
```

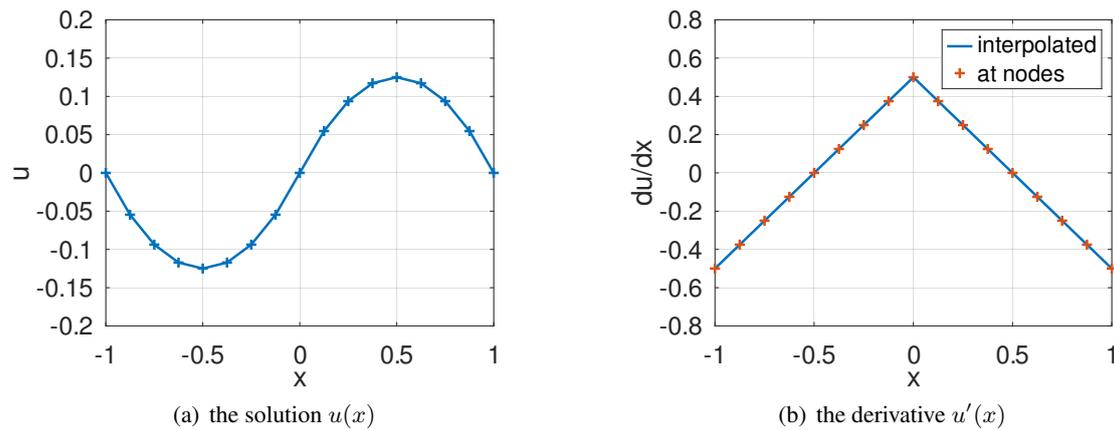


Figure 49: The solution and the first derivative, evaluated at the nodes and by interpolation

As a second BVP examine

$$-u''(x) + \text{sign}(x - 0.5) u'(x) = 1 \quad \text{for} \quad -1 < x < +1 \quad \text{and} \quad u(-1) = u(+1) = 0.$$

One might expect special behavior at  $x = 0.5$ .

- The approximate solution  $\vec{u}$  is determined by calling `BVP1D()`.
- The derivative is evaluated
  - by `pwquadinterp()` with a very high resolution.
  - by `FEM1DEvaluateDu()` at the nodes.

Then both are plotted and no difference between the two is visible in the full graph in Figure 50(a) i.e. the first derivative  $u'(x)$  seems to be continuous. Zooming in around  $x \approx 0.5$  leads to Figure 50(b) and a jump of the first derivative at  $x = 0.5$  is visible.

- Since there is a node at  $x = 0.5$  the code in `pwquadinterp()` returns the values of  $u'(x)$  for the subinterval to the left of  $x = 0.5$ , and then jumps to the values on the right subinterval. `FEM1DEvaluateDu()` returns the average value of the slopes of  $u(x)$  to the left and right of  $x = 0.5$ . This is visible in Figure 50(b).
- The size of the jump of  $u'(x)$  is smaller if more elements are used, e.g. by `n=2*8`.

The consequence: second order element are not  $C^1$  conforming. For 2D FEM algorithms the same is correct, visualized by Figure 45 on page 102.

#### Test\_C1conforming.m

```
n = 8; x = linspace(-1,1,n+1)';
[x,u] = BVP1D(x,1,@(x) sign(x-0.5),0,1,1,0,0);
figure(1); plot(x,u)
        xlabel('x'); ylabel('u')
x_fine = linspace(-1,1,10001)'; [u_fine,du_fine] = pwquadinterp(x,u,x_fine);
du = FEM1DEvaluateDu(x,u);
figure(2); plot(x_fine,du_fine,x,du,'+')
        xlabel('x'); ylabel('du/dx'); legend('interpolated','at nodes')
```

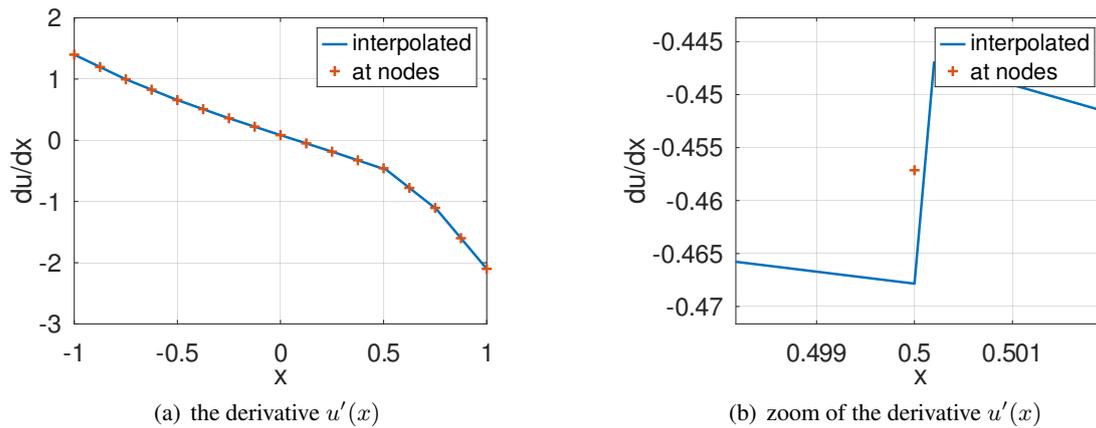


Figure 50: The first derivative, complete graph and zoomed in at  $x = 0.5$

The above example might lead to the suspicion that the jump in the first derivative is caused by the noncontinuous coefficient  $\text{sign}(x - 0.5)$  of  $u'(x)$ . This is not the case, even extremely smooth problems lead to jumps in the derivative. The boundary value problem

$$-u''(x) = -\exp(x) \quad \text{for} \quad -1 < x < +1 \quad \text{and} \quad u(-1) = \exp(-1) \quad \text{and} \quad u(+1) = \exp(+1)$$

has the exact solution  $u_{\text{exact}}(x) = \exp(x)$ . Solving the problem with only 2 subintervals leads to the errors of  $u(x)$  and  $u'(x)$  in Figure 51. In Figure 51(b) the sizable jump of the derivative at  $x = 0$  is clearly visible.

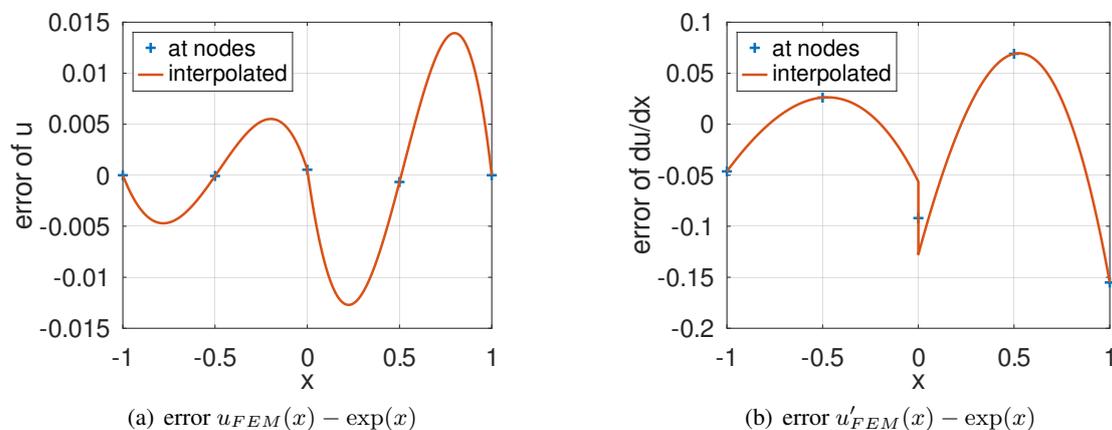


Figure 51: Differences of the FEM solution  $u_{FEM}(x)$  to the exact solution  $u_{\text{exact}}(x) = \exp(x)$

#### Test\_C1conforming.m

```
n = 2; x = linspace(-1,1,n+1)';
[x,u] = BVP1D(x,1,0,0,1,@(x) exp(x),exp(-1),exp(1));
figure(1); plot(x,u)
        xlabel('x'); ylabel('u')

x_fine = linspace(-1,1,10001)'; [u_fine,du_fine] = pwquadinterp(x,u,x_fine);
du = FEM1DEvaluateDu(x,u);
```

```

figure(2); plot(x_fine,du_fine,x,du,'+')
            xlabel('x'); ylabel('du/dx'); legend('interpolated','at nodes')
figure(3); plot(x,u,x_fine,u_fine,x_fine,exp(x_fine))
            xlabel('x'); ylabel('u');
            legend('at nodes','interpolated','exact','location','northwest');
figure(4); plot(x,u-exp(x),'+',x_fine,u_fine-exp(x_fine))
            xlabel('x'); ylabel('error of u')
            legend('at nodes','interpolated','location','northwest')
figure(5); plot(x,du-exp(x),'+',x_fine,du_fine-exp(x_fine))
            xlabel('x'); ylabel('error of du/dx')
            legend('at nodes','interpolated','location','northwest')

```

## 5.7 Superconvergence for a 1D BVP

Examine the solution of the 1D BVP

$$-u'' = \sin(x) \quad \text{for } 0 \leq x \leq \frac{\pi}{2} \quad \text{with } u(0) = 0 \quad \text{and } u'(\frac{\pi}{2}) = 0$$

with the exact solution  $u_{\text{exact}}(x) = \sin(x)$ . Use the command `BVP1D()` to find the numerical approximation  $u_{\text{FEM}}$  on a coarse grid, and then `pwquadinterp()` to evaluate the solution on a finer grid.

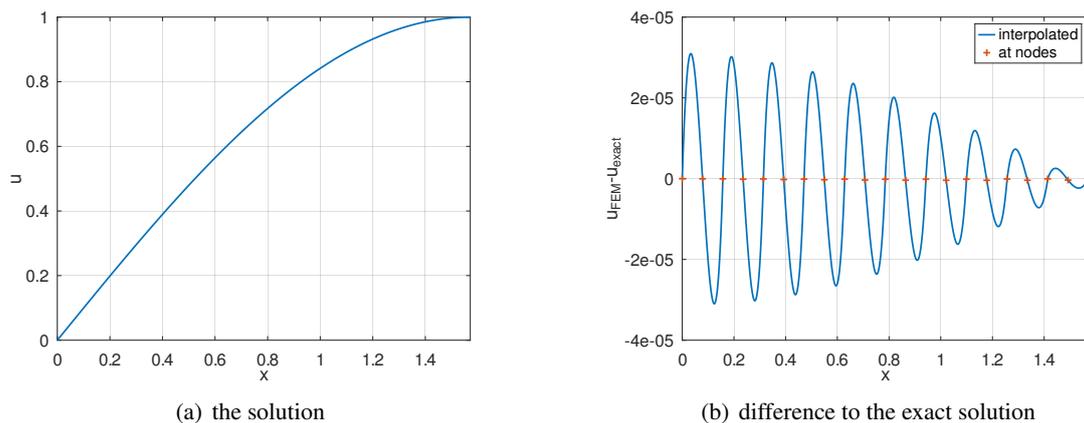


Figure 52: The solution and difference to the exact solution for a 1D BVP

Observe that

- the solution is rather accurate, even on a grid with few nodes, e.g. 21 nodes.
- The solution is considerably more accurate at the nodes, than in-between nodes. This effect is called superconvergence. You can not count on the effect of superconvergence, as it might not happen in your problem. The convergence in the  $L_2$  norm is given by the theoretical results in Section 6.7 on page 161.

### BVP1DSuperconvergence.m

```

N = 10; % number of elements, then 2*N+1 nodes
x = linspace(0,pi/2,N+1);
[xn,u] = BVP1D(x,1,0,0,1,@(x)-sin(x),0,1);

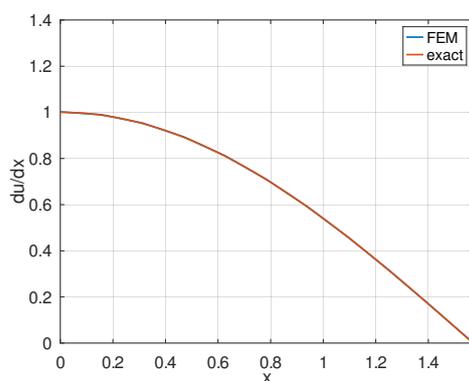
x_fine = linspace(0,pi/2,1001);

```

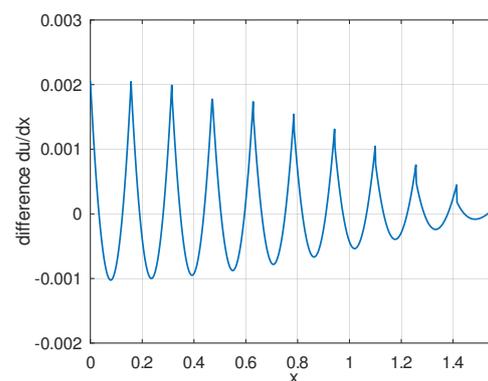
```
[u_fine du_fine, ddu_fine] = pwquadinterp(xn,u,x_fine);
figure(1); plot(x_fine,u_fine)
        xlabel('x'); ylabel('u')
figure(11); plot(x_fine,u_fine-sin(x_fine),xn,u-sin(xn),'+')
        xlabel('x'); ylabel('u_{FEM}-u_{exact}'); xlim([min(x_fine),max(x_fine)])
        legend('interpolated','at nodes')
```

The results generated by `pwquadinterp()` allow to display the derivatives and the deviation of the derivatives. Observe that

- the first derivative is piecewise linear, but it is hard to see without zooming in.
- the difference to the exact derivative oscillates on each subinterval.



(a) first derivative of the solution



(b) difference of the first derivatives

Figure 53: The derivative of the numerical solution and difference to the exact solution

#### BVP1DSuperconvergence.m

```
figure(12); plot(x_fine,du_fine-cos(x_fine))
        xlabel('x'); ylabel('difference du/dx'); xlim([min(x_fine),max(x_fine)])
figure(3); plot(x_fine,ddu_fine,x_fine,-sin(x_fine))
        xlabel('x'); ylabel('d^2u/dx^2'); xlim([min(x_fine),max(x_fine)])
        legend('FEM','exact')
figure(13); plot(x_fine,ddu_fine+sin(x_fine))
        xlabel('x'); ylabel('difference d^2u/dx^2'); xlim([min(x_fine),max(x_fine)])
```

The results generated by `pwquadinterp()` allow to display the second derivatives and the deviation of the derivatives. Observe that

- the second derivative is piecewise constant. This should be no surprise, since piecewise quadratic functions are used to generate the FEM solution.
- the difference to the exact second derivative changes the sign on each subinterval.

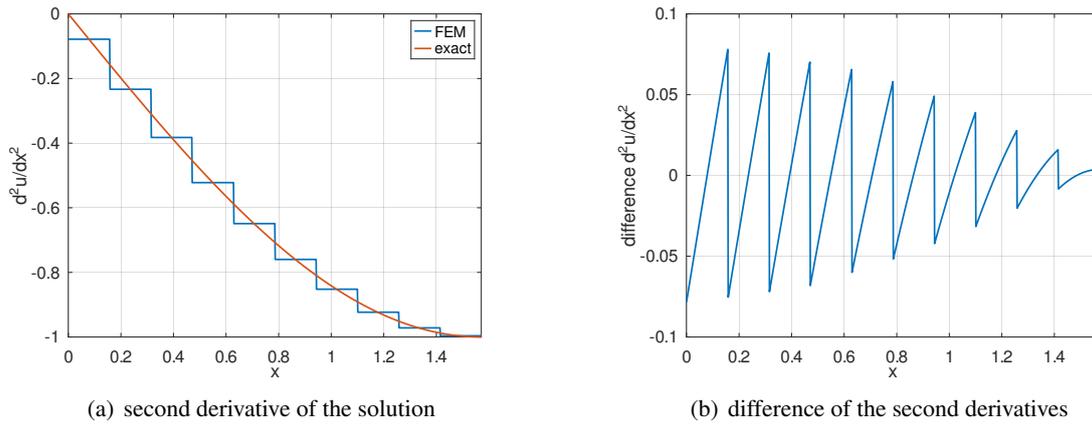


Figure 54: The second derivative of the numerical solution and difference to the exact solution

### 5.8 Stability of time steppers, or lack thereof

Examine the initial boundary value problem

$$\begin{aligned}
 \frac{\partial}{\partial t} u(x, t) &= \frac{\partial^2}{\partial x^2} u(x, t) && \text{for } 0 < x < 1 \\
 \frac{\partial}{\partial x} u(0, t) &= 0 && \text{for } t > 0 \\
 u(1, t) &= 1 && \text{for } t > 0 \\
 u(x, 0) &= 0 && \text{for } 0 < x < 1
 \end{aligned}$$

Observe that the initial condition  $u(x, 0) = 0$  does not satisfy the boundary condition  $u(1, t) = 1$ . As a consequence the behavior of solution will be critical for  $x$  close to 1. Examine the FEM solution generated by different time stepping algorithms. On the interval  $[0, 1]$  with 14 elements of order 2 use 10 time steps to find the solution at time  $t = 2$ .

- The implicit time stepper is unconditionally stable and L–stable. Thus expect convergence of the solution, but slow convergence, since the scheme is consistent of order 1. Find the result in Figure 55.
- The Crank–Nicolson time stepper is unconditionally stable, but not L–stable and consistent of order 2. Thus expect convergence of the solution, but CN will have a hard time to deal with the inconsistent initial condition. Find the result in Figure 56 and the problem around  $x = 1$  is obvious. The solution at  $x = 1$  oscillates wildly from step to step. This is caused by the stability function  $g(z) = \frac{2-z}{2+z}$  for the Crank–Nicolson stepper (see page 176). For very large  $z$  the stability condition  $|g(z)| < 1$  is satisfied, but  $g(z)$  is very close to  $-1$ . Thus instead of getting very small the corresponding contribution (eigen mode) will almost keep its amplitude, but flip its sign at each step.
- The explicit time stepper is only conditionally stable. Thus expect serious trouble for large time steps. Taking only 10 time steps leads to a large  $\Delta t$ . Find the (obviously invalid) result in Figure 57(a).
- The implicit Runge–Kutta time stepper is unconditionally stable, L–stable and consistent of order 2. Thus expect convergence of the solution, and it will take care of the inconsistent initial condition. Find the result in Figure 57(b).

Computations with finer grids and smaller time steps show that the value of  $u(x, 2) \approx 0.991$  is a good approximation of the true value.

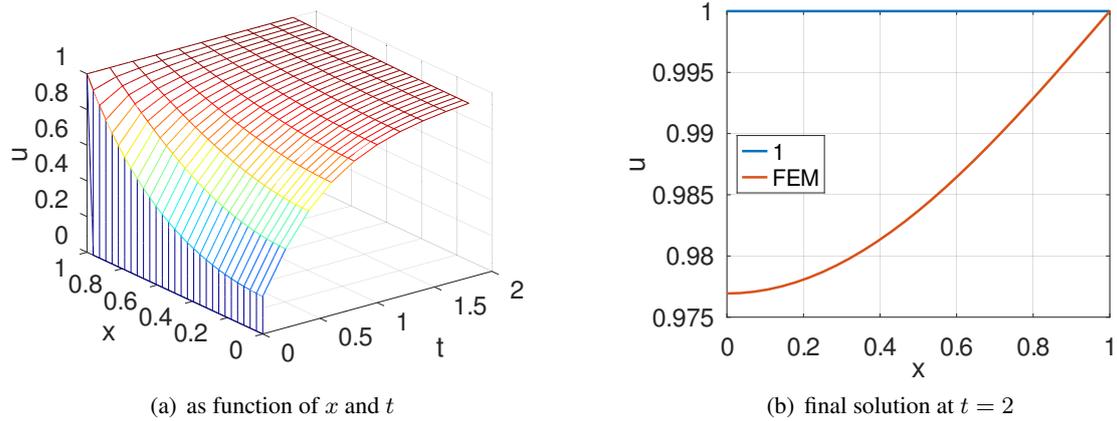


Figure 55: The solution generated by the implicit time stepper

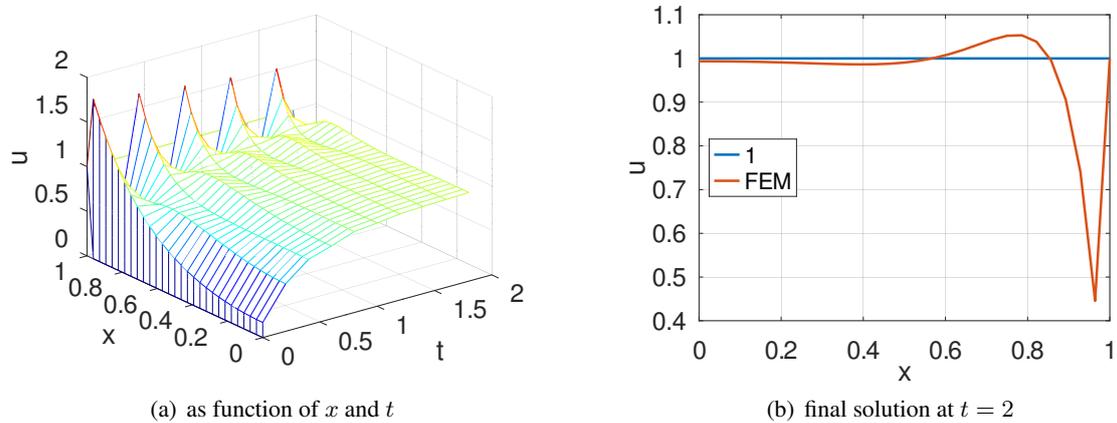


Figure 56: The solution generated by the Crank–Nicolson time stepper

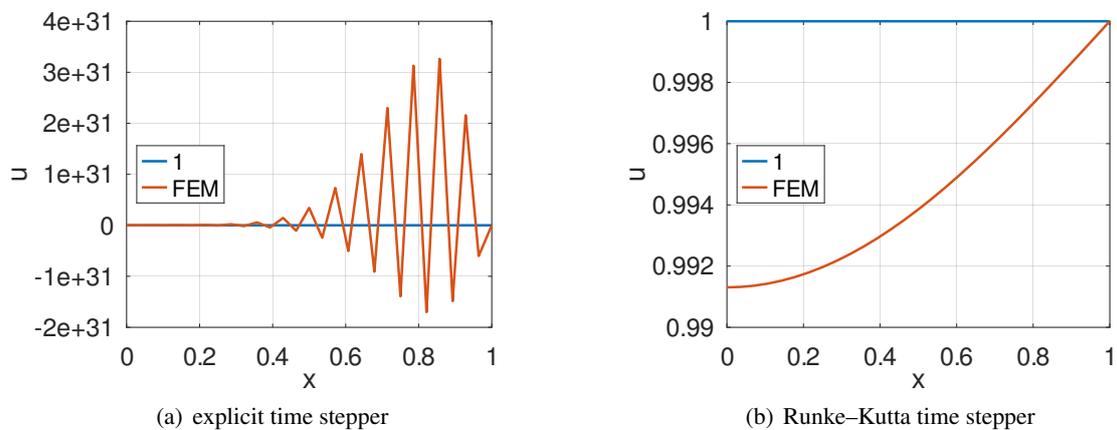


Figure 57: The solution at time  $t = 2$  generated by the explicit and Runge–Kutta time steppers

## Stability1D\_dynamic.m

```

BCleft = [0,0]; BCright = 1; xMax = 1;
interval = linspace(0,xMax,15);
t0 = 0 ; tend = 2; steps = [10,1];
w = 1; a = 1; b = 0; c = 0; d = 1; f = 0;
solver = 'implicit'; %% select one solver
%% solver = 'explicit';
%% solver = 'CN';
%% solver = 'RK';
[x,u,t] = IBVP1D(interval,w,a,b,c,d,f,BCleft,BCright,u0,t0,tend,steps,'solver',solver);

figure(1); mesh(t,x,u); ylim([0,xMax])
           xlabel('t'); ylabel('x'); zlabel('u')
u_at_0 = u(1,end)

u_ones = ones(size(x));
figure(2); plot(x,u_ones,x,u(:,end)); xlim([0,xMax])
           xlabel('x'); ylabel('u'); legend('1', 'FEM', 'location','west')

```

### 5.9 The shear locking effect by linear elements

Examine a domain  $\Omega = [-\frac{L}{2}, \frac{L}{2}] \times [-\frac{H}{2}, \frac{H}{2}] \subset \mathbb{R}^2$  with  $L = H = 0.1$  and apply a horizontal deformation  $u_1$  on the left and right edges at  $x = \pm\frac{L}{2}$  of size  $\pm c y = \pm 5 \cdot 10^{-4} y$ . The upper and lower edge are force free. Use the material parameters  $E = 100 \cdot 10^9$  and  $\nu = 0$ . Find the original and deformed domain in Figure 58. One can

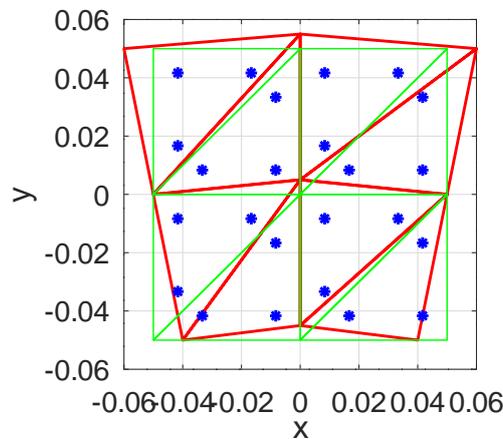


Figure 58: Original and deformed domain and the Gauss integration points for linear elements

verify<sup>11</sup> that an exact solution of the boundary value problem is given by

$$u_1(x, y) = \frac{2c}{L} x y \quad , \quad u_2(x, y) = \frac{c}{L} \left( \frac{L^2}{4} - x^2 \right) \quad \text{where} \quad c = 5 \cdot 10^{-4} .$$

<sup>11</sup>E.g. use [Stah08, §5] with  $\nu = 0$  and  $u_1 = x y$  and  $u_2 = -\frac{x^2}{2}$  to arrive at

$$\begin{aligned}
0 &\stackrel{?}{=} \frac{\partial^2 u_1}{\partial x^2} + \frac{\partial^2 u_1}{\partial y^2} + \frac{\partial}{\partial x} \left( \frac{\partial u_1}{\partial x} + \frac{\partial u_2}{\partial y} \right) = +0 + 0 + \frac{\partial}{\partial x} (y + 0) && \text{OK} \\
0 &\stackrel{?}{=} \frac{\partial^2 u_2}{\partial x^2} + \frac{\partial^2 u_2}{\partial y^2} + \frac{\partial}{\partial y} \left( \frac{\partial u_1}{\partial x} + \frac{\partial u_2}{\partial y} \right) = -1 + 0 + \frac{\partial}{\partial y} (y + 0) && \text{OK}
\end{aligned}$$

This exact solution leads to the strains  $\varepsilon_{xx} = \frac{2c}{L} y$  and  $\varepsilon_{yy} = \varepsilon_{xy} = 0$  and thus the elastic energy (use the elastic energy density (20) with  $\nu = 0$ )

$$\begin{aligned}
 U_{elast} &= U_{\varepsilon_{xx}} + U_{\varepsilon_{yy}} + U_{\varepsilon_{xy}} \\
 &= \frac{E}{2} \iint_{\Omega} \varepsilon_{xx}^2 dA + \frac{E}{2} \iint_{\Omega} \varepsilon_{yy}^2 dA + \frac{E}{2} \iint_{\Omega} 2\varepsilon_{xy}^2 dA \\
 &= \frac{E}{2} \int_{-H/2}^{+H/2} \int_{-L/2}^{+L/2} \frac{4c^2}{L^2} y^2 dx dy + 0 + 0 = \frac{E}{2} \frac{4c^2}{L^2} L \frac{2H^3}{38} = \frac{125}{3} \approx 41.667
 \end{aligned}$$

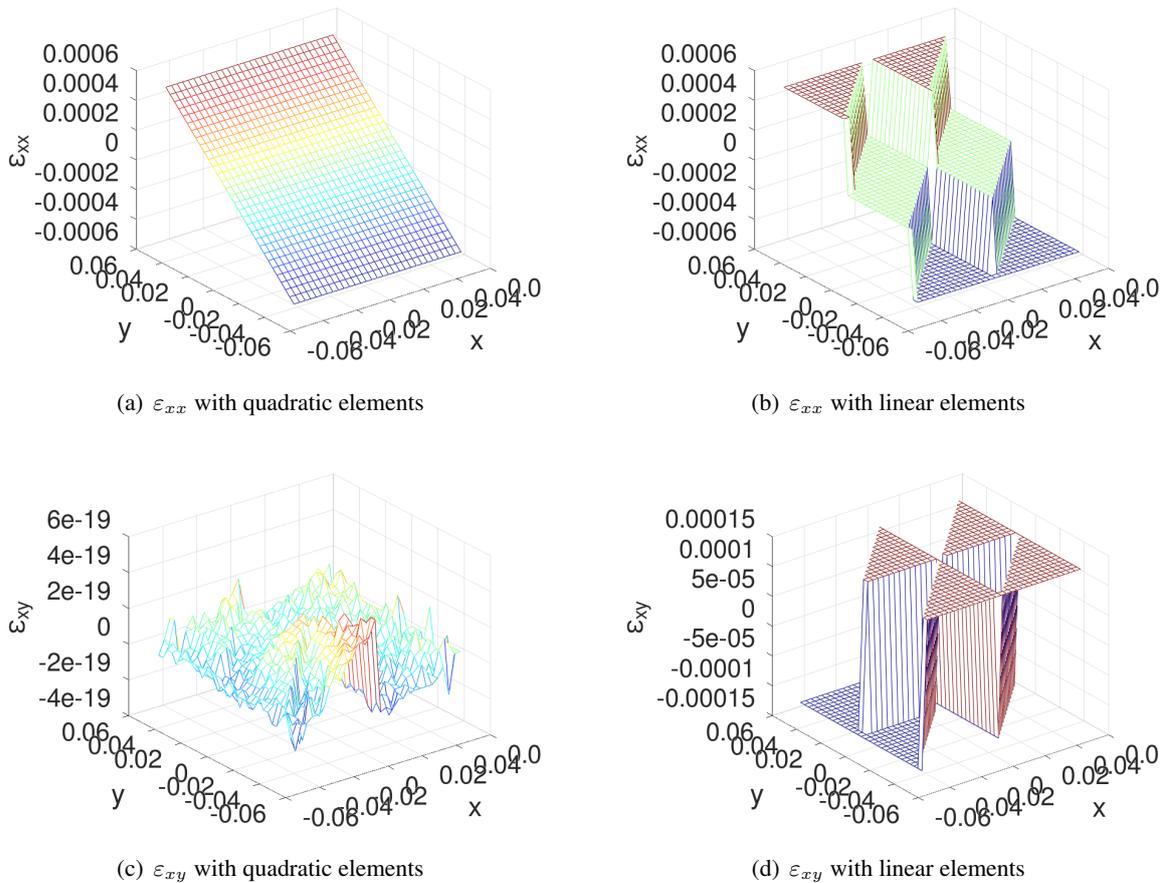


Figure 59: The strains  $\varepsilon_{xx}$  and  $\varepsilon_{xy}$  with two layers in each direction for linear and quadratic elements

Determine approximate solutions of this plane stress problem with  $NH=NL$  layers in either direction and using either linear or quadratic elements. Then use these solutions  $\vec{u}_1$  and  $\vec{u}_2$  and the function `FEMgriddata()` to evaluate the strains  $\varepsilon_{xx}$ ,  $\varepsilon_{yy}$  and  $\varepsilon_{xy}$  on a fine  $xy$ -grid. Find the results for two layers ( $NL=NH=2$ ) in Figure 59. Observe that the strains obtained by quadratic elements are very close to the strains of the exact solution. The strains based on linear elements show some surprising features:

- The strains are piecewise constant! This should be no surprise, since a partial derivative of order one of a piecewise linear function leads to a piecewise constant strain function. For this reason first order triangular elements are also called **Constant Strain Triangles**, or short **CST** elements.

- The piecewise constant approximation of the normal strain  $\varepsilon_{xx}$  is as good as can be, since only 8 triangular elements are used with this mesh.
- The piecewise approximation of the shearing strain  $\varepsilon_{xy}$  is drastically different from the exact value 0. This is caused by the two contributions to  $\varepsilon_{xy} = \frac{1}{2} \left( \frac{\partial u_1}{\partial y} + \frac{\partial u_2}{\partial x} \right)$ , which do not cancel out on the piecewise constant sections. The approximation based on second order elements is quite good, since  $10^{-19} \approx 0$ .

To examine the stiffness of the deformed body compute the elastic energy put into the body by the deformation. To arrive at reliable values a couple of steps are performed:

1. Generate a rather fine grid on the domain  $\Omega$ , using the command `meshgrid()`.
2. Evaluate the partial derivatives of  $u_1$  and  $u_2$  on the grid with the help of `FEMgriddata()`. Then compute the three strains.
3. Use the command `mesh()` to visualize a few strains, leading to Figure 59.
4. With the strains use the expressions for the elastic energy density to evaluate the different contributions.
5. Use an iterated trapezoidal rule (`trapz()`) to perform the numerical integration for the three contributions to the elastic energy.

The above is performed for different numbers of layers of first and second order elements. Find the results in Table 10. This table shows a few, possibly surprising, results.

element type	# of layers	$U_{elast}$	$U_{\varepsilon_{xx}}$	$U_{\varepsilon_{yy}}$	$U_{\varepsilon_{xy}}$
exact		41.666	41.666	0	0
quadratic	NL=NH=1	41.759	41.759	0	0
quadratic	NL=NH=2	41.759	41.759	0	0
quadratic	NL=NH=5	41.759	41.759	0	0
linear	NL=NH=1	187.5	125	0	62.5
linear	NL=NH=2	78.472	62.847	0	15.625
linear	NL=NH=5	48.122	45.622	0	2.500
linear	NL=NH=10	43.850	43.225	0	0.625

Table 10: Elastic energy contributions for shearing

- The results generated by second order elements are very accurate, even for one layer only. This is caused by the fact that the exact solution is a polynomial of degree 2 and thus can be represented exactly by second order elements. The remaining, small difference can be made smaller by using a better integration scheme. See the remarks below (page 119), where an exact result is obtained.
- The results based on linear elements are severely different. The elastic energy is considerably too high and thus the solid is considered to be much stiffer than it actually is. There are two contributions to this no-desirable effect:
  1. The piecewise constant patches lead to larger integrals.
  2. The shearing contribution by  $\varepsilon_{xy}$  does not vanish. The effect is often called **shear locking**.

For a small number of layers the effect is drastic, for larger number of layers the effect becomes smaller.

## ShearLocking.m

```

L = 0.1; H = 0.1; E = 100e9; nu = 0;
%% shearing of elements by applied displacement
NL = 2; %% elements along length L
NH = NL; %% elements along height H
Order = 1; %% order of elements, either 1 or 2

FEMmesh = CreateMeshRect([-L/2:L/NL:L/2],[-H/2:H/NH:+H/2],[-22,-22,-11,-11]);
if Order==2
    FEMmesh = MeshUpgrade(FEMmesh,'quadratic');
endif

function res = gD1(xy)
    Disp = 0.01;
    res = Disp*xy(:,1).*xy(:,2);
endfunction

[u1,u2] = PlaneStress(FEMmesh,E,nu,{0,0},{ 'gD1',0},{0,0});
figure(2); FEMtrimesh(FEMmesh,u1); xlabel('x'); ylabel('y'); zlabel('u1')
figure(3); FEMtrimesh(FEMmesh,u2); xlabel('x'); ylabel('y'); zlabel('u2')

figure(1); factor = 4e2;
trimesh(FEMmesh.elem,FEMmesh.nodes(:,1)+factor*u1,FEMmesh.nodes(:,2)+factor*u2,...
        'color','red','linewidth',2);
hold on ;
trimesh(FEMmesh.elem,FEMmesh.nodes(:,1),FEMmesh.nodes(:,2),...
        'color','green','linewidth',1);
plot(FEMmesh.GP(:,1),FEMmesh.GP(:,2),'b*');
hold off; xlabel('x'); ylabel('y'); xlim([-0.06,+0.06]); ylim([-0.06,+0.06]); axis equal

%% generate the data on the grid
x = linspace(-L/2,L/2,31); y = linspace(-H/2,+H/2,31); [xx,yy] = meshgrid(x,y);
[u1i,eps_xxi,eps_xy1i] = FEMgriddata(FEMmesh,u1,xx,yy);
[u2i,eps_xy2i,eps_yyi] = FEMgriddata(FEMmesh,u2,xx,yy);
eps_xyi = (eps_xy1i+eps_xy2i)/2;

figure(12); mesh(xx,yy,eps_xxi); xlabel('x'); ylabel('y'); zlabel('\epsilon_{xx}')
figure(13); mesh(xx,yy,eps_yyi); xlabel('x'); ylabel('y'); zlabel('\epsilon_{yy}')
figure(14); mesh(xx,yy,eps_xyi); xlabel('x'); ylabel('y'); zlabel('\epsilon_{xy}')

Wi = 0.5*E/(1-nu^2)*(eps_xxi.^2 + eps_yyi.^2+2*nu*eps_xxi.*eps_yyi+2*(1-nu)*eps_xyi.^2);
Wxxi = 0.5*E/(1-nu^2)*(eps_xxi.^2);
Wyyi = 0.5*E/(1-nu^2)*(eps_yyi.^2);
Wxxyyi = 0.5*E/(1-nu^2)*(2*nu*eps_xxi.*eps_yyi);
Wxyi = 0.5*E/(1-nu^2)*(2*(1-nu)*eps_xyi.^2);

figure(15); mesh(xx,yy,Wi);xlabel('x'); ylabel('y'); title('energy density')

EnergiesGrid = [trapz(x,trapz(y,Wi)),trapz(x,trapz(y,Wxxi)),...
                trapz(x,trapz(y,Wyyi)),trapz(x,trapz(y,Wxyi))]

```

The evaluation on a fine grid might seem unnecessary, since FEMoctave provides `EvaluateStrain()` to determine the values of the strains at the nodes. Then determine the contributions to the energy densities and integrate using `FEMIntegrate()`.

- The results for second order meshes seem reasonable.
- The results based on linear meshes are off, values and graphics. This is caused by the algorithms used:
  1. EvaluateStrain() returns values at the nodes. For the derivatives the average value of the neighboring elements are used, not the values inside the elements.
  2. FEMIntegrate() will then take those values at the nodes and (for linear elements) apply a piecewise linear interpolation, followed by a Gauss integration. Thus the values used for the integration are drastically different from the values used when the equation was solved.

#### ShearLocking.m

```

%%% evaluate at the nodes
[eps_xx,eps_yy,eps_xy] = EvaluateStrain(FEMmesh,u1,u2);

W = 0.5*E/(1-nu^2)*(eps_xx.^2 + eps_yy.^2+2*nu*eps_xx.*eps_yy+2*(1-nu)*eps_xy.^2);
Wxx = 0.5*E/(1-nu^2)*(eps_xx.^2);
Wyy = 0.5*E/(1-nu^2)*(eps_yy.^2);
Wxy = 0.5*E/(1-nu^2)*(2*(1-nu)*eps_xy.^2);

%% integration results are not reliable
EnergiesFEMIntegrate = [FEMIntegrate(FEMmesh,W),FEMIntegrate(FEMmesh,Wxx),...
                        FEMIntegrate(FEMmesh,Wyy),FEMIntegrate(FEMmesh,Wxy)];
figure(4); FEMtrimesh(FEMmesh,W);
                xlabel('x'); ylabel('y'); title('energy density, on nodes'); view([-50,20])

```

The above problem can be removed by evaluating the partial derivatives at the Gauss points, instead of the nodes. Use FEMEvaluateGP() to determine the contributions to the elastic energy density. Then integrate with FEMIntegrate().

#### ShearLocking.m

```

%% integrate by evaluation at the Gauss points
[u1G,gradU1] = FEMEvaluateGP(FEMmesh,u1);
[u2G,gradU2] = FEMEvaluateGP(FEMmesh,u2);
eps_xxG = gradU1(:,1); eps_yyG = gradU2(:,2); eps_xyG = (gradU1(:,2)+gradU2(:,1))/2;
W = 0.5*E/(1-nu^2)*(eps_xxG.^2 + eps_yyG.^2+2*nu*eps_xxG.*eps_yyG+2*(1-nu)*eps_xyG.^2);
Wxx = 0.5*E/(1-nu^2)*(eps_xxG.^2);
Wyy = 0.5*E/(1-nu^2)*(eps_yyG.^2);
Wxxyy = 0.5*E/(1-nu^2)*(2*nu*eps_xxG.*eps_yyG);
Wxy = 0.5*E/(1-nu^2)*(2*(1-nu)*eps_xyG.^2);
EnergiesFEMIntegrateGauss = [FEMIntegrate(FEMmesh,W),FEMIntegrate(FEMmesh,Wxx),...
                             FEMIntegrate(FEMmesh,Wyy),FEMIntegrate(FEMmesh,Wxy)];

```

Below find the results for two layers NL=NH=2 and first and second order elements. Shown are in that order

$$\begin{aligned}
 \iint_{\Omega} W &= \iint_{\Omega} W_{xx} + W_{yy} + W_{xy} + W_{xxyy} \\
 \iint_{\Omega} W_{xx} &= \frac{E}{2(1-\nu^2)} \iint_{\Omega} \varepsilon_{xx}^2 dA \\
 \iint_{\Omega} W_{yy} &= \frac{E}{2(1-\nu^2)} \iint_{\Omega} \varepsilon_{yy}^2 dA \\
 \iint_{\Omega} W_{xy} &= \frac{E}{2(1-\nu^2)} \iint_{\Omega} 2(1-\nu) \varepsilon_{xy}^2 dA
 \end{aligned}$$

- first order elements

EnergiesGrid	=	78.4722	62.8472	0	15.6250
EnergiesFEMIntegrate	=	67.4913	58.5938	0	8.8976
EnergiesFEMIntegrateGauss	=	78.1250	62.5000	0	15.6250

- second order elements

EnergiesGrid	=	4.1759e+01	4.1759e+01	6.8171e-30	8.2941e-30
EnergiesFEMIntegrate	=	4.1667e+01	4.1667e+01	6.6145e-30	2.7413e-30
EnergiesFEMIntegrateGauss	=	4.1667e+01	4.1667e+01	6.5378e-30	8.5890e-30

Observe that the results based on the integration with the Gauss points yields the same numbers as the exact formula.

### 5.10 Bending of an Euler beam

A plate of length  $L = 1$ , width  $W = 1$  and height  $H = 0.1$  is attached at the left edge and an upward force of  $F = 100$  is applied on the right side. Use the material parameters  $E = 100 \cdot 10^9$  and  $\nu = 0$ . Based on the Euler beam theory conclude

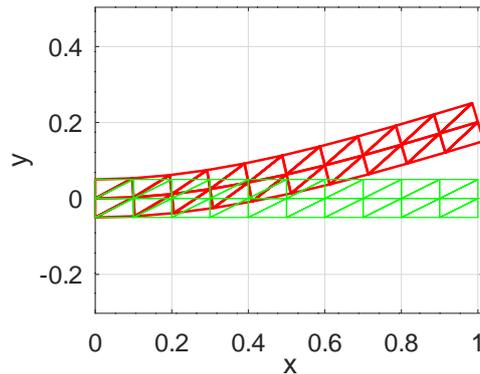


Figure 60: The original shape of the a beam and its (exaggerated) deformed shape, using two layers of elements

$$\begin{aligned} \frac{\partial^2}{\partial x^2} u_2(x, y) &= \frac{M}{EI} = \frac{F}{EI} (L - x) \quad , \quad \frac{\partial}{\partial x} u_2(x, y) = \frac{F}{EI} (Lx - \frac{1}{2} x^2) \\ u_2(x, y) &= \frac{F}{EI} (\frac{L}{2} x^2 - \frac{1}{6} x^3) \\ u_1(x, y) &= -y \frac{\partial}{\partial x} u_2(x, y) = -\frac{F}{EI} (Lx - \frac{1}{2} x^2) y \\ \varepsilon_{xx}(x, y) &= \frac{\partial u_1(x, y)}{\partial x} = -\frac{F}{EI} (L - x) y \quad , \quad \varepsilon_{yy}(x, y) = \frac{\partial u_2(x, y)}{\partial y} = 0 \\ \varepsilon_{xy}(x, y) &= \frac{1}{2} \left( \frac{\partial u_1}{\partial y} + \frac{\partial u_2}{\partial x} \right) = \frac{F}{EI} \left( -(Lx - \frac{1}{2} x^2) + (\frac{L}{2} 2x - \frac{1}{2} x^2) \right) = 0 \end{aligned}$$

For the above parameters with the second moment  $I = \frac{WH^3}{12}$  of the cross section obtain the following maximal values.

$$\begin{aligned} u_2(L, y) &= \frac{F}{3EI} L^3 = 4 \frac{F}{EWH^3} L^3 = 4 \cdot 10^{-6} \\ u_1(L, -H/2) &= \frac{F}{4EI} LH = 3 \frac{F}{EWH^2} L = 3 \cdot 10^{-7} \\ \varepsilon_{xx}(0, -H/2) &= \frac{F}{2EI} LH = 6 \frac{F}{EWH^2} L = 6 \cdot 10^{-7} \end{aligned}$$

Use these results to verify the accuracy of the numerical approximations.

To examine the performance of the FEM algorithms use a rectangular mesh with NL sections along the horizontal  $x$ -axis and NH layers in the vertical  $y$ -direction. The code is using first, second or third order elements. In Figure 61 find the mesh and the corresponding integration points for meshes with NL=10 and just one layer, i.e. NH=1. Observe that the figure uses different scaling, all triangles have height and width 0.1, which is usually recommended for good quality meshes. The code was run with NL=10 horizontal sections and NH=1 or 5 vertical sections. The elastic energy density  $W_{stress}$  is computed and displayed in Figure 62. Observe the piecewise constant energy density for linear elements, i.e. CST elements.

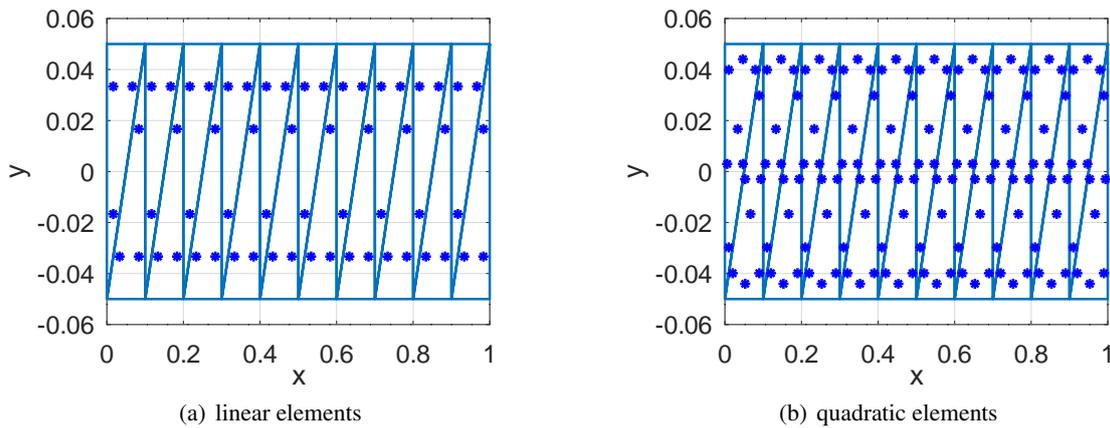


Figure 61: Meshes for linear and quadratic elements with one layer, with the integration points

Multiple runs of the code `BendingBeam.m` lead to the results in Table 11. The values for the elastic energy are computed with the help of the strain values at the Gauss points. Observe that second and third order elements generate rather accurate results, even for a very coarse grid. With a coarse grid of linear elements the effect of shear locking is clearly visible. But even for a  $80 \times 8$  grid the results are not very accurate.

#### BendingBeam.m

```
% bending of beam by applied force
L = 1; H = 0.1; E = 100e9; nu = 0; Force = 100;

NL = 20; %% number of elements along length L
NH = NL/10; %% number of elements along height H
Order = 2; %% order of elements, either 1 or 2
FEMmesh = CreateMeshRect([0:L/NL:L], [-H/2:H/NH:H/2], -22, -22, -11, -33);

figure(1); FEMtrimesh(FEMmesh); %% axis equal;
```

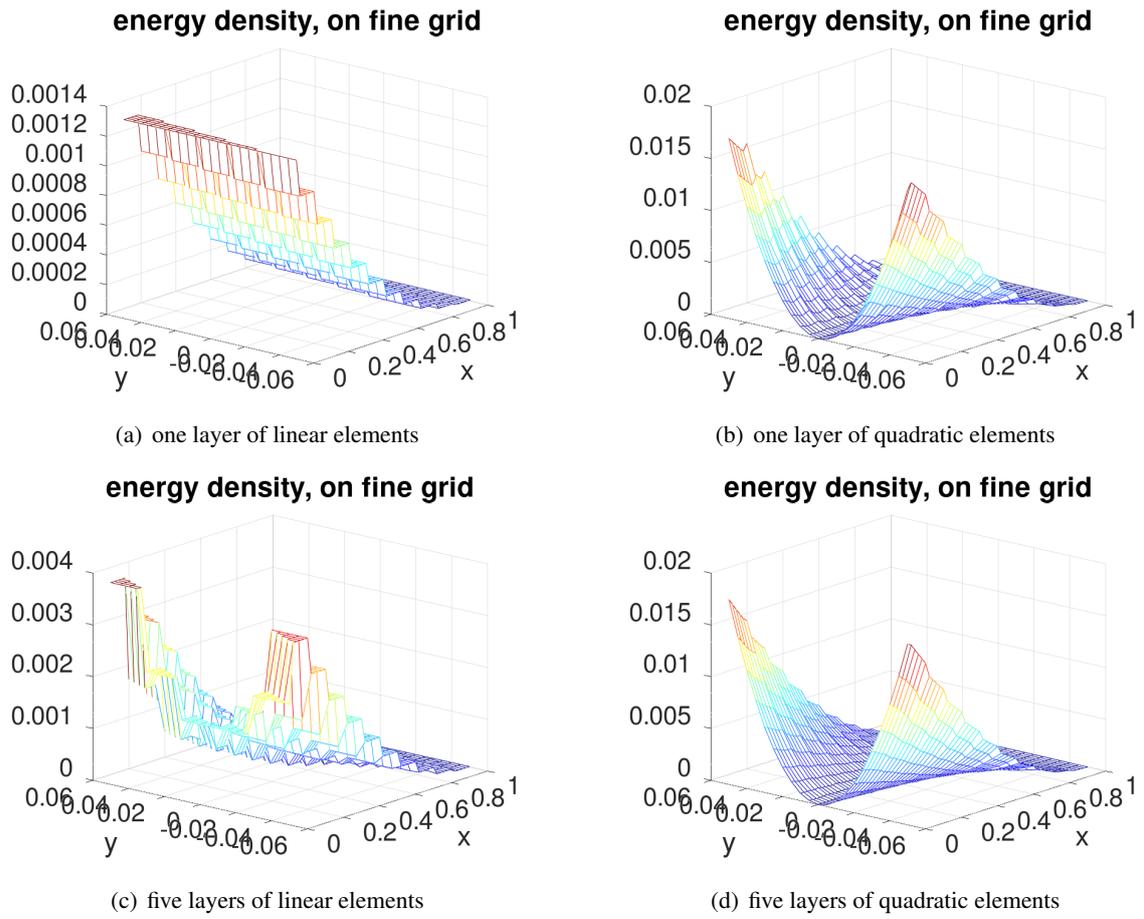


Figure 62: The elastic energy density of the bending beam with one or five layers

element order	NL	NH	$\max\{u_2\}$	$\max\{\varepsilon_{xx}\}$	energy
Euler beam, exact			$4 \cdot 10^{-6}$	$6 \cdot 10^{-7}$	$2 \cdot 10^{-4}$
first	80	8	$3.8159 \cdot 10^{-6}$	$5.7295 \cdot 10^{-7}$	$1.9079 \cdot 10^{-4}$
first	40	4	$3.3036 \cdot 10^{-6}$	$4.8900 \cdot 10^{-7}$	$1.6517 \cdot 10^{-4}$
first	20	2	$2.1525 \cdot 10^{-6}$	$3.1248 \cdot 10^{-7}$	$1.0762 \cdot 10^{-4}$
first	10	1	$0.9079 \cdot 10^{-6}$	$1.2718 \cdot 10^{-7}$	$0.4539 \cdot 10^{-4}$
second	80	8	$4.0243 \cdot 10^{-6}$	$6.1660 \cdot 10^{-7}$	$2.0120 \cdot 10^{-4}$
second	40	4	$4.0242 \cdot 10^{-6}$	$6.1101 \cdot 10^{-7}$	$2.0120 \cdot 10^{-4}$
second	20	2	$4.0235 \cdot 10^{-6}$	$6.0326 \cdot 10^{-7}$	$2.0117 \cdot 10^{-4}$
second	10	1	$4.0162 \cdot 10^{-6}$	$5.8832 \cdot 10^{-7}$	$2.0081 \cdot 10^{-4}$
third	80	8	$4.0244 \cdot 10^{-6}$	$6.2131 \cdot 10^{-7}$	$2.0122 \cdot 10^{-4}$
third	40	4	$4.0243 \cdot 10^{-6}$	$6.1708 \cdot 10^{-7}$	$2.0122 \cdot 10^{-4}$
third	20	2	$4.0243 \cdot 10^{-6}$	$6.1176 \cdot 10^{-7}$	$2.0121 \cdot 10^{-4}$
third	10	1	$4.0241 \cdot 10^{-6}$	$6.0592 \cdot 10^{-7}$	$2.0120 \cdot 10^{-4}$

Table 11: Different values for the deformation of a bending beam, depending on the size of the grid

```

hold on; plot(FEMmesh.GP(:,1),FEMmesh.GP(:,2),'b*'); hold off
xlabel('x'); ylabel('y')
switch Order
case 2
    FEMmesh = MeshUpgrade(FEMmesh,'quadratic');
case 3
    FEMmesh = MeshUpgrade(FEMmesh,'cubic');
endswitch

[u1,u2] = PlaneStress(FEMmesh,E,nu,{0,0},{0,0},{0,Force/H});
figure(2); FEMtrimesh(FEMmesh,u1); xlabel('x'); ylabel('y'); zlabel('u1')
figure(3); FEMtrimesh(FEMmesh,u2); xlabel('x'); ylabel('y'); zlabel('u2')

FEMoctave_u2Max = max(u2);
EulerBeam = 4*Force*L^3/(E*H^3);
MaximalDisplacements = [EulerBeam, FEMoctave_u2Max]
[eps_xx,eps_yy,eps_xy] = EvaluateStrain(FEMmesh,u1,u2);
figure(12); FEMtrimesh(FEMmesh,eps_xx); xlabel('x'); ylabel('y'); zlabel('eps_{xx}')
Results_Maxu1_Maxeps_xx = [max(abs(u1)), max(abs(eps_xx))]
W = 0.5*E/(1-nu^2)*(eps_xx.^2 + eps_yy.^2+2*nu*eps_xx.*eps_yy+2*(1-nu)*eps_xy.^2);

EnergyByForce = [Force*EulerBeam/2, Force*max(u2)/2]

figure(4);FEMtrimesh(FEMmesh,W); xlabel('x'); ylabel('y');
    title('energy density, on nodes'); view([-50,20])
figure(5);clf;FEMtricontour(FEMmesh,W); xlabel('x'); title('energy density')

%% integrate by evaluation at the Gauss points
[u1G,gradU1] = FEMEvaluateGP(FEMmesh,u1);
[u2G,gradU2] = FEMEvaluateGP(FEMmesh,u2);
eps_xxG = gradU1(:,1); eps_yyG = gradU2(:,2); eps_xyG = (gradU1(:,2)+gradU2(:,1))/2;

```

```

W = 0.5*E/(1-nu^2)*(eps_xxG.^2+eps_yyG.^2+2*nu*eps_xxG.*eps_yyG+2*(1-nu)*eps_xyG.^2);
EnergiesFEMIntegrateGauss = FEMIntegrate(FEMmesh,W)

[xx,yy] = meshgrid(linspace(0,L,101),linspace(-H/2,+H/2,51));
[u1i,eps_xxi,eps_xyli] = FEMgriddata(FEMmesh,u1,xx,yy);
[u2i,eps_xy2i,eps_yyi] = FEMgriddata(FEMmesh,u2,xx,yy);
eps_xyi = (eps_xyli+eps_xy2i)/2;

Wi = 0.5*E/(1-nu^2)*(eps_xxi.^2+eps_yyi.^2+2*nu*eps_xxi.*eps_yyi+2*(1-nu)*eps_xyi.^2);

figure(14); mesh(xx,yy,Wi);xlabel('x'); ylabel('y');
        title('energy density, on fine grid'); view([-50,20])

%% show deformed domain
factor = 1e5/2;
figure(100); trimesh(FEMmesh.elem,FEMmesh.nodes(:,1)+factor*u1,...
        FEMmesh.nodes(:,2)+factor*u2,'color','red','linewidth',2);
hold on; trimesh(FEMmesh.elem,FEMmesh.nodes(:,1),...
        FEMmesh.nodes(:,2),'color','green','linewidth',1);
hold off; xlabel('x'); ylabel('y'); axis equal

```

### 5.11 Missing boundary constraints and null spaces

Examine a domain  $\Omega \subset \mathbb{R}^2$  and minimize the elastic energy given by equation (21) with  $\vec{f} = \vec{g}_N = \vec{0}$

$$U(\vec{u}) = \iint_{\Omega} \frac{1}{2} \frac{E}{(1-\nu^2)} \left\langle \begin{bmatrix} 1 & \nu & 0 \\ \nu & 1 & 0 \\ 0 & 0 & 2(1-\nu) \end{bmatrix} \cdot \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix}, \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix} \right\rangle dA,$$

i.e. no external force and all of the boundary is free to move. Since the strains depend on derivatives of the displacement, the energy  $U(\vec{u})$  will not change for constant displacement vectors  $\vec{u}$ . In addition the domain can be rotated without deformation, i.e. without adding elastic energy. This leads to a three dimensional subspace on which the elastic energy vanishes. No reliable FEM algorithm will be able to solve the corresponding problem of minimizing the energy, since there is no unique minimum. To obtain a solution constraints have to be introduced, preventing the solid from moving in the  $x$  and  $y$  direction and prevent rotations.

As a consequence the global stiffness matrix  $\mathbf{A}$  should have a three dimensional null space, describing constant displacements and rotations. Thus expect three very small eigenvalues. Constant displacements vectors  $\vec{u}$  in  $x$  or  $y$  direction and rotations should satisfy

$$\mathbf{A} \vec{u} = \vec{0} \quad \text{and} \quad \frac{1}{2} \langle \vec{u}, \mathbf{A} \vec{u} \rangle = 0,$$

where the second expression corresponds to the elastic energy. These expressions are examined in the code below. The code verifies this on a trapezoidal domain with corners at  $(0, 0)$ ,  $(1, 0)$ ,  $(1, 2)$  and  $(0, 1)$ .

- There are three eigenvalues very close to zero. Due to the finite accuracy of the arithmetic on the CPU the values are not exactly zero<sup>12</sup>. The fourth eigenvalue is considerably larger. This confirms the three dimensional null space of the matrix  $\mathbf{A}$ .
- The vector `shift_x` implements a translation of the solid in  $x$  direction. Since  $\mathbf{A} * \text{shift}_x$  is approximately 0 the vector is in the null space of  $\mathbf{A}$ .

<sup>12</sup>The operation  $\mathbf{A} = (\mathbf{A} + \mathbf{A}')/2$  assures that the matrix is symmetric, to get around a problem in the implementation of `eigs()` in *Octave*.

- With the vector `shift_y` the behavior in  $y$  direction is examined.
- The displacement vector `rot_vec` examines a rotation of the solid, verifying that the energy is not increased by this rotation.
- The null space of the matrix **A** is spanned by the above three vectors.
- With the vector `rand_vec` examine an arbitrary displacement and observe that this vector is not in the null space and the energy is clearly increased.

#### TestNullSpace.m

```

Mesh = CreateMeshTriangle('test',[1 0 -22;2 0 -22;2 2 -22; 1 1 -22],0.01);

E = 1e9; nu = 0.3; f = {0,0}; gD = {0,0}; gN = {0,0}; %% set the parameters
if 0 %% plane stress
    [A,g] = PStressEquationM(Mesh,E,nu,f,gD,gD); %% determine matrix A
else %% axially symmetric
    [A,g] = AxiStressEquationM(Mesh,E,nu,f,gD,gD); %% determine matrix A
endif
A = (A+A')/2; %% assure that matrix is symmetric, it should be, but rounding errors
EigenValues = eigs(A,6,'sa') %% find the smallest eigenvalues

n = size(A,1)/2;
shift_x = [ones(n,1);zeros(n,1)]; %% constant shift in x direction
shift_x = shift_x/norm(shift_x);
Norm_Shift_x = [norm(A*shift_x),shift_x'*A*shift_x/2]

shift_y = [zeros(n,1);ones(n,1)]; %% constant shift in y direction
shift_y = shift_y/norm(shift_y);
Norm_Shift_y = [norm(A*shift_y),shift_y'*A*shift_y/2]

x = Mesh.nodes(:,1); y = Mesh.nodes(:,2); %% at point [x,y] add displacement [-y,x]
rot_vec = [-y;x]; %% a rotation
rot_vec = rot_vec/norm(rot_vec);
Norm_Rotation = [norm(A*rot_vec),rot_vec'*A*rot_vec/2]

rand_vec = [x.*y;y]; %% an arbitrary displacement vector
rand_vec = rand_vec/norm(rand_vec);
Norm_Random = [norm(A*rand_vec),rand_vec'*A*rand_vec/2]
-->
EigenValues = -5.9793e-07 -3.8610e-07 4.6269e-07 8.6494e+06 2.7384e+07 3.1225e+07

Norm_Shift_x = 3.1718e-07 7.8932e-10
Norm_Shift_y = 2.8492e-07 -1.9276e-08
Norm_Rotation = 3.3145e-07 2.2532e-09
Norm_Random = 5.1146e+07 5.3317e+06

```

The situation changes for axially symmetric problems, i.e. the domain in the  $xz$ -plane is rotated about the  $z$ -axis to obtain the object in the space  $\mathbb{R}^3$ . Instead of `PStressEquationM()` use `AxiStressEquationM()` to generate the stiffness matrix **A**. In the above code change the switch in the third line to obtain the results below.

- For axially symmetric problems moving the object up (in  $z$ -direction) does not lead to a deformation, thus there is at least a one-dimensional nullspace.

- Moving the intersection of the object with the plane  $y = 0$  in radial direction (in the  $x$ -direction) does deform the 3D object and thus increases the elastic energy.
- Rotating the intersection of the object with the plane  $y = 0$  does deform the 3D object and thus increases the elastic energy.

As a consequence there is only one eigenvalue (close to) zero, which is confirmed by the results below.

```
EigenValues = 1.7398e-06 4.0145e+06 7.3985e+06 1.8776e+07 5.0384e+07 5.4062e+07
```

```
Norm_Shift_x = 4.0504e+07 4.8423e+06
```

```
Norm_Shift_y = 6.0413e-07 -4.3348e-09
```

```
Norm_Rotation = 2.3078e+07 1.0892e+06
```

```
Norm_Random = 1.5364e+08 1.8131e+07
```

## 6 The Mathematics of the Algorithms for 2D FEM

In this section the mathematical background for the FEM method applied to the problems in Section 2 is explained. Most of the theory is used to solve the second order elliptic boundary value problem (1). The explanations are certainly not complete, but should provide enough information to ease the understanding of the code. For in-depth coverage consult one of the many books on FEM and/or numerical analysis. The starting point for this presentation are the lecture notes [Stah08]. Find a list of books on FEM in [Stah08, §0].

The organization of this section is as follows:

- 6.1 The definition of classical and weak solutions of boundary value problems is given and the connection, including calculus of variations, is shown.
- 6.2 The most often used triangular elements are shown.
- 6.3 The method of interpolation on triangles and Gauss integration is explained.
- 6.4 Element stiffness matrices for triangular elements of order 1 are carefully derived. The construction of the global stiffness matrix is explained. The integration of the different contributions is performed.
- 6.5 Element stiffness matrices for triangular elements of order 2 are carefully derived. The integration of the different contributions is performed.
- 6.6 Element stiffness matrices for triangular elements of order 3 are carefully derived. The integration of the different contributions is performed.
- 6.7 The theoretical convergence result is shown.
- 6.8 The algorithms to solve dynamic initial boundary value problems are presented.
  - In Section 6.8.1 dynamic heat equations are examined, using the Crank–Nicolson approach.
  - In Section 6.8.2 using eigenvalues to solve dynamic heat equations is explained.
  - In Section 6.8.3 dynamic wave equations are examined, using an implicit approximation.
  - In Section 6.8.4 using eigenvalues to solve dynamic wave equations is explained.
- 6.9 A few remarks on using the inverse power iteration to determine eigenvalues.

### 6.1 Classical solutions and weak solutions

A function  $u = u(x, y)$  is called a **classical solution** of the the BVP (1) iff it is twice differentiable and

$$\begin{aligned} -\nabla \cdot (a \nabla u - u \vec{b}) + b_0 u &= f & \text{for } (x, y) \in \Omega \\ u &= g_1 & \text{for } (x, y) \in \Gamma_1 \\ \vec{n} \cdot (a \nabla u - u \vec{b}) &= g_2 + g_3 u & \text{for } (x, y) \in \Gamma_2 \end{aligned}$$

Multiply this equation with a smooth function  $\phi$ , vanishing on  $\Gamma_1$ , and integrate over the domain  $\Omega$  to arrive at

$$\begin{aligned} 0 &= -\nabla \cdot (a \nabla u - u \vec{b}) + b_0 u - f \\ 0 &= \iint_{\Omega} \phi \left( -\nabla \cdot (a \nabla u - u \vec{b}) + b_0 u - f \right) dA \\ 0 &= \iint_{\Omega} \nabla \phi \cdot (a \nabla u - u \vec{b}) + \phi (b_0 u - f) dA - \int_{\Gamma} \phi \left( a \nabla u - u \vec{b} \right) \cdot \vec{n} ds \\ 0 &= \iint_{\Omega} \nabla \phi \cdot (a \nabla u - u \vec{b}) + \phi (b_0 u - f) dA - \int_{\Gamma_2} \phi (g_2 + g_3 u) ds. \end{aligned} \quad (27)$$

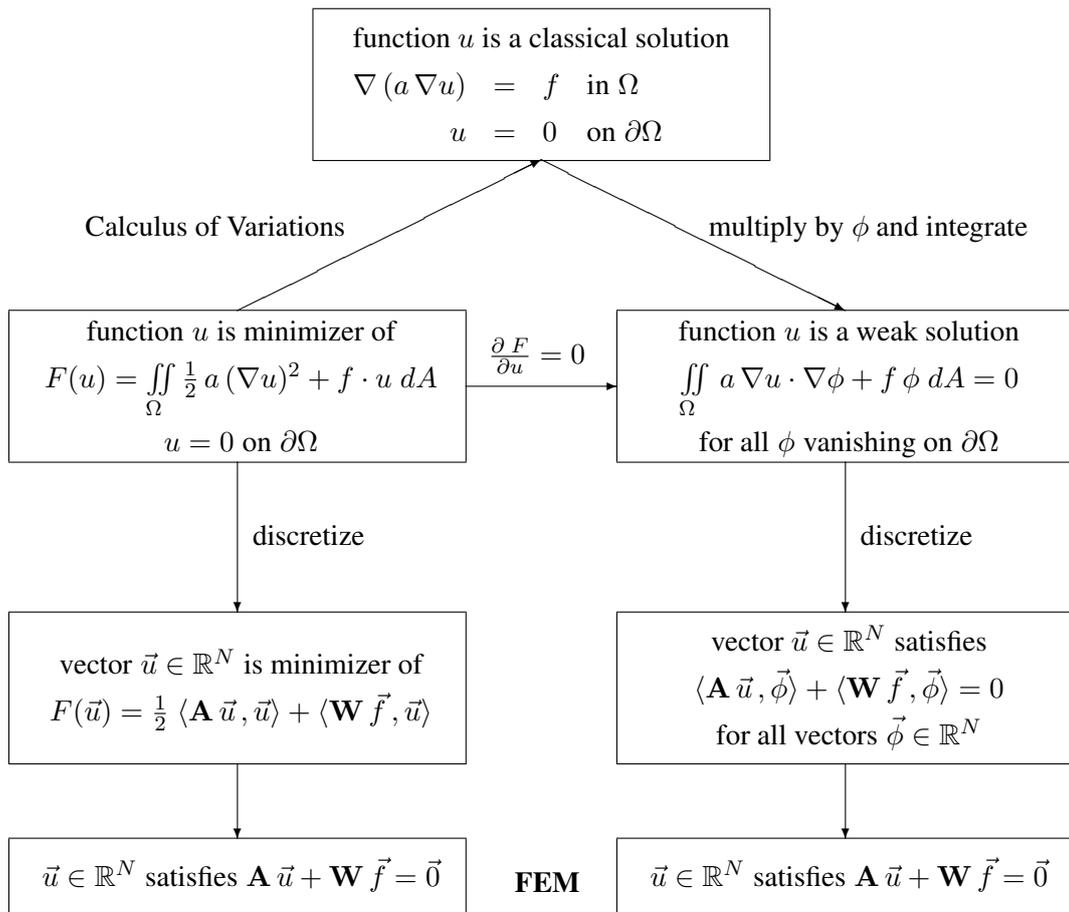


Figure 63: Classical and weak solutions, minimizers and FEM

If a function  $u$  satisfies (27) it is called a **weak solution** of the above BVP. If there is no convection term ( $\vec{b} = \vec{0}$ ) and some sign conditions for  $a$  and  $b_0$  are satisfied, the above is equivalent to minimizing the functional

$$F(u) = \iint_{\Omega} \frac{1}{2} a (\nabla u)^2 + \frac{1}{2} b_0 u^2 + f \cdot u \, dA - \int_{\Gamma_2} g_2 u + \frac{1}{2} g_3 u^2 \, ds$$

among all functions  $u$  satisfying the boundary condition  $u = g_1$  on  $\Gamma_1$ . Figure 63 shows connections between classical solutions, weak solutions and the resulting system of (linear) equations for the finite element approach. The left branch in Figure 63 illustrates the usage of minimization and calculus of variations in the context of FEM algorithms.

In the above equation integrals over the domain  $\Omega \subset \mathbb{R}^2$  have to be computed. To discretize this process use a triangularization of the domain, using grid points  $(x_i, y_i) \in \Omega$ ,  $1 \leq i \leq n$ . On each triangle  $T_k$  we replace the function  $u$  by polynomials of degree 1 (or 2, or 3). These polynomials are completely determined by their values at the three corners of the triangle (or corners and some points on the edges). Integrals over the full domain  $\Omega$  are split up into integrals over each triangle and then a summation

$$\iint_{\Omega} \dots dA = \sum_k \iint_{T_k} \dots dA.$$

The gradients of  $u$  and  $\phi$  are replaced by the gradients of the piecewise polynomials. Each contribution has to be written in the form

$$\iint_{T_k} \dots dA = \langle \mathbf{A}_k \vec{u}_k, \vec{\phi}_k \rangle + \langle \mathbf{W}_k \vec{f}_k, \vec{\phi}_k \rangle,$$

where  $\mathbf{A}_k$  is the **element stiffness matrix**.

The above integral will be rewritten as sum of the above integrations of the triangles, leading to the condition

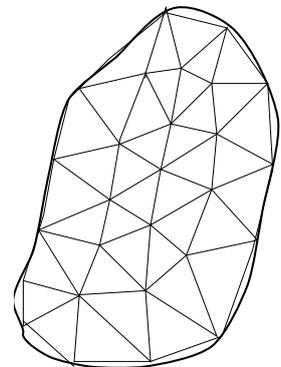
$$\langle \mathbf{A} \vec{u} + \mathbf{W} \vec{f}, \vec{\phi} \rangle = 0 \quad \text{for all } \vec{\phi} \in \mathbb{R}^N.$$

This condition is satisfied if  $\vec{u}$  solves the linear system  $\mathbf{A} \vec{u} = -\mathbf{W} \vec{f}$ . The matrix  $\mathbf{A}$  is called **global stiffness matrix**. It is this system of linear equations that will be solved to obtain an approximate solution of the boundary value problem (1).

## 6.2 A few triangular elements

There are different methods to construct finite elements on triangles. In Figure 64 find a graphical representation of a few commonly used elements.

- A solid dot at a position indicates that the value at this point is used as a DOF.
- A circle around a solid dot at a position indicates that the values of the first order partial derivatives are used as DOFs, e.g. in the Hermite elements.
- A double circle around a solid dot at a position indicates that the values, first and second order partial derivatives are used as DOFs, e.g. in the Argyris elements.
- A short line at a position indicates that the value of the normal derivative at this point is used as a DOF, e.g. in the Morley and Argyris elements.



An element is called

- $C^0$  conforming if the resulting solutions are continuous across element boundaries.
- $C^1$  conforming if the resulting solutions and the first order derivatives are continuous across element boundaries.

The codes in FEMoctave only use linear, quadratic and cubic elements.

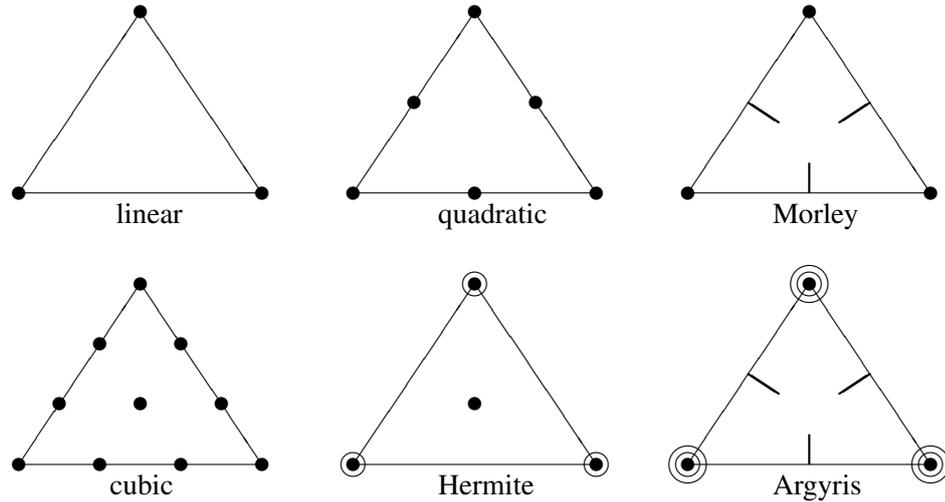


Figure 64: A few triangular elements

	linear	quadratic	Morley	cubic	Hermite	Argyris
degrees of freedom DOF	3	6	6	10	10	21
polynomial basis	$\mathbb{P}_1$	$\mathbb{P}_2$	$\mathbb{P}_2$	$\mathbb{P}_3$	$\mathbb{P}_3$	$\mathbb{P}_5$
$C^0$ conforming	yes	yes	no	yes	yes	yes
$C^1$ conforming	no	no	quasi	no	quasi	yes

Table 12: Properties of triangular elements

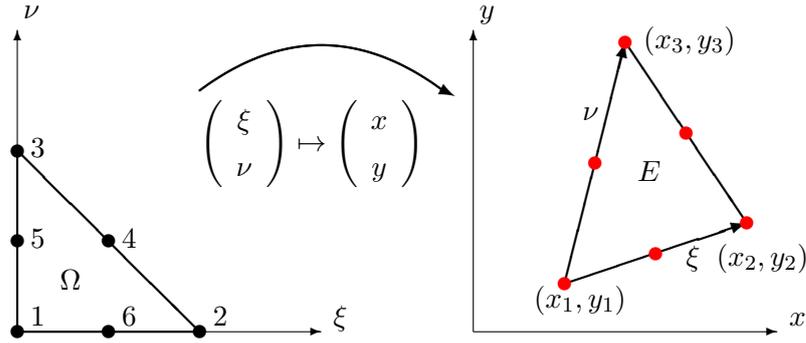
### 6.3 Transformation, interpolation and Gauss integration

From the above it is obvious that integration over general triangles is important for the development of FEM algorithms. It turns out to be convenient to find integration methods for a standard triangle and then consider the general triangle by appropriate coordinate transformations.

#### 6.3.1 Transformation of coordinates and integration over a general triangle

All of the necessary integrals for the FEM method are integrals over general triangles  $E$ . These can be written as images of a standard triangle in a  $(\xi, \nu)$ -plane, according to Figure 65. The transformation is given by

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + \xi \begin{pmatrix} x_2 - x_1 \\ y_2 - y_1 \end{pmatrix} + \nu \begin{pmatrix} x_3 - x_1 \\ y_3 - y_1 \end{pmatrix}$$

Figure 65: Transformation of the standard triangle  $\Omega$  to a general triangle  $E$ 

$$= \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + \begin{bmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{bmatrix} \cdot \begin{pmatrix} \xi \\ \nu \end{pmatrix} = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + \mathbf{T} \cdot \begin{pmatrix} \xi \\ \nu \end{pmatrix}$$

with the transformation matrix

$$\mathbf{T} = \begin{bmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{bmatrix}.$$

By using  $0 < \xi, \nu < 1$  with  $\xi + \nu < 1$  the standard triangle  $\Omega$  is mapped onto the general triangle  $E \subset \mathbb{R}^2$ . If the coordinates  $(x, y)$  are given find the values of  $(\xi, \nu)$  with the help of

$$\begin{pmatrix} \xi \\ \nu \end{pmatrix} = \mathbf{T}^{-1} \cdot \begin{pmatrix} x - x_1 \\ y - y_1 \end{pmatrix} = \frac{1}{\det(\mathbf{T})} \begin{bmatrix} y_3 - y_1 & -x_3 + x_1 \\ -y_2 + y_1 & x_2 - x_1 \end{bmatrix} \cdot \begin{pmatrix} x - x_1 \\ y - y_1 \end{pmatrix}.$$

If a function  $f(x, y)$  is to be integrated over the triangle  $E$  use the transformation

$$\iint_E f \, dA = \iint_{\Omega} f(\vec{x}(\xi, \nu)) \left| \det \left( \frac{\partial(x, y)}{\partial(\xi, \nu)} \right) \right| d\xi d\nu = |\det(\mathbf{T})| \int_0^1 \left( \int_0^{\nu} f(\vec{x}(\xi, \nu)) d\xi \right) d\nu. \quad (28)$$

The Jaccobi determinant is given by

$$\left| \det \left( \frac{\partial(x, y)}{\partial(\xi, \nu)} \right) \right| = |\det(\mathbf{T})| = |(x_2 - x_1)(y_3 - y_1) - (x_3 - x_1)(y_2 - y_1)|$$

If the orientation of the triangle is positive, then  $\det(\mathbf{T})$  will be positive. Since the area of the standard triangle  $\Omega$  equals  $\frac{1}{2}$  find

$$\text{area of } E = \frac{1}{2} |\det \mathbf{T}|.$$

For an efficient numerical integration over the standard triangle  $\Omega$  choose integration points  $\vec{g}_j \in \Omega$  and corresponding weights  $w_j$  for  $j = 1, 2, \dots, m$  and then work with the values of the function at those points, i.e.

$$\iint_{\Omega} f(\vec{\xi}) \, dA \approx \sum_{j=1}^m w_j f(\vec{g}_j). \quad (29)$$

The integration points and weights have to be chosen, such that the approximation error is as small as possible. Required are two essential conditions for the integration method:

- If a sample point is used in a Gauss integration, then all other points obtainable by permuting the three corners of the triangle must appear and with identical weight.
- All sample points must be inside the triangle (or on the triangle boundary)
- All weights  $w_i$  must be positive.

### 6.3.2 Gauss integration on the standard triangle with 3 Gauss points

In Figure 66 consider the three points at  $\vec{g}_1 = \frac{1}{2}(\lambda, \lambda)$ ,  $\vec{g}_2 = (1 - \lambda, \lambda/2)$  and  $\vec{g}_3 = (\lambda/2, 1 - \lambda)$ . Find optimal values for the parameters  $\lambda$  and  $w$  such that polynomials of degree as high as possible are integrated exactly by

$$\iint_{\Delta} f \, dA \approx w (f(\vec{g}_1) + f(\vec{g}_2) + f(\vec{g}_3)) .$$

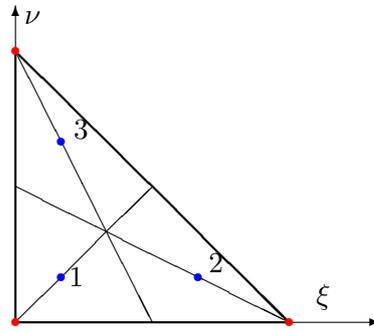


Figure 66: Gauss integration of order 2 on the standard triangle, using 3 integration points

To determine the optimal values determine a solution of a nonlinear system of 2 equations for the unknowns  $\lambda$  and  $w$ . Require that  $\xi^k$  for  $0 \leq k \leq 2$  be integrated exactly. This leads to the solution  $\lambda = 1/3$  and the weight  $w = 1/6$ . This approximate integration yields the exact results for polynomials  $f$  up to degree 2. Thus for a single triangle with diameter  $h$ , i.e. an area of the order  $h^2$ , the integration error for smooth functions is of the order  $h^3 \cdot h^2 = h^5$ . When dividing a large domain in sub-triangles of size  $h$  this leads to a total integration error of the order  $h^3$ .

The Gauss points and weights are given by

$$\mathbf{G} = \begin{bmatrix} 1/6 & 1/6 \\ 2/3 & 1/6 \\ 1/2 & 2/3 \end{bmatrix} \quad \text{and} \quad w = \frac{1}{6} .$$

For a general triangle the Gauss points are located at

$$\mathbf{X}_G = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + \begin{bmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{bmatrix} \cdot \mathbf{G}^T = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + \mathbf{T} \cdot \mathbf{G}^T .$$

This integration scheme will be used for linear elements.<sup>13</sup>

<sup>13</sup>One might be tempted to add the center of the triangle as a fourth point, but the resulting weight will be negative. This would lead to stiffness matrices that are not positive definite.

### 6.3.3 Gauss integration on the standard triangle with 7 Gauss points

As a second method use the points  $g_1 = (\lambda_1, \lambda_1)$  and  $g_4 = (\lambda_2, \lambda_2)$  along the diagonal  $\xi = \nu$ . Similarly use two more points along each connecting straight line from a corner of the triangle to the midpoint of the opposite edge. This leads to a total of 6 integration points where groups of 3 have the same weight. Finally add the midpoint with weight  $w_3$ . This is illustrated in Figure 67. The result is a  $7 \times 2$  matrix  $\mathbf{G}$  containing in each row the coordinates of one integration point  $\vec{g}_j$  and a vector  $\vec{w}$  with the corresponding integration weights. To determine the optimal

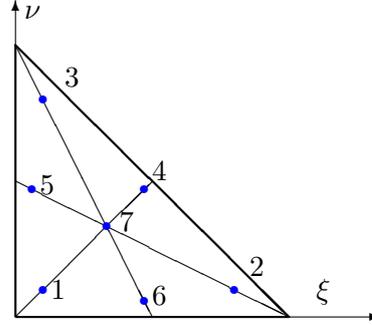


Figure 67: Gauss integration of order 5 on the standard triangle, using 7 integration points

values solve a nonlinear system of 5 equations for the unknowns  $\lambda_1$ ,  $\lambda_2$ ,  $w_1$ ,  $w_2$  and  $w_3$ . Require that  $\xi^k$  for  $0 \leq k \leq 5$  be integrated exactly. Find details in [Stah08]. Pick a solution of the resulting nonlinear system with  $0 < \lambda_1 < \lambda_2 < 1$  (points inside the triangle) and positive weights  $w_1$ ,  $w_2$  and  $w_3$ .

This approximate integration yields the exact results for polynomials  $f$  up to degree 5. Thus for one triangle with diameter  $h$  and an area of the order  $h^2$  the integration error for smooth functions is of the order  $h^6 \cdot h^2 = h^8$ . When dividing a large domain in sub-triangles of size  $h$  this leads to a total integration error of the order  $h^6$ . For most problems this error will be considerably smaller than the approximation error of the FEM method and it is reasonably safe to ignore the error.

The optimal choice of Gauss points and integration weights is given by<sup>14</sup>

$$\mathbf{G} = \begin{bmatrix} \lambda_1/2 & \lambda_1/2 \\ 1 - \lambda_1 & \lambda_1/2 \\ \lambda_1/2 & 1 - \lambda_1 \\ \lambda_2/2 & \lambda_2/2 \\ 1 - \lambda_2 & \lambda_2/2 \\ \lambda_2/2 & 1 - \lambda_2 \\ 1/3 & 1/3 \end{bmatrix} \approx \begin{bmatrix} 0.101287 & 0.101287 \\ 0.797427 & 0.101287 \\ 0.101287 & 0.797427 \\ 0.470142 & 0.470142 \\ 0.059716 & 0.470142 \\ 0.470142 & 0.059716 \\ 0.333333 & 0.333333 \end{bmatrix} \quad \text{and} \quad \vec{w} = \begin{pmatrix} w_1 \\ w_1 \\ w_1 \\ w_2 \\ w_2 \\ w_2 \\ w_3 \end{pmatrix} \approx \begin{pmatrix} 0.0629696 \\ 0.0629696 \\ 0.0629696 \\ 0.0661971 \\ 0.0661971 \\ 0.0661971 \\ 0.1125000 \end{pmatrix}. \quad (30)$$

Using the transformation results in this section compute the coordinates  $\mathbf{X}_G$  for the Gauss integration in a general triangle by

$$\mathbf{X}_G = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + \begin{bmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{bmatrix} \cdot \mathbf{G}^T = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + \mathbf{T} \cdot \mathbf{G}^T. \quad (31)$$

This notation is used to compute the Gauss points for a given triangulation of the domain, i.e. for the mesh.

<sup>14</sup>The exact values are  $\lambda_1 = (12 - 2\sqrt{15})/21$ ,  $\lambda_2 = (12 + 2\sqrt{15})/21$ ,  $w_1 = (155 - \sqrt{15})/2400$ ,  $w_4 = (155 + \sqrt{15})/2400$  and  $w_7 = 9/80$ .

### 6.4 Construction of first order elements

Assume that the function  $u$  is linear on each triangle  $T_k$ , thus determined by the values at the three corners. Then all integrals in expression (27) have to be examined. For the linear elements use the integration with 3 Gauss nodes in the triangle, as described in Section 6.3.2. All contributions in (27)

$$0 = \iint_{\Omega} \nabla \phi \cdot (a \nabla u - u \vec{b}) + \phi (b_0 u - f) dA - \int_{\Gamma_2} \phi (g_2 + g_3 u) ds$$

have to be transformed into

$$0 = \langle \vec{\phi}, \mathbf{A}\vec{u} + \mathbf{W}\vec{f} \rangle. \tag{32}$$

By integration over one triangle  $E$  find

$$\iint_E \nabla \phi \cdot (a \nabla u - u \vec{b}) + \phi (b_0 u - f) dA \approx \left\langle \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \end{pmatrix}, \mathbf{A}_E \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \right\rangle + \left\langle \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \end{pmatrix}, \mathbf{W}_E \vec{f}_E \right\rangle.$$

The matrix  $\mathbf{A}_E$  is the **element stiffness matrix** and  $\mathbf{W}_E \vec{f}_E$  the corresponding vector. These entries have to be added in the correct rows and columns of the global stiffness matrix. For this examine the local and global numbering of nodes in Figure 68. In each triangle the three corners are numbered by 1,2 and 3, but in the global mesh (consisting of many triangles) they are numbered by  $i,k$  and  $j$ . Thus the entries in the element stiffness matrix  $\mathbf{A}_E$  have to be added to rows/columns  $i,k$  and  $j$  in the global stiffness matrix  $\mathbf{A}$ .

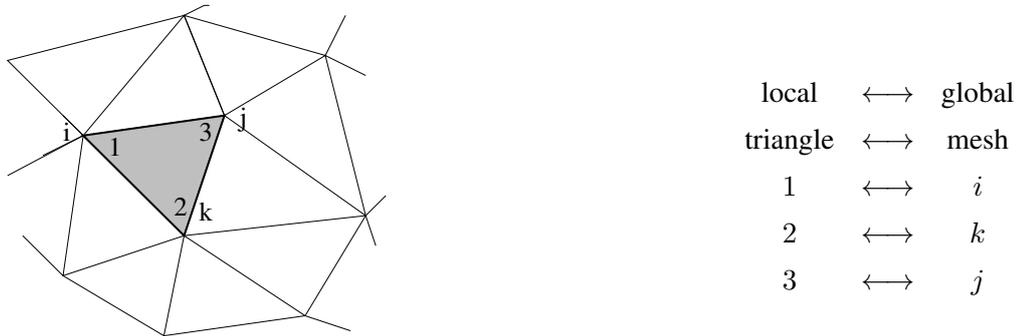


Figure 68: Local and global numbering of nodes

$$\mathbf{A}_E = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \longrightarrow \mathbf{A} = \mathbf{A} + \begin{matrix} \text{row } i \\ \text{row } j \\ \text{row } k \end{matrix} \begin{bmatrix} \text{col } i & \text{col } j & \text{col } k \\ \cdots & a_{11} & \cdots & a_{13} & \cdots & a_{12} & \cdots \\ \cdots & a_{31} & \cdots & a_{33} & \cdots & a_{32} & \cdots \\ \cdots & a_{21} & \cdots & a_{23} & \cdots & a_{22} & \cdots \end{bmatrix}$$

Similar procedures have to be applied to the vectors.

### 6.4.1 Linear interpolation on a triangle

If the values of the function  $\phi(x, y)$  at the three corners are given by  $\phi_1, \phi_2$  and  $\phi_3$  then the values  $\phi(\vec{g}_i)$  are given by

$$\begin{aligned}\phi(\vec{g}_1) &= \frac{2}{3}\phi_1 + \frac{1}{6}\phi_2 + \frac{1}{6}\phi_3 \\ \phi(\vec{g}_2) &= \frac{1}{6}\phi_1 + \frac{2}{3}\phi_2 + \frac{1}{6}\phi_3 \\ \phi(\vec{g}_3) &= \frac{1}{6}\phi_1 + \frac{1}{6}\phi_2 + \frac{2}{3}\phi_3\end{aligned}$$

or using a matrix notation

$$\begin{pmatrix} \phi(\vec{g}_1) \\ \phi(\vec{g}_2) \\ \phi(\vec{g}_3) \end{pmatrix} = \frac{1}{6} \begin{bmatrix} 4 & 1 & 1 \\ 1 & 4 & 1 \\ 1 & 1 & 4 \end{bmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \end{pmatrix} = \mathbf{M} \vec{\phi}.$$

This interpolation of the values from the nodes of the triangle to the Gauss points  $\vec{g}_i$  is independent of shape and size of the triangle.

For second order elements the construction of this interpolation matrix is performed using the basis functions (see Section 6.5.1). For the linear case use the simpler basis functions

$$\vec{\Phi}(\xi, \nu) = \begin{pmatrix} \Phi_1(\xi, \nu) \\ \Phi_2(\xi, \nu) \\ \Phi_3(\xi, \nu) \end{pmatrix} = \begin{pmatrix} 1 - \xi - \nu \\ \xi \\ \nu \end{pmatrix}$$

and a linear interpolation of a function given at the nodes is given by

$$f(\xi, \nu) = \sum_{i=1}^3 f_i \Phi_i(\xi, \nu).$$

Since

$$\frac{\partial}{\partial \xi} \vec{\Phi}(\xi, \nu) = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} \quad \text{and} \quad \frac{\partial}{\partial \nu} \vec{\Phi}(\xi, \nu) = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}$$

observe that the gradient does not depend on the position within the triangle.

### 6.4.2 Integration of $f \phi$

Examine different methods to give the function  $f$ : either by providing the values at the Gauss points, or by using the values at the nodes.

- If the values of the function  $f$  at the Gauss points  $\vec{g}_i$  are denoted by  $f_i$  then this integral is approximated by

$$\begin{aligned}\iint_E f \phi \, dA &\approx w \, 2 \, \text{area}(E) (f_1 \phi(\vec{g}_1) + f_2 \phi(\vec{g}_2) + f_3 \phi(\vec{g}_3)) \\ &= \frac{2 \, \text{area}(E)}{6} \langle \mathbf{M} \vec{\phi}, \vec{f} \rangle = \frac{\text{area}(E)}{3} \langle \vec{\phi}, \mathbf{M}^T \vec{f} \rangle.\end{aligned}$$

Thus find one contribution to (32).

- If the values of the function  $f$  at the nodes are denoted by  $f_i$  then first determine the values at the Gauss points by a linear interpolation. Then integrate as above, leading to the approximation

$$\iint_E f \phi \, dA \approx \frac{2 \text{area}(E)}{6} \langle \mathbf{M} \vec{\phi}, \mathbf{M} \vec{f} \rangle = \frac{\text{area}(E)}{3} \langle \vec{\phi}, \mathbf{M}^T \mathbf{M} \vec{f} \rangle.$$

The matrix

$$\mathbf{M}^T \mathbf{M} = \frac{1}{36} \begin{bmatrix} 18 & 9 & 9 \\ 9 & 18 & 9 \\ 9 & 9 & 18 \end{bmatrix} = \frac{1}{4} \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$$

is independent on the shape and size of the element (triangle). Thus find one contribution to (32).

#### 6.4.3 Integration of $b_0 u \phi$

Since the values of the functions  $u$  and  $\phi$  are known at the nodes interpolate both functions and then use the values of the function  $b_0(x, y)$  at the Gauss nodes to find

$$\begin{aligned} \iint_E b_0 u \phi \, dA &\approx w \, 2 \text{area}(E) \sum_{i=1}^3 b_0(\vec{g}_i) u(\vec{g}_i) \phi(\vec{g}_i) \\ &= \frac{2 \text{area}(E)}{6} \langle \mathbf{M} \vec{\phi}, \text{diag}(\vec{b}) \mathbf{M} \vec{u} \rangle = \frac{\text{area}(E)}{3} \langle \vec{\phi}, \mathbf{M}^T \text{diag}(\vec{b}_0) \mathbf{M} \vec{u} \rangle, \end{aligned}$$

where

$$\text{diag} \vec{b}_0 = \begin{bmatrix} b_0(\vec{g}_1) & 0 & 0 \\ 0 & b_0(\vec{g}_2) & 0 \\ 0 & 0 & b_0(\vec{g}_3) \end{bmatrix}.$$

If  $b_0(x, y)$  happens to be a constant, then the above may be simplified to

$$\iint_E b_0 u \phi \, dA \approx b_0 \frac{\text{area}(E)}{12} \langle \vec{\phi}, \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix} \vec{u} \rangle.$$

Thus find another contribution to (32).

#### 6.4.4 Integration of $a \nabla u \cdot \nabla \phi$

Since the functions  $u$  and  $\phi$  are linear on each triangle, we use the fact that the gradients are constant on each triangle. The gradient may be determined with the help of a normal vector of the plane passing through the three points

$$\begin{pmatrix} x_1 \\ y_1 \\ u_1 \end{pmatrix}, \quad \begin{pmatrix} x_2 \\ y_2 \\ u_2 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} x_3 \\ y_3 \\ u_3 \end{pmatrix}.$$

A normal vector  $\vec{n}$  is given by the vector product

$$\vec{n} = \begin{pmatrix} x_2 - x_1 \\ y_2 - y_1 \\ u_2 - u_1 \end{pmatrix} \times \begin{pmatrix} x_3 - x_1 \\ y_3 - y_1 \\ u_3 - u_1 \end{pmatrix} = \begin{pmatrix} +(y_2 - y_1) \cdot (u_3 - u_1) - (u_2 - u_1) \cdot (y_3 - y_1) \\ -(x_2 - x_1) \cdot (u_3 - u_1) + (u_2 - u_1) \cdot (x_3 - x_1) \\ +(x_2 - x_1) \cdot (y_3 - y_1) - (y_2 - y_1) \cdot (x_3 - x_1) \end{pmatrix}.$$

The third component of this vector equals twice the oriented<sup>15</sup> area of the triangle. To obtain the gradient in the first two components the vector has to be normalized, such that the third component equals  $-1$ . Find

$$\nabla u = \begin{pmatrix} \frac{du}{dx} \\ \frac{du}{dy} \end{pmatrix} = \frac{-1}{2 \text{area}(E)} \begin{pmatrix} +(y_2 - y_1) \cdot (u_3 - u_1) - (u_2 - u_1) \cdot (y_3 - y_1) \\ -(x_2 - x_1) \cdot (u_3 - u_1) + (u_2 - u_1) \cdot (x_3 - x_1) \end{pmatrix}.$$

This formula can be written in the form

$$\nabla u = \frac{-1}{2 \text{area}(E)} \begin{bmatrix} (y_3 - y_2) & (y_1 - y_3) & (y_2 - y_1) \\ (x_2 - x_3) & (x_3 - x_1) & (x_1 - x_2) \end{bmatrix} \cdot \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \frac{-1}{2 \text{area}(E)} \mathbf{G} \cdot \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}. \quad (33)$$

and thus

$$\langle \nabla \phi, \nabla u \rangle = \frac{1}{4 \text{area}(E)^2} \left\langle \mathbf{G} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \end{pmatrix}, \mathbf{G} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \right\rangle = \frac{1}{4 \text{area}(E)^2} \left\langle \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \end{pmatrix}, \mathbf{G}^T \cdot \mathbf{G} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \right\rangle.$$

If  $a_i$  are the values of the function  $a(x, y)$  at the Gauss points  $\vec{g}_i$  find

$$\iint_E a \nabla \phi \cdot \nabla u \, dA \approx \frac{a_1 + a_2 + a_3}{12 \text{area}(E)} \left\langle \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \end{pmatrix}, \mathbf{G}^T \cdot \mathbf{G} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \right\rangle.$$

As an exercise one can verify that the matrix  $\mathbf{G}^T \cdot \mathbf{G}$  is symmetric and positive semi-definite. The expression vanishes for constant vectors, i.e. for vanishing gradients.

#### 6.4.5 Integration of $u \vec{b} \cdot \nabla \phi$

Since the gradient of  $\phi$  is constant on each of the triangles use

$$\begin{pmatrix} \phi_x \\ \phi_y \end{pmatrix} = \nabla \phi = \frac{-1}{2 \text{area}(E)} \mathbf{G} \cdot \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \end{pmatrix} = \frac{-1}{2 \text{area}(E)} \begin{bmatrix} \mathbf{G}_x \\ \mathbf{G}_y \end{bmatrix} \cdot \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \end{pmatrix},$$

where

$$\mathbf{G}_x = \begin{bmatrix} y_3 - y_2 & y_1 - y_3 & y_2 - y_1 \end{bmatrix} \quad \text{and} \quad \mathbf{G}_y = \begin{bmatrix} x_2 - x_3 & x_3 - x_1 & x_1 - x_2 \end{bmatrix}.$$

Let  $b_{1,i}$  be the values of the first component of  $\vec{b}$  at the Gauss nodes and find

$$\begin{aligned} \iint_E u b_1 \phi_x \, dA &\approx \frac{\text{area}(E)}{3} \sum_{i=1}^3 u(\vec{g}_i) b_{1,i} \phi_{x,i} \\ &= \frac{-\text{area}(E)}{3 \cdot 2 \text{area}(E)} \left\langle \begin{bmatrix} \mathbf{G}_x \\ \mathbf{G}_x \\ \mathbf{G}_x \end{bmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \end{pmatrix}, \begin{bmatrix} b_{1,1} & 0 & 0 \\ 0 & b_{1,2} & 0 \\ 0 & 0 & b_{1,3} \end{bmatrix} \mathbf{M} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \right\rangle \end{aligned}$$

<sup>15</sup>It is quietly assumed that the third component of  $\vec{n}$  is positive. Since only the square of the gradient is used the influence of this ignorance will disappear. Generate meshes with triangles with a positive orientation also allow to assure  $n_3 > 0$ .

$$\begin{aligned}
&= \frac{-1}{6} \left\langle \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \end{pmatrix}, \left[ \mathbf{G}_x^T \quad \mathbf{G}_x^T \quad \mathbf{G}_x^T \right] \begin{bmatrix} b_{1,1} & 0 & 0 \\ 0 & b_{1,2} & 0 \\ 0 & 0 & b_{1,3} \end{bmatrix} \mathbf{M} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \right\rangle \\
&= \frac{-1}{6} \left\langle \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \end{pmatrix}, \mathbf{G}_x^T \begin{bmatrix} b_{1,1} & b_{1,2} & b_{1,3} \end{bmatrix} \mathbf{M} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \right\rangle \\
&= \frac{-1}{6} \left\langle \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \end{pmatrix}, \begin{bmatrix} b_{1,1}(y_3 - y_2) & b_{1,2}(y_3 - y_2) & b_{1,3}(y_3 - y_2) \\ b_{1,1}(y_1 - y_3) & b_{1,2}(y_1 - y_3) & b_{1,3}(y_1 - y_3) \\ b_{1,1}(y_2 - y_1) & b_{1,2}(y_2 - y_1) & b_{1,3}(y_2 - y_1) \end{bmatrix} \mathbf{M} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \right\rangle.
\end{aligned}$$

If the values of the second component of  $\vec{b}$  at the Gauss nodes are given by  $b_{2,i}$  find by similar computations

$$\begin{aligned}
\iint_E u b_2 \phi_y dA &\approx \frac{-\text{area}(E)}{3} \sum_{i=1}^3 u(\vec{g}_i) b_{2,i} \phi_{y,i} \\
&= \frac{-1}{6} \left\langle \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \end{pmatrix}, \mathbf{G}_y^T \begin{bmatrix} b_{2,1} & b_{2,2} & b_{2,3} \end{bmatrix} \mathbf{M} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \right\rangle \\
&= \frac{-1}{6} \left\langle \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \end{pmatrix}, \begin{bmatrix} b_{2,1}(x_2 - x_3) & b_{2,2}(x_2 - x_3) & b_{2,3}(x_2 - x_3) \\ b_{2,1}(x_3 - x_1) & b_{2,2}(x_3 - x_1) & b_{2,3}(x_3 - x_1) \\ b_{2,1}(x_1 - x_2) & b_{2,2}(x_1 - x_2) & b_{2,3}(x_1 - x_2) \end{bmatrix} \mathbf{M} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \right\rangle.
\end{aligned}$$

This leads to two more contributions to (32).

#### 6.4.6 Integration over boundary segments

In expression (27) compute integrals over the boundary

$$\int_{\Gamma_2} \phi (g_2 + g_3 u) ds.$$

For triangular domains the boundary consists of straight line segments. Replace the integral by a sum of line integrals and use a Gauss integration. Based on the two endpoints  $\vec{x}_1$  and  $\vec{x}_2$  use the values at the two Gauss integration points<sup>16</sup>

$$\begin{aligned}
\vec{p}_1 &= \frac{1}{2} (\vec{x}_1 + \vec{x}_2) - \frac{1}{2\sqrt{3}} (\vec{x}_2 - \vec{x}_1) \\
\vec{p}_2 &= \frac{1}{2} (\vec{x}_1 + \vec{x}_2) + \frac{1}{2\sqrt{3}} (\vec{x}_2 - \vec{x}_1).
\end{aligned}$$

<sup>16</sup>To derive the formula integrate 1,  $t$ ,  $t^2$  and  $t^3$  over the interval  $[-1, 1]$ .

$$\begin{aligned}
\int_{-1}^{+1} f(t) dt &= w_1 f(-\xi) + w_1 f(+\xi) \\
\int_{-1}^{+1} 1 dt = 2 &= w_1 1 + w_1 1 \implies w_1 = 1 \\
\int_{-1}^{+1} t dt = 0 &= -w_1 \xi + w_1 \xi = 0 \\
\int_{-1}^{+1} t^2 dt = \frac{2}{3} &= +w_1 \xi^2 + w_1 \xi^2 \implies \xi = \sqrt{1/3} \\
\int_{-1}^{+1} t^3 dt = 0 &= -w_1 \xi^3 + w_1 \xi^3 = 0
\end{aligned}$$

Thus  $t^4$  is not integrated exactly and the error is proportional to  $h^4$ .

Polynomials up to degree 3 are integrated exactly, thus the error is proportional to  $h^4$ . By linear interpolation between the points  $\vec{x}_1$  and  $\vec{x}_2$  find the values of the function  $u$  at the Gauss points to be

$$\begin{aligned} u(\vec{p}_1) &= (1 - \alpha) u_1 + \alpha u_2 \\ u(\vec{p}_2) &= \alpha u_1 + (1 - \alpha) u_2 \end{aligned}$$

or

$$\begin{pmatrix} u(\vec{p}_1) \\ u(\vec{p}_2) \end{pmatrix} = \begin{bmatrix} (1 - \alpha) & \alpha \\ \alpha & (1 - \alpha) \end{bmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix},$$

where  $\alpha = \frac{1-1/\sqrt{3}}{2} \approx 0.211325$ . Using the length  $L = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$  this leads to the approximations

$$\begin{aligned} \int \phi g_2 ds &\approx \frac{L}{2} \left\langle \begin{bmatrix} (1 - \alpha) & \alpha \\ \alpha & (1 - \alpha) \end{bmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}, \begin{pmatrix} g_2(\vec{p}_1) \\ g_2(\vec{p}_2) \end{pmatrix} \right\rangle \\ &= \frac{L}{2} \left\langle \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}, \begin{bmatrix} (1 - \alpha) & \alpha \\ \alpha & (1 - \alpha) \end{bmatrix} \begin{pmatrix} g_2(\vec{p}_1) \\ g_2(\vec{p}_2) \end{pmatrix} \right\rangle \\ \int \phi g_3 u ds &\approx \frac{L}{2} \left\langle \begin{bmatrix} (1 - \alpha) & \alpha \\ \alpha & (1 - \alpha) \end{bmatrix} \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}, \begin{bmatrix} g_3(\vec{p}_1) & 0 \\ 0 & g_3(\vec{p}_2) \end{bmatrix} \begin{bmatrix} (1 - \alpha) & \alpha \\ \alpha & (1 - \alpha) \end{bmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \right\rangle \\ &= \frac{L}{2} \left\langle \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}, \begin{bmatrix} (1 - \alpha) & \alpha \\ \alpha & (1 - \alpha) \end{bmatrix} \begin{bmatrix} (1 - \alpha) g_3(\vec{p}_1) & \alpha g_3(\vec{p}_1) \\ \alpha g_3(\vec{p}_2) & (1 - \alpha) g_3(\vec{p}_2) \end{bmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \right\rangle \\ &= \frac{L}{2} \left\langle \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}, \begin{bmatrix} (1 - \alpha)^2 g_3(\vec{p}_1) + \alpha^2 g_3(\vec{p}_2) & (1 - \alpha) \alpha (g_3(\vec{p}_1) + g_3(\vec{p}_2)) \\ (1 - \alpha) \alpha (g_3(\vec{p}_1) + g_3(\vec{p}_2)) & \alpha^2 g_3(\vec{p}_1) + (1 - \alpha)^2 g_3(\vec{p}_2) \end{bmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \right\rangle. \end{aligned}$$

The first expression will lead to a contribution to the RHS vector of the linear system to be solved, while the second expression will lead to entries in the matrix. These approximate integrations lead to the exact result if the function to be integrated is a polynomial of degree 3, or less. If  $h$  is the typical length of an edge then the error is of the order  $h^5$  for one line segment and thus of order  $h^4$  for the total boundary. This boundary integration is used for first order elements.

The second expression is of the form

$$\int \phi g_3 u ds \approx \langle \vec{\phi}, \mathbf{B} \vec{u} \rangle = \left\langle \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}, \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \right\rangle \quad (34)$$

and its effect on the linear system  $\mathbf{A} \vec{u} + \mathbf{W} \vec{f} = \vec{0}$  to be solved depends on nodes being on the Dirichlet part of the boundary.

- If  $u_1$  and  $u_2$  are both free, i.e. not on the Dirichlet section, then all entries of the matrix  $\mathbf{B}$  have to be added to the global stiffness matrix  $\mathbf{A}$ .
- If  $u_1$  and  $u_2$  are on the Dirichlet section, then nothing has to be added to  $\mathbf{A}$  and  $\vec{f}$ .
- If  $u_1$  is free and  $u_2$  is on the Dirichlet section, then only the first expression

$$b_{11} u_1 + b_{12} u_2 = b_{11} u_1 + b_{12} d_2$$

has to be added.  $d_2$  is the Dirichlet value at the position of  $u_2$ . Then  $b_{11}$  has to be taken into account in  $\mathbf{A}$  and  $b_{12} d_2$  has to be added to  $\mathbf{W} \vec{f}$ .

- If  $u_2$  is free and  $u_1$  is on the Dirichlet section, then only the second expression  $b_{21} u_1 + b_{22} u_2 = b_{21} d_1 + b_{22} u_2$  has to be added.  $d_1$  is the Dirichlet value at the position of  $u_1$ . Then  $b_{22}$  has to be taken into account in  $\mathbf{A}$  and  $b_{12} d_1$  has to be added to  $\mathbf{W} \vec{f}$ .

## 6.5 Construction of second order elements

In this section the construction of the element stiffness matrix and vector for triangular elements of order 2 is examined. The ideas are very similar to Section 6.4 for linear basis functions, but using a bit more mathematics is required. Again all contributions in (27)

$$0 = \iint_{\Omega} \nabla \phi \cdot (a \nabla u - u \vec{b}) + \phi (b_0 u - f) dA - \int_{\Gamma_2} \phi (g_2 + g_3 u) ds$$

have to be transformed into

$$0 = \langle \vec{\phi}, \mathbf{A} \vec{u} + \mathbf{W} \vec{f} \rangle.$$

For second order element a general quadratic function is used on each of the triangles in the mesh. There are 6 linearly independent polynomials of degree 2 or less, namely 1,  $x$ ,  $y$ ,  $x^2$ ,  $y^2$  and  $x \cdot y$ .

### 6.5.1 The basis functions for a second order element and quadratic interpolation

Examine the standard triangle  $\Omega$  in Figure 65 with the values of a function  $f(\xi, \nu)$  at the corners and at the midpoints of the edges. Use the numbering as shown in Figure 65. The parameters  $\xi$  and  $\nu$  at the nodes are given by Table 13. Construct polynomials  $\phi_i(\xi, \nu)$  of degree 2, such that

$$\Phi_i(\xi_j, \nu_j) = \delta_{i,j} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

i.e. each basis function is equal to 1 at one of the nodes and vanishes on all other nodes. These basis polynomials are given by

node $i$	1	2	3	4	5	6
$\xi_i$	0	1	0	$\frac{1}{2}$	0	$\frac{1}{2}$
$\nu_i$	0	0	1	$\frac{1}{2}$	$\frac{1}{2}$	0

Table 13: Coordinates of the nodes in the standard quadratic triangle

$$\vec{\Phi}(\xi, \nu) = \begin{pmatrix} \Phi_1(\xi, \nu) \\ \Phi_2(\xi, \nu) \\ \Phi_3(\xi, \nu) \\ \Phi_4(\xi, \nu) \\ \Phi_5(\xi, \nu) \\ \Phi_6(\xi, \nu) \end{pmatrix} = \begin{pmatrix} (1 - \xi - \nu)(1 - 2\xi - 2\nu) \\ \xi(2\xi - 1) \\ \nu(2\nu - 1) \\ 4\xi\nu \\ 4\nu(1 - \xi - \nu) \\ 4\xi(1 - \xi - \nu) \end{pmatrix} \quad (35)$$

and find their graphs in Figure 69.

Any quadratic polynomial  $f$  on the standard triangle  $\Omega$  can be written as linear combination of the basis functions by using

$$f(\xi, \nu) = \sum_{i=1}^6 f(\xi_i, \nu_i) \Phi_i(\xi, \nu) = \sum_{i=1}^6 f_i \Phi_i(\xi, \nu). \quad (36)$$

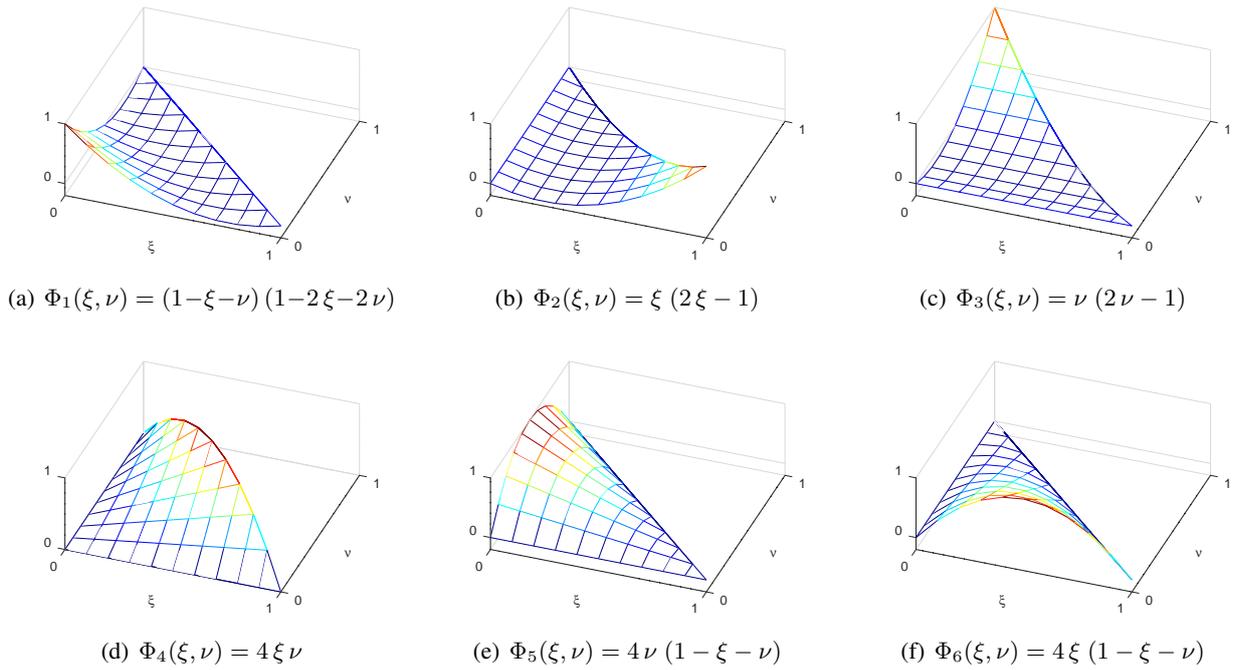


Figure 69: Basis functions for second order triangular elements

This is the formula to apply a quadratic interpolation on the triangle, using the values  $f_i$  of the function at the nodes. To use this interpolation for a given point  $(x, y)$  in the triangle  $E$  in Figure 65 determine the correct values of the parameters  $\xi$  and  $\nu$ , i.e. solve

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + \xi \begin{pmatrix} x_2 - x_1 \\ y_2 - y_1 \end{pmatrix} + \nu \begin{pmatrix} x_3 - x_1 \\ y_3 - y_1 \end{pmatrix}.$$

This is equivalent to the linear system

$$\mathbf{T} \begin{pmatrix} \xi \\ \nu \end{pmatrix} = \begin{bmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{bmatrix} \begin{pmatrix} \xi \\ \nu \end{pmatrix} = \begin{pmatrix} x - x_1 \\ y - y_1 \end{pmatrix}.$$

Since the  $2 \times 2$  matrix  $\mathbf{T}$  is invertible find

$$\begin{pmatrix} \xi \\ \nu \end{pmatrix} = \mathbf{T}^{-1} \cdot \begin{pmatrix} x - x_1 \\ y - y_1 \end{pmatrix} = \frac{1}{\det(\mathbf{T})} \begin{bmatrix} y_3 - y_1 & -x_3 + x_1 \\ -y_2 + y_1 & x_2 - x_1 \end{bmatrix} \cdot \begin{pmatrix} x - x_1 \\ y - y_1 \end{pmatrix}.$$

### 6.5.2 Determine values at the Gauss points and apply Gauss integration

Use equation (31) to determine the coordinates of the seven Gauss points. Then a function to be integrated can be evaluated at these Gauss points. Computing the values of the basis functions  $\Phi_i(\xi, \nu)$  at the Gauss points  $\vec{g}_j$  by  $m_{j,i} = \Phi_i(\vec{g}_j)$  and write

$$f(\vec{g}_j) = \sum_{i=1}^6 f_i \Phi_i(\vec{g}_j) = \sum_{i=1}^6 m_{j,i} f_i$$

or using a matrix notation  $\mathbf{M} \in \mathbb{R}^{7 \times 6}$

$$\begin{pmatrix} f(\vec{g}_1) \\ f(\vec{g}_2) \\ \vdots \\ f(\vec{g}_7) \end{pmatrix} = \begin{bmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,6} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,6} \\ \vdots & \vdots & \ddots & \vdots \\ m_{7,1} & m_{7,2} & \cdots & m_{7,6} \end{bmatrix} \cdot \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_6 \end{pmatrix} = \mathbf{M} \cdot \vec{f} \quad (37)$$

$$\approx \begin{bmatrix} +0.474353 & -0.080769 & -0.080769 & 0.041036 & 0.323074 & 0.323074 \\ -0.080769 & +0.474353 & -0.080769 & 0.323074 & 0.041036 & 0.323074 \\ -0.080769 & -0.080769 & +0.474353 & 0.323074 & 0.323074 & 0.041036 \\ -0.052584 & -0.028075 & -0.028075 & 0.884134 & 0.112300 & 0.112300 \\ -0.028075 & -0.052584 & -0.028075 & 0.112300 & 0.884134 & 0.112300 \\ -0.028075 & -0.028075 & -0.052584 & 0.112300 & 0.112300 & 0.884134 \\ -0.111111 & -0.111111 & -0.111111 & 0.444444 & 0.444444 & 0.444444 \end{bmatrix} \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \\ f_6 \end{pmatrix}$$

The Gauss integration can be written in the form

$$\iint_{\Omega} f(\xi, \nu) dA \approx \sum_{j=1}^7 w_j f(\vec{g}_j) = \langle \vec{w}, \mathbf{M} \cdot \vec{f} \rangle.$$

To integrate over the general triangle  $E$  use the transformation (28), i.e.

$$\iint_E f dA = \iint_{\Omega} f(\vec{x}(\xi, \nu)) \left| \det \left( \frac{\partial(x, y)}{\partial(\xi, \nu)} \right) \right| d\xi d\nu \approx |\det \mathbf{T}| \langle \vec{w}, \mathbf{M} \cdot \vec{f} \rangle.$$

Now all the tools to approximate the integrals required for the element stiffness matrix are available.

### 6.5.3 Integration of $f \phi$

The test function  $\phi$  is given by its values  $\vec{\phi}$  at the nodes, i.e. the corners of the triangle and the midpoints of the sides. Examine different methods to give the function  $f$ : either by providing the values at the Gauss points, or by using the values at the nodes.

- If the values of the function  $f$  at the Gauss points  $\vec{g}_i$  are denoted by  $f_i$  then this integral is approximated by

$$\begin{aligned} \iint_E f \phi dA &\approx |\det(\mathbf{T})| \sum_{j=1}^7 w_j f_j \phi(\vec{g}_j) = |\det(\mathbf{T})| \langle \text{diag}(\vec{w}) \vec{f}, \mathbf{M} \vec{\phi} \rangle \\ &= |\det(\mathbf{T})| \langle \mathbf{M}^T \text{diag}(\vec{w}) \vec{f}, \vec{\phi} \rangle, \end{aligned}$$

Thus find one contribution to (32).

- If the values of the function  $f$  at the nodes are denoted by  $f_i$  then first determine the values at the Gauss points by a quadratic interpolation. Then integrate as above, leading to the approximation

$$\iint_E f \phi dA \approx |\det(\mathbf{T})| \langle \text{diag}(\vec{w}) \mathbf{M} \vec{f}, \mathbf{M} \vec{\phi} \rangle = |\det(\mathbf{T})| \langle \mathbf{M}^T \text{diag}(\vec{w}) \mathbf{M} \vec{f}, \vec{\phi} \rangle.$$

The matrices  $\mathbf{M}^T \text{diag}(\vec{w})$  and  $\mathbf{M}^T \text{diag}(\vec{w}) \mathbf{M}$  are independent on the triangle  $E$ .

### 6.5.4 Integration of $b_0 u \phi$

Since the values of the functions  $u$  and  $\phi$  are known at the nodes use an interpolation and then the function  $b_0(x, y)$  at the Gauss nodes to find

$$\begin{aligned} \iint_E b_0 u \phi dA &\approx |\det(\mathbf{T})| \sum_{j=1}^7 w_j b_0(g_j) u(g_j) \phi(g_j) = |\det(\mathbf{T})| \langle \text{diag}(\vec{w}) \text{diag}(\vec{b}_0) \mathbf{M} \vec{u}, \mathbf{M} \vec{\phi} \rangle \\ &= |\det(\mathbf{T})| \langle \mathbf{M}^T \text{diag}(\vec{w}) \text{diag}(\vec{b}_0) \mathbf{M} \vec{u}, \vec{\phi} \rangle, \end{aligned}$$

where  $\text{diag}(\vec{b}_0) = \text{diag}(b_0(\vec{g}_1), b_0(\vec{g}_2), b_0(\vec{g}_3), \dots, b_0(\vec{g}_7))$ .

### 6.5.5 Transformation of the gradient to the standard triangle

To examine the contributions containing  $\nabla u$  or  $\nabla \phi$  requires considerably more tools than the ones used in Section 6.4.4 for linear elements. For linear elements the gradients are constant on each of the triangles. For quadratic elements the gradients are linear functions and thus not constant. First examine how the gradient behave under the transformation to the standard triangle, only then use the above integration methods.

According to Section 6.3.1 the coordinates  $(\xi, \nu)$  of the standard triangle are connected to the global coordinates  $(x, y)$  by

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + \begin{bmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{bmatrix} \cdot \begin{pmatrix} \xi \\ \nu \end{pmatrix} = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + \mathbf{T} \cdot \begin{pmatrix} \xi \\ \nu \end{pmatrix}$$

or equivalently

$$\begin{pmatrix} \xi \\ \nu \end{pmatrix} = \mathbf{T}^{-1} \cdot \begin{pmatrix} x - x_1 \\ y - y_1 \end{pmatrix} = \frac{1}{\det(\mathbf{T})} \begin{bmatrix} y_3 - y_1 & -x_3 + x_1 \\ -y_2 + y_1 & x_2 - x_1 \end{bmatrix} \cdot \begin{pmatrix} x - x_1 \\ y - y_1 \end{pmatrix}.$$

If a function  $f(x, y)$  is given on the general triangle  $E$  can pull it back to the standard triangle by

$$g(\xi, \nu) = f(x(\xi, \nu), y(\xi, \nu))$$

and then compute the gradient of  $g(\xi, \nu)$  with respect to its independent variables  $\xi$  and  $\nu$ . The result will depend on the partial derivatives of  $f$  with respect to  $x$  and  $y$ . The standard chain rule implies

$$\begin{aligned} \frac{\partial}{\partial \xi} g(\xi, \nu) &= \frac{\partial}{\partial \xi} f(x(\xi, \nu), y(\xi, \nu)) = \frac{\partial f(x, y)}{\partial x} \frac{\partial x}{\partial \xi} + \frac{\partial f(x, y)}{\partial y} \frac{\partial y}{\partial \xi} \\ &= \frac{\partial f(x, y)}{\partial x} (x_2 - x_1) + \frac{\partial f(x, y)}{\partial y} (y_2 - y_1) \\ \frac{\partial}{\partial \nu} g(\xi, \nu) &= \frac{\partial}{\partial \nu} f(x(\xi, \nu), y(\xi, \nu)) = \frac{\partial f(x, y)}{\partial x} \frac{\partial x}{\partial \nu} + \frac{\partial f(x, y)}{\partial y} \frac{\partial y}{\partial \nu} \\ &= \frac{\partial f(x, y)}{\partial x} (x_3 - x_1) + \frac{\partial f(x, y)}{\partial y} (y_3 - y_1). \end{aligned}$$

This can be written with the help of matrices in the form

$$\begin{pmatrix} \frac{\partial g}{\partial \xi} \\ \frac{\partial g}{\partial \nu} \end{pmatrix} = \begin{bmatrix} (x_2 - x_1) & (y_2 - y_1) \\ (x_3 - x_1) & (y_3 - y_1) \end{bmatrix} \cdot \begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{pmatrix} = \mathbf{T}^T \cdot \begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{pmatrix}$$

or equivalently

$$\left( \frac{\partial g}{\partial \xi}, \frac{\partial g}{\partial \nu} \right) = \left( \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right) \cdot \mathbf{T}. \quad (38)$$

This implies

$$\left( \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right) = \left( \frac{\partial g}{\partial \xi}, \frac{\partial g}{\partial \nu} \right) \cdot \mathbf{T}^{-1} = \frac{1}{\det \mathbf{T}} \left( \frac{\partial g}{\partial \xi}, \frac{\partial g}{\partial \nu} \right) \cdot \begin{bmatrix} y_3 - y_1 & -x_3 + x_1 \\ -y_2 + y_1 & x_2 - x_1 \end{bmatrix}$$

or by transposition

$$\begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{pmatrix} = \frac{1}{\det \mathbf{T}} \begin{bmatrix} y_3 - y_1 & -y_2 + y_1 \\ -x_3 + x_1 & x_2 - x_1 \end{bmatrix} \begin{pmatrix} \frac{\partial g}{\partial \xi} \\ \frac{\partial g}{\partial \nu} \end{pmatrix}. \quad (39)$$

Let  $g$  be a function on the standard triangle  $\Omega$  given as a linear combination of the basis functions, i.e.

$$g(\xi, \nu) = \sum_{i=1}^6 g_i \Phi_i(\xi, \nu)$$

where the basis function  $\Phi_i(\xi, \nu)$  are given by (35). Then its gradient with respect to  $\xi$  and  $\nu$  can be determined with the help of elementary partial derivatives applied to the expressions in (35). The result is

$$\text{grad } \vec{\Phi} = \begin{bmatrix} -3 + 4\xi + 4\nu & -3 + 4\xi + 4\nu \\ 4\xi - 1 & 0 \\ 0 & 4\nu - 1 \\ 4\nu & 4\xi \\ -4\nu & 4 - 4\xi - 8\nu \\ 4 - 8\xi - 4\nu & -4\xi \end{bmatrix} = \begin{bmatrix} \vec{\Phi}_\xi(\xi, \nu) & \vec{\Phi}_\nu(\xi, \nu) \end{bmatrix}. \quad (40)$$

Thus find on the standard triangle  $\Omega$

$$\left( \frac{\partial g}{\partial \xi}, \frac{\partial g}{\partial \nu} \right) = (g_1, g_2, g_3, g_4, g_5, g_6) \cdot \begin{bmatrix} \vec{\Phi}_\xi(\xi, \nu) & \vec{\Phi}_\nu(\xi, \nu) \end{bmatrix} = \vec{g}^T \cdot \begin{bmatrix} \vec{\Phi}_\xi(\xi, \nu) & \vec{\Phi}_\nu(\xi, \nu) \end{bmatrix}.$$

If the function  $\varphi(x, y)$  is given on the general triangle  $E$  as linear combination of the basis functions on  $E$  find

$$\varphi(x, y) = \sum_{i=1}^6 \varphi_i \Phi_i(\xi(x, y), \nu(x, y)).$$

Now combine the results in this section to conclude

$$\left( \frac{\partial \varphi}{\partial x}, \frac{\partial \varphi}{\partial y} \right) = \left( \frac{\partial \varphi}{\partial \xi}, \frac{\partial \varphi}{\partial \nu} \right) \cdot \mathbf{T}^{-1} = \vec{\varphi}^T \cdot \begin{bmatrix} \vec{\Phi}_\xi & \vec{\Phi}_\nu \end{bmatrix} \cdot \mathbf{T}^{-1}$$

or by transposition

$$\begin{pmatrix} \frac{\partial \varphi}{\partial x} \\ \frac{\partial \varphi}{\partial y} \end{pmatrix} = (\mathbf{T}^{-1})^T \cdot \begin{bmatrix} \vec{\Phi}_\xi^T \\ \vec{\Phi}_\nu^T \end{bmatrix} \cdot \vec{\varphi} = \frac{1}{\det(\mathbf{T})} \begin{bmatrix} +y_3 - y_1 & -y_2 + y_1 \\ -x_3 + x_1 & +x_2 - x_1 \end{bmatrix} \cdot \begin{bmatrix} \vec{\Phi}_\xi^T \\ \vec{\Phi}_\nu^T \end{bmatrix} \cdot \vec{\varphi}$$

and the same identities can be spelled out for the two components independently.

$$\frac{\partial \varphi}{\partial x} = \frac{1}{\det(\mathbf{T})} \left[ (+y_3 - y_1) \vec{\Phi}_\xi^T + (-y_2 + y_1) \vec{\Phi}_\nu^T \right] \cdot \vec{\varphi}, \quad (41)$$

$$\frac{\partial \varphi}{\partial y} = \frac{1}{\det(\mathbf{T})} \left[ (-x_3 + x_1) \vec{\Phi}_\xi^T + (+x_2 - x_1) \vec{\Phi}_\nu^T \right] \cdot \vec{\varphi} \quad (42)$$

For the numerical integration use the values of the gradients at the Gauss integration points  $\vec{g}_j = (\xi_j, \nu_j)$ . The values of the function  $\varphi$  at the Gauss points can be computed with the help of the interpolation matrix  $\mathbf{M}$  by

$$\begin{pmatrix} \varphi(\vec{g}_1) \\ \varphi(\vec{g}_2) \\ \vdots \\ \varphi(\vec{g}_7) \end{pmatrix} = \mathbf{M} \cdot \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_6 \end{pmatrix}.$$

Similarly we define the interpolation matrices for the partial derivatives. Using

$$\mathbf{M}_\xi = \begin{bmatrix} -3 + 4\xi_1 + 4\nu_1 & 4\xi_1 - 1 & 0 & 4\nu_1 & -4\nu_1 & 4 - 8\xi_1 - 4\nu_1 \\ -3 + 4\xi_2 + 4\nu_2 & 4\xi_2 - 1 & 0 & 4\nu_2 & -4\nu_2 & 4 - 8\xi_2 - 4\nu_2 \\ \vdots & & & & & \vdots \\ -3 + 4\xi_7 + 4\nu_7 & 4\xi_7 - 1 & 0 & 4\nu_7 & -4\nu_7 & 4 - 8\xi_7 - 4\nu_7 \end{bmatrix}$$

$$\approx \begin{bmatrix} -2.18971 & -0.59485 & 0.00000 & 0.40515 & -0.40515 & 2.78456 \\ 0.59485 & 2.18971 & 0.00000 & 0.40515 & -0.40515 & -2.78456 \\ 0.59485 & -0.59485 & 0.00000 & 3.18971 & -3.18971 & 0.00000 \\ 0.76114 & 0.88057 & 0.00000 & 1.88057 & -1.88057 & -1.64170 \\ -0.88057 & -0.76114 & 0.00000 & 1.88057 & -1.88057 & 1.64170 \\ -0.88057 & 0.88057 & 0.00000 & 0.23886 & -0.23886 & 0.00000 \\ -0.33333 & 0.33333 & 0.00000 & 1.33333 & -1.33333 & 0.00000 \end{bmatrix}$$

find

$$\begin{pmatrix} \varphi_\xi(\vec{g}_1) \\ \varphi_\xi(\vec{g}_2) \\ \vdots \\ \varphi_\xi(\vec{g}_7) \end{pmatrix} = \mathbf{M}_\xi \cdot \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_6 \end{pmatrix}.$$

Similarly write

$$\mathbf{M}_\nu = \begin{bmatrix} -3 + 4\xi_1 + 4\nu_1 & 0 & 4\nu_1 - 1 & 4\xi_1 & 4 - 4\xi_1 - 8\nu_1 & -4\xi_1 \\ -3 + 4\xi_2 + 4\nu_2 & 0 & 4\nu_2 - 1 & 4\xi_2 & 4 - 4\xi_2 - 8\nu_2 & -4\xi_2 \\ \vdots & & & & & \vdots \\ -3 + 4\xi_7 + 4\nu_7 & 0 & 4\nu_7 - 1 & 4\xi_7 & 4 - 4\xi_7 - 8\nu_7 & -4\xi_7 \end{bmatrix}$$

$$\approx \begin{bmatrix} -2.18971 & 0.00000 & -0.59485 & 0.40515 & 2.78456 & -0.40515 \\ 0.59485 & 0.00000 & -0.59485 & 3.18971 & 0.00000 & -3.18971 \\ 0.59485 & 0.00000 & 2.18971 & 0.40515 & -2.78456 & -0.40515 \\ 0.76114 & 0.00000 & 0.88057 & 1.88057 & -1.64170 & -1.88057 \\ -0.88057 & 0.00000 & 0.88057 & 0.23886 & 0.00000 & -0.23886 \\ -0.88057 & 0.00000 & -0.76114 & 1.88057 & 1.64170 & -1.88057 \\ -0.33333 & 0.00000 & 0.33333 & 1.33333 & 0.00000 & -1.33333 \end{bmatrix}$$

and

$$\begin{pmatrix} \varphi_\nu(\vec{g}_1) \\ \varphi_\nu(\vec{g}_2) \\ \vdots \\ \varphi_\nu(\vec{g}_7) \end{pmatrix} = \mathbf{M}_\nu \cdot \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_6 \end{pmatrix}.$$

The matrices  $\mathbf{M}_\xi$  and  $\mathbf{M}_\nu$  allow to compute the values of the partial derivatives at the Gauss points in the standard triangle  $\Omega$  and they are independent on the general triangle  $E$ .

Combining the above two computations use the notation

$$\vec{x}_i = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + \mathbf{T} \cdot \begin{pmatrix} \xi_i \\ \nu_i \end{pmatrix} \quad \text{for } i = 1, 2, 3, \dots, 7$$

and find for the first component  $\varphi_x = \frac{\partial \varphi}{\partial x}$  of the gradient at the Gauss points

$$\begin{pmatrix} \varphi_x(\vec{x}_1) \\ \varphi_x(\vec{x}_2) \\ \vdots \\ \varphi_x(\vec{x}_7) \end{pmatrix} = \frac{1}{\det(\mathbf{T})} \left[ (+y_3 - y_1) \mathbf{M}_\xi^T + (-y_2 + y_1) \mathbf{M}_\nu^T \right] \cdot \vec{\phi}$$

and for the second component of the gradient

$$\begin{pmatrix} \varphi_y(\vec{x}_1) \\ \varphi_y(\vec{x}_2) \\ \vdots \\ \varphi_y(\vec{x}_7) \end{pmatrix} = \frac{1}{\det(\mathbf{T})} \left[ (-x_3 + x_1) \mathbf{M}_\xi^T + (+x_2 - x_1) \mathbf{M}_\nu^T \right] \cdot \vec{\phi}.$$

The above results for  $\mathbf{M}_\xi$  and  $\mathbf{M}_\nu$  can be coded in *Octave* and then used to compute the element stiffness matrix.

### 6.5.6 Partial derivatives at the nodes

For post processing one also needs the partial derivatives of the function at the nodes. On the standard triangle  $\Omega$  use the formulas for the partial derivatives of the basis functions in expression (40) to find them at the nodes,

given by the  $(\xi, \nu)$  coordinates in Table 13 for quadratic elements.

$$\begin{pmatrix} \varphi_\xi(\xi_1, \nu_1) \\ \varphi_\xi(\xi_2, \nu_2) \\ \varphi_\xi(\xi_3, \nu_3) \\ \varphi_\xi(\xi_4, \nu_4) \\ \varphi_\xi(\xi_5, \nu_5) \\ \varphi_\xi(\xi_6, \nu_6) \end{pmatrix} = \begin{bmatrix} -3 & 1 & 1 & 1 & -1 & -1 \\ -1 & 3 & -1 & 1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 2 & 2 & 0 \\ 0 & 0 & -4 & -2 & -2 & 0 \\ 4 & -4 & 0 & -2 & 2 & 0 \end{bmatrix} \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \varphi_3 \\ \varphi_4 \\ \varphi_5 \\ \varphi_6 \end{pmatrix} = \mathbf{N}_\xi \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \varphi_3 \\ \varphi_4 \\ \varphi_5 \\ \varphi_6 \end{pmatrix}$$

and

$$\begin{pmatrix} \varphi_\nu(\xi_1, \nu_1) \\ \varphi_\nu(\xi_2, \nu_2) \\ \varphi_\nu(\xi_3, \nu_3) \\ \varphi_\nu(\xi_4, \nu_4) \\ \varphi_\nu(\xi_5, \nu_5) \\ \varphi_\nu(\xi_6, \nu_6) \end{pmatrix} = \begin{bmatrix} -3 & 1 & 1 & 1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 3 & 1 & 1 & -1 \\ 0 & 4 & 0 & 2 & 0 & 2 \\ 4 & 0 & -4 & -2 & 0 & 2 \\ 0 & -4 & 0 & -2 & 0 & -2 \end{bmatrix} \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \varphi_3 \\ \varphi_4 \\ \varphi_5 \\ \varphi_6 \end{pmatrix} = \mathbf{N}_\nu \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \varphi_3 \\ \varphi_4 \\ \varphi_5 \\ \varphi_6 \end{pmatrix}$$

Now use the transformation formulas (41) and (42) to determine the gradient of a function on the general triangle

$$\varphi(x, y) = \sum_{i=1}^6 \varphi_i \Phi_i(\xi(x, y), \nu(x, y))$$

at the nodes  $(x_i, y_i)$  in the general triangle  $E$ , leading to

$$\begin{pmatrix} \varphi_x(x_1, y_1) \\ \varphi_x(x_2, y_2) \\ \vdots \\ \varphi_x(x_6, y_6) \end{pmatrix} = \frac{1}{\det(\mathbf{T})} \left[ (+y_3 - y_1) \mathbf{N}_\xi^T + (-y_2 + y_1) \mathbf{N}_\nu^T \right] \cdot \vec{\varphi},$$

$$\begin{pmatrix} \varphi_y(x_1, y_1) \\ \varphi_y(x_2, y_2) \\ \vdots \\ \varphi_y(x_6, y_6) \end{pmatrix} = \frac{1}{\det(\mathbf{T})} \left[ (-x_3 + x_1) \mathbf{N}_\xi^T + (+x_2 - x_1) \mathbf{N}_\nu^T \right] \cdot \vec{\varphi}.$$

These results are useful to evaluate the gradient at the nodes. Observe that the results depends on the triangle used for the interpolation and a node is typically member of more than one triangle.

### 6.5.7 Integration of $u \vec{b} \cdot \nabla \phi$ and $a \nabla u \cdot \nabla \phi$

The vector function  $\vec{b}(\vec{x})$  has to be evaluated at the Gauss integration points  $\vec{g}_j$ . Then the integration of

$$\iint_E u \vec{b} \cdot \nabla \phi \, dA = \iint_E u b_1 \frac{\partial \phi}{\partial x} \, dA + \iint_E u b_2 \frac{\partial \phi}{\partial y} \, dA$$

is approximated by

$$\begin{aligned} \iint_E u b_1 \frac{\partial \phi}{\partial x} dA &\approx \frac{|\det \mathbf{T}|}{\det \mathbf{T}} \langle ((y_3 - y_1) \mathbf{M}_\xi^T + (-y_2 + y_1) \mathbf{M}_\nu^T) \cdot \text{diag}(\vec{wb}_1) \cdot \mathbf{M} \cdot \vec{u}, \vec{\phi} \rangle \\ \iint_E u b_2 \frac{\partial \phi}{\partial y} dA &\approx \frac{|\det \mathbf{T}|}{\det \mathbf{T}} \langle ((-x_3 + x_1) \mathbf{M}_\xi^T + (x_2 - x_1) \mathbf{M}_\nu^T) \cdot \text{diag}(\vec{wb}_2) \cdot \mathbf{M} \cdot \vec{u}, \vec{\phi} \rangle. \end{aligned}$$

The function  $a \nabla u \cdot \nabla \phi = a \left( \frac{\partial u}{\partial x} \frac{\partial \phi}{\partial x} + \frac{\partial u}{\partial y} \frac{\partial \phi}{\partial y} \right)$  has to be evaluated at the Gauss integration points  $\vec{g}_j$ , then multiplied by the Gauss weights  $w_i$  and added up. Use the vector  $\vec{wa}$  with the values of the function  $a(x_i, y_i)$  and the weights  $w_i$  at the Gauss points to obtain

$$\begin{aligned} \iint_E a \frac{\partial u(\vec{x})}{\partial x} \frac{\partial \phi(\vec{x})}{\partial x} dA &= |\det \mathbf{T}| \int_{\Omega} a(\vec{x}(\xi, \nu)) \frac{\partial u(\vec{x}(\xi, \nu))}{\partial x} \frac{\partial \phi(\vec{x}(\xi, \nu))}{\partial x} d\xi d\nu \\ &\approx \frac{|\det \mathbf{T}|}{(\det \mathbf{T})^2} \langle \mathbf{A}_x \cdot \vec{u}, \vec{\phi} \rangle = \frac{1}{|\det \mathbf{T}|} \langle \mathbf{A}_x \cdot \vec{u}, \vec{\phi} \rangle \\ \iint_E a \frac{\partial u(\vec{x})}{\partial y} \frac{\partial \phi(\vec{x})}{\partial y} dA &= |\det \mathbf{T}| \int_{\Omega} a(\vec{x}(\xi, \nu)) \frac{\partial u(\vec{x}(\xi, \nu))}{\partial y} \frac{\partial \phi(\vec{x}(\xi, \nu))}{\partial y} d\xi d\nu \\ &\approx \frac{|\det \mathbf{T}|}{(\det \mathbf{T})^2} \langle \mathbf{A}_y \cdot \vec{u}, \vec{\phi} \rangle = \frac{1}{|\det \mathbf{T}|} \langle \mathbf{A}_y \cdot \vec{u}, \vec{\phi} \rangle \end{aligned}$$

where

$$\begin{aligned} \mathbf{A}_x &= \left[ (+y_3 - y_1) \mathbf{M}_\xi + (-y_2 + y_1) \mathbf{M}_\nu \right]^T \cdot \text{diag}(\vec{wa}) \cdot \left[ (+y_3 - y_1) \mathbf{M}_\xi + (-y_2 + y_1) \mathbf{M}_\nu \right] \\ \mathbf{A}_y &= \left[ (-x_3 + x_1) \mathbf{M}_\xi + (+x_2 - x_1) \mathbf{M}_\nu \right]^T \cdot \text{diag}(\vec{wa}) \cdot \left[ (-x_3 + x_1) \mathbf{M}_\xi + (+x_2 - x_1) \mathbf{M}_\nu \right]. \end{aligned}$$

### 6.5.8 Integration over boundary segments

In expression (27) we have to compute integrals over the boundary

$$\int_{\Gamma_2} \phi (g_2 + g_3 u) ds.$$

For triangular domains the boundary consists of straight line segments. Thus replace the integral by a sum of line integrals and use a Gauss integration. Based on the two endpoints  $\vec{x}_1$  and  $\vec{x}_3$  and the midpoint  $\vec{x}_2 = \frac{1}{2} (\vec{x}_1 + \vec{x}_3)$  use the values at three Gauss integration points. Based on<sup>17</sup>

$$\int_{-h/2}^{h/2} f(x) dx \approx \frac{h}{18} \left( 5 f\left(-\frac{\sqrt{3}}{2\sqrt{5}} h\right) + 8 f(0) + 5 f\left(\frac{\sqrt{3}}{2\sqrt{5}} h\right) \right)$$

<sup>17</sup>To derive the 3 point Gauss integration scheme use

$$\begin{aligned} \int_{-1}^{+1} f(t) dt &= w_1 f(-\xi) + w_0 f(0) + w_1 f(+\xi) \\ \int_{-1}^{+1} 1 dt = 2 &= w_1 1 + w_0 1 + w_1 1 \\ \int_{-1}^{+1} t dt = 0 &= -w_1 \xi + w_0 0 + w_1 \xi = 0 \\ \int_{-1}^{+1} t^2 dt = \frac{2}{3} &= +w_1 \xi^2 + w_1 \xi^2 \\ \int_{-1}^{+1} t^3 dt = 0 &= -w_1 \xi^3 + w_1 \xi^3 = 0 \\ \int_{-1}^{+1} t^4 dt = \frac{2}{5} &= +w_1 \xi^4 + w_1 \xi^4 \\ \int_{-1}^{+1} t^5 dt = 0 &= -w_1 \xi^5 + w_1 \xi^5 = 0 \end{aligned}$$

polynomials up to degree 5 are integrated exactly, thus the error on one interval is proportional to  $h^7$ . To evaluate a function at the Gauss points

$$\begin{aligned}\vec{p}_1 &= \frac{1}{2} (\vec{x}_1 + \vec{x}_3) - \frac{\sqrt{3}}{2\sqrt{5}} (\vec{x}_3 - \vec{x}_1) \\ \vec{p}_2 &= \vec{x}_2 = \frac{1}{2} (\vec{x}_1 + \vec{x}_3) \\ \vec{p}_3 &= \frac{1}{2} (\vec{x}_1 + \vec{x}_3) + \frac{\sqrt{3}}{2\sqrt{5}} (\vec{x}_3 - \vec{x}_1)\end{aligned}$$

use a quadratic interpolation of a function with  $f_- = f(-h/2)$ ,  $f_0 = f(0)$  and  $f_+ = f(+h/2)$ . Since<sup>18</sup>

$$f(x) = f_0 + \frac{f_+ - f_-}{h} x + 2 \frac{f_- - 2f_0 + f_+}{h^2} x^2$$

the quadratic interpolation result at  $\pm\alpha h$  is

$$\begin{aligned}f(\pm\alpha h) &= f_0 \pm (f_+ - f_-) \alpha + 2 (f_- - 2f_0 + f_+) \alpha^2 \\ &= f_- (\pm\alpha + 2\alpha^2) + f_0 (1 - 4\alpha^2) + f_+ (\mp\alpha + 2\alpha^2)\end{aligned}$$

where  $\alpha = \frac{\sqrt{3}}{2\sqrt{5}} = \frac{\sqrt{15}}{10} \approx 0.316$ . If a function  $u$  is given at the two endpoints by  $u_1$  and  $u_3$  and at the midpoint by  $u_2$  obtain

$$\begin{aligned}\begin{pmatrix} u(\vec{p}_1) \\ u(\vec{p}_2) \\ u(\vec{p}_3) \end{pmatrix} &= \begin{bmatrix} +\alpha + 2\alpha^2 & 1 - 4\alpha^2 & -\alpha + 2\alpha^2 \\ 0 & 1 & 0 \\ -\alpha + 2\alpha^2 & 1 - 4\alpha^2 & +\alpha + 2\alpha^2 \end{bmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \\ &= \mathbf{M}_B \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \approx \begin{bmatrix} +0.68730 & 0.4 & -0.08730 \\ 0 & 1 & 0 \\ -0.08730 & 0.4 & +0.68730 \end{bmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}\end{aligned}\quad (43)$$

With the length  $L = \sqrt{(x_3 - x_1)^2 + (y_3 - y_1)^2}$  of the segment this leads to the approximations

$$\begin{aligned}\int_{\text{edge}} \phi g_2 ds &\approx \frac{L}{18} \left\langle \mathbf{M}_B \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \end{pmatrix}, \begin{pmatrix} 5 g_2(\vec{p}_1) \\ 8 g_2(\vec{p}_2) \\ 5 g_2(\vec{p}_3) \end{pmatrix} \right\rangle = \frac{L}{18} \left\langle \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \end{pmatrix}, \mathbf{M}_B^T \begin{pmatrix} 5 g_2(\vec{p}_1) \\ 8 g_2(\vec{p}_2) \\ 5 g_2(\vec{p}_3) \end{pmatrix} \right\rangle \\ \int_{\text{edge}} \phi g_3 u ds &\approx \frac{L}{18} \left\langle \mathbf{M}_B \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \end{pmatrix}, \begin{bmatrix} 5 g_3(\vec{p}_1) & 0 & 0 \\ 0 & 8 g_3(\vec{p}_2) & 0 \\ 0 & 0 & 5 g_3(\vec{p}_3) \end{bmatrix} \mathbf{M}_B \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \right\rangle \\ &= \frac{L}{18} \left\langle \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \end{pmatrix}, \mathbf{M}_B^T \begin{bmatrix} 5 g_3(\vec{p}_1) & 0 & 0 \\ 0 & 8 g_3(\vec{p}_2) & 0 \\ 0 & 0 & 5 g_3(\vec{p}_3) \end{bmatrix} \mathbf{M}_B \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \right\rangle.\end{aligned}$$

Thus  $t^6$  is not integrated exactly and the error is proportional to  $h^6$ . The system to be solved is

$$\begin{cases} w_0 + 2w_1 = 2 \\ 2w_1 \xi^2 = \frac{2}{3} \\ 2w_1 \xi^4 = \frac{2}{5} \end{cases} \implies \xi^2 = \frac{3}{5}, w_1 = \frac{5}{9}, w_0 = \frac{8}{9}.$$

<sup>18</sup>To verify use  $f(0) = f_0$  and

$$f(\pm h/2) = f_0 \pm \frac{f_+ - f_-}{h} \frac{h}{2} + 2 \frac{f_- - 2f_0 + f_+}{h^2} \frac{h^2}{4} = f_0 \pm \frac{1}{2}(f_+ - f_-) + \frac{1}{2}(f_- - 2f_0 + f_+).$$

The first expression will lead to a contribution to the RHS vector of the linear system to be solved, while the second expression will lead to entries in the matrix. These approximate integrations lead to the exact result if the function to be integrated is a polynomial of degree 5, or less. If  $h$  is the typical length of an edge then the error is of the order  $h^7$  for one line segment and thus of order  $h^6$  for the total boundary. This boundary integration is used for the second order elements.

The second expression is of the form

$$\int \phi g_3 u ds \approx \langle \vec{\phi}, \mathbf{B} \vec{u} \rangle = \left\langle \begin{pmatrix} \phi_2 \\ \phi_2 \\ \phi_3 \end{pmatrix}, \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} \right\rangle$$

and its effect on the linear system  $\mathbf{A} \vec{u} + \mathbf{W} \vec{f} = \vec{0}$  depends on nodes being on the Dirichlet part of the boundary.

- If  $u_1$  and  $u_3$  are both free, i.e. not on the Dirichlet section, then  $u_2$  is free too. All entries of the matrix  $\mathbf{B}$  have to be added to the global stiffness matrix  $\mathbf{A}$ .
- If  $u_1$  and  $u_3$  are on the Dirichlet section, then nothing has to be added to  $\mathbf{A}$  and  $\vec{f}$ .
- If  $u_1$  and  $u_2$  are free and  $u_3$  is on the Dirichlet section, then only the first two expressions

$$\begin{aligned} b_{11} u_1 + b_{12} u_2 + b_{13} u_3 &= b_{11} u_1 + b_{12} u_2 + b_{13} d_3 \\ b_{21} u_1 + b_{22} u_2 + b_{23} u_3 &= b_{21} u_1 + b_{22} u_2 + b_{23} d_3 \end{aligned}$$

have to be added.  $d_3$  is the Dirichlet value at the position of  $u_3$ .  $b_{13} g_3$  and  $b_{23} d_3$  have to be added to  $\mathbf{W} \vec{f}$ , the other expression to  $\mathbf{A}$ .

- If  $u_2$  and  $u_3$  are free and  $u_1$  is on the Dirichlet section, then only the second and third expressions

$$\begin{aligned} b_{21} u_1 + b_{22} u_2 + b_{23} u_3 &= b_{21} d_1 + b_{22} u_2 + b_{23} u_3 \\ b_{31} u_1 + b_{32} u_2 + b_{33} u_3 &= b_{31} d_1 + b_{32} u_2 + b_{33} u_3 \end{aligned}$$

have to be added.  $d_1$  is the Dirichlet value at the position of  $u_1$ .  $b_{21} g_1$  and  $b_{31} d_1$  have to be added to  $\mathbf{W} \vec{f}$ , the other expression to  $\mathbf{A}$ .

- If  $u_1$  and  $u_3$  are free, then  $u_2$  has to be free too, since it is the midpoint of a Neumann section of the boundary.

## 6.6 Construction of third order elements

In this section the construction of the element stiffness matrix and vector for triangular elements of order 3 is examined. The ideas are extremely similar to Section 6.5 for quadratic functions. Again all contributions in (27)

$$0 = \iint_{\Omega} \nabla \phi \cdot (a \nabla u - u \vec{b}) + \phi (b_0 u - f) dA - \int_{\Gamma_2} \phi (g_2 + g_3 u) ds$$

have to be transformed into

$$0 = \langle \vec{\phi}, \mathbf{A} \vec{u} + \mathbf{W} \vec{f} \rangle. \quad (44)$$

For third order elements a general cubic function is used on each of the triangles in the mesh. There are 10 linearly independent polynomials of degree 3 or less, namely  $1, x, y, x^2, xy, y^2, x^3, x^2y, xy^2$  and  $y^3$ .

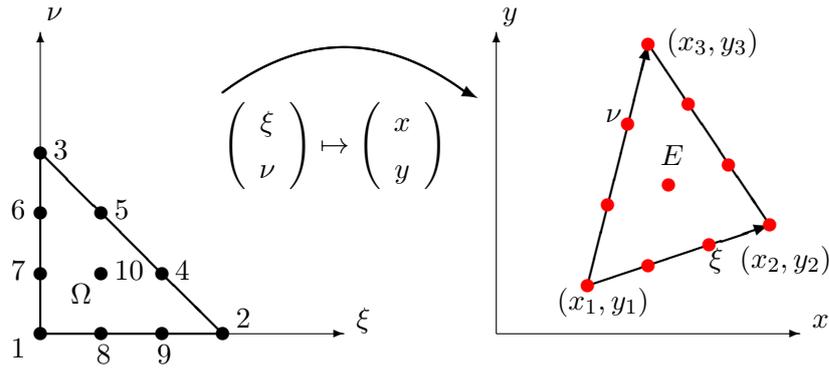


Figure 70: Transformation of the cubic standard triangle  $\Omega$  to a general triangle  $E$

**6.6.1 The basis functions for a third order element and cubic interpolation**

Examine the standard triangle  $\Omega$  in Figure 70 with the values of a function  $f(\xi, \nu)$  at the corners, the points on the edges and the midpoint. Use the numbering as shown in Figure 70. The parameters  $\xi$  and  $\nu$  at the nodes are given by Table 14. Construct polynomials  $\phi_i(\xi, \nu)$  of degree 3, such that

$$\Phi_i(\xi_j, \nu_j) = \delta_{i,j} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

i.e. each basis function is equal to 1 at one of the nodes and vanishes on all other nodes. These basis polynomials are given by<sup>19</sup>

node $i$	1	2	3	4	5	6	7	8	9	10
$\xi_i$	0	1	0	$\frac{2}{3}$	$\frac{1}{3}$	0	0	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{1}{3}$
$\nu_i$	0	0	1	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	0	0	$\frac{1}{3}$

Table 14: Coordinates of the nodes in the standard cubic triangle

$$\vec{\Phi}(\xi, \nu) = \begin{pmatrix} \Phi_1(\xi, \nu) \\ \Phi_2(\xi, \nu) \\ \Phi_3(\xi, \nu) \\ \Phi_4(\xi, \nu) \\ \Phi_5(\xi, \nu) \\ \Phi_6(\xi, \nu) \\ \Phi_7(\xi, \nu) \\ \Phi_8(\xi, \nu) \\ \Phi_9(\xi, \nu) \\ \Phi_{10}(\xi, \nu) \end{pmatrix} = \begin{pmatrix} (1 - (\xi + \nu)) (1 - 3(\xi + \nu)) (1 - \frac{3}{2}(\xi + \nu)) \\ \xi (3\xi - 1) (\frac{3}{2}\xi - 1) \\ \nu (3\nu - 1) (\frac{3}{2}\nu - 1) \\ \frac{9}{2} \xi \nu (3\xi - 1) \\ \frac{9}{2} \xi \nu (3\nu - 1) \\ \frac{9}{2} \nu (1 - (\xi + \nu)) (3\nu - 1) \\ 9 \nu (1 - (\xi + \nu)) (1 - \frac{3}{2}(\xi + \nu)) \\ 9 \xi (1 - \frac{3}{2}(\xi + \nu)) (1 - (\xi + \nu)) \\ \frac{9}{2} \xi (3\xi - 1) (1 - (\xi + \nu)) \\ 27 \xi \nu (1 - (\xi + \nu)) \end{pmatrix} \tag{45}$$

<sup>19</sup>Use that the level curves of the functions  $\xi$ ,  $\nu$  and  $1 - (\xi + \nu)$  at the levels  $0, \frac{1}{3}, \frac{2}{3}$  and 1 are straight lines through the nodes. For each node use these functions to write down a polynomial vanishing at all other nodes, then choose the leading factor such that at the node the value equals 1.

$$= \begin{pmatrix} 1 - \frac{11}{2} \xi - \frac{11}{2} \nu + 9 \xi^2 + 18 \xi \nu + 9 \nu^2 - \frac{9}{2} \xi^3 - \frac{27}{2} \xi^2 \nu - \frac{27}{2} \xi \nu^2 - \frac{9}{2} \nu^3 \\ \xi - \frac{9}{2} \xi^2 + \frac{9}{2} \xi^3 \\ \nu - \frac{9}{2} \nu^2 + \frac{9}{2} \nu^3 \\ -\frac{9}{2} \xi \nu + \frac{27}{2} \xi^2 \nu \\ -\frac{9}{2} \xi \nu + \frac{27}{2} \xi \nu^2 \\ -\frac{9}{2} \nu + \frac{9}{2} \xi \nu + 18 \nu^2 - \frac{27}{2} \xi \nu^2 - \frac{27}{2} \nu^3 \\ 9 \nu - \frac{45}{2} \xi \nu - \frac{45}{2} \nu^2 + \frac{27}{2} \xi^2 \nu + 27 \xi \nu^2 + \frac{27}{2} \nu^3 \\ 9 \xi - \frac{45}{2} \xi^2 - \frac{45}{2} \xi \nu + \frac{27}{2} \xi^3 + 27 \xi^2 \nu + \frac{27}{2} \xi \nu^2 \\ -\frac{9}{2} \xi + 18 \xi^2 + \frac{9}{2} \xi \nu - \frac{27}{2} \xi^3 - \frac{27}{2} \xi^2 \nu \\ 27 \xi \nu - 27 \xi^2 \nu - 27 \xi \nu^2 \end{pmatrix} \quad (46)$$

and find their graphs in Figure 71.

Any cubic polynomial  $f$  on the standard triangle  $\Omega$  can be written as linear combination of the 10 basis functions by using

$$f(\xi, \nu) = \sum_{i=1}^{10} f(\xi_i, \nu_i) \Phi_i(\xi, \nu) = \sum_{i=1}^{10} f_i \Phi_i(\xi, \nu). \quad (47)$$

This is the formula to apply a cubic interpolation on the triangle, using the values  $f_i = f(\xi_i, \nu_i)$  of the function at the nodes. To use this interpolation for a given point  $(x, y)$  in the triangle  $E$  in Figure 70. The transformation from the standard triangle  $\Omega$  to the general triangle  $E$  is identical to the second order elements, i.e.

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + \begin{bmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{bmatrix} \begin{pmatrix} \xi \\ \nu \end{pmatrix} = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + \mathbf{T} \begin{pmatrix} \xi \\ \nu \end{pmatrix}$$

and

$$\begin{pmatrix} \xi \\ \nu \end{pmatrix} = \mathbf{T}^{-1} \cdot \begin{pmatrix} x - x_1 \\ y - y_1 \end{pmatrix} = \frac{1}{\det(\mathbf{T})} \begin{bmatrix} y_3 - y_1 & -x_3 + x_1 \\ -y_2 + y_1 & x_2 - x_1 \end{bmatrix} \cdot \begin{pmatrix} x - x_1 \\ y - y_1 \end{pmatrix}.$$

### 6.6.2 Determine values at the Gauss points and apply Gauss integration

Use equation (31) (page 132) to determine the coordinates of the seven Gauss points. Then a function to be integrated can be evaluated at these Gauss points. Determine the values of the basis functions  $\Phi_i(\xi, \nu)$  at the Gauss points  $\vec{g}_j$  by  $m_{j,i} = \Phi_i(\vec{g}_j)$  and write

$$f(\vec{g}_j) = \sum_{i=1}^{10} f_i \Phi_i(\vec{g}_j) = \sum_{i=1}^{10} m_{j,i} f_i$$

or using a matrix notation with  $\mathbf{M} \in \mathbb{R}^{7 \times 10}$

$$\begin{pmatrix} f(\vec{g}_1) \\ f(\vec{g}_2) \\ \vdots \\ f(\vec{g}_7) \end{pmatrix} = \begin{bmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,10} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,10} \\ \vdots & \vdots & \ddots & \vdots \\ m_{7,1} & m_{7,2} & \cdots & m_{7,10} \end{bmatrix} \cdot \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_{10} \end{pmatrix} = \mathbf{M} \cdot \vec{f} \approx \quad (48)$$

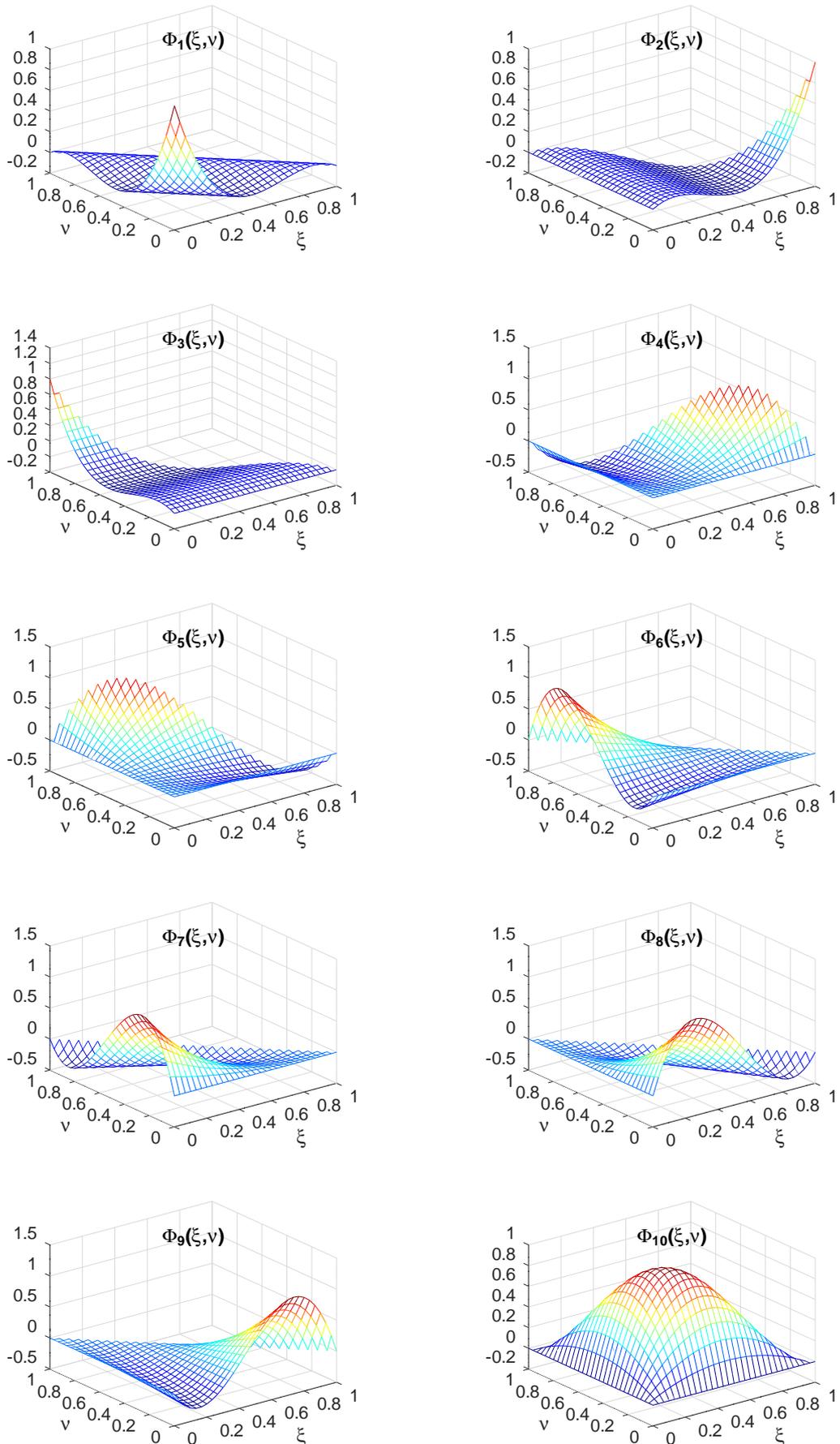


Figure 71: The 10 basis functions for third order triangular elements

$$\approx \begin{bmatrix} +0.22 & +0.06 & +0.06 & -0.03 & -0.03 & -0.25 & +0.51 & +0.51 & -0.25 & +0.22 \\ +0.06 & +0.22 & +0.06 & +0.51 & -0.25 & -0.03 & -0.03 & -0.25 & +0.51 & +0.22 \\ +0.06 & +0.06 & +0.22 & -0.25 & +0.51 & +0.51 & -0.25 & -0.03 & -0.03 & +0.22 \\ +0.04 & -0.06 & -0.06 & +0.41 & +0.41 & +0.05 & -0.10 & -0.10 & +0.05 & +0.36 \\ -0.06 & +0.04 & -0.06 & -0.10 & +0.05 & +0.41 & +0.41 & +0.05 & -0.10 & +0.36 \\ -0.06 & -0.06 & +0.04 & +0.05 & -0.10 & -0.10 & +0.05 & +0.41 & +0.41 & +0.36 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ \vdots \\ f_9 \\ f_{10} \end{pmatrix}$$

The Gauss integration can be written in the form

$$\iint_{\Omega} f(\xi, \nu) dA \approx \sum_{j=1}^7 w_j f(\vec{g}_j) = \langle \vec{w}, \mathbf{M} \cdot \vec{f} \rangle.$$

To integrate over the general triangle  $E$  use the transformation (28), i.e.

$$\iint_E f dA = \iint_{\Omega} f(\vec{x}(\xi, \nu)) \left| \det \left( \frac{\partial(x, y)}{\partial(\xi, \nu)} \right) \right| d\xi d\nu \approx |\det \mathbf{T}| \langle \vec{w}, \mathbf{M} \cdot \vec{f} \rangle.$$

Now all the tools to approximate the integrals required for the element stiffness matrix are available.

### 6.6.3 Integration of $f \phi$ and $b_0 u \phi$

These integrations are identical to the case of quadratic elements. The test function  $\phi$  is given by its values  $\vec{\phi}$  at the nodes, i.e. the corners of the triangle and the two points on each side.

- If the values of the function  $f$  at the Gauss points  $\vec{g}_i$  are denoted by  $f_i$  then this integral is approximated by

$$\iint_E f \phi dA \approx |\det(\mathbf{T})| \sum_{j=1}^7 w_j f_j \phi(g_j) = |\det(\mathbf{T})| \langle \text{diag}(\vec{w}) \vec{f}, \mathbf{M} \vec{\phi} \rangle = |\det(\mathbf{T})| \langle \mathbf{M}^T \text{diag}(\vec{w}) \vec{f}, \vec{\phi} \rangle.$$

Thus find one contribution to (44).

- If the values of the function  $f$  at the nodes are denoted by  $f_i$  then first determine the values at the Gauss points by a cubic interpolation. Then integrate as above, leading to

$$\iint_E f \phi dA \approx |\det(\mathbf{T})| \langle \text{diag}(\vec{w}) \mathbf{M} \vec{f}, \mathbf{M} \vec{\phi} \rangle = |\det(\mathbf{T})| \langle \mathbf{M}^T \text{diag}(\vec{w}) \mathbf{M} \vec{f}, \vec{\phi} \rangle.$$

- Since the values of the functions  $u$  and  $\phi$  are known at the nodes use an interpolation and then the function  $b_0(x, y)$  at the Gauss nodes to find

$$\begin{aligned} \iint_E b_0 u \phi dA &\approx |\det(\mathbf{T})| \sum_{j=1}^7 w_j b_0(g_j) u(g_j) \phi(g_j) = |\det(\mathbf{T})| \langle \text{diag}(\vec{w}) \text{diag}(\vec{b}_0) \mathbf{M} \vec{u}, \mathbf{M} \vec{\phi} \rangle \\ &= |\det(\mathbf{T})| \langle \mathbf{M}^T \text{diag}(\vec{w}) \text{diag}(\vec{b}_0) \mathbf{M} \vec{u}, \vec{\phi} \rangle. \end{aligned}$$

The matrices  $\mathbf{M}^T \text{diag}(\vec{w})$  and  $\mathbf{M}^T \text{diag}(\vec{w}) \mathbf{M}$  are again independent on the triangle  $E$ , but different from the case of quadratic elements.

### 6.6.4 Transformation of the gradient to the standard triangle

Computing the partial derivatives is again very similar to the case of quadratic elements. If a function  $f(x, y)$  is given on the general triangle  $E$  can pull it back to the standard triangle by

$$g(\xi, \nu) = f(x(\xi, \nu), y(\xi, \nu))$$

and then compute the gradient of  $g(\xi, \nu)$  with respect to its independent variables  $\xi$  and  $\nu$ . The result is This can be written with the help of matrices in the form

$$\begin{pmatrix} \frac{\partial g}{\partial \xi} \\ \frac{\partial g}{\partial \nu} \end{pmatrix} = \begin{bmatrix} (x_2 - x_1) & (y_2 - y_1) \\ (x_3 - x_1) & (y_3 - y_1) \end{bmatrix} \cdot \begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{pmatrix} = \mathbf{T}^T \cdot \begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{pmatrix}$$

or equivalently

$$\begin{pmatrix} \frac{\partial g}{\partial \xi} \\ \frac{\partial g}{\partial \nu} \end{pmatrix} = \begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{pmatrix} \cdot \mathbf{T},$$

or

$$\begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{pmatrix} = \frac{1}{\det \mathbf{T}} \begin{bmatrix} y_3 - y_1 & -y_2 + y_1 \\ -x_3 + x_1 & x_2 - x_1 \end{bmatrix} \begin{pmatrix} \frac{\partial g}{\partial \xi} \\ \frac{\partial g}{\partial \nu} \end{pmatrix}.$$

Let  $u$  be a function on the standard triangle  $\Omega$  given as a linear combination of the basis functions, i.e.  $u(\xi, \nu) = \sum_{i=1}^{10} u_i \Phi_i(\xi, \nu)$ , where the basis function  $\Phi_i(\xi, \nu)$  are given by (46). Then its gradient with respect to  $\xi$  and  $\nu$  can be determined with the help of elementary partial derivatives applied to the expressions in (46). The results are

$$\vec{\Phi}_\xi(\xi, \nu) = \frac{\partial}{\partial \xi} \vec{\Phi}(\xi, \nu) = \begin{pmatrix} -\frac{11}{2} + 18\xi + 18\nu - \frac{27}{2}\xi^2 - 27\xi\nu - \frac{27}{2}\nu^2 \\ 1 - 9\xi + \frac{27}{2}\xi^2 \\ 0 \\ -\frac{9}{2}\nu + 27\xi\nu \\ -\frac{9}{2}\nu + \frac{27}{2}\nu^2 \\ \frac{9}{2}\nu - \frac{27}{2}\nu^2 \\ -\frac{45}{2}\nu + 27\xi\nu + 27\nu^2 \\ 9 - 45\xi - \frac{45}{2}\nu + \frac{81}{2}\xi^2 + 54\xi\nu + \frac{27}{2}\nu^2 \\ -\frac{9}{2} + 36\xi + \frac{9}{2}\nu - \frac{81}{2}\xi^2 - 27\xi\nu \\ 27\nu - 54\xi\nu - 27\nu^2 \end{pmatrix} \quad (49)$$

and

$$\vec{\Phi}_\nu(\xi, \nu) = \frac{\partial}{\partial \nu} \vec{\Phi}(\xi, \nu) = \begin{pmatrix} -\frac{11}{2} + 18\xi + 18\nu - \frac{27}{2}\xi^2 - 27\xi\nu - \frac{27}{2}\nu^2 \\ 0 \\ 1 - 9\nu + \frac{27}{2}\nu^2 \\ -\frac{9}{2}\xi + \frac{27}{2}\xi^2 \\ -\frac{9}{2}\xi + 27\xi\nu \\ -\frac{9}{2} + \frac{9}{2}\xi + 36\nu - 27\xi\nu - \frac{81}{2}\nu^2 \\ 9 - \frac{45}{2}\xi - 45\nu + \frac{27}{2}\xi^2 + 54\xi\nu + \frac{81}{2}\nu^2 \\ -\frac{45}{2}\xi + 27\xi^2 + 27\xi\nu \\ +\frac{9}{2}\xi - \frac{27}{2}\xi^2 \\ 27\xi - 27\xi^2 - 54\xi\nu \end{pmatrix}. \quad (50)$$

Thus find on the standard triangle  $\Omega$

$$\left( \frac{\partial u}{\partial \xi}, \frac{\partial u}{\partial \nu} \right) = (u_1, u_2, \dots, u_{10}) \cdot \left[ \vec{\Phi}_\xi(\xi, \nu) \quad \vec{\Phi}_\nu(\xi, \nu) \right] = \vec{u}^T \cdot \left[ \vec{\Phi}_\xi(\xi, \nu) \quad \vec{\Phi}_\nu(\xi, \nu) \right].$$

For a function  $\varphi(x, y) = \sum_{i=1}^{10} \varphi_i \Phi_i(\xi(x, y), \nu(x, y))$  use the above to conclude

$$\begin{pmatrix} \frac{\partial \varphi}{\partial x} \\ \frac{\partial \varphi}{\partial y} \end{pmatrix} = \frac{1}{\det(\mathbf{T})} \begin{bmatrix} +y_3 - y_1 & -y_2 + y_1 \\ -x_3 + x_1 & +x_2 - x_1 \end{bmatrix} \cdot \begin{bmatrix} \vec{\Phi}_\xi^T \\ \vec{\Phi}_\nu^T \end{bmatrix} \cdot \vec{\varphi}$$

or spelled out for the two components independently

$$\begin{aligned} \frac{\partial \varphi}{\partial x} &= \frac{1}{\det(\mathbf{T})} \left[ (+y_3 - y_1) \vec{\Phi}_\xi^T + (-y_2 + y_1) \vec{\Phi}_\nu^T \right] \cdot \vec{\varphi}, \\ \frac{\partial \varphi}{\partial y} &= \frac{1}{\det(\mathbf{T})} \left[ (-x_3 + x_1) \vec{\Phi}_\xi^T + (+x_2 - x_1) \vec{\Phi}_\nu^T \right] \cdot \vec{\varphi}. \end{aligned}$$

For the numerical integration use the values of the gradients at the Gauss integration points  $\vec{g}_j = (\xi_j, \nu_j)$ . Using expression (46) the values of the function  $\varphi$  at the Gauss points can be computed with the help of the interpolation matrix  $\mathbf{M}$  byc

$$\begin{pmatrix} \varphi(\vec{g}_1) \\ \varphi(\vec{g}_2) \\ \vdots \\ \varphi(\vec{g}_7) \end{pmatrix} = \mathbf{M} \cdot \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_{10} \end{pmatrix}.$$

Similarly, using (49) and (50), define the interpolation matrices for the partial derivatives.

$$\frac{\partial}{\partial \xi} \begin{pmatrix} \varphi(\vec{g}_1) \\ \varphi(\vec{g}_2) \\ \vdots \\ \varphi(\vec{g}_7) \end{pmatrix} = \mathbf{M}_\xi \cdot \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_{10} \end{pmatrix} \quad \text{and} \quad \frac{\partial}{\partial \nu} \begin{pmatrix} \varphi(\vec{g}_1) \\ \varphi(\vec{g}_2) \\ \vdots \\ \varphi(\vec{g}_7) \end{pmatrix} = \mathbf{M}_\nu \cdot \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \vdots \\ \varphi_{10} \end{pmatrix}. \quad (51)$$

Approximate values are

$$\mathbf{M}_\xi \approx \begin{bmatrix} -2.408 & 0.227 & 0 & -0.179 & -0.317 & 0.317 & -1.725 & 3.271 & -1.090 & 1.904 \\ -0.227 & 2.408 & 0 & 1.725 & -0.317 & 0.317 & 0.179 & 1.090 & -3.271 & -1.904 \\ -0.227 & 0.227 & 0 & -1.408 & 4.996 & -4.996 & 1.408 & -0.138 & 0.138 & 0 \\ -0.511 & -0.247 & 0 & 3.852 & 0.868 & -0.868 & 1.358 & 1.137 & -0.379 & -5.210 \\ 0.247 & 0.511 & 0 & -1.358 & 0.868 & -0.868 & -3.852 & 0.379 & -1.137 & 5.210 \\ 0.247 & -0.247 & 0 & 0.489 & -0.221 & 0.221 & -0.489 & -2.984 & 2.984 & 0 \\ 0.500 & -0.500 & 0 & 1.500 & 0 & 0 & -1.500 & -1.500 & 1.500 & 0 \end{bmatrix}$$

and

$$\mathbf{M}_\nu \approx \begin{bmatrix} -2.269 & 0 & 0.227 & -0.317 & -0.179 & -1.090 & 3.271 & -1.725 & 0.317 & 1.904 \\ 0.863 & 0 & 0.227 & 4.996 & -1.408 & 0.138 & -0.138 & 1.408 & -4.996 & 0 \\ 0.863 & 0 & 2.408 & -0.317 & 1.725 & -3.271 & 1.090 & 0.179 & 0.317 & -1.904 \\ 2.473 & 0 & -0.247 & 0.868 & 3.852 & -0.379 & 1.137 & 1.358 & -0.868 & -5.210 \\ 0.626 & 0 & -0.247 & -0.221 & 0.489 & 2.984 & -2.984 & -0.489 & 0.221 & 0 \\ 0.626 & 0 & 0.511 & 0.868 & -1.358 & -1.137 & 0.379 & -3.852 & -0.868 & 5.210 \\ 2.000 & 0 & -0.500 & 0 & 1.500 & 1.500 & -1.500 & -1.500 & 0 & 0 \end{bmatrix}.$$

The matrices  $\mathbf{M}_\xi$  and  $\mathbf{M}_\nu$  allow to compute the values of the partial derivatives at the Gauss points in the standard triangle  $\Omega$  and they are independent on the general triangle  $E$ .

Combining the above two computations use the notation

$$\vec{x}_i = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + \mathbf{T} \cdot \begin{pmatrix} \xi_i \\ \nu_i \end{pmatrix} \quad \text{for } i = 1, 2, 3, \dots, 7$$

and find for the first component  $\varphi_x = \frac{\partial \varphi}{\partial x}$  of the gradient at the Gauss points

$$\begin{pmatrix} \varphi_x(\vec{x}_1) \\ \varphi_x(\vec{x}_2) \\ \vdots \\ \varphi_x(\vec{x}_7) \end{pmatrix} = \frac{1}{\det(\mathbf{T})} \left[ (+y_3 - y_1) \mathbf{M}_\xi^T + (-y_2 + y_1) \mathbf{M}_\nu^T \right] \cdot \vec{\phi}$$

and for the second component of the gradient

$$\begin{pmatrix} \varphi_y(\vec{x}_1) \\ \varphi_y(\vec{x}_2) \\ \vdots \\ \varphi_y(\vec{x}_7) \end{pmatrix} = \frac{1}{\det(\mathbf{T})} \left[ (-x_3 + x_1) \mathbf{M}_\xi^T + (+x_2 - x_1) \mathbf{M}_\nu^T \right] \cdot \vec{\phi}.$$

The above results for  $\mathbf{M}_\xi$  and  $\mathbf{M}_\nu$  can be coded in *Octave* and then used to compute the element stiffness matrix.

### 6.6.5 Integration of $u \vec{b} \cdot \nabla \phi$ and $a \nabla u \cdot \nabla \phi$

The vector function  $\vec{b}(\vec{x})$  has to be evaluated at the Gauss integration points  $\vec{g}_j$ . Then the integration of

$$\iint_E u \vec{b} \cdot \nabla \phi \, dA = \iint_E u b_1 \frac{\partial \phi}{\partial x} \, dA + \iint_E u b_2 \frac{\partial \phi}{\partial y} \, dA$$

is approximated by

$$\begin{aligned} \iint_E u b_1 \frac{\partial \phi}{\partial x} \, dA &\approx \frac{|\det \mathbf{T}|}{\det \mathbf{T}} \langle ((y_3 - y_1) \mathbf{M}_\xi^T + (-y_2 + y_1) \mathbf{M}_\nu^T) \cdot \text{diag}(\vec{wb}_1) \cdot \mathbf{M} \cdot \vec{u}, \vec{\phi} \rangle \\ \iint_E u b_2 \frac{\partial \phi}{\partial y} \, dA &\approx \frac{|\det \mathbf{T}|}{\det \mathbf{T}} \langle ((-x_3 + x_1) \mathbf{M}_\xi^T + (x_2 - x_1) \mathbf{M}_\nu^T) \cdot \text{diag}(\vec{wb}_2) \cdot \mathbf{M} \cdot \vec{u}, \vec{\phi} \rangle. \end{aligned}$$

The function  $a \nabla u \cdot \nabla \phi = a \left( \frac{\partial u}{\partial x} \frac{\partial \phi}{\partial x} + \frac{\partial u}{\partial y} \frac{\partial \phi}{\partial y} \right)$  has to be evaluated at the Gauss integration points  $\vec{g}_j$ , then multiplied by the Gauss weights  $w_i$  and added up. Use the vector  $\vec{w}\vec{a}$  with the values of the function  $a(x_i, y_i)$  and the weights  $w_i$  at the Gauss points to obtain

$$\begin{aligned} \iint_E a \frac{\partial u(\vec{x})}{\partial x} \frac{\partial \phi(\vec{x})}{\partial x} dA &= |\det \mathbf{T}| \iint_{\Omega} a(\vec{x}(\xi, \nu)) \frac{\partial u(\vec{x}(\xi, \nu))}{\partial x} \frac{\partial \phi(\vec{x}(\xi, \nu))}{\partial x} d\xi d\nu \\ &\approx \frac{|\det \mathbf{T}|}{(\det \mathbf{T})^2} \langle \mathbf{A}_x \cdot \vec{u}, \vec{\phi} \rangle = \frac{1}{|\det \mathbf{T}|} \langle \mathbf{A}_x \cdot \vec{u}, \vec{\phi} \rangle \\ \iint_E a \frac{\partial u(\vec{x})}{\partial y} \frac{\partial \phi(\vec{x})}{\partial y} dA &= |\det \mathbf{T}| \iint_{\Omega} a(\vec{x}(\xi, \nu)) \frac{\partial u(\vec{x}(\xi, \nu))}{\partial y} \frac{\partial \phi(\vec{x}(\xi, \nu))}{\partial y} d\xi d\nu \\ &\approx \frac{|\det \mathbf{T}|}{(\det \mathbf{T})^2} \langle \mathbf{A}_y \cdot \vec{u}, \vec{\phi} \rangle = \frac{1}{|\det \mathbf{T}|} \langle \mathbf{A}_y \cdot \vec{u}, \vec{\phi} \rangle \end{aligned}$$

where

$$\begin{aligned} \mathbf{A}_x &= \left[ (+y_3 - y_1) \mathbf{M}_\xi + (-y_2 + y_1) \mathbf{M}_\nu \right]^T \cdot \text{diag}(\vec{w}\vec{a}) \cdot \left[ (+y_3 - y_1) \mathbf{M}_\xi + (-y_2 + y_1) \mathbf{M}_\nu \right] \\ \mathbf{A}_y &= \left[ (-x_3 + x_1) \mathbf{M}_\xi + (+x_2 - x_1) \mathbf{M}_\nu \right]^T \cdot \text{diag}(\vec{w}\vec{a}) \cdot \left[ (-x_3 + x_1) \mathbf{M}_\xi + (+x_2 - x_1) \mathbf{M}_\nu \right]. \end{aligned}$$

### 6.6.6 Partial derivatives at the nodes

For post processing one also needs the partial derivatives of the function at the nodes. On the standard triangle  $\Omega$  use the formulas for the partial derivatives of the basis functions in expressions (49) and (50) to find them at the nodes, given by the  $(\xi, \nu)$  coordinates in Table 14 for cubic elements.

$$\begin{pmatrix} \varphi_\xi(\xi_1, \nu_1) \\ \varphi_\xi(\xi_2, \nu_2) \\ \varphi_\xi(\xi_3, \nu_3) \\ \varphi_\xi(\xi_4, \nu_4) \\ \varphi_\xi(\xi_5, \nu_5) \\ \varphi_\xi(\xi_6, \nu_6) \\ \varphi_\xi(\xi_7, \nu_7) \\ \varphi_\xi(\xi_8, \nu_8) \\ \varphi_\xi(\xi_9, \nu_9) \\ \varphi_\xi(\xi_{10}, \nu_{10}) \end{pmatrix} = \begin{bmatrix} \frac{-11}{2} & 1 & 0 & 0 & 0 & 0 & 0 & 9 & \frac{-9}{2} & 0 \\ -1 & \frac{11}{2} & 0 & 0 & 0 & 0 & 0 & \frac{9}{2} & -9 & 0 \\ -1 & 1 & 0 & \frac{-9}{2} & 9 & -9 & \frac{9}{2} & 0 & 0 & 0 \\ -1 & 1 & 0 & \frac{9}{2} & 0 & 0 & \frac{3}{2} & 3 & -3 & -6 \\ -1 & \frac{-1}{2} & 0 & 3 & 3 & -3 & 3 & \frac{3}{2} & 0 & -6 \\ \frac{1}{2} & 1 & 0 & -3 & 3 & -3 & -3 & 0 & \frac{-3}{2} & 6 \\ -1 & 1 & 0 & \frac{-3}{2} & 0 & 0 & \frac{-9}{2} & 3 & -3 & 6 \\ -1 & \frac{-1}{2} & 0 & 0 & 0 & 0 & 0 & \frac{-3}{2} & 3 & 0 \\ \frac{1}{2} & 1 & 0 & 0 & 0 & 0 & 0 & -3 & \frac{3}{2} & 0 \\ \frac{1}{2} & \frac{-1}{2} & 0 & \frac{3}{2} & 0 & 0 & \frac{-3}{2} & \frac{-3}{2} & \frac{3}{2} & 0 \end{bmatrix} \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \varphi_3 \\ \varphi_4 \\ \varphi_5 \\ \varphi_6 \\ \varphi_7 \\ \varphi_8 \\ \varphi_9 \\ \varphi_{10} \end{pmatrix} = \mathbf{N}_\xi \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \varphi_3 \\ \varphi_4 \\ \varphi_5 \\ \varphi_6 \\ \varphi_7 \\ \varphi_8 \\ \varphi_9 \\ \varphi_{10} \end{pmatrix}$$

and

$$\begin{pmatrix} \varphi_\nu(\xi_1, \nu_1) \\ \varphi_\nu(\xi_2, \nu_2) \\ \varphi_\nu(\xi_3, \nu_3) \\ \varphi_\nu(\xi_4, \nu_4) \\ \varphi_\nu(\xi_5, \nu_5) \\ \varphi_\nu(\xi_6, \nu_6) \\ \varphi_\nu(\xi_7, \nu_7) \\ \varphi_\nu(\xi_8, \nu_8) \\ \varphi_\nu(\xi_9, \nu_9) \\ \varphi_\nu(\xi_{10}, \nu_{10}) \end{pmatrix} = \begin{bmatrix} \frac{-11}{2} & 0 & 1 & 0 & 0 & \frac{-9}{2} & 9 & 0 & 0 & 0 \\ -1 & 0 & 1 & 9 & \frac{-9}{2} & 0 & 0 & \frac{9}{2} & -9 & 0 \\ -1 & 0 & \frac{11}{2} & 0 & 0 & -9 & \frac{9}{2} & 0 & 0 & 0 \\ -1 & 0 & \frac{-1}{2} & 3 & 3 & 0 & \frac{3}{2} & 3 & -3 & -6 \\ -1 & 0 & 1 & 0 & \frac{9}{2} & -3 & 3 & \frac{3}{2} & 0 & -6 \\ \frac{1}{2} & 0 & 1 & 0 & 0 & \frac{3}{2} & -3 & 0 & 0 & 0 \\ -1 & 0 & \frac{-1}{2} & 0 & 0 & 3 & \frac{-3}{2} & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & \frac{-3}{2} & -3 & 3 & \frac{-9}{2} & 0 & 6 \\ \frac{1}{2} & 0 & 1 & 3 & -3 & \frac{-3}{2} & 0 & -3 & -3 & 6 \\ \frac{1}{2} & 0 & \frac{-1}{2} & 0 & \frac{3}{2} & \frac{3}{2} & \frac{-3}{2} & \frac{-3}{2} & 0 & 0 \end{bmatrix} \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \varphi_3 \\ \varphi_4 \\ \varphi_5 \\ \varphi_6 \\ \varphi_7 \\ \varphi_8 \\ \varphi_9 \\ \varphi_{10} \end{pmatrix} = \mathbf{N}_\nu \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \varphi_3 \\ \varphi_4 \\ \varphi_5 \\ \varphi_6 \\ \varphi_7 \\ \varphi_8 \\ \varphi_9 \\ \varphi_{10} \end{pmatrix}$$

Now use the transformation formulas (41) and (42) to determine the gradient of a function on the general triangle

$$\varphi(x, y) = \sum_{i=1}^{10} \varphi_i \Phi_i(\xi(x, y), \nu(x, y))$$

at the nodes  $(x_i, y_i)$  in the general triangle  $E$ , leading to

$$\begin{pmatrix} \varphi_x(x_1, y_1) \\ \varphi_x(x_2, y_2) \\ \vdots \\ \varphi_x(x_{10}, y_{10}) \end{pmatrix} = \frac{1}{\det(\mathbf{T})} \left[ (+y_3 - y_1) \mathbf{N}_\xi^T + (-y_2 + y_1) \mathbf{N}_\nu^T \right] \cdot \vec{\varphi},$$

$$\begin{pmatrix} \varphi_y(x_1, y_1) \\ \varphi_y(x_2, y_2) \\ \vdots \\ \varphi_y(x_{10}, y_{10}) \end{pmatrix} = \frac{1}{\det(\mathbf{T})} \left[ (-x_3 + x_1) \mathbf{N}_\xi^T + (+x_2 - x_1) \mathbf{N}_\nu^T \right] \cdot \vec{\varphi}.$$

These results are useful to evaluate the gradient at the nodes. Observe that the results depends on the triangle used for the interpolation and a node is typically member of more than one triangle.

### 6.6.7 Integration over boundary segments

In expression (27) integrals over the section  $\Gamma_2$  of the boundary are required.

$$\int_{\Gamma_2} \phi (g_2 + g_3 u) ds$$

For triangular domains the boundary consists of straight line segments. Thus replace the integral by a sum of line integrals and use a Gauss integration. Based on the two endpoints  $\vec{x}_1$  and  $\vec{x}_3$  and the midpoint  $\vec{x}_2 = \frac{1}{2} (\vec{x}_1 + \vec{x}_3)$  use the values at three Gauss integration points.

Based on

$$\begin{aligned} \int_{-h/2}^{h/2} f(x) dx &\approx \frac{h}{18} \left( 5 f\left(-\frac{\sqrt{3}}{2\sqrt{5}} h\right) + 8 f(0) + 5 f\left(\frac{\sqrt{3}}{2\sqrt{5}} h\right) \right) \\ &= \frac{h}{18} \left( 5 f\left(-\frac{\sqrt{15}}{10} h\right) + 8 f(0) + 5 f\left(\frac{\sqrt{15}}{10} h\right) \right) \end{aligned}$$

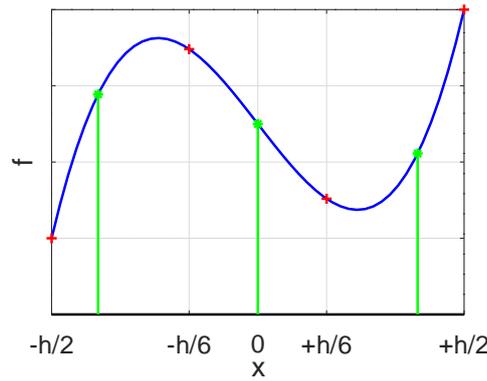


Figure 72: The interpolation from four nodes to three Gauss points on an interval  $[-\frac{h}{2}, +\frac{h}{2}]$

polynomials up to degree 5 are integrated exactly, thus the error on one interval is proportional to  $h^7$ . To evaluate a function at the Gauss points

$$\begin{aligned}\vec{p}_1 &= \frac{1}{2} (\vec{x}_1 + \vec{x}_4) - \frac{\sqrt{3}}{2\sqrt{5}} (\vec{x}_4 - \vec{x}_1) \\ \vec{p}_2 &= \frac{1}{2} (\vec{x}_1 + \vec{x}_4) \\ \vec{p}_3 &= \frac{1}{2} (\vec{x}_1 + \vec{x}_4) + \frac{\sqrt{3}}{2\sqrt{5}} (\vec{x}_4 - \vec{x}_1)\end{aligned}$$

use a cubic interpolation of a function with  $f_{-2} = f(-h/2)$ ,  $f_{-1} = f(-h/6)$ ,  $f_{+1} = f(+h/6)$  and  $f_{+2} = f(+h/2)$ . Required are the values at  $x = 0$  and  $x = \pm \frac{\sqrt{15}}{10} h \approx \pm 0.387 h$ . This is illustrated in Figure 72 with the values of the function  $f(x)$  indicated by red spots and the interpolation position and values in green. The computations are tedious shown at the end of this section in Subsection 6.6.8 and lead to

$$\begin{pmatrix} u(\vec{p}_1) \\ u(\vec{p}_2) \\ u(\vec{p}_3) \end{pmatrix} = \mathbf{M}_B \begin{pmatrix} f_{-2} \\ f_{-1} \\ f_{+1} \\ f_{+2} \end{pmatrix} \approx \begin{bmatrix} 0.4880 & 0.7479 & -0.2979 & 0.06199 \\ -0.0625 & 0.5625 & 0.5625 & -0.0625 \\ 0.06199 & -0.2979 & 0.7479 & 0.4880 \end{bmatrix} \begin{pmatrix} f_{-2} \\ f_{-1} \\ f_{+1} \\ f_{+2} \end{pmatrix}$$

With the length  $L = \sqrt{(x_4 - x_1)^2 + (y_4 - y_1)^2}$  of the segment this leads to the approximations

$$\begin{aligned}\int_{\text{edge}} \phi g_2 ds &\approx \frac{L}{18} \left\langle \mathbf{M}_B \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \end{pmatrix}, \begin{pmatrix} 5 g_2(\vec{p}_1) \\ 8 g_2(\vec{p}_2) \\ 5 g_2(\vec{p}_3) \end{pmatrix} \right\rangle = \frac{L}{18} \left\langle \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \end{pmatrix}, \mathbf{M}_B^T \begin{pmatrix} 5 g_2(\vec{p}_1) \\ 8 g_2(\vec{p}_2) \\ 5 g_2(\vec{p}_3) \end{pmatrix} \right\rangle \\ \int_{\text{edge}} \phi g_3 u ds &\approx \frac{L}{18} \left\langle \mathbf{M}_B \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \end{pmatrix}, \begin{bmatrix} 5 g_3(\vec{p}_1) & 0 & 0 \\ 0 & 8 g_3(\vec{p}_2) & 0 \\ 0 & 0 & 5 g_3(\vec{p}_3) \end{bmatrix} \mathbf{M}_B \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} \right\rangle\end{aligned}$$

$$= \frac{L}{18} \left\langle \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \end{pmatrix}, \mathbf{M}_B^T \begin{bmatrix} 5g_3(\vec{p}_1) & 0 & 0 \\ 0 & 8g_3(\vec{p}_2) & 0 \\ 0 & 0 & 5g_3(\vec{p}_3) \end{bmatrix} \mathbf{M}_B \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} \right\rangle.$$

The first expression will lead to a contribution to the RHS vector of the linear system to be solved, while the second expression will lead to entries in the matrix. These approximate integrations lead to the exact result if the function to be integrated is a polynomial of degree 5, or less. If  $h$  is the typical length of an edge then the error is of the order  $h^7$  for one line segment and thus of order  $h^6$  for the total boundary. This boundary integration is used for third order elements. The second expression is of the form

$$\int \phi g_3 u ds \approx \langle \vec{\phi}, \mathbf{B} \vec{u} \rangle = \left\langle \begin{pmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \end{pmatrix}, \begin{bmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \\ b_{31} & b_{32} & b_{33} & b_{34} \\ b_{41} & b_{42} & b_{43} & b_{44} \end{bmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} \right\rangle$$

and its effect on the linear system  $\mathbf{A} \vec{u} + \mathbf{W} \vec{f} = \vec{0}$  to be solved depends on nodes being on the Dirichlet section of the boundary or the Neumann section.

- If  $u_1$  and  $u_4$  are free, i.e. not on the Dirichlet section, then  $u_2$  and  $u_3$  are free too. All entries of the matrix  $\mathbf{B}$  have to be added to the global stiffness matrix  $\mathbf{A}$ .
- If  $u_1$  and  $u_4$  are on the Dirichlet section, then  $u_2$  and  $u_3$  are on the Dirichlet section too. Nothing has to be added to  $\mathbf{A}$  and  $\vec{f}$ .
- If  $u_1, u_2$  and  $u_3$  are free and  $u_4$  is on the Dirichlet section, then only the first three expressions

$$\begin{aligned} b_{11} u_1 + b_{12} u_2 + b_{13} u_3 + b_{14} u_4 &= b_{11} u_1 + b_{12} u_2 + b_{13} u_3 + b_{14} d_4 \\ b_{21} u_1 + b_{22} u_2 + b_{23} u_3 + b_{24} u_4 &= b_{21} u_1 + b_{22} u_2 + b_{23} u_3 + b_{24} d_4 \\ b_{31} u_1 + b_{32} u_2 + b_{33} u_3 + b_{34} u_4 &= b_{31} u_1 + b_{32} u_2 + b_{33} u_3 + b_{34} d_4 \end{aligned}$$

have to be taken into account.  $d_4$  is the Dirichlet value at the position of  $u_4$ . The contributions  $b_{14} d_4$ ,  $b_{24} d_4$  and  $b_{34} d_4$  have to be added to  $\mathbf{W} \vec{f}$ , the other expressions to  $\mathbf{A}$ .

- If  $u_2, u_3$  and  $u_4$  are free and  $u_1$  is on the Dirichlet section, then only the least three expressions

$$\begin{aligned} b_{21} u_1 + b_{22} u_2 + b_{23} u_3 + b_{24} u_4 &= b_{21} d_1 + b_{22} u_2 + b_{23} u_3 + b_{24} u_4 \\ b_{31} u_1 + b_{32} u_2 + b_{33} u_3 + b_{34} u_4 &= b_{31} d_1 + b_{32} u_2 + b_{33} u_3 + b_{34} u_4 \\ b_{41} u_1 + b_{42} u_2 + b_{43} u_3 + b_{44} u_4 &= b_{41} d_1 + b_{42} u_2 + b_{43} u_3 + b_{44} u_4 \end{aligned}$$

have to be taken into account.  $d_1$  is the Dirichlet value at the position of  $u_1$ . The contributions  $b_{21} d_1$ ,  $b_{31} d_1$  and  $b_{41} d_1$  have to be added to  $\mathbf{W} \vec{f}$ , the other expressions to  $\mathbf{A}$ .

### 6.6.8 From a polynomial interpolation to the Gauss integration points

For an interval  $[-h/2, +h/2]$  use a polynomial  $p(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3$ , leading to

$$f_{-2} = p(-h/2) = c_0 - \frac{1}{2} h c_1 + \frac{1}{4} h^2 c_2 - \frac{1}{8} h^3 c_3$$

$$\begin{aligned}
f_{-1} = p(-h/6) &= c_0 - \frac{1}{6} h c_1 + \frac{1}{36} h^2 c_2 - \frac{1}{216} h^3 c_3 \\
f_{+1} = p(+h/6) &= c_0 + \frac{1}{6} h c_1 + \frac{1}{36} h^2 c_2 + \frac{1}{216} h^3 c_3 \\
f_{+2} = p(+h/2) &= c_0 + \frac{1}{2} h c_1 + \frac{1}{4} h^2 c_2 + \frac{1}{8} h^3 c_3
\end{aligned}$$

or with a matrix notation

$$\begin{bmatrix} +1 & -\frac{1}{2} & +\frac{1}{4} & -\frac{1}{8} \\ +1 & -\frac{1}{6} & +\frac{1}{36} & -\frac{1}{216} \\ +1 & +\frac{1}{6} & +\frac{1}{36} & +\frac{1}{216} \\ +1 & +\frac{1}{2} & +\frac{1}{4} & +\frac{1}{8} \end{bmatrix} \begin{pmatrix} c_0 \\ h c_1 \\ h^2 c_2 \\ h^3 c_3 \end{pmatrix} = \begin{pmatrix} f_{-2} \\ f_{-1} \\ f_{+1} \\ f_{+2} \end{pmatrix}.$$

The corresponding inverse matrix leads to

$$\begin{pmatrix} c_0 \\ h c_1 \\ h^2 c_2 \\ h^3 c_3 \end{pmatrix} = \frac{1}{16} \begin{bmatrix} -1 & +9 & +9 & -1 \\ +2 & -54 & +54 & -2 \\ +36 & -36 & -36 & +36 \\ -72 & +216 & -216 & +72 \end{bmatrix} \begin{pmatrix} f_{-2} \\ f_{-1} \\ f_{+1} \\ f_{+2} \end{pmatrix}.$$

With  $\lambda = \frac{\sqrt{15}}{10} \approx 0.3873$  and  $p(\lambda h) = c_0 + \lambda c_1 h + \lambda^2 c_2 h^2 + \lambda^3 c_3 h^3$  obtain

$$\begin{aligned}
p(\lambda h) &= \frac{1}{16} \begin{bmatrix} 1 & \lambda & \lambda^2 & \lambda^3 \end{bmatrix} \begin{bmatrix} -1 & +9 & +9 & -1 \\ +2 & -54 & +54 & -2 \\ +36 & -36 & -36 & +36 \\ -72 & +216 & -216 & +72 \end{bmatrix} \begin{pmatrix} f_{-2} \\ f_{-1} \\ f_{+1} \\ f_{+2} \end{pmatrix} \\
\begin{pmatrix} p(-\lambda h) \\ p(0) \\ p(+\lambda h) \end{pmatrix} &= \frac{1}{16} \begin{bmatrix} 1 & -\lambda & \lambda^2 & -\lambda^3 \\ 1 & 0 & 0 & 0 \\ 1 & +\lambda & \lambda^2 & +\lambda^3 \end{bmatrix} \begin{bmatrix} -1 & +9 & +9 & -1 \\ +2 & -54 & +54 & -2 \\ +36 & -36 & -36 & +36 \\ -72 & +216 & -216 & +72 \end{bmatrix} \begin{pmatrix} f_{-2} \\ f_{-1} \\ f_{+1} \\ f_{+2} \end{pmatrix} \\
&= \mathbf{M}_B \begin{pmatrix} f_{-2} \\ f_{-1} \\ f_{+1} \\ f_{+2} \end{pmatrix} \approx \begin{bmatrix} 0.4880 & 0.7479 & -0.2979 & 0.06199 \\ -0.0625 & 0.5625 & 0.5625 & -0.0625 \\ 0.06199 & -0.2979 & 0.7479 & 0.4880 \end{bmatrix} \begin{pmatrix} f_{-2} \\ f_{-1} \\ f_{+1} \\ f_{+2} \end{pmatrix}
\end{aligned}$$

This matrix  $\mathbf{M}_B$  is used on the segments of the boundary to determine the values at the Gauss integration points, given the values of the four nodes of a third order element.

## 6.7 Convergence of the approximate solutions $u_h$ to the exact solution $u$

A key feature of a good FEM algorithm is a rapid convergence. As the diameter  $h$  of the triangles converges to 0, the approximate solution  $u_h(x, y)$  should converge to the exact solution  $u(x, y)$ . The statements below are correct for very smooth exact solutions and “nice” domains. Find more information in books on the mathematical background of FEM, e.g. [AxelBark84] or consult [Stah08].

It is convenient to state the approximation results using two norms on the function space  $L_2(\Omega)$  and the Sobolev space  $V = H^1(\Omega) = W^{1,2}(\Omega)$ . The norms are given by

$$\|u\|_2^2 = \iint_{\Omega} u^2(x, y) \, dA \quad \text{and} \quad \|u\|_V^2 = \iint_{\Omega} u^2(x, y) + \|\nabla u(x, y)\|^2 \, dA.$$

The convergence results assume that the meshes are well defined, e.g. satisfy a minimal angle condition.

- If the solutions  $u_h$  are generated by first order, triangular elements, i.e. piecewise linear functions, then

$$\|u_h - u\|_V \leq C h \quad \text{and} \quad \|u_h - u\|_2 \leq C_1 h^2$$

for some constants  $C$  and  $C_1$  independent on  $h$ . A short formulation is

- $u_h$  converges to  $u$  with an error proportional to  $h^2$  as  $h \rightarrow 0$ .
  - $\nabla u_h$  converges to  $\nabla u$  with an error proportional to  $h$  as  $h \rightarrow 0$ .
- If the solutions  $u_h$  are generated by second order, triangular elements, i.e. piecewise quadratic functions, then

$$\|u_h - u\|_V \leq C h^2 \quad \text{and} \quad \|u_h - u\|_2 \leq C_1 h^3$$

for some constants  $C$  and  $C_1$  independent on  $h$ . A short formulation is

- $u_h$  converges to  $u$  with an error proportional to  $h^3$  as  $h \rightarrow 0$ .
  - $\nabla u_h$  converges to  $\nabla u$  with an error proportional to  $h^2$  as  $h \rightarrow 0$ .
- If the solutions  $u_h$  are generated by third order, triangular elements, i.e. piecewise cubic functions, then

$$\|u_h - u\|_V \leq C h^3 \quad \text{and} \quad \|u_h - u\|_2 \leq C_1 h^4$$

for some constants  $C$  and  $C_1$  independent on  $h$ . A short formulation is

- $u_h$  converges to  $u$  with an error proportional to  $h^4$  as  $h \rightarrow 0$ .
- $\nabla u_h$  converges to  $\nabla u$  with an error proportional to  $h^3$  as  $h \rightarrow 0$ .

Observe that the convergence results are about the integral of differences, and not point-wise estimates. In addition the exact solution  $u$  is assumed to be smooth. Thus one has to be careful when using the estimates for problems with limited regularity of the type in Section 9.4.

## 6.8 Dynamic problems

There are two distinct classes of dynamic problems:

- Parabolic problems with the heat equation  $\dot{u} = \Delta u$  as the typical example.
- Hyperbolic problems with the wave equation  $\ddot{u} = \Delta u$  as the typical example.

For both types the following sections will present unconditionally stable, consistent time stepping algorithms.

### 6.8.1 Dynamic problems of the heat equation type

Examine an IBVP (4) of parabolic type.

$$\begin{aligned} \rho \frac{\partial}{\partial t} u - \nabla \cdot (a \nabla u - u \vec{b}) + b_0 u &= f & \text{for } (x, y, t) \in \Omega \times (0, T] \\ u &= g_1 & \text{for } (x, y, t) \in \Gamma_1 \times (0, T] \\ \vec{n} \cdot (a \nabla u - u \vec{b}) &= g_2 + g_3 u & \text{for } (x, y, t) \in \Gamma_2 \times (0, T] \\ u &= u_0 & \text{on } \Omega \text{ at } t = 0 \end{aligned}$$

First the problem is reduced to a new problem with homogeneous boundary conditions, i.e.  $g_1 = g_2 = 0$ . Solve the static problem with nonhomogeneous boundary conditions.

$$\begin{aligned} -\nabla \cdot (a \nabla u_B - u_B \vec{b}) + b_0 u_B &= 0 & \text{for } (x, y, t) \in \Omega \\ u_B &= g_1 & \text{for } (x, y) \in \Gamma_1 \\ \vec{n} \cdot (a \nabla u_B + u_B \vec{b}) &= g_2 + g_3 u_B & \text{for } (x, y, t) \in \Gamma_2 \end{aligned} \quad (52)$$

Then the new function  $v(x, y, t) = u(x, y, t) - u_B(x, y)$  is a solution of an initial boundary value problem with no constant boundary contributions, i.e.  $g_1 = g_2 = 0$ .

$$\begin{aligned} \rho \frac{\partial}{\partial t} v - \nabla \cdot (a \nabla v - v \vec{b}) + b_0 v &= f & \text{for } (x, y, t) \in \Omega \times (0, T] \\ v &= 0 & \text{for } (x, y, t) \in \Gamma_1 \times (0, T] \\ \vec{n} \cdot (a \nabla v + v \vec{b}) &= g_3 v & \text{for } (x, y, t) \in \Gamma_2 \times (0, T] \\ v &= u_0 - u_B & \text{on } \Omega \text{ at } t = 0 \end{aligned}$$

This equation is transformed to a system of ordinary differential equations.

$$\mathbf{W} \frac{d}{dt} \vec{v}(t) + \mathbf{A} \vec{v}(t) = \vec{f}(t) \quad \text{with } \vec{v}(0) = \vec{v}_0. \quad (53)$$

The implementation assumes that the coefficient functions  $\rho$ ,  $a$ ,  $b_0$ ,  $\vec{b}$  and  $g_i$  depend on  $(x, y)$ , while  $f$  may depend on time  $t$  and the position  $(x, y)$ . Then use a Crank–Nicolson<sup>20</sup> approximation to advance the solution from time  $t$  to  $t + \Delta t$ .

$$\begin{aligned} \mathbf{W} \frac{\vec{v}(t + \Delta t) - \vec{v}(t)}{\Delta t} &= -\mathbf{A} \frac{\vec{v}(t + \Delta t) + \vec{v}(t)}{2} + \vec{f}(t + \Delta t/2) \\ \left( \mathbf{W} + \frac{\Delta t}{2} \mathbf{A} \right) \vec{v}(t + \Delta t) &= + \left( \mathbf{W} - \frac{\Delta t}{2} \mathbf{A} \right) \vec{v}(t) + \Delta t \vec{f}(t + \Delta t/2) \end{aligned}$$

For each time step such a system has to be solved. Observe that the matrix on the left does not change as time advances. Using an sparsity preserving LU factorization of the matrix on the left, these systems can be solved efficiently. The matrices  $\mathbf{P}$  and  $\mathbf{Q}$  are permutation matrices with  $\mathbf{P}^{-1} = \mathbf{P}^T$ . A substantial amount of time has to be used to perform the LU factorization, but then the time stepping is fast.

$$\begin{aligned} \mathbf{P} \left( \mathbf{W} + \frac{\Delta t}{2} \mathbf{A} \right) \mathbf{Q} &= \mathbf{L} \mathbf{U} & \text{LU factorization} \\ \left( \mathbf{W} + \frac{\Delta t}{2} \mathbf{A} \right) \vec{v} &= \vec{b} & \text{system to be solved} \\ \mathbf{P} \left( \mathbf{W} + \frac{\Delta t}{2} \mathbf{A} \right) \mathbf{Q} \mathbf{Q}^{-1} \vec{v} &= \mathbf{P} \vec{b} \\ \mathbf{L} \mathbf{U} \mathbf{Q}^{-1} \vec{v} &= \mathbf{P} \vec{b} \\ \vec{v} &= \mathbf{Q} (\mathbf{U} \setminus (\mathbf{L} \setminus (\mathbf{P} \vec{b}))) & \text{in the Octave code} \end{aligned}$$

<sup>20</sup>This is a standard choice and unconditionally stable, see e.g. [Stah08, §4]. For the 1D dynamic problems with IBVP1D () (Section 7.6 on page 176) more time stepping algorithms are available.

With the computed  $\vec{v}(t)$  then find the solution  $\vec{u}(t) = \vec{v}(t) + \vec{u}_B$  of the original problem.

If the matrix  $\mathbf{A}$  is symmetric and positive definite one can use Cholesky factorization with row and column permutations to preserve the sparsity, as much as possible. This should be faster than a LU factorization.

$$\begin{aligned} \mathbf{Q}^T (\mathbf{W} + \frac{\Delta t}{2} \mathbf{A}) \mathbf{Q} &= \mathbf{R}^T \mathbf{R} && \text{Cholesky factorization} \\ (\mathbf{W} + \frac{\Delta t}{2} \mathbf{A}) \vec{v} &= \vec{b} && \text{system to be solved} \\ \mathbf{Q}^T (\mathbf{W} + \frac{\Delta t}{2} \mathbf{A}) \mathbf{Q} \mathbf{Q}^T \vec{v} &= \mathbf{Q}^T \vec{b} \\ \mathbf{R}^T \mathbf{R} \mathbf{Q}^T \vec{v} &= \mathbf{Q}^T \vec{b} \\ \vec{v} &= \mathbf{Q} (\mathbf{R} \setminus (\mathbf{R}^T \setminus (\mathbf{Q}^T \vec{b}))) && \text{in the Octave code} \end{aligned}$$

The *Octave* manual claims that a lower Cholesky factorization is often faster.

$$\begin{aligned} \mathbf{Q}^T (\mathbf{W} + \frac{\Delta t}{2} \mathbf{A}) \mathbf{Q} &= \mathbf{L} \mathbf{L}^T && \text{lower Cholesky factorization} \\ (\mathbf{W} + \frac{\Delta t}{2} \mathbf{A}) \vec{v} &= \vec{b} && \text{system to be solved} \\ \mathbf{Q}^T (\mathbf{W} + \frac{\Delta t}{2} \mathbf{A}) \mathbf{Q} \mathbf{Q}^T \vec{v} &= \mathbf{Q}^T \vec{b} \\ \mathbf{L} \mathbf{L}^T \mathbf{Q}^T \vec{v} &= \mathbf{Q}^T \vec{b} \\ \vec{v} &= \mathbf{Q} (\mathbf{L}^T \setminus (\mathbf{L} \setminus (\mathbf{Q}^T \vec{b}))) && \text{in the Octave code} \end{aligned}$$

### 6.8.2 Using eigenvalues for dynamic problems of the heat equation type

With equation (53) for  $\vec{f} = \vec{0}$

$$\mathbf{W} \frac{d}{dt} \vec{v}(t) + \mathbf{A} \vec{v}(t) = \vec{0} \quad \text{with} \quad \vec{v}(0) = \vec{v}_0$$

observe that a generalized eigenvalue  $\lambda$  with eigenvector  $\vec{v}$ , i.e.

$$\mathbf{A} \vec{v} = \lambda \mathbf{W} \vec{v}$$

leads to a solution  $\vec{u}(t) = c \exp(-\lambda t) \vec{v}$ , since

$$\begin{aligned} \mathbf{W} \frac{d}{dt} \vec{u}(t) &= -\lambda \mathbf{W} \vec{v} \exp(-\lambda t) \\ \mathbf{A} \vec{u}(t) &= +\lambda \mathbf{W} \vec{v} \exp(-\lambda t) \end{aligned}$$

Thus for  $\lambda > 0$  find an exponentially decaying solution of the IBVP.

### 6.8.3 Dynamic problems of the wave equation type

Examine an IBVP (6) of hyperbolic type.

$$\begin{aligned} \rho \frac{\partial^2}{\partial t^2} u + 2\alpha \frac{\partial}{\partial t} u - \nabla \cdot (a \nabla u - u \vec{b}) + b_0 u &= f && \text{for } (x, y, t) \in \Omega \times (0, T] \\ u &= g_1 && \text{for } (x, y, t) \in \Gamma_1 \times (0, T] \\ \vec{n} \cdot (a \nabla u - u \vec{b}) &= g_2 + g_3 u && \text{for } (x, y, t) \in \Gamma_2 \times (0, T] \\ u &= u_0 && \text{on } \Omega \text{ at } t = 0 \\ \frac{\partial}{\partial t} u &= v_0 && \text{on } \Omega \text{ at } t = 0 \end{aligned}$$

First the problem is reduced to a new problem with homogeneous boundary conditions, i.e.  $g_1 = g_2 = 0$ , using (52). Then the new function  $v(x, y, t) = u(x, y, t) - u_B(x, y)$  is a solution of an initial boundary value problem with no constant boundary contributions, i.e.  $g_1 = g_2 = 0$ .

$$\begin{aligned} \rho \frac{\partial^2}{\partial t^2} v + 2\alpha \frac{\partial}{\partial t} v(t) - \nabla \cdot (a \nabla u - v \vec{b}) + b_0 v &= f & \text{for } (x, y, t) \in \Omega \times (0, T] \\ v &= 0 & \text{for } (x, y, t) \in \Gamma_1 \times (0, T] \\ \vec{n} \cdot (a \nabla v - v \vec{b}) &= g_3 v & \text{for } (x, y, t) \in \Gamma_2 \times (0, T] \\ v &= u_0 - u_B & \text{on } \Omega \text{ at } t = 0 \\ \frac{\partial}{\partial t} v &= v_0 & \text{on } \Omega \text{ at } t = 0 \end{aligned}$$

This equation is transformed to a system of ordinary differential equations.

$$\mathbf{W} \frac{d^2}{dt^2} \vec{v}(t) + 2\mathbf{D} \frac{d}{dt} \vec{v}(t) + \mathbf{A} \vec{v}(t) = \vec{f}(t) \quad \text{with} \quad \vec{v}(0) = \vec{u}_0 - \vec{u}_B, \quad \frac{d}{dt} \vec{v}(0) = \vec{v}_0 \quad (54)$$

The implementation assumes that the coefficient functions  $\rho$ ,  $\alpha$ ,  $a$ ,  $b_0$ ,  $\vec{b}$  and  $g_i$  depend on  $(x, y)$ , while  $f$  may depend on time  $t$  and the position  $(x, y)$ . Then use an implicit approximation<sup>21</sup> to advance the solution from time  $t - \Delta t$  and  $t$  to  $t + \Delta t$ .

$$\begin{aligned} \mathbf{W} \frac{d^2}{dt^2} \vec{v}(t) &= -2\mathbf{D} \frac{d}{dt} \vec{v}(t) - \mathbf{A} \vec{v}(t) + \vec{f}(t) \\ \mathbf{W} \frac{\vec{v}(t - \Delta t) - 2\vec{v}(t) + \vec{v}(t + \Delta t)}{(\Delta t)^2} &= -2\mathbf{D} \frac{\vec{v}(t + \Delta t) - \vec{v}(t - \Delta t)}{2\Delta t} - \\ &\quad - \mathbf{A} \frac{\vec{v}(t - \Delta t) + 2\vec{v}(t) + \vec{v}(t + \Delta t)}{4} + \vec{f}(t) \\ \left( +\mathbf{W} + \Delta t \mathbf{D} + \frac{(\Delta t)^2}{4} \mathbf{A} \right) \vec{v}(t + \Delta t) &= - \left( \mathbf{W} - \Delta t \mathbf{D} + \frac{(\Delta t)^2}{4} \mathbf{A} \right) \vec{v}(t - \Delta t) + \\ &\quad + \left( 2\mathbf{W} - \frac{(\Delta t)^2}{2} \mathbf{A} \right) \vec{v}(t) + (\Delta t)^2 \vec{f}(t) \end{aligned}$$

This scheme is unconditionally stable and consistent of order 2. Observe that the matrices do not change as time advances. Thus use again a sparsity preserving LU factorization for the time stepping. The above scheme is unconditionally stable, at least for constant coefficients<sup>22</sup>.

- To construct the solution at the initial time  $\Delta t$  use the initial value  $u_0$  and initial velocity  $v_0$  and a scheme with the same order of consistency, with respect to time. An explicit scheme for the first step leads to

$$\begin{aligned} \frac{d}{dt} \vec{v}(0) &= \vec{v}_0 \approx \frac{\vec{v}(\Delta t) - \vec{v}(-\Delta t)}{2\Delta t} \implies \vec{v}(-\Delta t) \approx \vec{v}(\Delta t) - 2\Delta t \vec{v}_0 \\ \mathbf{W} \frac{d^2}{dt^2} \vec{v}(t) &= -2\mathbf{D} \frac{d}{dt} \vec{v}(t) - \mathbf{A} \vec{v}(t) + \vec{f}(t) \\ \mathbf{W} \frac{\vec{v}(t - \Delta t) - 2\vec{v}(t) + \vec{v}(t + \Delta t)}{(\Delta t)^2} &= -2\mathbf{D} \frac{\vec{v}(t + \Delta t) - \vec{v}(t - \Delta t)}{2\Delta t} - \mathbf{A} \vec{v}(t) + \vec{f}(t) \\ (\mathbf{W} + \Delta t \mathbf{D}) \vec{v}(t + \Delta t) &= -(\mathbf{W} - \Delta t \mathbf{D}) \vec{v}(t - \Delta t) + 2\mathbf{W} \vec{v}(t) + (\Delta t)^2 (-\mathbf{A} \vec{v}(t) + \vec{f}(t)) \\ (\mathbf{W} + \Delta t \mathbf{D}) \vec{v}(\Delta t) &= -(\mathbf{W} - \Delta t \mathbf{D}) (\vec{v}(\Delta t) - 2\Delta t \vec{v}_0) + \\ &\quad + 2\mathbf{W} (\vec{u}_0 - \vec{u}_B) + (\Delta t)^2 (-\mathbf{A} (\vec{u}_0 - \vec{u}_B) + \vec{f}(0)) \end{aligned}$$

<sup>21</sup>This is a standard choice and unconditionally stable, see e.g. [Stah08, §4].

<sup>22</sup>I have a proof in `WaveStability.tex`.

$$\begin{aligned}
2 \mathbf{W} \vec{v}(\Delta t) &= +2 (\mathbf{W} - \Delta t \mathbf{D}) \Delta t \vec{v}_0 + \\
&\quad +2 \mathbf{W} (\vec{u}_0 - \vec{u}_B) + (\Delta t)^2 (-\mathbf{A} (\vec{u}_0 - \vec{u}_B) + \vec{f}(0)) \\
\mathbf{W} \vec{v}(\Delta t) &= (\mathbf{W} - \Delta t \mathbf{D}) \Delta t \vec{v}_0 + \\
&\quad + \mathbf{W} (\vec{u}_0 - \vec{u}_B) + \frac{1}{2} (\Delta t)^2 (-\mathbf{A} (\vec{u}_0 - \vec{u}_B) + \vec{f}(0)).
\end{aligned}$$

This is currently implemented. The conditional stability for this single step should not cause a major problem.

- One could also use  $\vec{v}(-\Delta t) \approx \vec{v}(\Delta t) - 2 \Delta t \vec{v}_0$  in the implicit scheme at  $t = 0$ .

$$\begin{aligned}
\left( +\mathbf{W} + \Delta t \mathbf{D} + \frac{(\Delta t)^2}{4} \mathbf{A} \right) \vec{v}(t + \Delta t) &= - \left( \mathbf{W} - \Delta t \mathbf{D} + \frac{(\Delta t)^2}{4} \mathbf{A} \right) \vec{v}(t - \Delta t) + \\
&\quad + \left( 2 \mathbf{W} - \frac{(\Delta t)^2}{2} \mathbf{A} \right) \vec{v}(t) + (\Delta t)^2 \vec{f}(t) \\
\left( +\mathbf{W} + \Delta t \mathbf{D} + \frac{(\Delta t)^2}{4} \mathbf{A} \right) \vec{v}(\Delta t) &= - \left( \mathbf{W} - \Delta t \mathbf{D} + \frac{(\Delta t)^2}{4} \mathbf{A} \right) (\vec{v}(\Delta t) - 2 \Delta t \vec{v}_0) + \\
&\quad + \left( 2 \mathbf{W} - \frac{(\Delta t)^2}{2} \mathbf{A} \right) \vec{v}(0) + (\Delta t)^2 \vec{f}(0) \\
\left( +2 \mathbf{W} + 2 \frac{(\Delta t)^2}{4} \mathbf{A} \right) \vec{v}(\Delta t) &= +2 \Delta t \left( \mathbf{W} - \Delta t \mathbf{D} + \frac{(\Delta t)^2}{4} \mathbf{A} \right) \vec{v}_0 + \\
&\quad + \left( 2 \mathbf{W} - \frac{(\Delta t)^2}{2} \mathbf{A} \right) \vec{v}(0) + (\Delta t)^2 \vec{f}(0)
\end{aligned}$$

This initial step requires solving a new system of linear equations. If there is no damping term ( $\mathbf{D} = 0$ ) it is the same system as for the time stepping, thus should be used.

#### 6.8.4 Using eigenvalues for dynamic problems of the wave equation type

With equation (54) for  $\vec{f} = \vec{0}$  and a damping factor  $D \mathbf{W}$  with a constant  $D \geq 0$  (instead of the matrix  $\mathbf{D}$ )

$$\mathbf{W} \frac{d^2}{dt^2} \vec{v}(t) + 2 D \mathbf{W} \frac{d}{dt} \vec{v}(t) + \mathbf{A} \vec{v}(t) = \vec{0} \quad (55)$$

observe that a generalized eigenvalue  $\lambda > 0$  with eigenvector  $\vec{v}$ , i.e.  $\mathbf{A} \vec{v} = \lambda \mathbf{W} \vec{v}$  and weak damping  $0 \leq D < \sqrt{\lambda}$  leads to a solution  $\vec{u}(t) = \exp(\mu t) \vec{v}$  with  $\mu \in \mathbb{C}$ , since

$$\begin{aligned}
\vec{0} &= \mu^2 \mathbf{W} \vec{v} \exp(\mu t) + \mu 2 D \mathbf{W} \vec{v} \exp(\mu t) + \lambda \mathbf{W} \vec{v} \exp(\mu t) \\
0 &= \mu^2 + \mu 2 D + \lambda \\
\mu_{1,2} &= \frac{1}{2} \left( -2 D \pm \sqrt{4 D^2 - 4 \lambda} \right) = -D \pm i \sqrt{\lambda - D^2} \in \mathbb{C}
\end{aligned}$$

Thus the real solutions are of the form

$$\vec{u}(t) = \exp(-D t) \left( \vec{v}_1 \cos(\sqrt{\lambda - D^2} t) + \vec{v}_2 \sin(\sqrt{\lambda - D^2} t) \right).$$

The angular velocity of the exponentially decaying oscillations is given by  $\omega = \sqrt{\lambda - D^2}$ .

- For the case of strong damping  $D > \sqrt{\lambda}$  use

$$\mu_{1,2} = -D \pm \sqrt{D^2 - \lambda} \in \mathbb{R}$$

to find two exponentially decaying solutions

$$\vec{u}(t) = c_1 \exp(-D + \sqrt{D^2 - \lambda} t) \vec{v}_1 + c_2 \exp(-D - \sqrt{D^2 - \lambda} t) \vec{v}_2.$$

- If the damping term is not in the special form  $D \mathbf{W} \frac{d}{dt} \vec{v}(t)$  the above, simple approach does not work. Instead replace equation (55) by the first order system

$$\frac{d}{dt} \begin{pmatrix} v(t) \\ \mathbf{W} \frac{d}{dt} \vec{v}(t) \end{pmatrix} = \begin{pmatrix} \frac{d}{dt} \vec{v}(t) \\ -2 \mathbf{D} \frac{d}{dt} \vec{v}(t) - \mathbf{A} \vec{v}(t) \end{pmatrix}$$

or with a matrix notation

$$\frac{d}{dt} \begin{bmatrix} \mathbb{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{W} \end{bmatrix} \begin{pmatrix} v(t) \\ \frac{d}{dt} \vec{v}(t) \end{pmatrix} = \begin{bmatrix} \mathbf{0} & \mathbb{I} \\ -\mathbf{A} & -2 \mathbf{D} \end{bmatrix} \begin{pmatrix} v(t) \\ \frac{d}{dt} \vec{v}(t) \end{pmatrix}.$$

Thus the generalized eigenvalues of

$$\begin{bmatrix} \mathbf{0} & \mathbb{I} \\ -\mathbf{A} & -2 \mathbf{D} \end{bmatrix} \vec{x} = \lambda \begin{bmatrix} \mathbb{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{W} \end{bmatrix} \vec{x}$$

provide information on the behavior of the solutions of the wave equation. This is not implemented in FEMoctave.

## 6.9 Inverse power iteration or `eigs()` to determine small eigenvalues of positive definite matrices

The algorithm to solve the generalized eigenvalue problem

$$\mathbf{A} \vec{x} = \lambda \mathbf{B} \vec{x}$$

for given, positive definite matrices  $\mathbf{A}$  and  $\mathbf{B}$  is based on inverse power iteration. A small number of the smallest eigenvalues can be estimated with reasonable efficiency. This algorithm imposes some restrictions though:

- Both matrices  $\mathbf{A}$  and  $\mathbf{B}$  have to be symmetric and strictly positive definite.
- Only very few eigenvalues and eigenvectors should be computed. The convergence rate for too many eigenvalues is unacceptable.
- There are obvious improvements possible, but I hope for an *Octave* implementation of the command `eigs()`. **This is the case now, thus I use `eigs()`.** Thus some of the notes on eigenvalues do not apply any more.

The algorithm is presented in [GoluVanLoan96] and some more details are worked out in [VarFEM], available at [andreasstahel.github.io/Notes/VarFEM.pdf](https://andreasstahel.github.io/Notes/VarFEM.pdf).

To determine the first  $m$  eigenvalues proceed as follows.

- Create an  $n \times m$  matrix  $\mathbf{V}_0$  with the initial vectors  $\vec{v}_{j,0}$  as its columns.
- repeat until desired precision is reached
  - solve the matrix equation  $\mathbf{A} \cdot \mathbf{V}_k = \mathbf{B} \cdot \mathbf{V}_{k-1}$  or  $\mathbf{V}_k = \mathbf{A}^{-1} \cdot \mathbf{B} \cdot \mathbf{V}_{k-1}$
  - ortho-normalize the columns of  $\mathbf{V}_k$ , using a generalized Gram-Schmidt algorithm. The resulting columns of  $\mathbf{V}_k$  are orthonormal with respect to the scalar product  $\langle \vec{x}, \mathbf{B} \vec{y} \rangle$ .
- for  $j = 1, 2 \dots m$  compute  $\beta_j = \langle \mathbf{V}(:, j), \mathbf{A} \cdot \mathbf{V}(:, j) \rangle$ . Then  $\beta_j$  should be good approximations to the eigenvalues.

The error estimates are based on results in [Demm97]. For a normalized, approximate eigenvector  $\vec{v}_i$  and the corresponding approximate eigenvalue  $\beta_i$  compute the residual  $\vec{r} = \mathbf{A} \vec{v}_i - \beta_i \mathbf{B} \vec{v}_i$ . Then the estimates

$$\min_{\lambda_j \in \sigma(\mathbf{A})} |\beta_i - \lambda_j| \leq \sqrt{\langle \vec{r}, \mathbf{B}^{-1} \vec{r} \rangle} \quad \text{and} \quad |\beta_i - \lambda_j| \leq \frac{\langle \vec{r}, \mathbf{B}^{-1} \vec{r} \rangle}{\text{gap}} \quad (56)$$

are valid. The denominator gap measures the distance to the next eigenvalue.

$$\text{gap} = \min\{|\beta_i - \lambda_j| : \lambda_j \in \sigma(\mathbf{A}), j \neq i\}.$$

Without the exact values of the eigenvalues  $\lambda_i$  there is no way to compute gap exactly. Thus use the approximate values. Expect the error estimate to have its problems at multiple eigenvalues. For the largest, computed eigenvalue one can not estimate gap reliably, since no information on the next eigenvalue is available.

## 7 The Algorithms for 1D FEM

In this section the algorithm for FEM algorithms for problems with one independent variable are presented, for sake of completeness. Only second order elements will be used. These notes are based on the presentation in the class room notes [Stah08, §6.9].

- 7.1 The 1D problems to be examined are shown.
- 7.2 The element stiffness matrix is constructed.
- 7.3 The boundary conditions are taken into account.
- 7.5 The tools to evaluate the solutions of 1D problems are introduced.
- 7.6 Dynamic problems of order 1 with respect to time are examined. Four different time steppers are introduced and compared.
- 7.7 Dynamic problems of order 2 with respect to time are examined. An implicit time stepper is used.
- 7.8 An algorithm to solve nonlinear boundary value problems is presented.

### 7.1 The problems to be examined

The ordinary differential equation to be examined is of the form

$$- (a(x) u'(x))' + b(x) u'(x) + c(x) u(x) = d(x) f(x) \quad (57)$$

with some boundary conditions. Multiplying (57) by a smooth test function  $\phi(x)$  and an integration by parts leads to

$$\begin{aligned} 0 &= \int_{x_0}^{x_n} \left( - (a(x) u'(x))' + b(x) u'(x) + c(x) u(x) - d(x) f(x) \right) \phi(x) dx = \\ &= -a(x) u'(x) \phi(x) \Big|_{x=x_0}^{x_n} + \int_{x_0}^{x_n} a(x) u'(x) \phi'(x) + (b(x) u'(x) + c(x) u(x) - d(x) f(x)) \phi(x) dx \end{aligned} \quad (58)$$

The ODE (57) has to be supplemented with boundary conditions at the two endpoints  $x = x_0$  and  $x = x_n$ .

$$\begin{aligned} u(x_i) &= g_D && \text{Dirichlet} \\ a(x_i) u'(x_i) &= g_{N1} + g_{N2} u(x_i) && \text{Neumann} \end{aligned} \quad (59)$$

If the contribution  $b(x)u'(x)$  vanishes in the ODE (57), solving the ODE is closely related to minimizing the “energy” expression

$$F(u) = \int_{x_0}^{x_n} \frac{1}{2} a(x) (u'(x))^2 + \frac{1}{2} c(x) u^2(x) - d(x) f(x) u(x) dx, \quad (60)$$

respecting the boundary conditions. For a Neumann condition at the right endpoint add the contribution

$$-g_{N1} u(x_n) - \frac{1}{2} g_{N2} u^2(x_n) \quad (61)$$

to the above functional  $F(u)$  and similar at the left endpoint add

$$+g_{N1} u(x_0) + \frac{1}{2} g_{N2} u^2(x_0)$$

Using FEM this equation will be discretized, leading to the global stiffness matrix  $\mathbf{A}$  and the global weight matrix  $\mathbf{M}$ , such that  $\langle \mathbf{A}\vec{u} - \mathbf{M}\vec{f}, \vec{\phi} \rangle = 0$  for all vectors  $\vec{\phi}$ . This then leads to the linear system  $\mathbf{A}\vec{u} = \mathbf{M}\vec{f}$  to be solved for the vector  $\vec{u}$ .

The corresponding first order dynamic equation is given by

$$w(x) \frac{\partial}{\partial t} u(x, t) - (a(x) u'(x, t))' + b(x) u'(x, t) + c(x) u(x, t) = d(x) f(x, t) \quad (62)$$

with the initial condition  $u(x, 0) = u_0(x)$  and the boundary conditions, either Dirichlet or Neumann.

The initial value problem of order 2 is

$$w_2(x) \frac{\partial^2}{\partial t^2} u(x, t) + 2w_1(x) \frac{\partial}{\partial t} u(x, t) - (a(x) u'(x, t))' + b(x) u'(x, t) + c(x) u(x, t) = d(x) f(x, t) \quad (63)$$

again with the corresponding boundary conditions.

A nonlinear boundary value problem is of the form

$$- (a(x, u(x), u'(x)) u'(x))' + b(x) u'(x) + c(x) u(x) = d(x) f(x, u(x), u'(x)) \quad (64)$$

with the corresponding linear boundary conditions.

## 7.2 Interpolation, Gauss integration and the element stiffness matrices

In a first step the code will extend the provided interval  $[x_0, x_1, x_2, \dots, x_n]$  and add the midpoints  $x_{i+0.5} = \frac{x_i + x_{i+1}}{2}$  to the interval, i.e. the new discretization will consist of  $2n + 1$  points. The nodes are at  $\vec{x} = [x_0, x_{0.5}, x_1, x_{1.5}, \dots, x_{n-1}, x_{n-0.5}, x_n] \in \mathbb{R}^{2n+1}$ .

To generate a finite element formulation first examine an subinterval  $x_i \leq x \leq x_{i+1}$ . The nodes for the FEM algorithm are the two endpoints  $x_i, x_{i+1}$  and the midpoint  $x_{i+0.5} = \frac{x_i + x_{i+1}}{2}$ . For given coefficient functions  $a(x), b(x), c(x)$  and  $d(x)$  and the values of the functions  $u(x), f(x)$  and  $\phi(x)$  at the three nodes. Then use a quadratic interpolation to construct the functions  $u(x), f(x)$  and  $\phi(x)$  on the interval. Four integrals have to be examined.

$$\begin{aligned} I_f &= \int_{x_i}^{x_{i+1}} d(x) f(x) \phi(x) dr, & I_0 &= \int_{x_i}^{x_{i+1}} c(x) u(x) \phi(x) dr, \\ I_1 &= \int_{x_i}^{x_{i+1}} b(x) u'(x) \phi(x) dr & \text{and} & I_2 = \int_{x_i}^{x_{i+1}} a(x) u'(x) \phi'(x) dr \end{aligned}$$

To compute these integrals use the very efficient 3–point Gauss integration on a standard interval  $[-\frac{h}{2}, \frac{h}{2}]$  of length  $h$ .

- On the interval  $-\frac{h}{2} \leq x \leq +\frac{h}{2}$  the 3-point Gauss integration formula is given by

$$\int_{-h/2}^{h/2} u(x) dx \approx \frac{h}{18} \left( 5 u\left(-\sqrt{\frac{3}{5}} \frac{h}{2}\right) + 8 u(0) + 5 u\left(+\sqrt{\frac{3}{5}} \frac{h}{2}\right) \right). \quad (65)$$

- The three values of a function  $u(x)$  at  $u(-h/2) = u_-$ ,  $u(0) = u_0$  and  $u(h/2) = u_+$  determine a quadratic interpolating polynomial<sup>23</sup>

$$u(x) = u_0 + \frac{u_+ - u_-}{h} x + \frac{u_+ - 2u_0 + u_-}{h^2} 2x^2.$$

Use  $x = 0$  and  $x = \pm\sqrt{\frac{3}{5}} \frac{h}{2}$  to determine the values of  $u(x)$  at the Gauss points by<sup>24</sup>

$$\begin{aligned} \begin{pmatrix} u\left(-\sqrt{\frac{3}{5}} \frac{h}{2}\right) \\ u(0) \\ u\left(+\sqrt{\frac{3}{5}} \frac{h}{2}\right) \end{pmatrix} &= \begin{bmatrix} \frac{3}{10} + \frac{\sqrt{\frac{3}{5}}}{2} & \frac{4}{10} & \frac{3}{10} - \frac{\sqrt{\frac{3}{5}}}{2} \\ 0 & 1 & 0 \\ \frac{3}{10} - \frac{\sqrt{\frac{3}{5}}}{2} & \frac{4}{10} & \frac{3}{10} + \frac{\sqrt{\frac{3}{5}}}{2} \end{bmatrix} \cdot \begin{pmatrix} u_- \\ u_0 \\ u_+ \end{pmatrix} \\ &= \frac{1}{10} \begin{bmatrix} 3 + \sqrt{15} & 4 & 3 - \sqrt{15} \\ 0 & 10 & 0 \\ 3 - \sqrt{15} & 4 & 3 + \sqrt{15} \end{bmatrix} \cdot \begin{pmatrix} u_- \\ u_0 \\ u_+ \end{pmatrix} = \mathbf{G}_0 \cdot \begin{pmatrix} u_- \\ u_0 \\ u_+ \end{pmatrix}. \end{aligned}$$

Use this Gaussian interpolation matrix to compute the values of the function at the Gauss integration points, using the values at the nodes.

- The matrix  $\mathbf{G}_0$  is basis to construct a matrix to interpolate from the nodes of a FEM grid to the Gauss points. As example consider a grid with three elements of order 2 and thus  $3 \cdot 2 + 1 = 7$  nodes and there are  $3 \cdot 3 = 9$  Gauss points. Each block in the matrix is given by  $\mathbf{G}_0$ . This leads to the  $9 \times 7$  matrix which

<sup>23</sup>To verify the formula use  $u(0) = u_0$  and for  $x = \pm\frac{h}{2}$

$$u\left(\pm\frac{h}{2}\right) = u_0 \pm \frac{u_+ - u_-}{h} \frac{h}{2} + \frac{u_+ - 2u_0 + u_-}{h^2} \frac{2h^2}{4} = u_0(1-1) + u_+ \left(\pm\frac{1}{2} + \frac{1}{2}\right) - u_- \left(\pm\frac{1}{2} - \frac{1}{2}\right).$$

<sup>24</sup>E.g. to obtain the last row in  $\mathbf{G}_0$  evaluate  $u\left(+\sqrt{\frac{3}{5}} \frac{h}{2}\right)$ .

$$\begin{aligned} u\left(+\sqrt{\frac{3}{5}} \frac{h}{2}\right) &= u_0 + \frac{u_+ - u_-}{h} \sqrt{\frac{3}{5}} \frac{h}{2} + \frac{u_+ - 2u_0 + u_-}{h^2} 2 \frac{3}{5} \frac{h^2}{2^2} \\ &= \left(-\sqrt{\frac{3}{5}} \frac{1}{2} + \frac{3}{10}\right) u_- + \left(1 - \frac{3}{5}\right) u_0 + \left(+\sqrt{\frac{3}{5}} \frac{1}{2} + \frac{3}{10}\right) u_+ \\ &= \left(\frac{3}{10} - \frac{\sqrt{\frac{3}{5}}}{2}\right) u_- + \frac{4}{10} u_0 + \left(\frac{3}{10} + \frac{\sqrt{\frac{3}{5}}}{2}\right) u_+ \end{aligned}$$



The above notation leads to the required integrals. With  $\Delta x_i = x_{i+1} - x_i$  obtain

$$\begin{aligned} I_f &= \int_{x_i}^{x_{i+1}} -d(x) f(x) \phi(x) dx \approx \Delta x_i \langle \mathbf{W} \mathbf{d} \mathbf{G}_0 \vec{f}, \mathbf{G}_0 \vec{\phi} \rangle = -\Delta x_i \langle \mathbf{G}_0^T \mathbf{W} \mathbf{d} \mathbf{G}_0 \vec{f}, \vec{\phi} \rangle \\ I_0 &= \int_{x_i}^{x_{i+1}} c(x) u(x) \phi(x) dx \approx \Delta x_i \langle \mathbf{W} \mathbf{c} \mathbf{G}_0 \vec{u}, \mathbf{G}_0 \vec{\phi} \rangle = \Delta x_i \langle \mathbf{G}_0^T \mathbf{W} \mathbf{c} \mathbf{G}_0 \vec{u}, \vec{\phi} \rangle \\ I_1 &= \int_{x_i}^{x_{i+1}} b(x) u'(x) \phi(x) dx \approx \frac{\Delta x_i}{\Delta x_i} \langle \mathbf{W} \mathbf{b} \mathbf{G}_1 \vec{u}, \mathbf{G}_0 \vec{\phi} \rangle = \langle \mathbf{G}_0^T \mathbf{W} \mathbf{b} \mathbf{G}_1 \vec{u}, \vec{\phi} \rangle \\ I_2 &= \int_{x_i}^{x_{i+1}} a(x) u'(x) \phi'(x) dx \approx \frac{\Delta x_i}{(\Delta x_i)^2} \langle \mathbf{W} \mathbf{a} \mathbf{G}_1 \vec{u}, \mathbf{G}_1 \vec{\phi} \rangle = \frac{1}{\Delta x_i} \langle \mathbf{G}_1^T \mathbf{W} \mathbf{a} \mathbf{G}_1 \vec{u}, \vec{\phi} \rangle \end{aligned}$$

Apply the above integrals to full interval  $I = [x_0, x_n]$ , discretized by  $x_0 < x_1 < x_2 < \dots < x_{n-1} < x_n$ . For the second order elements the midpoints  $x_{i+0.5} = \frac{x_i + x_{i+1}}{2}$  of the intervals will be used too, leading to the nodes at  $x_0 < x_{0.5} < x_1 < x_{1.5} < x_2 < \dots < x_{n-1} < x_{n-0.5} < x_n$ , i.e.  $\vec{x} \in \mathbb{R}^{2n+1}$ . Examine the discrete version of the weak solution, thus integrals (or sums) of the type

$$\begin{aligned} I &= \int_{x_i}^{x_r} a(x) u'(x) \phi'(x) + b(x) u'(x) \phi(x) + c(x) u(x) \phi(x) - d(x) f(x) \phi(x) dx \\ &= \sum_{i=1}^n \int_{x_i}^{x_{i+1}} a(x) u'(x) \phi'(x) + b(x) u'(x) \phi(x) + c(x) u(x) \phi(x) - d(x) f(x) \phi(x) dx \\ &\approx \sum_{i=1}^n \left( \frac{1}{\Delta x_i} \langle \mathbf{G}_1^T \mathbf{W} \mathbf{a}_i \mathbf{G}_1 \vec{u}_i, \vec{\phi}_i \rangle + \langle \mathbf{G}_0^T \mathbf{W} \mathbf{b}_i \mathbf{G}_1 \vec{u}_i, \vec{\phi}_i \rangle + \right. \\ &\quad \left. + \Delta x_i \langle \mathbf{G}_0^T \mathbf{W} \mathbf{c}_i \mathbf{G}_0 \vec{u}_i, \vec{\phi}_i \rangle - \Delta x_i \langle \mathbf{G}_0^T \mathbf{W} \mathbf{d}_i \mathbf{G}_0 \vec{f}_i, \vec{\phi}_i \rangle \right) \\ &= \langle \mathbf{A} \vec{u} - \mathbf{M} \vec{f}, \vec{\phi} \rangle \quad \text{for all vectors } \vec{\phi} \in \mathbb{R}^{2n+1}. \end{aligned}$$

The stiffness matrix  $\mathbf{A}$  and the weight matrix  $\mathbf{M}$  are both of size  $(2n+1) \times (2n+1)$ , but the boundary conditions are not taken into account yet. This has to be done with some care, since the differential equation (57) has a unique solution only if boundary conditions are taken into account.

### 7.3 Taking boundary conditions into account

Ignoring the boundary conditions the linear system to be solved for  $\vec{u} \in \mathbb{R}^{2n+1}$  is

$$\mathbf{A} \vec{u} = \mathbf{M} \vec{f}, \quad (67)$$

where  $\vec{f} \in \mathbb{R}^{2n+1}$  contains the values of  $f(x)$  at the nodes  $\vec{x} \in \mathbb{R}^{2n+1}$ . The contribution in (58) by boundary terms is

$$a(x) u'(x) \phi(x) \Big|_{x=x_0}^{x_n} = a(x_n) u'(x_n) \phi(x_n) - a(x_0) u'(x_0) \phi(x_0).$$

This leads to different algorithms to take Dirichlet or Neumann conditions into account. There are four possible combinations of Dirichlet (D) and Neumann (N) boundary conditions: DD, DN, ND and NN.

DD Dirichlet conditions at both endpoints.

The first component of the vector  $\vec{u}$  equals  $u(x_0) = g_{D1}$  and the for the last component use  $u(x_n) = g_{D2}$ .

Equation (67) reads as

$$\mathbf{A} \begin{pmatrix} g_{D1} \\ u_{0.5} \\ u_1 \\ \vdots \\ u_{2n-1} \\ u_{2n-0.5} \\ g_{D2} \end{pmatrix} = \mathbf{M} \begin{pmatrix} f_0 \\ f_{0.5} \\ f_1 \\ \vdots \\ f_{n-1} \\ f_{n-0.5} \\ f_n \end{pmatrix}$$

Remove the first and last row in the matrix  $\mathbf{A}$ . Split off the first column  $\vec{a}_f$  and the last column  $\vec{a}_l$  form the matrix  $[\vec{a}_f, \mathbf{A}_r, \vec{a}_l]$  and in the matrix  $\mathbf{M} \in \mathbb{R}^{(2n+1) \times (2n+1)}$  remove the first and last row, leading to  $\mathbf{M}_r \in \mathbb{R}^{(2n-1) \times (2n+1)}$ . Then examine

$$\mathbf{A} \vec{u} = \mathbf{M} \vec{f} \quad \longrightarrow \quad \left[ \vec{a}_f \quad \mathbf{A}_r \quad \vec{a}_l \right] \begin{pmatrix} g_{D1} \\ \vec{u}_r \\ g_{D2} \end{pmatrix} = \mathbf{M}_r \vec{f} \quad \longrightarrow \quad \mathbf{A}_r \vec{u}_r = \mathbf{M}_r \vec{f} - \vec{a}_f g_{D1} - \vec{a}_l g_{D2},$$

where  $\mathbf{A}_r \in \mathbb{R}^{(2n-1) \times (2n-1)}$ ,  $\vec{u}_r \in \mathbb{R}^{2n-1}$ ,  $\mathbf{M}_r \in \mathbb{R}^{(2n-1) \times (2n+1)}$  and  $\vec{f} \in \mathbb{R}^{2n+1}$ . The solution is then given by  $[g_{D1}, \vec{u}_r, g_{D2}] \in \mathbb{R}^{2n+1}$ .

DN Dirichlet at  $x_0$  and Neumann at  $x_n$

The Dirichlet condition  $u(x_0) = g_{D1}$  leads to the modifications of the first rows and columns of  $\mathbf{A}$  and  $\mathbf{M}$ , just as above. For the boundary condition  $a(x_n) u'(x_n) = g_{N1} + g_{N2} u(x_n)$  the correct type of contributions will have to be taken into account. The additional contributions in (61) to the functional  $F(u)$  in (60) lead to

$$\mathbf{A} \begin{pmatrix} g_{D1} \\ u_{0.5} \\ u_1 \\ \vdots \\ u_{n-1} \\ u_{n-0.5} \\ u_n \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ g_{N2} u_n \end{pmatrix} = \mathbf{M} \begin{pmatrix} f_0 \\ f_{0.5} \\ f_1 \\ \vdots \\ f_{n-1} \\ f_{n-0.5} \\ f_n \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ g_{N1} \end{pmatrix}.$$

Thus the matrix  $\mathbf{A}$  and the RHS have to be modified accordingly. From the last entry on the diagonal of the matrix  $\mathbf{A}$  subtract  $g_{N2}$ .

$$\left[ \vec{a}_f \quad \mathbf{A}_r \right] \begin{pmatrix} g_{D1} \\ \vec{u}_r \end{pmatrix} = \mathbf{M}_r \vec{f} + \vec{g}_{N1} \quad \longrightarrow \quad \mathbf{A}_r \vec{u}_r = \mathbf{M}_r \vec{f} - \vec{a}_f g_{D1} + \vec{g}_{N1},$$

where  $\mathbf{A}_r \in \mathbb{R}^{2n \times 2n}$ ,  $\vec{u}_r \in \mathbb{R}^{2n}$ ,  $\mathbf{M}_r \in \mathbb{R}^{2n \times (2n+1)}$  and  $\vec{f} \in \mathbb{R}^{2n+1}$ . The solution is then given by  $[g_{D1}, \vec{u}_r] \in \mathbb{R}^{2n+1}$ .

ND Neumann at  $x_0$  and Dirichlet at  $x_n$

The Dirichlet condition  $u(x_n) = g_{D2}$  leads to the modifications of the last rows and columns, just as above.

For the boundary condition  $a(x_0) u'(x_0) = g_{N1} + g_{N2} u(x_0)$  the correct type of contributions will have to be taken into account. The additional contributions in (61) to the functional  $F(u)$  leads to

$$\mathbf{A} \begin{pmatrix} u_0 \\ u_{0.5} \\ u_1 \\ \vdots \\ u_{n-0.5} \\ g_{D2} \end{pmatrix} + \begin{pmatrix} g_{N2} u_0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} = \mathbf{M} \begin{pmatrix} f_0 \\ f_{0.5} \\ f_1 \\ \vdots \\ f_{n-0.5} \\ f_n \end{pmatrix} - \begin{pmatrix} g_{N1} \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}.$$

Observe the different signs of the Neumann contribution. Thus the matrix  $\mathbf{A}$  and the RHS have to be modified accordingly. To the first entry on the diagonal of the matrix  $\mathbf{A}$  add  $g_{N2}$ .

$$\left[ \mathbf{A}_r \quad \vec{a}_l \right] \begin{pmatrix} \vec{u}_r \\ g_{D2} \end{pmatrix} = \mathbf{M}_r \vec{f} + \vec{g}_{N1} \quad \longrightarrow \quad \mathbf{A}_r \vec{u}_r = \mathbf{M}_r \vec{f} - \vec{a}_l g_{D1} - \vec{g}_{N1},$$

where  $\mathbf{A}_r \in \mathbb{R}^{2n \times 2n}$ ,  $\vec{u}_r \in \mathbb{R}^{2n}$ ,  $\mathbf{M}_r \in \mathbb{R}^{2n \times (2n+1)}$  and  $\vec{f} \in \mathbb{R}^{2n+1}$ . The solution is then given by  $[\vec{u}_r, g_{D2}] \in \mathbb{R}^{2n+1}$ .

NN Neumann conditions at both endpoints.

The modifications are given by

$$\mathbf{A} \begin{pmatrix} u_0 \\ u_{0.5} \\ u_1 \\ \vdots \\ u_{n-0.5} \\ u_n \end{pmatrix} + \begin{pmatrix} +g_{N2\_left} u_0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ -g_{N2\_right} u_n \end{pmatrix} = \mathbf{M} \begin{pmatrix} f_0 \\ f_{0.5} \\ f_1 \\ \vdots \\ f_{n-0.5} \\ f_n \end{pmatrix} + \begin{pmatrix} -g_{N1\_left} \\ 0 \\ 0 \\ \vdots \\ 0 \\ +g_{N1\_right} \end{pmatrix}.$$

The size of the matrices  $\mathbf{A}$  and  $\mathbf{M}$  remains unchanged. To the first entry on the diagonal of the matrix  $\mathbf{A}$  add  $g_{N2\_left}$  and from the last entry subtract  $g_{N2\_right}$ .

$$\mathbf{A}_r \vec{u} = \mathbf{M} \vec{f} + \vec{g}_{N1},$$

where  $\mathbf{A}_r \in \mathbb{R}^{(2n+1) \times (2n+1)}$ ,  $\vec{u} \in \mathbb{R}^{2n+1}$ ,  $\mathbf{M} \in \mathbb{R}^{(2n+1) \times (2n+1)}$  and  $\vec{f} \in \mathbb{R}^{2n+1}$ .

## 7.4 Solving the BVP with a system of linear equations

Using the above algorithms leads to the linear system

$$\mathbf{A}_r \vec{u}_r - \mathbf{M}_r \vec{f} = \vec{0} \quad \text{or} \quad \vec{u} = \mathbf{A}_r^{-1} \mathbf{M}_r \vec{f} = \mathbf{A}_r \setminus \mathbf{M}_r \vec{f}.$$

The resulting matrix  $\mathbf{A}_r$  is symmetric if  $b(x)$  vanishes identically and it has a band structure with semi-bandwidth 3, i.e. in each row there are up to 5 nonzero entries around the diagonal.

### 7.5 Evaluation of the solution between nodes and evaluation of the derivative

The above algorithms lead to the values of the solution at the nodes. Between nodes the algorithm is based on piece-wise quadratic interpolation. Thus it is advisable to use a quadratic interpolation if the values at more points are required. The *Octave* function `pwquadinterp()` does just that, and it can be used to evaluate first and second derivatives too.

To evaluate the derivatives at the nodes use the quadratic interpolation. For a quadratic function with  $u_- = u(-\frac{h}{2})$ ,  $u(0) = u_0$  and  $u_+ = u(+\frac{h}{2})$  the derivative

$$u'(x) = \frac{u_+ - u_-}{h} + \frac{u_+ - 2u_0 + u_-}{h^2} 4x$$

leads to values at the end- and mid-points.

$$\begin{aligned} u'(-\frac{h}{2}) &= \frac{u_+ - u_-}{h} - \frac{u_+ - 4u_0 + u_-}{h^2} 2h = \frac{1}{h} (-3u_- + 2u_0 - 1u_+) \\ u'(0) &= \frac{1}{h} (-1u_- + 0u_0 + 1u_+) \\ u'(+\frac{h}{2}) &= \frac{u_+ - u_-}{h} + \frac{u_+ - 4u_0 + u_-}{h^2} 2h = \frac{1}{h} (+1u_- - 2u_0 + 3u_+) \end{aligned}$$

With a matrix notation write this in the form

$$\begin{pmatrix} u'(-\frac{h}{2}) \\ u'(0) \\ u'(+\frac{h}{2}) \end{pmatrix} = \frac{1}{h} \begin{bmatrix} -3 & +4 & -1 \\ -1 & 0 & +1 \\ +1 & -4 & +3 \end{bmatrix} \begin{pmatrix} u_- \\ u_0 \\ u_+ \end{pmatrix}.$$

The second derivative is constant on each subinterval and given by

$$u''(x) = 4 \frac{u_- - 2u_0 + u_+}{h^2}.$$

This is used in the code `FEM1DEvaluateDu()` to determine the derivatives at the nodes, generated by `BVP1D()`. At the endpoints of the subintervals the average of derivatives in the two neighboring intervals is used.

#### FEM1DEvaluateDu.m

```
function [du, ddu] = FEM1DEvaluateDu(x, u)
%% du = FEM1DEvaluateDu(x, u)
%% evaluate first and second derivatives at the nodes x
%% requires the interval x and u to be generated by BVP1D()
n = (length(x)-1)/2; %% number of subintervals
du = zeros(size(x)); ddu = du;
M = 0.5*[-3, 4, -1; -2, 0, 2; 1, -4, 3]; %% matrix to determine first derivatives
h = diff(x); h = h(1:2:end)*2;
for jj = 1:n
    range = [2*jj-1:2*jj+1];
    du(range) += M*u(range)/h(jj);
    ddu(range) += 2*[1; 2; 1]*[1 -2 1]*u(range)/h(jj)^2;
endfor
du([1; end]) *= 2; ddu([1; end]) *= 2;
endfunction
```

## 7.6 The first order dynamic problem

The approach to solve the dynamic problem (62)

$$w(x) \frac{\partial}{\partial t} u(x, t) - (a(x) u'(x, t))' + b(x) u'(x, t) + c(x) u(x, t) = d(x) f(x, t)$$

is similar to the 2D approach presented in Section 6.8.1. The first step is to solve the static problem

$$- (a(x) u'_B(x))' + b(x) u'_B(x) + c(x) u_B(x) = 0 \quad \text{with} \quad u_B(x_i) = g_D \quad \text{or} \quad a(x_i) u'_B(x_i) = g_{N1} + g_{N2} u_B(x_i)$$

Then the new function  $v(x, t) = u(x, t) - u_B(x)$  is a solution of an initial boundary value problem with no constant boundary contributions, i.e. at the boundary points  $x_0$  and  $x_n$

$$v(x_i, t) = 0 \quad \text{or} \quad a(x_i) v'(x_i, t) = 0 + g_{N2}(x_i) v(x_i, t)$$

and the initial condition  $v(x, 0) = u_0(x) - u_B(x)$ . This problem is then discretized with respect to  $x$ , leading to a system of ordinary differential equations, similar to equation (53)

$$\mathbf{W} \frac{d}{dt} \vec{v}(t) + \mathbf{A} \vec{v}(t) = \mathbf{M} \vec{f}(t) \quad \text{with} \quad \vec{v}(0) = \vec{v}_0 \quad (68)$$

There are many algorithms available to perform a time step from  $\vec{v}(t)$  to  $\vec{v}(t + \Delta t)$ . Table 15 shows a few key properties of the solvers used in `IBVP1D()`.

- order of consistency: order of consistency of the time stepping algorithm.
- A–stability: a solver is called A–stable if for ODEs of the type  $\frac{d}{dt} \vec{x}(t) = \mathbf{A} \vec{x}(t)$  with negative eigenvalues the approximate solutions remain bounded, independent on the stepsize  $\Delta t$ .
- L–stability: a solver is called L–stable if the above approximate solutions converge to zero rapidly for large eigenvalues  $\lambda < 0$ .<sup>25</sup>
- # of systems to solve: the number of linear systems that have to be solved for one time step.

algorithm	order of consistency	A–stability	L–stability	# of systems to solve
explicit	1	conditional	no	1
implicit	1	unconditional	yes	1
Crank–Nicolson	2	unconditional	no	1
Runge–Kutta, L–stable	2	unconditional	yes	2

Table 15: Properties of the ODE solvers used in `IBVP1D()`

`FEMoctave` uses a very simple implementation of the time steppers, i.e. no stepsize control or adaptation is performed. The stepsize  $\Delta t$  is constant on the time interval to be examined. One can specify the number of time slices to be returned as result and how many steps to take between the returned times.

<sup>25</sup> For an ODE  $\dot{x}(t) = -\lambda x(t)$  applying one time step is implemented by multiplying with a factor  $g(\lambda \Delta t) = g(z)$ . The function  $g(z)$  depends on the algorithm.

- explicit:  $\frac{u_{i+1} - u_i}{\Delta t} = -\lambda u_i$ , thus  $u_{i+1} = (1 - \lambda \Delta t) u_i$  and  $g(z) = 1 - z$ .
- implicit:  $\frac{u_{i+1} - u_i}{\Delta t} = -\lambda u_{i+1}$ , thus  $u_{i+1} = \frac{1}{1 + \lambda \Delta t} u_i$  and  $g(z) = \frac{1}{1+z}$ .
- Crank–Nicolson:  $\frac{u_{i+1} - u_i}{\Delta t} = -\lambda \frac{u_{i+1} + u_i}{2}$ , thus  $u_{i+1} = \frac{2 - \lambda \Delta t}{2 + \lambda \Delta t} u_i$  and  $g(z) = \frac{2-z}{2+z}$ .

If  $|g(z)| < 1$  for  $\text{Re}(z) > 0$  then the algorithm is A–stable. If  $\lim_{z \rightarrow +\infty} |g(z)| = 0$  then the algorithm is L–stable.

### 7.6.1 An explicit time step

To solve (68) with an explicit solver with time step  $\Delta t$  use

$$\begin{aligned}\mathbf{W} \frac{\vec{v}(t + \Delta t) - \vec{v}(t)}{\Delta t} + \mathbf{A} \vec{v}(t) &= \mathbf{M} \vec{f}(t) \\ \mathbf{W} \vec{v}(t + \Delta t) &= \mathbf{W} \vec{v}(t) + \Delta t \left( -\mathbf{A} \vec{v}(t) + \mathbf{M} \vec{f}(t) \right) \\ \vec{v}(t + \Delta t) &= \vec{v}(t) + \Delta t \mathbf{W}^{-1} \left( -\mathbf{A} \vec{v}(t) + \mathbf{M} \vec{f}(t) \right)\end{aligned}$$

### 7.6.2 An implicit time step

To solve (68) with an implicit solver with time step  $\Delta t$  use

$$\begin{aligned}\mathbf{W} \frac{\vec{v}(t + \Delta t) - \vec{v}(t)}{\Delta t} + \mathbf{A} \vec{v}(t + \Delta t) &= \mathbf{M} \vec{f}(t + \Delta t) \\ (\mathbf{W} + \Delta t \mathbf{A}) \vec{v}(t + \Delta t) &= \mathbf{W} \vec{v}(t) + \Delta t \mathbf{M} \vec{f}(t + \Delta t) \\ \vec{v}(t + \Delta t) &= (\mathbf{W} + \Delta t \mathbf{A})^{-1} \left( \mathbf{W} \vec{v}(t) + \Delta t \mathbf{M} \vec{f}(t + \Delta t) \right)\end{aligned}$$

### 7.6.3 A Crank–Nicolson time step

In Section 6.8.1 (page 163) the algorithm of Crank–Nicolson is presented as time stepper. With above notation a time step is given by

$$\begin{aligned}\mathbf{W} \frac{\vec{v}(t + \Delta t) - \vec{v}(t)}{\Delta t} + \mathbf{A} \frac{\vec{v}(t + \Delta t) + \vec{v}(t)}{2} &= \mathbf{M} \vec{f}(t + \Delta t/2) \\ \left( \mathbf{W} + \frac{\Delta t}{2} \mathbf{A} \right) \vec{v}(t + \Delta t) &= \left( \mathbf{W} - \frac{\Delta t}{2} \mathbf{A} \right) \vec{v}(t) + \Delta t \mathbf{M} \vec{f}(t + \Delta t/2)\end{aligned}$$

This is the default time stepper used in FEMoctave.

### 7.6.4 An L–stable Runge–Kutta solver, DIRK

This is a *diagonally implicit Runge–Kutta* method or a DIRK method. Find more details in [Butc03, §361]. The Butcher table of the algorithm is given by

$$\begin{array}{c|cc} \theta & \theta & 0 \\ \hline 1 & 1 - \theta & \theta \\ y_{n+1} & 1 - \theta & \theta \end{array} \quad \text{with } \theta = 1 - \frac{1}{\sqrt{2}}, \text{ i.e.} \quad \begin{array}{c|cc} 1 - \frac{1}{\sqrt{2}} & 1 - \frac{1}{\sqrt{2}} & 0 \\ \hline 1 & \frac{1}{\sqrt{2}} & 1 - \frac{1}{\sqrt{2}} \\ y_{n+1} & \frac{1}{\sqrt{2}} & 1 - \frac{1}{\sqrt{2}} \end{array} .$$

To apply this algorithm to the ODE  $\mathbf{W} \frac{d}{dt} \vec{u}(t) = -\mathbf{A} \vec{u}(t) + \mathbf{M} \vec{f}(t)$  use the above Butcher table.

$$\begin{aligned}\mathbf{W} \vec{k}_1 &= -\mathbf{A} (\vec{u}_n + \theta \Delta t \vec{k}_1) + \mathbf{M} \vec{f}(t_n + \theta \Delta t) \\ \mathbf{W} \vec{k}_2 &= -\mathbf{A} (\vec{u}_n + \Delta t ((1 - \theta) \vec{k}_1 + \theta \vec{k}_2)) + \mathbf{M} \vec{f}(t_n + \Delta t) \\ \vec{u}_{n+1} &= \vec{u}_n + \Delta t ((1 - \theta) \vec{k}_1 + \theta \vec{k}_2)\end{aligned}$$

With  $\frac{1}{2} - \theta = \theta(1 - \theta) = \frac{\theta}{\sqrt{2}}$  and tedious algebra (spelled out in [Stah08, §4.5.8]) this leads to two linear systems to be solved, with the same matrix  $\mathbf{W} + \theta \Delta t \mathbf{A}$ . Thus only one LU factorization is necessary for the time stepper.

$$(\mathbf{W} + \theta \Delta t \mathbf{A}) \vec{k}_1 = -\mathbf{A} \vec{u}_n + \mathbf{M} \vec{f}(t_n + \theta \Delta t) \quad (69)$$

$$\begin{aligned}(\mathbf{W} + \theta \Delta t \mathbf{A}) \vec{u}_{n+1} &= (\mathbf{W} - \Delta t \frac{1}{\sqrt{2}} \mathbf{A}) \vec{u}_n - (\Delta t)^2 (\frac{1}{2} - \theta) \mathbf{A} \vec{k}_1 + \\ &+ \Delta t \mathbf{M} \left( (1 - \theta) \vec{f}(t_n + \theta \Delta t) + \theta \vec{f}(t_n + \Delta t) \right)\end{aligned} \quad (70)$$

## 7.7 The second order dynamic problem

The approach to solve the dynamic problem (63)

$$w_2(x) \frac{\partial^2}{\partial t^2} u(x, t) + 2w_1(x) \frac{\partial}{\partial t} u(x, t) - (a(x) u'(x, t))' + b(x) u'(x, t) + c(x) u(x, t) + d(x) f(x, t) = 0$$

is very similar to the 2D approach presented in Section 6.8.3. The first step is to solve the static problem

$$- (a(x) u'_B(x))' + b(x) u'_B(x) + c(x) u_B(x) = 0 \quad \text{with} \quad u_B(x_i) = g_D \quad \text{or} \quad a(x_i) u'_B(x_i) = g_{N1} + g_{N2} u_B(x_i)$$

Then the new function<sup>26</sup>  $v(x, t) = u(x, t) - u_B(x)$  is a solution of an initial boundary value problem with no constant boundary contributions, i.e. at the boundary points  $x_i$

$$v(x_i, t) = 0 \quad \text{or} \quad a(x_i) v'(x_i, t) = 0 + g_{N2}(x_i) v(x_i, t)$$

and the initial condition  $v(x, 0) = u_0(x) - u_B(x)$ . This problem is then discretized with respect to  $x$ , leading to a system of ordinary differential equations of order 2, similar to equation (54).

$$\mathbf{W}_2 \frac{d^2}{dt^2} \vec{v}(t) + 2\mathbf{W}_1 \frac{d}{dt} \vec{v}(t) + \mathbf{A} \vec{v}(t) = \mathbf{M} \vec{f}(t) \quad \text{with} \quad \vec{v}(0) = \vec{u}_0 - \vec{u}_B \quad \text{and} \quad \frac{d}{dt} v(0) = \vec{v}_0 \quad (71)$$

The implementation assumes that the coefficient functions  $w_2$ ,  $w_1$ ,  $a$ ,  $b$ ,  $c$  and  $d$  depend on  $x$ , while  $f$  may depend on time  $t$  and position  $x$ . Then use an implicit approximation to advance the solution from time  $t - \Delta t$  and  $t$  to  $t + \Delta t$ .

$$\begin{aligned} \mathbf{W}_2 \frac{d^2}{dt^2} \vec{v}(t) &= -2\mathbf{W}_1 \frac{d}{dt} \vec{v}(t) - \mathbf{A} \vec{v}(t) + \mathbf{M} \vec{f}(t) \\ \mathbf{W}_2 \frac{\vec{v}(t - \Delta t) - 2\vec{v}(t) + \vec{v}(t + \Delta t)}{(\Delta t)^2} &= -2\mathbf{W}_1 \frac{\vec{v}(t + \Delta t) - \vec{v}(t - \Delta t)}{2\Delta t} - \\ &\quad - \mathbf{A} \frac{\vec{v}(t - \Delta t) + 2\vec{v}(t) + \vec{v}(t + \Delta t)}{4} + \mathbf{M} \vec{f}(t) \\ \left( +\mathbf{W}_2 + \Delta t \mathbf{W}_1 + \frac{(\Delta t)^2}{4} \mathbf{A} \right) \vec{v}(t + \Delta t) &= - \left( \mathbf{W}_2 - \Delta t \mathbf{W}_1 + \frac{(\Delta t)^2}{4} \mathbf{A} \right) \vec{v}(t - \Delta t) + \\ &\quad + \left( 2\mathbf{W}_2 - \frac{(\Delta t)^2}{2} \mathbf{A} \right) \vec{v}(t) + (\Delta t)^2 \mathbf{M} \vec{f}(t) \end{aligned}$$

This scheme is unconditionally stable and consistent of order 2. Observe that the matrices do not change as time advances. Thus use again a sparsity preserving LU factorization for the time stepping.

To construct the solution at the initial time  $\Delta t$  use the initial value  $u_0$  and initial velocity  $v_0$  and a scheme with the same order of consistency. with respect to time. An explicit scheme<sup>27</sup> for the first step leads to

$$\begin{aligned} \frac{d}{dt} \vec{v}(0) &= \vec{v}_0 \approx \frac{\vec{v}(\Delta t) - \vec{v}(-\Delta t)}{2\Delta t} \implies \vec{v}(-\Delta t) \approx \vec{v}(\Delta t) - 2\Delta t \vec{v}_0 \\ \mathbf{W}_2 \frac{d^2}{dt^2} \vec{v}(0) &= -2\mathbf{W}_1 \frac{d}{dt} \vec{v}(0) - \mathbf{A} \vec{v}(0) + \mathbf{M} \vec{f}(0) \\ \mathbf{W}_2 \frac{\vec{v}(-\Delta t) - 2\vec{v}(0) + \vec{v}(\Delta t)}{(\Delta t)^2} &= -2\mathbf{W}_1 \frac{\vec{v}(+\Delta t) - \vec{v}(-\Delta t)}{2\Delta t} - \mathbf{A} \vec{v}(0) + \mathbf{M} \vec{f}(0) \\ (\mathbf{W}_2 + \Delta t \mathbf{W}_1) \vec{v}(+\Delta t) &= -(\mathbf{W}_2 - \Delta t \mathbf{W}_1) \vec{v}(-\Delta t) + 2\mathbf{W}_2 \vec{v}(0) + (\Delta t)^2 (-\mathbf{A} \vec{v}(0) + \mathbf{M} \vec{f}(0)) \\ &= -(\mathbf{W}_2 - \Delta t \mathbf{W}_1) (\vec{v}(\Delta t) - 2\Delta t \vec{v}_0) + \end{aligned}$$

<sup>26</sup>Observe that  $v(x, t)$  is **not** the velocity, sorry for the inconvenient notation, but it is consistent with other parts of these notes.

<sup>27</sup>Very similar to Section 6.8.3.

$$\begin{aligned}
& +2 \mathbf{W}_2 (\vec{u}_0 - \vec{u}_B) + (\Delta t)^2 (-\mathbf{A} (\vec{u}_0 - \vec{u}_B) + \mathbf{M} \vec{f}(0)) \\
2 \mathbf{W}_2 \vec{v}(\Delta t) & = +2 (\mathbf{W}_2 - \Delta t \mathbf{W}_1) \Delta t \vec{v}_0 + \\
& +2 \mathbf{W}_2 (\vec{u}_0 - \vec{u}_B) + (\Delta t)^2 (-\mathbf{A} (\vec{u}_0 - \vec{u}_B) + \mathbf{M} \vec{f}(0)) \\
\mathbf{W}_2 \vec{v}(\Delta t) & = (\mathbf{W}_2 - \Delta t \mathbf{W}_1) \Delta t \vec{v}_0 + \\
& + \mathbf{W}_2 (\vec{u}_0 - \vec{u}_B) + \frac{1}{2} (\Delta t)^2 (-\mathbf{A} (\vec{u}_0 - \vec{u}_B) + \mathbf{M} \vec{f}(0)).
\end{aligned}$$

## 7.8 Nonlinear boundary value problems, Newton's method and partial substitution

For smooth functions  $a(x, u, u')$  and  $f(x, u, u')$  examine a nonlinear boundary value problem of the form (64)

$$-(a(x, u(x), u'(x)) u'(x))' + b(x) u'(x) + c(x) u(x) = d(x) f(x, u(x), u'(x)), \quad (72)$$

with linear, constant boundary conditions, Dirichlet or Neumann. The essential tool is Newton's method, combined with a partial substitution. For this use a linear Taylor approximation of the nonlinear function  $f(x, u, u')$

$$f(x, u + \phi, u' + \phi') \approx f(x, u, u') + f_u(x, u, u') \phi + f_{u'}(x, u, u') \phi'$$

with the notations  $f_u = \frac{\partial f}{\partial u}$  and  $f_{u'} = \frac{\partial f}{\partial u'}$ . For an approximate solution  $u_n(x)$  search a solution of the form  $u_n(x) + \phi(x)$ . Examine the linear boundary value problem for the perturbation  $\phi$ .

$$\begin{aligned}
-(a u_n' + \phi')' + b(u_n' + \phi') + c(u_n + \phi) & = d \cdot (f(\cdot, u_n, u_n') + \\
& + f_u(\cdot, u_n, u_n') \phi + f_{u'}(\cdot, u_n, u_n') \phi') \\
-(a \phi')' + (b - d \cdot f_{u'}(\cdot, u_n, u_n')) \phi' + (c - d \cdot f_u(\cdot, u_n, u_n')) \phi & = + (a u_n')' - b u_n' - c u_n + d \cdot f(\cdot, u_n, u_n')
\end{aligned}$$

and then update the solution to  $u_{n+1} = u_n + \phi$ . For the perturbation  $\phi$  use zero boundary conditions, assuming that  $u_0$  satisfied the boundary conditions. If the initial guess  $u_0(x)$  is close enough to an isolated solution the algorithm will converge.

To implement the above algorithm in FEMoctave start out with an initial function  $u(x) = u_0(x)$ , hopefully close to the true solution.

- Start with the function  $u_0(x)$  and evaluate (if necessary)  $f_0 = f(x, u_0(x))$ , or  $f_0 = f(x, u_0(x), u_0'(x))$  at the nodes. Evaluate at the Gauss points (if necessary)

$$\begin{aligned}
a_0 & = a(x, u_0(x), u_0'(x)) \\
b_0 & = b(x) \\
c_0 & = c(x)
\end{aligned}$$

- Solve the boundary value problem for  $u_n$

$$-(a_0 u_n'(x))' + b_0 u_n'(x) + c_0 u_n(x) = d \cdot f_0$$

with the correct boundary conditions.

- Repeat

- Store the current solution  $u_{old} = u_n$ .
- If  $a$  depends on  $u$  or  $u'$  evaluate

$$a_n = a(x, u_n(x), u_n'(x)) \quad \text{at the Gauss points.}$$

If  $a$  does not depend on  $u$  and  $u'$ , reuse  $a_n = a_0$ .

- If  $f$  depends on  $u$  or  $u'$  evaluate

$$f_n = f(x, u_n(x), u'_n(x)) \quad \text{at the nodes}$$

$$f_u = \frac{\partial}{\partial u} f(x, u_n(x), u'_n(x)) \quad \text{at the Gauss points}$$

$$f_{u'} = \frac{\partial}{\partial u'} f(x, u_n(x), u'_n(x)) \quad \text{at the Gauss points}$$

- Evaluate

$$\text{RHS}_n = -(a_n u'_n)' + b_0 u'_n + c_0 u_n - d \cdot f_n$$

This can be done with a matrix multiplication.

- Solve the boundary value problem for the perturbation  $\phi$

$$-(a_n \phi'(x))' + (b_0 - d \cdot f_{u'}) \phi'(x) + (c_0 - d \cdot f_u) \phi(x) = -\text{RHS}_n$$

with homogeneous boundary conditions.

- Update  $u_n \rightarrow u_n + \phi$ .
- If  $f$  depends on  $u$  or  $u'$  evaluate at the nodes.

$$f_n = f(x, u_n(x), u'_n(x))$$

- If  $a$  depends on  $u$  or  $u'$  evaluate

$$a_n = a(x, u_n(x), u'_n(x)) \quad \text{at the Gauss points.}$$

If  $a$  does not depend on  $u$ , reuse  $a_n = a_0$ .

- Solve the boundary value problem for  $u_n$

$$-(a_0 u'_n(x))' + b_0 u'(x) + c_0 u_n(x) = d \cdot f_n$$

with the correct boundary conditions.

- until  $\|u_n - u_{old}\|$  small enough or too many iterations.

This is **not a pure Newton's approach**, but a combination with a (partial) substitution method.

## 8 Plane Elasticity and Axially Symmetric Elasticity

Find the description of the plane elasticity problems in Section 2.13, starting on page 12.

### 8.1 The plane stress problem

For a plane stress problem it is assumed that there are no stresses in  $z$ -direction, i.e.  $\sigma_z = \tau_{xz} = \tau_{yz} = 0$ . The elastic energy density is given by equation (20), i.e.

$$W_{stress} = \frac{E}{2(1-\nu^2)} (\varepsilon_{xx}^2 + \varepsilon_{yy}^2 + 2\nu\varepsilon_{xx}\varepsilon_{yy} + 2(1-\nu)\varepsilon_{xy}^2).$$

With FEMoctave examine plane stress deformations with (only) three types of boundary conditions.

$$\begin{aligned} \vec{u} &= \vec{g}_D && \text{on Dirichlet boundary } \Gamma_1, \text{ i.e. prescribed displacement} \\ \text{force density} &= \vec{g}_N && \text{on Neumann boundary } \Gamma_2, \text{ i.e. prescribed force density} \\ \text{force density} &= \vec{0} && \text{on free boundary } \Gamma_3 \end{aligned} \quad (73)$$

With this the total energy of a plane stress problem can be written in the form<sup>28</sup>

$$\begin{aligned} U(\vec{u}) &= U_{elast} + U_{Vol} + U_{Surf} \\ &= \iint_{\Omega} \frac{1}{2} \frac{E}{(1-\nu^2)} \left\langle \begin{bmatrix} 1 & \nu & 0 \\ \nu & 1 & 0 \\ 0 & 0 & 2(1-\nu) \end{bmatrix} \cdot \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix}, \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix} \right\rangle dA - \\ &\quad - \iint_{\Omega} \vec{f} \cdot \vec{u} dA - \int_{\Gamma_2} \vec{g}_N \cdot \vec{u} ds. \end{aligned} \quad (74)$$

Using the Bernoulli principle this energy has to be minimized. A discretization of the displacements  $u_1(x, y)$  and  $u_2(x, y)$  leads to a vector  $\vec{u} = (\vec{u}_1, \vec{u}_2)$  and the above total energy has to be written in the form

$$\frac{1}{2} \langle \vec{u}, \mathbf{A}\vec{u} \rangle + \langle \vec{u}, \mathbf{W}\vec{f} \rangle.$$

Then the approximate minimizer is given as solution of the linear system  $\mathbf{A}\vec{u} = -\mathbf{W}\vec{f}$ . This setup is very similar to Figure 63 on page 127.

Another approach is to use perturbed displacements  $u_1 + \phi_1$  and  $u_2 + \phi_2$  and dropping higher order contributions in  $\phi_i$ . Use the approximations

$$\begin{aligned} \varepsilon_{xx} &= \frac{\partial(u_1 + \phi_1)}{\partial x} = \frac{\partial u_1}{\partial x} + \frac{\partial \phi_1}{\partial x} \\ \varepsilon_{xx}^2 &= \left(\frac{\partial(u_1 + \phi_1)}{\partial x}\right)^2 \approx \left(\frac{\partial u_1}{\partial x}\right)^2 + 2\left(\frac{\partial u_1}{\partial x}\right)\left(\frac{\partial \phi_1}{\partial x}\right) \\ \varepsilon_{yy}^2 &\approx \left(\frac{\partial u_2}{\partial y}\right)^2 + 2\left(\frac{\partial u_2}{\partial y}\right)\left(\frac{\partial \phi_2}{\partial y}\right) \\ \varepsilon_{xx}\varepsilon_{yy} &\approx \frac{\partial u_1}{\partial x} \frac{\partial u_2}{\partial y} + \frac{\partial u_1}{\partial x} \frac{\partial \phi_2}{\partial y} + \frac{\partial u_2}{\partial y} \frac{\partial \phi_1}{\partial x} \\ 2\varepsilon_{xy} &= \frac{\partial(u_1 + \phi_1)}{\partial y} + \frac{\partial(u_2 + \phi_2)}{\partial x} = \frac{\partial u_1}{\partial y} + \frac{\partial \phi_1}{\partial y} + \frac{\partial u_2}{\partial x} + \frac{\partial \phi_2}{\partial x} \end{aligned}$$

<sup>28</sup>We quietly dropped the constant thickness  $H$  from all expressions.

$$\begin{aligned}
4\varepsilon_{xy}^2 &\approx \left(\frac{\partial u_1}{\partial y}\right)^2 + \left(\frac{\partial u_2}{\partial x}\right)^2 + 2\frac{\partial u_1}{\partial y}\left(\frac{\partial \phi_1}{\partial y} + \frac{\partial \phi_2}{\partial x}\right) + 2\frac{\partial u_2}{\partial x}\left(\frac{\partial \phi_1}{\partial y} + \frac{\partial \phi_2}{\partial x}\right) \\
&= \left(\frac{\partial u_1}{\partial y} + \frac{\partial u_2}{\partial x}\right)^2 + 2\frac{\partial \phi_1}{\partial y}\left(\frac{\partial u_1}{\partial y} + \frac{\partial u_2}{\partial x}\right) + 2\frac{\partial \phi_2}{\partial x}\left(\frac{\partial u_1}{\partial y} + \frac{\partial u_2}{\partial x}\right)
\end{aligned}$$

Based on  $\frac{2(1-\nu^2)}{E} W(u) = \varepsilon_{xx}^2 + \varepsilon_{yy}^2 + 2\nu\varepsilon_{xx}\varepsilon_{yy} + 2(1-\nu)\varepsilon_{xy}^2$  conclude

$$\begin{aligned}
\frac{2(1-\nu^2)}{E} \left( W(\vec{u} + \vec{\phi}) - W(\vec{u}) \right) &\approx 2\frac{\partial u_1}{\partial x}\frac{\partial \phi_1}{\partial x} + 2\frac{\partial u_2}{\partial y}\frac{\partial \phi_2}{\partial y} + \\
&+ 2\nu\left(\frac{\partial u_1}{\partial x}\frac{\partial \phi_2}{\partial y} + \frac{\partial u_2}{\partial y}\frac{\partial \phi_1}{\partial x}\right) + \\
&+ \frac{4}{4}(1-\nu)\left(\frac{\partial \phi_1}{\partial y}\left(\frac{\partial u_1}{\partial y} + \frac{\partial u_2}{\partial x}\right) + \frac{\partial \phi_2}{\partial x}\left(\frac{\partial u_1}{\partial y} + \frac{\partial u_2}{\partial x}\right)\right)
\end{aligned}$$

and use

$$\begin{aligned}
\iint_{\Omega} \vec{f} \cdot (\vec{u} + \vec{\phi}) dA &= \iint_{\Omega} \vec{f} \cdot \vec{u} dA + \iint_{\Omega} f_1 \phi_1 + f_2 \phi_2 dA \\
\int_{\Gamma_2} \vec{g}_N \cdot (\vec{u} + \vec{\phi}) ds &= \int_{\Gamma_2} \vec{g}_N \cdot \vec{u} ds + \int_{\Gamma_2} g_1 \phi_1 + g_2 \phi_2 ds.
\end{aligned}$$

This leads to

$$\begin{aligned}
U(\vec{u} + \vec{\phi}) - U(\vec{u}) &\approx + \iint_{\Omega} \frac{E}{1-\nu^2} \left( \frac{\partial \phi_1}{\partial x} \left( \frac{\partial u_1}{\partial x} + \nu \frac{\partial u_2}{\partial y} \right) + \frac{1-\nu}{2} \frac{\partial \phi_1}{\partial y} \left( \frac{\partial u_1}{\partial y} + \frac{\partial u_2}{\partial x} \right) \right) - \phi_1 f_1 dA + \\
&+ \iint_{\Omega} \frac{E}{1-\nu^2} \left( \frac{\partial \phi_2}{\partial y} \left( \frac{\partial u_2}{\partial y} + \nu \frac{\partial u_1}{\partial x} \right) + \frac{1-\nu}{2} \frac{\partial \phi_2}{\partial x} \left( \frac{\partial u_1}{\partial y} + \frac{\partial u_2}{\partial x} \right) \right) - \phi_2 f_2 dA - \\
&- \int_{\Gamma_2} \phi_1 g_1 + \phi_2 g_2 ds. \tag{75}
\end{aligned}$$

Using Bernoulli's principle this expression should vanish for all perturbations  $\vec{\phi}$ . Use discrete approximations of the functions  $u_i$  and  $\phi_i$  to write the vanishing condition for expression (75) in the form

$$\langle \vec{\phi}, \mathbf{A}\vec{u} + \mathbf{W}\vec{f} \rangle = 0 \quad \text{for all } \vec{\phi}.$$

## 8.2 Construction of first order elements

The algorithm in this section is based on the results in Section 6.4 (p. 133), with the expressions in equation (75) to be integrated over a triangle  $T$ . The approximation consists of piecewise linear, triangular segments. Thus the first order partial derivatives are constant on each triangle. Consequently the strains are constant on each triangle. This is the reason for the name Constant Strain Triangle, short CST.

### 8.2.1 Integration of $f_1 \phi_1 + f_2 \phi_2$

- If the values of the functions  $f_1$  and  $f_2$  at the Gauss points are denoted by the vectors  $\vec{f}_1$  and  $\vec{f}_2$ , then use the approximation

$$\begin{aligned}
\iint_T f_1 \phi_1 + f_2 \phi_2 dA &\approx \frac{\text{area}(T)}{3} \left( \langle \mathbf{M} \vec{\phi}_1, \vec{f}_1 \rangle + \langle \mathbf{M} \vec{\phi}_2, \vec{f}_2 \rangle \right) \\
&= \frac{\text{area}(T)}{3} \left( \langle \vec{\phi}_1, \mathbf{M}^T \vec{f}_1 \rangle + \langle \vec{\phi}_2, \mathbf{M}^T \vec{f}_2 \rangle \right).
\end{aligned}$$

$\mathbf{M} \in \mathbb{R}^{3 \times 3}$  is the matrix for interpolation from the nodes to the Gauss points, given by

$$\mathbf{M} = \frac{1}{6} \begin{bmatrix} 4 & 1 & 1 \\ 1 & 4 & 1 \\ 1 & 1 & 4 \end{bmatrix}.$$

- If the values of the functions  $f_1$  and  $f_2$  at the nodes are denoted by the vectors  $\vec{f}_1$  and  $\vec{f}_2$ , then use the approximation

$$\begin{aligned} \iint_T f_1 \phi_1 + f_2 \phi_2 dA &\approx \frac{\text{area}(T)}{3} \left( \langle \mathbf{M} \vec{\phi}_1, \mathbf{M} \vec{f}_1 \rangle + \langle \mathbf{M} \vec{\phi}_2, \mathbf{M} \vec{f}_2 \rangle \right) \\ &= \frac{\text{area}(T)}{3} \left( \langle \vec{\phi}_1, \mathbf{M}^T \mathbf{M} \vec{f}_1 \rangle + \langle \vec{\phi}_2, \mathbf{M}^T \mathbf{M} \vec{f}_2 \rangle \right) \end{aligned}$$

Thus find one contribution to (75). With a block matrix notation the above can be written in the form

$$\frac{\text{area}(T)}{3} \left\langle \begin{pmatrix} \vec{\phi}_1 \\ \vec{\phi}_2 \end{pmatrix}, \begin{bmatrix} \mathbf{M}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{M}^T \end{bmatrix} \begin{pmatrix} \vec{f}_1 \\ \vec{f}_2 \end{pmatrix} \right\rangle \quad \text{or} \quad \frac{\text{area}(T)}{3} \left\langle \begin{pmatrix} \vec{\phi}_1 \\ \vec{\phi}_2 \end{pmatrix}, \begin{bmatrix} \mathbf{M}^T \mathbf{M} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}^T \mathbf{M} \end{bmatrix} \begin{pmatrix} \vec{f}_1 \\ \vec{f}_2 \end{pmatrix} \right\rangle.$$

### 8.2.2 Integration of the terms involving derivatives of $\phi_1$ and $\phi_2$

For linear elements the gradient of the functions  $u_i$  and  $\phi_i$  are constant and using equation (33) given by

$$\nabla u = \frac{-1}{2 \text{area}(T)} \begin{bmatrix} (y_3 - y_2) & (y_1 - y_3) & (y_2 - y_1) \\ (x_2 - x_3) & (x_3 - x_1) & (x_1 - x_2) \end{bmatrix} \cdot \vec{u} = \begin{bmatrix} \mathbf{G}_x \\ \mathbf{G}_y \end{bmatrix} \vec{u}.$$

Evaluate the coefficients  $E$  and  $\nu$  at the Gauss points  $\vec{g}_i$  and define the averaged values

$$a_1 = \frac{1}{3} \sum_{i=1}^3 \frac{E(\vec{g}_i)}{1 - \nu^2(\vec{g}_i)}, \quad a_2 = \frac{1}{3} \sum_{i=1}^3 \frac{\nu(\vec{g}_i) E(\vec{g}_i)}{1 - \nu^2(\vec{g}_i)} \quad \text{and} \quad a_3 = \frac{1}{3} \sum_{i=1}^3 \frac{E(\vec{g}_i)}{2(1 + \nu(\vec{g}_i))}.$$

This leads to the approximations

$$\begin{aligned} I_{\phi_1} &= \iint_T \frac{E}{1 - \nu^2} \left( \frac{\partial \phi_1}{\partial x} \left( \frac{\partial u_1}{\partial x} + \nu \frac{\partial u_2}{\partial y} \right) + \frac{1 - \nu}{2} \frac{\partial \phi_1}{\partial y} \left( \frac{\partial u_1}{\partial y} + \frac{\partial u_2}{\partial x} \right) \right) dA \\ &\approx a_1 \langle \mathbf{G}_x \vec{\phi}_1, \mathbf{G}_x \vec{u}_1 \rangle + a_2 \langle \mathbf{G}_x \vec{\phi}_1, \mathbf{G}_y \vec{u}_2 \rangle + a_3 \langle \mathbf{G}_y \vec{\phi}_1, \mathbf{G}_y \vec{u}_1 + \mathbf{G}_x \vec{u}_2 \rangle \\ &= a_1 \langle \vec{\phi}_1, \mathbf{G}_x^T \mathbf{G}_x \vec{u}_1 \rangle + a_2 \langle \vec{\phi}_1, \mathbf{G}_x^T \mathbf{G}_y \vec{u}_2 \rangle + a_3 \langle \vec{\phi}_1, \mathbf{G}_y^T \mathbf{G}_y \vec{u}_1 + \mathbf{G}_y^T \mathbf{G}_x \vec{u}_2 \rangle \\ I_{\phi_2} &= \iint_T \frac{E}{1 - \nu^2} \left( \frac{\partial \phi_2}{\partial y} \left( \frac{\partial u_2}{\partial y} + \nu \frac{\partial u_1}{\partial x} \right) + \frac{1 - \nu}{2} \frac{\partial \phi_2}{\partial x} \left( \frac{\partial u_1}{\partial y} + \frac{\partial u_2}{\partial x} \right) \right) dA \\ &\approx a_1 \langle \mathbf{G}_y \vec{\phi}_2, \mathbf{G}_y \vec{u}_2 \rangle + a_2 \langle \mathbf{G}_y \vec{\phi}_2, \mathbf{G}_x \vec{u}_1 \rangle + a_3 \langle \mathbf{G}_x \vec{\phi}_2, \mathbf{G}_y \vec{u}_1 + \mathbf{G}_x \vec{u}_2 \rangle \\ &= a_1 \langle \vec{\phi}_2, \mathbf{G}_y^T \mathbf{G}_y \vec{u}_2 \rangle + a_2 \langle \vec{\phi}_2, \mathbf{G}_y^T \mathbf{G}_x \vec{u}_1 \rangle + a_3 \langle \vec{\phi}_2, \mathbf{G}_x^T \mathbf{G}_y \vec{u}_1 + \mathbf{G}_x^T \mathbf{G}_x \vec{u}_2 \rangle \end{aligned}$$

With a block matrix notation write the above in the form

$$I_{\vec{\phi}} \approx \left\langle \begin{pmatrix} \vec{\phi}_1 \\ \vec{\phi}_2 \end{pmatrix}, \begin{bmatrix} a_1 \mathbf{G}_x^T \mathbf{G}_x + a_3 \mathbf{G}_y^T \mathbf{G}_y & a_2 \mathbf{G}_x^T \mathbf{G}_y + a_3 \mathbf{G}_y^T \mathbf{G}_x \\ a_2 \mathbf{G}_y^T \mathbf{G}_x + a_3 \mathbf{G}_x^T \mathbf{G}_y & a_1 \mathbf{G}_y^T \mathbf{G}_y + a_3 \mathbf{G}_x^T \mathbf{G}_x \end{bmatrix} \begin{pmatrix} \vec{u}_1 \\ \vec{u}_2 \end{pmatrix} \right\rangle =: \left\langle \begin{pmatrix} \vec{\phi}_1 \\ \vec{\phi}_2 \end{pmatrix}, \mathbf{G} \begin{pmatrix} \vec{u}_1 \\ \vec{u}_2 \end{pmatrix} \right\rangle.$$

The symmetric  $6 \times 6$  matrix  $\mathbf{G} \in \mathbb{R}^{6 \times 6}$  is the element stiffness matrix for the triangle  $T$ , containing contributions to (75).

### 8.2.3 The boundary integral

The boundary integral is similar to (34) on page 138. With  $\alpha = \frac{1-1/\sqrt{3}}{2}$  use the symmetric interpolation matrix from nodes to Gauss points

$$\mathbf{M}_b = \begin{bmatrix} 1 - \alpha & \alpha \\ \alpha & 1 - \alpha \end{bmatrix}$$

and the length  $L$  of the edge segment for the approximate integral

$$\int_{\text{edge}} \vec{g}_N \cdot \vec{\phi} \, ds = \int_{\text{edge}} g_1 \phi_1 + g_2 \phi_2 \, ds \approx \frac{L}{2} \langle \vec{\phi}_1, \mathbf{M}_b \vec{g}_1 \rangle + \frac{L}{2} \langle \vec{\phi}_2, \mathbf{M}_b \vec{g}_2 \rangle,$$

where the functions  $g_1$  and  $g_2$  are evaluated at the Gauss points.

## 8.3 Construction of second order elements

The algorithm in this section is based on the results in Section 6.5 (p. 139), with the expressions in equation (75) to be integrated over a triangle  $T$ . The approximation consists of piecewise quadratic, triangular segments. Thus the first order partial derivatives are linear on each triangle.

### 8.3.1 Integration of $f_1 \phi_1 + f_2 \phi_2$

Use the Gauss weights  $\vec{w} \in \mathbb{R}^7$  from equation (30) on page 132 for the approximate integration over one triangle  $T$ .

- If the values of the functions  $f_1$  and  $f_2$  at the seven Gauss points are denoted by the vectors  $\vec{f}_1$  and  $\vec{f}_2 \in \mathbb{R}^7$ , then use the approximation

$$\begin{aligned} \iint_T f_1 \phi_1 + f_2 \phi_2 \, dA &\approx \text{area}(T) \left( \langle \mathbf{M} \vec{\phi}_1, \text{diag}(\vec{w}) \vec{f}_1 \rangle + \langle \mathbf{M} \vec{\phi}_2, \text{diag}(\vec{w}) \vec{f}_2 \rangle \right) \\ &= \text{area}(T) \left( \langle \vec{\phi}_1, \mathbf{M}^T \text{diag}(\vec{w}) \vec{f}_1 \rangle + \langle \vec{\phi}_2, \mathbf{M}^T \text{diag}(\vec{w}) \vec{f}_2 \rangle \right) \\ &= \text{area}(T) \left\langle \begin{pmatrix} \vec{\phi}_1 \\ \vec{\phi}_2 \end{pmatrix}, \begin{bmatrix} \mathbf{M}^T \text{diag}(\vec{w}) & \mathbf{0} \\ \mathbf{0} & \mathbf{M}^T \text{diag}(\vec{w}) \end{bmatrix} \begin{pmatrix} \vec{f}_1 \\ \vec{f}_2 \end{pmatrix} \right\rangle. \end{aligned}$$

$\mathbf{M} \in \mathbb{R}^{7 \times 6}$  is the matrix for interpolation from the nodes to the Gauss points, given in equation (37) on page 141.

- If the values of the functions  $f_1$  and  $f_2$  at the nodes are denoted by the vectors  $\vec{f}_1$  and  $\vec{f}_2 \in \mathbb{R}^6$ , then use the approximation

$$\begin{aligned} \iint_T f_1 \phi_1 + f_2 \phi_2 \, dA &\approx \text{area}(T) \left( \langle \mathbf{M} \vec{\phi}_1, \text{diag}(\vec{w}) \mathbf{M} \vec{f}_1 \rangle + \langle \mathbf{M} \vec{\phi}_2, \text{diag}(\vec{w}) \mathbf{M} \vec{f}_2 \rangle \right) \\ &= \text{area}(T) \left( \langle \vec{\phi}_1, \mathbf{M}^T \text{diag}(\vec{w}) \mathbf{M} \vec{f}_1 \rangle + \langle \vec{\phi}_2, \mathbf{M}^T \text{diag}(\vec{w}) \mathbf{M} \vec{f}_2 \rangle \right) \\ &= \text{area}(T) \left\langle \begin{pmatrix} \vec{\phi}_1 \\ \vec{\phi}_2 \end{pmatrix}, \begin{bmatrix} \mathbf{M}^T \text{diag}(\vec{w}) \mathbf{M} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}^T \text{diag}(\vec{w}) \mathbf{M} \end{bmatrix} \begin{pmatrix} \vec{f}_1 \\ \vec{f}_2 \end{pmatrix} \right\rangle. \end{aligned}$$

Thus find one contribution to (75). Observe that  $\mathbf{M}^T \text{diag}(\vec{w}) \mathbf{M}$  is a  $6 \times 6$  matrix, independent on the triangle  $T$ .

### 8.3.2 Integration of the terms involving derivatives of $\phi_1$ and $\phi_2$

Using the results from Section 6.5 the partial derivatives at the nodes of functions  $\phi$  given at the nodes find for the first component  $\varphi_x = \frac{\partial \phi}{\partial x}$  of the gradient at the Gauss points

$$\begin{pmatrix} \varphi_x(\vec{x}_1) \\ \varphi_x(\vec{x}_2) \\ \vdots \\ \varphi_x(\vec{x}_7) \end{pmatrix} = \frac{1}{\det(\mathbf{T})} \left[ (+y_3 - y_1) \mathbf{M}_\xi^T + (-y_2 + y_1) \mathbf{M}_\nu^T \right] \cdot \vec{\phi} =: \mathbf{G}_x \vec{\phi}$$

and for the second component of the gradient

$$\begin{pmatrix} \varphi_y(\vec{x}_1) \\ \varphi_y(\vec{x}_2) \\ \vdots \\ \varphi_y(\vec{x}_7) \end{pmatrix} = \frac{1}{\det(\mathbf{T})} \left[ (-x_3 + x_1) \mathbf{M}_\xi^T + (+x_2 - x_1) \mathbf{M}_\nu^T \right] \cdot \vec{\phi} =: \mathbf{G}_y \vec{\phi}.$$

Evaluate the coefficients  $E$  and  $\nu$  at the Gauss points  $g_i$  and multiply by the Gauss integration weights to obtain the three diagonal matrices

$$\mathbf{A}_1 = \text{diag} \begin{pmatrix} w_1 \frac{E(\vec{g}_1)}{1-\nu^2(\vec{g}_1)} \\ w_2 \frac{E(\vec{g}_2)}{1-\nu^2(\vec{g}_2)} \\ \vdots \\ w_7 \frac{E(\vec{g}_7)}{1-\nu^2(\vec{g}_7)} \end{pmatrix}, \quad \mathbf{A}_2 = \text{diag} \begin{pmatrix} w_1 \frac{\nu(\vec{g}_1)E(\vec{g}_1)}{1-\nu^2(\vec{g}_1)} \\ w_2 \frac{\nu(\vec{g}_2)E(\vec{g}_2)}{1-\nu^2(\vec{g}_2)} \\ \vdots \\ w_7 \frac{E(\nu(\vec{g}_7)\vec{g}_7)}{1-\nu^2(\vec{g}_7)} \end{pmatrix} \quad \text{and} \quad \mathbf{A}_3 = \text{diag} \begin{pmatrix} w_1 \frac{E(\vec{g}_1)}{2(1+\nu(\vec{g}_1))} \\ w_2 \frac{E(\vec{g}_2)}{2(1+\nu(\vec{g}_2))} \\ \vdots \\ w_7 \frac{E(\vec{g}_7)}{2(1+\nu(\vec{g}_7))} \end{pmatrix}.$$

This leads to the approximations

$$\begin{aligned} \frac{I_{\phi_1}}{\text{area}(T)} &= \frac{1}{\text{area}(T)} \iint_T \frac{E}{1-\nu^2} \left( \frac{\partial \phi_1}{\partial x} \left( \frac{\partial u_1}{\partial x} + \nu \frac{\partial u_2}{\partial y} \right) + \frac{1-\nu}{2} \frac{\partial \phi_1}{\partial y} \left( \frac{\partial u_1}{\partial y} + \frac{\partial u_2}{\partial x} \right) \right) dA \\ &\approx \langle \mathbf{A}_1 \mathbf{G}_x \vec{\phi}_1, \mathbf{G}_x \vec{u}_1 \rangle + \langle \text{diag } \mathbf{A}_2 \mathbf{G}_x \vec{\phi}_1, \mathbf{G}_y \vec{u}_2 \rangle + \langle \mathbf{A}_3 \mathbf{G}_y \vec{\phi}_1, \mathbf{G}_y \vec{u}_1 + \mathbf{G}_x \vec{u}_2 \rangle \\ &= \langle \vec{\phi}_1, \mathbf{G}_x^T \mathbf{A}_1 \mathbf{G}_x \vec{u}_1 \rangle + \langle \vec{\phi}_1, \mathbf{G}_x^T \mathbf{A}_2 \mathbf{G}_y \vec{u}_2 \rangle + \langle \vec{\phi}_1, \mathbf{G}_y^T \mathbf{A}_3 \mathbf{G}_y \vec{u}_1 + \mathbf{G}_y^T \mathbf{A}_3 \mathbf{G}_x \vec{u}_2 \rangle \\ \frac{I_{\phi_2}}{\text{area}(T)} &= \frac{1}{\text{area}(T)} \iint_T \frac{E}{1-\nu^2} \left( \frac{\partial \phi_2}{\partial y} \left( \frac{\partial u_2}{\partial y} + \nu \frac{\partial u_1}{\partial x} \right) + \frac{1-\nu}{2} \frac{\partial \phi_2}{\partial x} \left( \frac{\partial u_1}{\partial y} + \frac{\partial u_2}{\partial x} \right) \right) dA \\ &\approx \langle \mathbf{A}_1 \mathbf{G}_y \vec{\phi}_2, \mathbf{G}_y \vec{u}_2 \rangle + \langle \mathbf{A}_2 \mathbf{G}_y \vec{\phi}_2, \mathbf{G}_x \vec{u}_1 \rangle + \langle \mathbf{A}_3 \mathbf{G}_x \vec{\phi}_2, \mathbf{G}_y \vec{u}_1 + \mathbf{G}_x \vec{u}_2 \rangle \\ &= \langle \vec{\phi}_2, \mathbf{G}_y^T \mathbf{A}_1 \mathbf{G}_y \vec{u}_2 \rangle + \langle \vec{\phi}_2, \mathbf{G}_y^T \mathbf{A}_2 \mathbf{G}_x \vec{u}_1 \rangle + \langle \vec{\phi}_2, \mathbf{G}_x^T \mathbf{A}_3 \mathbf{G}_y \vec{u}_1 + \mathbf{G}_x^T \mathbf{A}_3 \mathbf{G}_x \vec{u}_2 \rangle. \end{aligned}$$

With a block matrix notation write the above in the form

$$\begin{aligned} I_{\vec{\phi}} = I_{\phi_1} + I_{\phi_2} &\approx \text{area}(T) \left\langle \begin{pmatrix} \vec{\phi}_1 \\ \vec{\phi}_2 \end{pmatrix}, \begin{bmatrix} \mathbf{G}_x^T \mathbf{A}_1 \mathbf{G}_x + \mathbf{G}_y^T \mathbf{A}_3 \mathbf{G}_y & \mathbf{G}_x^T \mathbf{A}_2 \mathbf{G}_y + \mathbf{G}_y^T \mathbf{A}_3 \mathbf{G}_x \\ \mathbf{G}_y^T \mathbf{A}_2 \mathbf{G}_x + \mathbf{G}_x^T \mathbf{A}_3 \mathbf{G}_y & \mathbf{G}_y^T \mathbf{A}_1 \mathbf{G}_y + \mathbf{G}_x^T \mathbf{A}_3 \mathbf{G}_x \end{bmatrix} \begin{pmatrix} \vec{u}_1 \\ \vec{u}_2 \end{pmatrix} \right\rangle \\ &=: \text{area}(T) \left\langle \begin{pmatrix} \vec{\phi}_1 \\ \vec{\phi}_2 \end{pmatrix}, \mathbf{G} \begin{pmatrix} \vec{u}_1 \\ \vec{u}_2 \end{pmatrix} \right\rangle. \end{aligned}$$

The symmetric  $12 \times 12$  matrix  $\mathbf{G} \in \mathbb{R}^{12 \times 12}$  is the element stiffness matrix for the triangle  $T$ , containing contributions to (75).

### 8.3.3 The boundary integral

The boundary integral is similar to (43) on page 148, i.e. based on

$$\int_{-h/2}^{h/2} f(x) dx \approx \frac{h}{18} \left( 5 f\left(-\frac{\sqrt{3}}{2\sqrt{5}} h\right) + 8 f(0) + 5 f\left(\frac{\sqrt{3}}{2\sqrt{5}} h\right) \right).$$

If the values of a function  $f$  at the two endpoints and the midpoint are denoted by  $(f_1, f_2, f_3)$  use a quadratic interpolation to find the values at the three Gauss integration points, given by

$$\begin{pmatrix} f(\vec{p}_1) \\ f(\vec{p}_2) \\ f(\vec{p}_3) \end{pmatrix} = \mathbf{M}_B \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix} \approx \begin{bmatrix} +0.68730 & 0.4 & -0.08730 \\ 0 & 1 & 0 \\ -0.08730 & 0.4 & +0.68730 \end{bmatrix} \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix}$$

and with the length  $L$  of the segment on the edge obtain the approximate integral

$$\begin{aligned} \int_{\text{edge}} g_1 \phi_1 + g_2 \phi_2 ds &\approx \frac{L}{18} \langle \mathbf{M}_B \vec{\phi}_1, \begin{pmatrix} 5 g_1(\vec{p}_1) \\ 8 g_1(\vec{p}_2) \\ 5 g_1(\vec{p}_3) \end{pmatrix} \rangle + \frac{L}{18} \langle \mathbf{M}_B \vec{\phi}_2, \begin{pmatrix} 5 g_2(\vec{p}_1) \\ 8 g_2(\vec{p}_2) \\ 5 g_2(\vec{p}_3) \end{pmatrix} \rangle \\ &= \frac{L}{18} \langle \vec{\phi}_1, \mathbf{M}_B^T \begin{pmatrix} 5 g_1(\vec{p}_1) \\ 8 g_1(\vec{p}_2) \\ 5 g_1(\vec{p}_3) \end{pmatrix} \rangle + \frac{L}{18} \langle \vec{\phi}_2, \mathbf{M}_B^T \begin{pmatrix} 5 g_2(\vec{p}_1) \\ 8 g_2(\vec{p}_2) \\ 5 g_2(\vec{p}_3) \end{pmatrix} \rangle \end{aligned}$$

The integration weights can be combined with the interpolation matrix  $\mathbf{M}_B$  by

$$\mathbf{M}_{BC} = \frac{1}{18} \mathbf{M}_B^T \begin{bmatrix} 5 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 5 \end{bmatrix} \approx \begin{bmatrix} 0.1909 & 0 & -0.0242 \\ 0.1111 & 0.4444 & 0.1111 \\ -0.0242 & 0 & 0.1909 \end{bmatrix}.$$

This matrix  $\mathbf{M}_{BC}$  does not depend on the current edge segment and now use

$$\int_{\text{edge}} g_1 \phi_1 + g_2 \phi_2 ds \approx L \langle \vec{\phi}_1, \mathbf{M}_{BC} \begin{pmatrix} g_1(\vec{p}_1) \\ g_1(\vec{p}_2) \\ g_1(\vec{p}_3) \end{pmatrix} \rangle + L \langle \vec{\phi}_2, \mathbf{M}_{BC} \begin{pmatrix} g_2(\vec{p}_1) \\ g_2(\vec{p}_2) \\ g_2(\vec{p}_3) \end{pmatrix} \rangle.$$

The effect of the boundary integral on the global stiffness matrix and the vector is very similar to the approach shown at the end of Section 6.5.8.

## 8.4 Construction of third order elements

The methods in this section are a combination of the tools used to construct third order elements for elliptic problems (Section 6.6) and the methods in the previous Section 8.3 to construct second order elements.

### 8.4.1 Integration of $f_1 \phi_1 + f_2 \phi_2$

Use the Gauss weights  $\vec{w} \in \mathbb{R}^7$  from equation (30) on page 132 for the approximate integration over one triangle  $T$ .

- If the values of the functions  $f_1$  and  $f_2$  at the seven Gauss points are denoted by the vectors  $\vec{f}_1$  and  $\vec{f}_2 \in \mathbb{R}^7$ , then use the approximation

$$\begin{aligned} \iint_T f_1 \phi_1 + f_2 \phi_2 dA &\approx \text{area}(T) \left( \langle \mathbf{M} \vec{\phi}_1, \text{diag}(\vec{w}) \vec{f}_1 \rangle + \langle \mathbf{M} \vec{\phi}_2, \text{diag}(\vec{w}) \vec{f}_2 \rangle \right) \\ &= \text{area}(T) \left( \langle \vec{\phi}_1, \mathbf{M}^T \text{diag}(\vec{w}) \vec{f}_1 \rangle + \langle \vec{\phi}_2, \mathbf{M}^T \text{diag}(\vec{w}) \vec{f}_2 \rangle \right) \\ &= \text{area}(T) \left\langle \begin{pmatrix} \vec{\phi}_1 \\ \vec{\phi}_2 \end{pmatrix}, \begin{bmatrix} \mathbf{M}^T \text{diag}(\vec{w}) & \mathbf{0} \\ \mathbf{0} & \mathbf{M}^T \text{diag}(\vec{w}) \end{bmatrix} \begin{pmatrix} \vec{f}_1 \\ \vec{f}_2 \end{pmatrix} \right\rangle. \end{aligned}$$

$\mathbf{M} \in \mathbb{R}^{7 \times 10}$  is the matrix for interpolation from the nodes to the Gauss points, given in equation (48) on page 151.

- If the values of the functions  $f_1$  and  $f_2$  at the nodes are denoted by the vectors  $\vec{f}_1$  and  $\vec{f}_2 \in \mathbb{R}^{10}$ , then use the approximation

$$\begin{aligned} \iint_T f_1 \phi_1 + f_2 \phi_2 dA &\approx \text{area}(T) \left( \langle \mathbf{M} \vec{\phi}_1, \text{diag}(\vec{w}) \mathbf{M} \vec{f}_1 \rangle + \langle \mathbf{M} \vec{\phi}_2, \text{diag}(\vec{w}) \mathbf{M} \vec{f}_2 \rangle \right) \\ &= \text{area}(T) \left( \langle \vec{\phi}_1, \mathbf{M}^T \text{diag}(\vec{w}) \mathbf{M} \vec{f}_1 \rangle + \langle \vec{\phi}_2, \mathbf{M}^T \text{diag}(\vec{w}) \mathbf{M} \vec{f}_2 \rangle \right). \end{aligned}$$

Thus find one contribution to (75). Observe that  $\mathbf{M}^T \text{diag}(\vec{w}) \mathbf{M}$  is a  $10 \times 10$  matrix, independent on the triangle  $T$ .

#### 8.4.2 Integration of the terms involving derivatives of $\phi_1$ and $\phi_2$

Using the results from Section 6.6 the partial derivatives at the nodes of functions  $\phi$  given at the nodes find for the first component  $\varphi_x = \frac{\partial \varphi}{\partial x}$  of the gradient at the Gauss points

$$\begin{pmatrix} \varphi_x(\vec{x}_1) \\ \varphi_x(\vec{x}_2) \\ \vdots \\ \varphi_x(\vec{x}_7) \end{pmatrix} = \frac{1}{\det(\mathbf{T})} \left[ (+y_3 - y_1) \mathbf{M}_\xi^T + (-y_2 + y_1) \mathbf{M}_\nu^T \right] \cdot \vec{\phi} =: \mathbf{G}_x \vec{\phi}$$

and for the second component of the gradient

$$\begin{pmatrix} \varphi_y(\vec{x}_1) \\ \varphi_y(\vec{x}_2) \\ \vdots \\ \varphi_y(\vec{x}_7) \end{pmatrix} = \frac{1}{\det(\mathbf{T})} \left[ (-x_3 + x_1) \mathbf{M}_\xi^T + (+x_2 - x_1) \mathbf{M}_\nu^T \right] \cdot \vec{\phi} =: \mathbf{G}_y \vec{\phi},$$

using the interpolation matrices  $\mathbf{M}_\xi$  and  $\mathbf{M}_\nu$  in equation (51) for the partial derivatives and the transformation rule (39) for the gradient. The matrices  $\mathbf{G}_x$  and  $\mathbf{G}_y$  are of size  $7 \times 10$  and depend on the actual element, i.e. the triangle  $T$ .

Evaluate the coefficients  $E$  and  $\nu$  at the Gauss points  $\vec{g}_i$  and multiply by the Gauss integration weights to obtain the three diagonal matrices

$$\mathbf{A}_1 = \text{diag} \begin{pmatrix} w_1 \frac{E(\vec{g}_1)}{1-\nu^2(\vec{g}_1)} \\ w_2 \frac{E(\vec{g}_2)}{1-\nu^2(\vec{g}_2)} \\ \vdots \\ w_7 \frac{E(\vec{g}_7)}{1-\nu^2(\vec{g}_7)} \end{pmatrix}, \quad \mathbf{A}_2 = \text{diag} \begin{pmatrix} w_1 \frac{\nu(\vec{g}_1)E(\vec{g}_1)}{1-\nu^2(\vec{g}_1)} \\ w_2 \frac{\nu(\vec{g}_2)E(\vec{g}_2)}{1-\nu^2(\vec{g}_2)} \\ \vdots \\ w_7 \frac{E(\nu(\vec{g}_7)\vec{g}_7)}{1-\nu^2(\vec{g}_7)} \end{pmatrix} \quad \text{and} \quad \mathbf{A}_3 = \text{diag} \begin{pmatrix} w_1 \frac{E(\vec{g}_1)}{2(1+\nu(\vec{g}_1))} \\ w_2 \frac{E(\vec{g}_2)}{2(1+\nu(\vec{g}_2))} \\ \vdots \\ w_7 \frac{E(\vec{g}_7)}{2(1+\nu(\vec{g}_7))} \end{pmatrix}.$$

This leads to the approximations

$$\begin{aligned} \frac{I_{\phi_1}}{\text{area}(T)} &= \frac{1}{\text{area}(T)} \iint_T \frac{E}{1-\nu^2} \left( \frac{\partial \phi_1}{\partial x} \left( \frac{\partial u_1}{\partial x} + \nu \frac{\partial u_2}{\partial y} \right) + \frac{1-\nu}{2} \frac{\partial \phi_1}{\partial y} \left( \frac{\partial u_1}{\partial y} + \frac{\partial u_2}{\partial x} \right) \right) dA \\ &\approx \langle \mathbf{A}_1 \mathbf{G}_x \vec{\phi}_1, \mathbf{G}_x \vec{u}_1 \rangle + \langle \text{diag } \mathbf{A}_2 \mathbf{G}_x \vec{\phi}_1, \mathbf{G}_y \vec{u}_2 \rangle + \langle \mathbf{A}_3 \mathbf{G}_y \vec{\phi}_1, \mathbf{G}_y \vec{u}_1 + \mathbf{G}_x \vec{u}_2 \rangle \\ &= \langle \vec{\phi}_1, \mathbf{G}_x^T \mathbf{A}_1 \mathbf{G}_x \vec{u}_1 \rangle + \langle \vec{\phi}_1, \mathbf{G}_x^T \mathbf{A}_2 \mathbf{G}_y \vec{u}_2 \rangle + \langle \vec{\phi}_1, \mathbf{G}_y^T \mathbf{A}_3 \mathbf{G}_y \vec{u}_1 + \mathbf{G}_y^T \mathbf{A}_3 \mathbf{G}_x \vec{u}_2 \rangle \\ \frac{I_{\phi_2}}{\text{area}(T)} &= \frac{1}{\text{area}(T)} \iint_T \frac{E}{1-\nu^2} \left( \frac{\partial \phi_2}{\partial y} \left( \frac{\partial u_2}{\partial y} + \nu \frac{\partial u_1}{\partial x} \right) + \frac{1-\nu}{2} \frac{\partial \phi_2}{\partial x} \left( \frac{\partial u_1}{\partial y} + \frac{\partial u_2}{\partial x} \right) \right) dA \\ &\approx \langle \mathbf{A}_1 \mathbf{G}_y \vec{\phi}_2, \mathbf{G}_y \vec{u}_2 \rangle + \langle \mathbf{A}_2 \mathbf{G}_y \vec{\phi}_2, \mathbf{G}_x \vec{u}_1 \rangle + \langle \mathbf{A}_3 \mathbf{G}_x \vec{\phi}_2, \mathbf{G}_y \vec{u}_1 + \mathbf{G}_x \vec{u}_2 \rangle \\ &= \langle \vec{\phi}_2, \mathbf{G}_y^T \mathbf{A}_1 \mathbf{G}_y \vec{u}_2 \rangle + \langle \vec{\phi}_2, \mathbf{G}_y^T \mathbf{A}_2 \mathbf{G}_x \vec{u}_1 \rangle + \langle \vec{\phi}_2, \mathbf{G}_x^T \mathbf{A}_3 \mathbf{G}_y \vec{u}_1 + \mathbf{G}_x^T \mathbf{A}_3 \mathbf{G}_x \vec{u}_2 \rangle. \end{aligned}$$

With a block matrix notation write the above in the form

$$\begin{aligned} I_{\vec{\phi}} = I_{\phi_1} + I_{\phi_2} &\approx \text{area}(T) \left\langle \begin{pmatrix} \vec{\phi}_1 \\ \vec{\phi}_2 \end{pmatrix}, \begin{bmatrix} \mathbf{G}_x^T \mathbf{A}_1 \mathbf{G}_x + \mathbf{G}_y^T \mathbf{A}_3 \mathbf{G}_y & \mathbf{G}_x^T \mathbf{A}_2 \mathbf{G}_y + \mathbf{G}_y^T \mathbf{A}_3 \mathbf{G}_x \\ \mathbf{G}_y^T \mathbf{A}_2 \mathbf{G}_x + \mathbf{G}_x^T \mathbf{A}_3 \mathbf{G}_y & \mathbf{G}_y^T \mathbf{A}_1 \mathbf{G}_y + \mathbf{G}_x^T \mathbf{A}_3 \mathbf{G}_x \end{bmatrix} \begin{pmatrix} \vec{u}_1 \\ \vec{u}_2 \end{pmatrix} \right\rangle \\ &=: \text{area}(T) \left\langle \begin{pmatrix} \vec{\phi}_1 \\ \vec{\phi}_2 \end{pmatrix}, \mathbf{G} \begin{pmatrix} \vec{u}_1 \\ \vec{u}_2 \end{pmatrix} \right\rangle. \end{aligned}$$

The symmetric  $20 \times 20$  matrix  $\mathbf{G} \in \mathbb{R}^{20 \times 20}$  is the element stiffness matrix for the triangle  $T$ , containing contributions to (75).

### 8.4.3 The boundary integral

The boundary integral is similar to (6.6.7) on page 158, i.e. based on

$$\int_{-h/2}^{h/2} f(x) dx \approx \frac{h}{18} \left( 5 f\left(-\frac{\sqrt{3}}{2\sqrt{5}} h\right) + 8 f(0) + 5 f\left(\frac{\sqrt{3}}{2\sqrt{5}} h\right) \right).$$

If the values of a function  $f$  at the two endpoints and the two points on the edge are denoted by  $(f_{-2}, f_{-1}, f_{+1}, f_{+2})$  use a cubic interpolation to find the values at the three Gauss integration points, given by

$$\begin{pmatrix} u(\vec{p}_1) \\ u(\vec{p}_2) \\ u(\vec{p}_3) \end{pmatrix} = \mathbf{M}_B \begin{pmatrix} f_{-2} \\ f_{-1} \\ f_{+1} \\ f_{+2} \end{pmatrix} \approx \begin{bmatrix} 0.4880 & 0.7479 & -0.2979 & 0.06199 \\ -0.0625 & 0.5625 & 0.5625 & -0.0625 \\ 0.06199 & -0.2979 & 0.7479 & 0.4880 \end{bmatrix} \begin{pmatrix} f_{-2} \\ f_{-1} \\ f_{+1} \\ f_{+2} \end{pmatrix}$$

and with the length  $L$  of the segment on the edge obtain the approximate integral

$$\begin{aligned} \int_{\text{edge}} g_1 \phi_1 + g_2 \phi_2 ds &\approx \frac{L}{18} \langle \mathbf{M}_B \vec{\phi}_1, \begin{pmatrix} 5 g_1(\vec{p}_1) \\ 8 g_1(\vec{p}_2) \\ 5 g_1(\vec{p}_3) \end{pmatrix} \rangle + \frac{L}{18} \langle \mathbf{M}_B \vec{\phi}_2, \begin{pmatrix} 5 g_2(\vec{p}_1) \\ 8 g_2(\vec{p}_2) \\ 5 g_2(\vec{p}_3) \end{pmatrix} \rangle \\ &= \frac{L}{18} \langle \vec{\phi}_1, \mathbf{M}_B^T \begin{pmatrix} 5 g_1(\vec{p}_1) \\ 8 g_1(\vec{p}_2) \\ 5 g_1(\vec{p}_3) \end{pmatrix} \rangle + \frac{L}{18} \langle \vec{\phi}_2, \mathbf{M}_B^T \begin{pmatrix} 5 g_2(\vec{p}_1) \\ 8 g_2(\vec{p}_2) \\ 5 g_2(\vec{p}_3) \end{pmatrix} \rangle. \end{aligned}$$

The integration weights can be combined with the interpolation matrix  $\mathbf{M}_B$  by

$$\mathbf{M}_{BC} = \frac{1}{18} \mathbf{M}_B^T \begin{bmatrix} 5 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 5 \end{bmatrix} \approx \begin{bmatrix} 0.1356 & -0.0278 & 0.0172 \\ 0.2077 & 0.2500 & -0.0827 \\ -0.0827 & 0.2500 & 0.2077 \\ 0.0172 & -0.0278 & 0.1356 \end{bmatrix}.$$

This matrix  $\mathbf{M}_{BC}$  does not depend on the current edge segment and leads to

$$\int_{\text{edge}} g_1 \phi_1 + g_2 \phi_2 ds \approx L \langle \vec{\phi}_1, \mathbf{M}_{BC} \begin{pmatrix} g_1(\vec{p}_1) \\ g_1(\vec{p}_2) \\ g_1(\vec{p}_3) \end{pmatrix} \rangle + L \langle \vec{\phi}_2, \mathbf{M}_{BC} \begin{pmatrix} g_2(\vec{p}_1) \\ g_2(\vec{p}_2) \\ g_2(\vec{p}_3) \end{pmatrix} \rangle.$$

The effect of the boundary integral on the global stiffness matrix and the vector is very similar to the approach shown at the end of Section 6.6.7.

## 8.5 The plane strain problem

For a plane strain problem it is assumed that there are no strains in  $z$ -direction, i.e.

$$\varepsilon_{xz} = \varepsilon_{yz} = \varepsilon_z = 0.$$

With the modified material parameters in equation (23)  $\nu^* = \frac{\nu}{1-\nu}$  and  $E^* = \frac{E}{1-\nu^2}$  this leads to a simplification of Hooke's law.

$$\begin{aligned} \begin{pmatrix} \sigma_x \\ \sigma_y \\ \tau_{xy} \end{pmatrix} &= \frac{E}{(1+\nu)(1-2\nu)} \begin{bmatrix} 1-\nu & \nu & 0 \\ \nu & 1-\nu & 0 \\ 0 & 0 & 1-2\nu \end{bmatrix} \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix} \\ &= \frac{E^*}{(1-\nu^*)(1+\nu^*)} \begin{bmatrix} 1 & \nu^* & 0 \\ \nu^* & 1 & 0 \\ 0 & 0 & 1-\nu^* \end{bmatrix} \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix} \\ \sigma_z &= \frac{E \nu (\varepsilon_{xx} + \varepsilon_{yy})}{(1+\nu)(1-2\nu)} \end{aligned}$$

This is very similar to Hooke's law (19) for the plane stress situation, but with  $E^*$  and  $\nu^*$  instead of  $E$  and  $\nu$ . The energy density is in this case given by

$$W_{\text{strain}} = \frac{1}{2} \frac{E}{(1+\nu)(1-2\nu)} \left\langle \begin{bmatrix} 1-\nu & \nu & 0 \\ \nu & 1-\nu & 0 \\ 0 & 0 & 2(1-2\nu) \end{bmatrix} \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix}, \begin{pmatrix} \varepsilon_{xx} \\ \varepsilon_{yy} \\ \varepsilon_{xy} \end{pmatrix} \right\rangle$$

$$\begin{aligned}
&= \frac{E(1-\nu)}{2(1+\nu)(1-2\nu)} \left( \varepsilon_{xx}^2 + \varepsilon_{yy}^2 + 2\frac{\nu}{1-\nu} \varepsilon_{xx}\varepsilon_{yy} + 2\frac{1-2\nu}{1-\nu} \varepsilon_{xy}^2 \right) \\
&= \frac{E^*}{2(1-(\nu^*)^2)} \left( \varepsilon_{xx}^2 + \varepsilon_{yy}^2 + 2\nu^* \varepsilon_{xx}\varepsilon_{yy} + 2(1-\nu^*) \varepsilon_{xy}^2 \right). \tag{76}
\end{aligned}$$

This is very similar to the elastic energy density (20) for plane stress problems.

As a consequence of the similarity of the plane strain and plane stress problem there is no need for extensive new codes for plane strain problems. It is sufficient to write a wrapper to modify the material parameters.

## 8.6 Elasticity for axially symmetric setups

Examine displacements

$$\begin{pmatrix} u_1(x, y, z) \\ u_2(x, y, z) \\ u_3(x, y, z) \end{pmatrix} = \begin{pmatrix} u_r(r, z) \cos \varphi \\ u_r(r, z) \sin \varphi \\ u_z(r, z) \end{pmatrix}.$$

and the total energy to be minimized is given by expression (25) on page 19.

$$\begin{aligned}
U(\vec{u}) &= U_{elast} + U_{Vol} + U_{Surf} \\
&= \iint_{\Omega} \frac{2\pi r E}{2(1+\nu)(1-2\nu)} \left( (1-\nu)(\varepsilon_{rr}^2 + \varepsilon_{zz}^2 + \frac{1}{r^2} u_r^2) + 2\nu(\varepsilon_{rr}\varepsilon_{zz} + \frac{1}{r} u_r(\varepsilon_{rr} + \varepsilon_{zz})) \right) dA + \\
&\quad + \iint_{\Omega} \frac{2\pi r E}{1+\nu} \varepsilon_{rz}^2 dA - \iint_{\Omega} 2\pi r \vec{f} \cdot \vec{u} dA - \int_{\Gamma_2} 2\pi r \vec{g}_N \cdot \vec{u} ds.
\end{aligned}$$

Use the strains

$$\begin{pmatrix} \varepsilon_{rr} \\ \varepsilon_{\phi\phi} \\ \varepsilon_{zz} \\ \varepsilon_{rz} \end{pmatrix} = \begin{pmatrix} \frac{\partial u_r}{\partial r} \\ \frac{1}{r} u_r \\ \frac{\partial u_z}{\partial z} \\ \frac{1}{2} \left( \frac{\partial u_r}{\partial z} + \frac{\partial u_z}{\partial r} \right) \end{pmatrix}$$

to rewrite the above with the displacement functions  $u_r$  and  $u_z$  in the form

$$\begin{aligned}
\frac{U(\vec{u})}{2\pi} &= \iint_{\Omega} \frac{r E}{2(1+\nu)(1-2\nu)} \left( (1-\nu) \left( \left( \frac{\partial u_r}{\partial r} \right)^2 + \left( \frac{\partial u_z}{\partial z} \right)^2 + \frac{1}{r^2} u_r^2 \right) + \right. \\
&\quad \left. + 2\nu \left( \left( \frac{\partial u_r}{\partial r} \right) \left( \frac{\partial u_z}{\partial z} \right) + \frac{1}{r} u_r \left( \frac{\partial u_r}{\partial r} + \frac{\partial u_z}{\partial z} \right) \right) \right) dA + \\
&\quad + \iint_{\Omega} \frac{r E}{1+\nu} \frac{1}{4} \left( \frac{\partial u_r}{\partial z} + \frac{\partial u_z}{\partial r} \right)^2 dA - \\
&\quad - \iint_{\Omega} r (f_r u_r + f_z u_z) dA - \int_{\Gamma_2} r (g_{Nr} u_r + g_{Nz} u_z) ds.
\end{aligned}$$

Expanding and ignoring quadratic terms of the perturbation  $\vec{\phi}$  leads to

$$\begin{aligned}
\frac{U(\vec{u} + \vec{\phi})}{2\pi} &\approx \frac{U(\vec{u})}{2\pi} + \iint_{\Omega} \frac{r E}{(1+\nu)(1-2\nu)} \left( (1-\nu) \left( \frac{\partial u_r}{\partial r} \frac{\partial \phi_r}{\partial r} + \frac{\partial u_z}{\partial z} \frac{\partial \phi_z}{\partial z} + \frac{1}{r^2} u_r \phi_r \right) + \right. \\
&\quad \left. + \nu \left( \left( \frac{\partial u_r}{\partial r} \right) \left( \frac{\partial \phi_z}{\partial z} \right) + \left( \frac{\partial \phi_r}{\partial r} \right) \left( \frac{\partial u_z}{\partial z} \right) + \frac{1}{r} u_r \left( \frac{\partial \phi_r}{\partial r} + \frac{\partial \phi_z}{\partial z} \right) + \frac{1}{r} \phi_r \left( \frac{\partial u_r}{\partial r} + \frac{\partial u_z}{\partial z} \right) \right) \right) dA +
\end{aligned}$$

$$\begin{aligned}
& + \iint_{\Omega} \frac{r E}{1 + \nu} \frac{1}{2} \left( \frac{\partial u_r}{\partial z} \frac{\partial \phi_r}{\partial z} + \frac{\partial u_z}{\partial r} \frac{\partial \phi_z}{\partial r} + \frac{\partial u_r}{\partial z} \frac{\partial \phi_z}{\partial r} + \frac{\partial \phi_r}{\partial z} \frac{\partial u_z}{\partial r} \right) dA - \\
& - \iint_{\Omega} r (f_r \phi_r + f_z \phi_z) dA - \int_{\Gamma_2} r (g_{Nr} \phi_r + g_{Nz} \phi_z) ds .
\end{aligned}$$

These integrals can be separated into contributions with  $\phi_r$ ,  $\phi_z$ ,  $f$  and  $g_N$ .

$$\frac{U(\vec{u} + \vec{\phi})}{2\pi} \approx \frac{U(\vec{u})}{2\pi} + I_{\phi_r} + I_{\phi_z} + I_f + I_g \quad (77)$$

$$\begin{aligned}
I_{\phi_r} = & \iint_{\Omega} \frac{r E}{(1 + \nu)(1 - 2\nu)} \left( (1 - \nu) \left( \frac{\partial u_r}{\partial r} \frac{\partial \phi_r}{\partial r} + \frac{1}{r^2} u_r \phi_r \right) + \right. \\
& + \nu \left( \frac{\partial u_z}{\partial z} \frac{\partial \phi_r}{\partial r} + \frac{1}{r} u_r \frac{\partial \phi_r}{\partial r} + \frac{1}{r} \phi_r \left( \frac{\partial u_r}{\partial r} + \frac{\partial u_z}{\partial z} \right) \right) \left. + \right. \\
& \left. + \frac{r E}{2(1 + \nu)} \left( \frac{\partial u_r}{\partial z} + \frac{\partial u_z}{\partial r} \right) \frac{\partial \phi_r}{\partial z} dA \right.
\end{aligned} \quad (78)$$

$$\begin{aligned}
I_{\phi_z} = & \iint_{\Omega} \frac{r E}{(1 + \nu)(1 - 2\nu)} \left( (1 - \nu) \frac{\partial u_z}{\partial z} \frac{\partial \phi_z}{\partial z} + \nu \left( \frac{\partial u_r}{\partial r} \frac{\partial \phi_z}{\partial z} + \frac{1}{r} u_r \frac{\partial \phi_z}{\partial z} \right) \right) + \\
& + \frac{r E}{2(1 + \nu)} \left( \frac{\partial u_r}{\partial z} + \frac{\partial u_z}{\partial r} \right) \frac{\partial \phi_z}{\partial r} dA
\end{aligned} \quad (79)$$

$$I_f = - \iint_{\Omega} r (f_r \phi_r + f_z \phi_z) dA \quad (80)$$

$$I_g = - \int_{\Gamma_2} r (g_{Nr} \phi_r + g_{Nz} \phi_z) ds \quad (81)$$

## 8.7 Construction of first order elements

This is similar to the computations in Section 8.2, starting in page 182. In this section the element stiffness matrix is constructed. Then use the procedure in Section 6.4 (page 133) to determine the global stiffness matrix.

### 8.7.1 Integration of $r (f_r \phi_r + f_z \phi_z)$

Evaluate the radius  $r$  at the three Gauss points of the triangle  $T$ , leading to the diagonal matrix  $\mathbf{R} = \text{diag}([r_1, r_2, r_3])$  and use the interpolation matrix from the corners to the Gauss points

$$\mathbf{M} = \frac{1}{6} \begin{bmatrix} 4 & 1 & 1 \\ 1 & 4 & 1 \\ 1 & 1 & 4 \end{bmatrix} .$$

- If the values of the functions  $f_r$  and  $f_z$  at the Gauss points are denoted by the vectors  $\vec{f}_r$  and  $\vec{f}_z$ , then use the approximation

$$\begin{aligned}
\iint_T r (f_r \phi_r + f_z \phi_z) dA & \approx \frac{\text{area}(T)}{3} \left( \langle \mathbf{M} \vec{\phi}_r, \mathbf{R} \vec{f}_r \rangle + \langle \mathbf{M} \vec{\phi}_z, \mathbf{R} \vec{f}_z \rangle \right) \\
& = \frac{\text{area}(T)}{3} \left( \langle \vec{\phi}_r, \mathbf{M}^T \mathbf{R} \vec{f}_r \rangle + \langle \vec{\phi}_z, \mathbf{M}^T \mathbf{R} \vec{f}_z \rangle \right) .
\end{aligned}$$

- If the values of the functions  $f_r$  and  $f_z$  at the nodes are denoted by the vectors  $\vec{f}_r$  and  $\vec{f}_z$ , then use the approximation

$$\begin{aligned} \iint_T r (f_r \phi_r + f_z \phi_z) dA &\approx \frac{\text{area}(T)}{3} \left( \langle \mathbf{M} \vec{\phi}_r, \mathbf{RM} \vec{f}_r \rangle + \langle \mathbf{M} \vec{\phi}_z, \mathbf{RM} \vec{f}_z \rangle \right) \\ &= \frac{\text{area}(T)}{3} \left( \langle \vec{\phi}_r, \mathbf{M}^T \mathbf{RM} \vec{f}_r \rangle + \langle \vec{\phi}_z, \mathbf{M}^T \mathbf{RM} \vec{f}_z \rangle \right) \end{aligned}$$

With the above the contributions in (80) for each element stiffness matrix can be determined.

### 8.7.2 Integration of the terms involving derivatives of $\phi_z$ and $\phi_r$

For linear elements the gradient of the functions  $u_i$  and  $\phi_i$  are constant and using equation (33) given by

$$\nabla u = \frac{-1}{2 \text{area}(T)} \begin{bmatrix} (z_3 - z_2) & (z_1 - z_3) & (z_2 - z_1) \\ (r_2 - r_3) & (r_3 - r_1) & (r_1 - r_2) \end{bmatrix} \cdot \vec{u} = \begin{bmatrix} \mathbf{G}_r \\ \mathbf{G}_z \end{bmatrix} \vec{u}.$$

Evaluate the coefficients  $E$  and  $\nu$  at the Gauss points  $\vec{g}_i$  and define the average values  $a_j$ , vector  $\vec{a}_4$  and the diagonal matrix  $\mathbf{A}_2$ . Since the derivatives of order one of the displacements are piecewise constant, some of the expressions require less computational effort to determine.

$$\begin{aligned} a_1 &= \frac{\text{area}(T)}{3} \sum_{i=1}^3 \frac{r(\vec{g}_i) E(\vec{g}_i) (1 - \nu(\vec{g}_i))}{(1 + \nu(\vec{g}_i)) (1 - 2\nu(\vec{g}_i))} \\ \mathbf{A}_2 &= \frac{\text{area}(T)}{3} \text{diag} \left( \frac{E(\vec{g}_i) (1 - \nu(\vec{g}_i))}{r(\vec{g}_i) (1 + \nu(\vec{g}_i)) (1 - 2\nu(\vec{g}_i))} \right) \\ a_3 &= \frac{\text{area}(T)}{3} \sum_{i=1}^3 \frac{r(\vec{g}_i) E(\vec{g}_i) \nu(\vec{g}_i)}{(1 + \nu(\vec{g}_i)) (1 - 2\nu(\vec{g}_i))} \\ (\vec{a}_4)_i &= \frac{\text{area}(T)}{3} \frac{E(\vec{g}_i) \nu(\vec{g}_i)}{(1 + \nu(\vec{g}_i)) (1 - 2\nu(\vec{g}_i))} \\ a_5 &= \frac{\text{area}(T)}{3} \sum_{i=1}^3 \frac{r(\vec{g}_i) E(\vec{g}_i)}{2(1 + \nu(\vec{g}_i))} \end{aligned}$$

This leads to the approximate integrals

$$\begin{aligned} I_{\phi_r} &= \iint_T \frac{r E}{(1 + \nu)(1 - 2\nu)} \left( (1 - \nu) \left( \frac{\partial u_r}{\partial r} \frac{\partial \phi_r}{\partial r} + \frac{1}{r^2} u_r \phi_r \right) + \right. \\ &\quad \left. + \nu \left( \frac{\partial u_z}{\partial z} \frac{\partial \phi_r}{\partial r} + \frac{1}{r} u_r \frac{\partial \phi_r}{\partial r} + \frac{1}{r} \phi_r \left( \frac{\partial u_r}{\partial r} + \frac{\partial u_z}{\partial z} \right) \right) \right) + \\ &\quad + \frac{r E}{2(1 + \nu)} \left( \frac{\partial u_r}{\partial z} + \frac{\partial u_z}{\partial r} \right) \frac{\partial \phi_r}{\partial z} dA \\ &\approx a_1 \langle \mathbf{G}_r \vec{\phi}_r, \mathbf{G}_r \vec{u}_r \rangle + \langle \mathbf{M} \vec{\phi}_r, \mathbf{A}_2 \mathbf{M} \vec{u}_r \rangle + a_3 \langle \mathbf{G}_r \vec{\phi}_r, \mathbf{G}_z \vec{u}_z \rangle + \\ &\quad + \langle \vec{a}_4 \mathbf{G}_r \vec{\phi}_r, \mathbf{M} \vec{u}_r \rangle + \langle \mathbf{M} \vec{\phi}_r, \vec{a}_4 (\mathbf{G}_r \vec{u}_r + \mathbf{G}_z \vec{u}_z) \rangle + a_5 \langle \mathbf{G}_z \vec{\phi}_r, (\mathbf{G}_z \vec{u}_r + \mathbf{G}_r \vec{u}_z) \rangle \\ &= a_1 \langle \vec{\phi}_r, \mathbf{G}_r^T \mathbf{G}_r \vec{u}_r \rangle + \langle \vec{\phi}_r, \mathbf{M}^T \mathbf{A}_2 \mathbf{M} \vec{u}_r \rangle + a_3 \langle \vec{\phi}_r, \mathbf{G}_r^T \mathbf{G}_z \vec{u}_z \rangle + \\ &\quad + \langle \vec{\phi}_r, (\vec{a}_4 \mathbf{G}_r)^T \mathbf{M} \vec{u}_r \rangle + \langle \vec{\phi}_r, \mathbf{M}^T \vec{a}_4 (\mathbf{G}_r \vec{u}_r + \mathbf{G}_z \vec{u}_z) \rangle + a_5 \langle \vec{\phi}_r, \mathbf{G}_z^T (\mathbf{G}_z \vec{u}_r + \mathbf{G}_r \vec{u}_z) \rangle \\ &= \langle a_1 \mathbf{G}_r^T \mathbf{G}_r \vec{u}_r + \mathbf{M}^T \mathbf{A}_2 \mathbf{M} \vec{u}_r + a_3 \mathbf{G}_r^T \mathbf{G}_z \vec{u}_z + (\vec{a}_4 \mathbf{G}_r)^T \mathbf{M} \vec{u}_r, \vec{\phi}_r \rangle + \end{aligned}$$

$$\begin{aligned}
& + \langle \mathbf{M}^T \vec{a}_4 (\mathbf{G}_r \vec{u}_r + \mathbf{G}_z \vec{u}_z) + a_5 \mathbf{G}_z^T (\mathbf{G}_z \vec{u}_r + \mathbf{G}_r \vec{u}_z), \vec{\phi}_r \rangle \\
= & \langle (a_1 \mathbf{G}_r^T \mathbf{G}_r + \mathbf{M}^T \mathbf{A}_2 \mathbf{M} + (\vec{a}_4 \mathbf{G}_r)^T \mathbf{M} + \mathbf{M}^T \vec{a}_4 \mathbf{G}_r + a_5 \mathbf{G}_z^T \mathbf{G}_z) \vec{u}_r, \vec{\phi}_r \rangle \\
& + \langle (a_3 \mathbf{G}_r^T \mathbf{G}_z + \mathbf{M}^T \vec{a}_4 \mathbf{G}_z + a_5 \mathbf{G}_z^T \mathbf{G}_r) \vec{u}_z, \vec{\phi}_r \rangle
\end{aligned}$$

and

$$\begin{aligned}
I_{\phi_z} &= \iint_T \frac{r E}{(1+\nu)(1-2\nu)} \left( (1-\nu) \frac{\partial u_z}{\partial z} \frac{\partial \phi_z}{\partial z} + \nu \left( \frac{\partial u_r}{\partial r} \frac{\partial \phi_z}{\partial z} + \frac{1}{r} u_r \frac{\partial \phi_z}{\partial z} \right) \right) + \\
& \quad + \frac{r E}{2(1+\nu)} \left( \frac{\partial u_r}{\partial z} + \frac{\partial u_z}{\partial r} \right) \frac{\partial \phi_z}{\partial r} dA \\
&\approx a_1 \langle \mathbf{G}_z \vec{\phi}_z, \mathbf{G}_z \vec{u}_z \rangle + a_3 \langle \mathbf{G}_z \vec{\phi}_z, \mathbf{G}_r \vec{u}_r \rangle + \langle \vec{a}_4 \mathbf{G}_z \vec{\phi}_z, \mathbf{M} \vec{u}_r \rangle + a_5 \langle \mathbf{G}_r \vec{\phi}_z, (\mathbf{G}_z \vec{u}_r + \mathbf{G}_r \vec{u}_z) \rangle \\
&= a_1 \langle \vec{\phi}_z, \mathbf{G}_z^T \mathbf{G}_z \vec{u}_z \rangle + a_3 \langle \vec{\phi}_z, \mathbf{G}_r^T \mathbf{G}_r \vec{u}_r \rangle + \langle \vec{\phi}_z, (\vec{a}_4 \mathbf{G}_z)^T \mathbf{M} \vec{u}_r \rangle + a_5 \langle \vec{\phi}_z, \mathbf{G}_r^T (\mathbf{G}_z \vec{u}_r + \mathbf{G}_r \vec{u}_z) \rangle \\
&= \langle a_1 \mathbf{G}_z^T \mathbf{G}_z \vec{u}_z + a_3 \mathbf{G}_r^T \mathbf{G}_r \vec{u}_r + (\vec{a}_4 \mathbf{G}_z)^T \mathbf{M} \vec{u}_r + a_5 \mathbf{G}_r^T (\mathbf{G}_z \vec{u}_r + \mathbf{G}_r \vec{u}_z), \vec{\phi}_z \rangle \\
&= \langle (a_3 \mathbf{G}_r^T \mathbf{G}_z + (\vec{a}_4 \mathbf{G}_z)^T \mathbf{M} + a_5 \mathbf{G}_z^T \mathbf{G}_r) \vec{u}_r, \vec{\phi}_z \rangle + \langle (a_1 \mathbf{G}_z^T \mathbf{G}_z + a_5 \mathbf{G}_r^T \mathbf{G}_r) \vec{u}_z, \vec{\phi}_z \rangle
\end{aligned}$$

All of the above contributions are of the form  $\langle \mathbf{A}_r \vec{u}_{r,z}, \vec{\phi}_r \rangle$  or  $\langle \mathbf{A}_z \vec{u}_{r,z}, \vec{\phi}_z \rangle$  and thus contributions to the element stiffness matrix  $\mathbf{A} \in \mathbb{M}^{6 \times 6}$ , where  $\mathbf{A}_{r,z} \in \mathbb{M}^{3 \times 3}$ . For sake of completeness here the  $6 \times 6$  element stiffness matrix.

$$\mathbf{A} = \begin{bmatrix} a_1 \mathbf{G}_r^T \mathbf{G}_r + \mathbf{M}^T \mathbf{A}_2 \mathbf{M} + (\vec{a}_4 \mathbf{G}_r)^T \mathbf{M} + \mathbf{M}^T \vec{a}_4 \mathbf{G}_r + a_5 \mathbf{G}_z^T \mathbf{G}_z & a_3 \mathbf{G}_r^T \mathbf{G}_z + \mathbf{M}^T \vec{a}_4 \mathbf{G}_z + a_5 \mathbf{G}_z^T \mathbf{G}_r \\ a_3 \mathbf{G}_z^T \mathbf{G}_r + (\vec{a}_4 \mathbf{G}_z)^T \mathbf{M} + a_5 \mathbf{G}_r^T \mathbf{G}_z & a_1 \mathbf{G}_z^T \mathbf{G}_z + a_5 \mathbf{G}_r^T \mathbf{G}_r \end{bmatrix}$$

### 8.7.3 The boundary integral

The boundary integral is similar to (34) on page 138. With  $\alpha = \frac{1-1/\sqrt{3}}{2}$  use the symmetric interpolation matrix from nodes to Gauss points

$$\mathbf{M}_b = \begin{bmatrix} 1-\alpha & \alpha \\ \alpha & 1-\alpha \end{bmatrix} = \mathbf{M}_b^T$$

to find the two Gauss points  $\vec{p}_1$  and  $\vec{p}_2$  and to evaluate the radius  $r$  at the Gauss points, leading to the diagonal matrix  $\mathbf{R} = \text{diag}([r(\vec{p}_1), r(\vec{p}_2)])$ . Then use the length  $L$  of the edge segment for the approximate integral

$$\begin{aligned}
\int_{\text{edge}} r (g_{Nr} \phi_r + g_{Nz} \phi_z) ds &\approx \frac{L}{2} \langle \mathbf{M}_b \vec{\phi}_r, \mathbf{R} \vec{g}_{Nr} \rangle + \frac{L}{2} \langle \mathbf{M}_b \vec{\phi}_z, \mathbf{R} \vec{g}_{Nz} \rangle \\
&= \frac{L}{2} \langle \vec{\phi}_r, \mathbf{M}_b, \mathbf{R} \vec{g}_{Nr} \rangle + \frac{L}{2} \langle \vec{\phi}_z, \mathbf{M}_b, \mathbf{R} \vec{g}_{Nz} \rangle,
\end{aligned}$$

where the functions  $g_{Nr}$  and  $g_{Nz}$  are evaluated at the Gauss points.

## 8.8 Construction of second order elements

To construct elements of order 2 combine procedures from Section 8.3 for second order elements for plane stress problems and the previous section 8.7 where first order elements are generated for axisymmetric problems.

### 8.8.1 Integration of $r (f_r \phi_r + f_z \phi_z)$

Use the Gauss weights  $\vec{w} \in \mathbb{R}^7$  from equation (30) on page 132 for the approximate integration over one triangle  $T$  and the vector  $\vec{r} = (r_1, r_2, \dots, r_7)^T$  of the radii at the Gauss points. With these construct the diagonal matrix

$$\mathbf{R}\mathbf{W} = \text{diag}([r_1 w_1, r_2 w_2, \dots, r_7 w_7]) \in \mathbb{M}^{7 \times 7}.$$

- If the values of the functions  $f_r$  and  $f_z$  at the seven Gauss points are denoted by the vectors  $\vec{f}_r$  and  $\vec{f}_z \in \mathbb{R}^7$ , then use the approximation

$$\begin{aligned} \iint_T r (f_r \phi_r + f_z \phi_z) dA &\approx \text{area}(T) \left( \langle \mathbf{M} \vec{\phi}_r, \text{diag}(\vec{w}) \vec{f}_r \rangle + \langle \mathbf{M} \vec{\phi}_z, \mathbf{R}\mathbf{W} \vec{f}_z \rangle \right) \\ &= \text{area}(T) \left( \langle \vec{\phi}_r, \mathbf{M}^T \mathbf{R}\mathbf{W} \vec{f}_r \rangle + \langle \vec{\phi}_z, \mathbf{M}^T \mathbf{R}\mathbf{W} \vec{f}_z \rangle \right) \\ &= \text{area}(T) \left\langle \begin{pmatrix} \vec{\phi}_r \\ \vec{\phi}_z \end{pmatrix}, \begin{bmatrix} \mathbf{M}^T \mathbf{R}\mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}^T \mathbf{R}\mathbf{W} \end{bmatrix} \begin{pmatrix} \vec{f}_r \\ \vec{f}_z \end{pmatrix} \right\rangle. \end{aligned}$$

$\mathbf{M} \in \mathbb{R}^{7 \times 6}$  is the matrix for interpolation from the nodes to the Gauss points, given in equation (37) on page 141.

- If the values of the functions  $f_r$  and  $f_z$  at the nodes are denoted by the vectors  $\vec{f}_r$  and  $\vec{f}_z \in \mathbb{R}^6$ , then use the approximation

$$\begin{aligned} \iint_T r (f_r \phi_r + f_z \phi_z) dA &\approx \text{area}(T) \left( \langle \mathbf{M} \vec{\phi}_r, \mathbf{R}\mathbf{W} \mathbf{M} \vec{f}_r \rangle + \langle \mathbf{M} \vec{\phi}_z, \mathbf{R}\mathbf{W} \mathbf{M} \vec{f}_z \rangle \right) \\ &= \text{area}(T) \left( \langle \vec{\phi}_r, \mathbf{M}^T \mathbf{R}\mathbf{W} \mathbf{M} \vec{f}_r \rangle + \langle \vec{\phi}_z, \mathbf{M}^T \mathbf{R}\mathbf{W} \mathbf{M} \vec{f}_z \rangle \right) \\ &= \text{area}(T) \left\langle \begin{pmatrix} \vec{\phi}_r \\ \vec{\phi}_z \end{pmatrix}, \begin{bmatrix} \mathbf{M}^T \mathbf{R}\mathbf{W} \mathbf{M} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}^T \mathbf{R}\mathbf{W} \mathbf{M} \end{bmatrix} \begin{pmatrix} \vec{f}_r \\ \vec{f}_z \end{pmatrix} \right\rangle. \end{aligned}$$

With the above the contributions in (80) for each element stiffness matrix can be determined. Observe that  $\mathbf{M}^T \mathbf{R}\mathbf{W} \mathbf{M}$  is a  $6 \times 6$  matrix, independent on the triangle  $T$ .

### 8.8.2 Integration of the terms involving derivatives of $\phi_r$ and $\phi_z$

Using the results from Section 6.5 the partial derivatives at the nodes of functions  $\phi$  given at the nodes find for the first component of the gradient at the Gauss points

$$\frac{\partial}{\partial r} \vec{\phi} = \frac{1}{\det(\mathbf{T})} \left[ (+z_3 - z_1) \mathbf{M}_\xi^T + (-z_2 + z_1) \mathbf{M}_\nu^T \right] \cdot \vec{\phi} =: \mathbf{G}_r \vec{\phi}$$

and for the second component of the gradient

$$\frac{\partial}{\partial z} \vec{\phi} = \frac{1}{\det(\mathbf{T})} \left[ (-r_3 + r_1) \mathbf{M}_\xi^T + (+r_2 - r_1) \mathbf{M}_\nu^T \right] \cdot \vec{\phi} =: \mathbf{G}_z \vec{\phi}.$$

Observe that the matrices  $\mathbf{G}_r$  and  $\mathbf{G}_z$  depend on the triangle  $T$ . Evaluate the coefficients  $E$  and  $\nu$  at the Gauss points  $\vec{g}_i$  and define diagonal matrices  $\mathbf{A}_j$ .

$$\begin{aligned} \mathbf{A}_1 &= \text{area}(T) \text{diag} \left( \frac{w_i r(\vec{g}_i) E(\vec{g}_i) (1 - \nu(\vec{g}_i))}{(1 + \nu(\vec{g}_i)) (1 - 2\nu(\vec{g}_i))} \right) \\ \mathbf{A}_2 &= \text{area}(T) \text{diag} \left( \frac{w_i E(\vec{g}_i) (1 - \nu(\vec{g}_i))}{r(\vec{g}_i) (1 + \nu(\vec{g}_i)) (1 - 2\nu(\vec{g}_i))} \right) \\ \mathbf{A}_3 &= \text{area}(T) \text{diag} \left( \frac{w_i r(\vec{g}_i) E(\vec{g}_i) \nu(\vec{g}_i)}{(1 + \nu(\vec{g}_i)) (1 - 2\nu(\vec{g}_i))} \right) \\ \mathbf{A}_4 &= \text{area}(T) \text{diag} \left( \frac{w_i E(\vec{g}_i) \nu(\vec{g}_i)}{(1 + \nu(\vec{g}_i)) (1 - 2\nu(\vec{g}_i))} \right) \\ \mathbf{A}_5 &= \text{area}(T) \text{diag} \left( \frac{w_i r(\vec{g}_i) E(\vec{g}_i)}{2(1 + \nu(\vec{g}_i))} \right) \end{aligned}$$

This leads to the approximate integrals

$$\begin{aligned}
I_{\phi_r} &= \iint_T \frac{r E}{(1+\nu)(1-2\nu)} \left( (1-\nu) \left( \frac{\partial u_r}{\partial r} \frac{\partial \phi_r}{\partial r} + \frac{1}{r^2} u_r \phi_r \right) + \right. \\
&\quad \left. + \nu \left( \frac{\partial u_z}{\partial z} \frac{\partial \phi_r}{\partial r} + \frac{1}{r} u_r \frac{\partial \phi_r}{\partial r} + \frac{1}{r} \phi_r \left( \frac{\partial u_r}{\partial r} + \frac{\partial u_z}{\partial z} \right) \right) \right) + \\
&\quad + \frac{r E}{2(1+\nu)} \left( \frac{\partial u_r}{\partial z} + \frac{\partial u_z}{\partial r} \right) \frac{\partial \phi_r}{\partial z} dA \\
&\approx \langle \mathbf{G}_r \vec{\phi}_r, \mathbf{A}_1 \mathbf{G}_r \vec{u}_r \rangle + \langle \mathbf{M} \vec{\phi}_r, \mathbf{A}_2 \mathbf{M} \vec{u}_r \rangle + \langle \mathbf{G}_r \vec{\phi}_r, \mathbf{A}_3 \mathbf{G}_z \vec{u}_z \rangle + \\
&\quad + \langle \mathbf{G}_r \vec{\phi}_r, \mathbf{A}_4 \mathbf{M} \vec{u}_r \rangle + \langle \mathbf{M} \vec{\phi}_r, \mathbf{A}_4 (\mathbf{G}_r \vec{u}_r + \mathbf{G}_z \vec{u}_z) \rangle + \langle \mathbf{G}_z \vec{\phi}_r, \mathbf{A}_5 (\mathbf{G}_z \vec{u}_r + \mathbf{G}_r \vec{u}_z) \rangle \\
&= \langle \vec{\phi}_r, \mathbf{G}_r^T \mathbf{A}_1 \mathbf{G}_r \vec{u}_r \rangle + \langle \vec{\phi}_r, \mathbf{M}^T \mathbf{A}_2 \mathbf{M} \vec{u}_r \rangle + \langle \vec{\phi}_r, \mathbf{G}_r^T \mathbf{A}_3 \mathbf{G}_z \vec{u}_z \rangle + \\
&\quad + \langle \vec{\phi}_r, \mathbf{G}_r^T \mathbf{A}_4 \mathbf{M} \vec{u}_r \rangle + \langle \vec{\phi}_r, \mathbf{M}^T \mathbf{A}_4 (\mathbf{G}_r \vec{u}_r + \mathbf{G}_z \vec{u}_z) \rangle + \langle \vec{\phi}_r, \mathbf{G}_z^T \mathbf{A}_5 (\mathbf{G}_z \vec{u}_r + \mathbf{G}_r \vec{u}_z) \rangle \\
&= \langle \mathbf{G}_r^T \mathbf{A}_1 \mathbf{G}_r \vec{u}_r + \mathbf{M}^T \mathbf{A}_2 \mathbf{M} \vec{u}_r + \mathbf{G}_r^T \mathbf{A}_3 \mathbf{G}_z \vec{u}_z + \mathbf{G}_r^T \mathbf{A}_4 \mathbf{M} \vec{u}_r, \vec{\phi}_r \rangle + \\
&\quad + \langle \mathbf{M}^T \mathbf{A}_4 (\mathbf{G}_r \vec{u}_r + \mathbf{G}_z \vec{u}_z) + \mathbf{G}_z^T \mathbf{A}_5 (\mathbf{G}_z \vec{u}_r + \mathbf{G}_r \vec{u}_z), \vec{\phi}_r \rangle \\
&= \langle (\mathbf{G}_r^T \mathbf{A}_1 \mathbf{G}_r + \mathbf{M}^T \mathbf{A}_2 \mathbf{M} + \mathbf{G}_r^T \mathbf{A}_4 \mathbf{M} + \mathbf{M}^T \mathbf{A}_4 \mathbf{G}_r + \mathbf{G}_z^T \mathbf{A}_5 \mathbf{G}_z) \vec{u}_r, \vec{\phi}_r \rangle \\
&\quad + \langle (\mathbf{G}_r^T \mathbf{A}_3 \mathbf{G}_z + \mathbf{M}^T \mathbf{A}_4 \mathbf{G}_z + \mathbf{G}_z^T \mathbf{A}_5 \mathbf{G}_r) \vec{u}_z, \vec{\phi}_r \rangle
\end{aligned}$$

and

$$\begin{aligned}
I_{\phi_z} &= \iint_T \frac{r E}{(1+\nu)(1-2\nu)} \left( (1-\nu) \frac{\partial u_z}{\partial z} \frac{\partial \phi_z}{\partial z} + \nu \left( \frac{\partial u_r}{\partial r} \frac{\partial \phi_z}{\partial z} + \frac{1}{r} u_r \frac{\partial \phi_z}{\partial z} \right) \right) + \\
&\quad + \frac{r E}{2(1+\nu)} \left( \frac{\partial u_r}{\partial z} + \frac{\partial u_z}{\partial r} \right) \frac{\partial \phi_z}{\partial r} dA \\
&\approx \langle \mathbf{G}_z \vec{\phi}_z, \mathbf{A}_1 \mathbf{G}_z \vec{u}_z \rangle + \langle \mathbf{G}_z \vec{\phi}_z, \mathbf{A}_3 \mathbf{G}_r \vec{u}_r \rangle + \langle \mathbf{A}_4 \mathbf{G}_z \vec{\phi}_z, \mathbf{M} \vec{u}_r \rangle + \langle \mathbf{G}_r \vec{\phi}_z, \mathbf{A}_5 (\mathbf{G}_z \vec{u}_r + \mathbf{G}_r \vec{u}_z) \rangle \\
&= \langle \vec{\phi}_z, \mathbf{G}_z^T \mathbf{A}_1 \mathbf{G}_z \vec{u}_z \rangle + \langle \vec{\phi}_z, \mathbf{G}_z^T \mathbf{A}_3 \mathbf{G}_r \vec{u}_r \rangle + \langle \vec{\phi}_z, \mathbf{G}_z^T \mathbf{A}_4 \mathbf{M} \vec{u}_r \rangle + \langle \vec{\phi}_z, \mathbf{G}_r^T \mathbf{A}_5 (\mathbf{G}_z \vec{u}_r + \mathbf{G}_r \vec{u}_z) \rangle \\
&= \langle \mathbf{G}_z^T \mathbf{A}_1 \mathbf{G}_z \vec{u}_z + \mathbf{G}_z^T \mathbf{A}_3 \mathbf{G}_r \vec{u}_r + \mathbf{G}_z^T \mathbf{A}_4 \mathbf{M} \vec{u}_r + \mathbf{G}_r^T \mathbf{A}_5 (\mathbf{G}_z \vec{u}_r + \mathbf{G}_r \vec{u}_z), \vec{\phi}_z \rangle \\
&= \langle (\mathbf{G}_z^T \mathbf{A}_3 \mathbf{G}_r + \mathbf{G}_z^T \mathbf{A}_4 \mathbf{M} + \mathbf{G}_r^T \mathbf{A}_5 \mathbf{G}_z) \vec{u}_r, \vec{\phi}_z \rangle + \langle (\mathbf{G}_z^T \mathbf{A}_1 \mathbf{G}_z + \mathbf{G}_r^T \mathbf{A}_5 \mathbf{G}_r) \vec{u}_z, \vec{\phi}_z \rangle
\end{aligned}$$

All of the above contributions are of the form  $\langle \mathbf{A}_r \vec{u}_{r,z}, \vec{\phi}_r \rangle$  or  $\langle \mathbf{A}_z \vec{u}_{r,z}, \vec{\phi}_z \rangle$  and thus contributions to the element stiffness matrix  $\mathbf{A} \in \mathbb{M}^{12 \times 12}$ , where  $\mathbf{A}_{r,z} \in \mathbb{M}^{6 \times 6}$ . For sake of completeness here the  $12 \times 12$  element stiffness matrix  $\mathbf{A}$ .

$$\begin{bmatrix} \mathbf{G}_r^T \mathbf{A}_1 \mathbf{G}_r + \mathbf{M}^T \mathbf{A}_2 \mathbf{M} + \mathbf{G}_r^T \mathbf{A}_4 \mathbf{M} + \mathbf{M}^T \mathbf{A}_4 \mathbf{G}_r + \mathbf{G}_z^T \mathbf{A}_5 \mathbf{G}_z & \mathbf{G}_r^T \mathbf{A}_3 \mathbf{G}_z + \mathbf{M}^T \mathbf{A}_4 \mathbf{G}_z + \mathbf{G}_z^T \mathbf{A}_5 \mathbf{G}_r \\ \mathbf{G}_z^T \mathbf{A}_3 \mathbf{G}_r + \mathbf{G}_z^T \mathbf{A}_4 \mathbf{M} + \mathbf{G}_r^T \mathbf{A}_5 \mathbf{G}_z & \mathbf{G}_z^T \mathbf{A}_1 \mathbf{G}_z + \mathbf{G}_r^T \mathbf{A}_5 \mathbf{G}_r \end{bmatrix}$$

### 8.8.3 The boundary integral

The boundary integral is constructed similar to the procedures in Section 8.3.3, i.e. building on

$$\int_{-h/2}^{h/2} f(x) dx \approx \frac{h}{18} \left( 5 f\left(-\frac{\sqrt{3}}{2\sqrt{5}} h\right) + 8 f(0) + 5 f\left(\frac{\sqrt{3}}{2\sqrt{5}} h\right) \right).$$

If the values of a function  $f$  at the two endpoints and the midpoint are denoted by  $\vec{f} = (f_1, f_2, f_3)^T$  use a quadratic interpolation to find the values at the three Gauss integration points, given by  $\mathbf{M}_B \vec{f}$  and evaluate the

radii  $r_i$  at the Gauss points. With the length  $L$  of the segment on the edge obtain the approximate integral

$$\begin{aligned} \int_{\text{edge}} r (g_r \phi_r + g_z \phi_z) ds &\approx \frac{L}{18} \langle \mathbf{M}_B \vec{\phi}_r, \begin{pmatrix} 5 r_1 g_r(\vec{p}_1) \\ 8 r_2 g_r(\vec{p}_2) \\ 5 r_3 g_r(\vec{p}_3) \end{pmatrix} \rangle + \frac{L}{18} \langle \mathbf{M}_B \vec{\phi}_z, \begin{pmatrix} 5 r_1 g_z(\vec{p}_1) \\ 8 r_2 g_z(\vec{p}_2) \\ 5 r_3 g_z(\vec{p}_3) \end{pmatrix} \rangle \\ &= \frac{L}{18} \langle \vec{\phi}_r, \mathbf{M}_B^T \begin{pmatrix} 5 r_1 g_r(\vec{p}_1) \\ 8 r_2 g_r(\vec{p}_2) \\ 5 r_3 g_r(\vec{p}_3) \end{pmatrix} \rangle + \frac{L}{18} \langle \vec{\phi}_z, \mathbf{M}_B^T \begin{pmatrix} 5 r_1 g_z(\vec{p}_1) \\ 8 r_2 g_z(\vec{p}_2) \\ 5 r_3 g_z(\vec{p}_3) \end{pmatrix} \rangle \end{aligned}$$

The integration weights can be combined with the interpolation matrix  $\mathbf{M}_B$  by

$$\mathbf{M}_{BC} = \frac{1}{18} \mathbf{M}_B^T \begin{bmatrix} 5 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 5 \end{bmatrix} \approx \begin{bmatrix} 0.1909 & 0 & -0.0242 \\ 0.1111 & 0.4444 & 0.1111 \\ -0.0242 & 0 & 0.1909 \end{bmatrix}.$$

This matrix  $\mathbf{M}_{BC}$  does not depend on the current edge segment and leads to

$$\int_{\text{edge}} r (g_r \phi_r + g_z \phi_z) ds \approx L \langle \vec{\phi}_r, \mathbf{M}_{BC} \begin{pmatrix} r_1 g_r(\vec{p}_1) \\ r_2 g_r(\vec{p}_2) \\ r_3 g_r(\vec{p}_3) \end{pmatrix} \rangle + L \langle \vec{\phi}_z, \mathbf{M}_{BC} \begin{pmatrix} r_1 g_z(\vec{p}_1) \\ r_2 g_z(\vec{p}_2) \\ r_3 g_z(\vec{p}_3) \end{pmatrix} \rangle.$$

The effect of the boundary integral on the global stiffness matrix and the vector is very similar to the approach shown at the end of Section 6.5.8.

## 8.9 Construction of third order elements

To construct elements of order 3 combine procedures from Section 8.4 for third order elements for plane stress problems and the previous section 8.8 where second order elements are generated for axisymmetric problems.

### 8.9.1 Integration of $r (f_r \phi_r + f_z \phi_z)$

The computations are identical to Section 8.8.1 for second order elements. The only difference is the interpolation matrix  $\mathbf{M} \in \mathbb{M}^{7 \times 10}$ , which has to interpolate from the 10 nodes to the 7 Gauss points. See equation (48) on page 151. The contributions in (80) for each element stiffness matrix can be determined. The matrix  $\mathbf{M}^T \mathbf{R} \mathbf{W} \mathbf{M}$  is a  $10 \times 10$  matrix, independent on the triangle  $T$ .

### 8.9.2 Integration of the terms involving derivatives of $\phi_z$ and $\phi_r$

The algorithm is extremely similar to Section 8.8.2 for second order elements, but the matrices  $\mathbf{M}_\xi$  and  $\mathbf{M}_\nu$  are of size  $7 \times 10$ . This leads to the matrices  $\mathbf{G}_x$  and  $\mathbf{G}_y \in \mathbb{M}^{7 \times 10}$  to evaluate the partial derivatives at the nodes, using the values of the function at the nodes. The resulting matrices  $\mathbf{A}_r$  and  $\mathbf{A}_z$  are of size  $10 \times 10$ , leading to the element stiffness matrix  $\mathbf{A} \in \mathbb{M}^{20 \times 20}$ .

### 8.9.3 The boundary integral

The algorithm is extremely similar to Section 8.8.3 for second order elements. The effect of the boundary integral on the global stiffness matrix and the vector is very similar to the approach shown at the end of Section 6.6.7.

## 9 Examples, Examples, Examples

### 9.1 An elliptic problem with variable coefficients

The elliptic BVP in Section 5.5 is

$$\begin{aligned} -\nabla \cdot ((1+x^2)\nabla u(x,y)) &= -4(1+x^2)\exp(-2y) && \text{for } (x,y) \in \Omega \\ \frac{\partial u(y,0)}{\partial x} &= 0 && \text{for } 1 \leq y \leq 2 \\ u(x,y) &= \exp(-2y) && \text{on other sections of the boundary} \end{aligned}$$

on the domain shown in Figure 73(a). The exact solution is given by  $u_e(x,y) = \exp(-2y)$ . To solve this BVP with FEMoctave use the following steps:

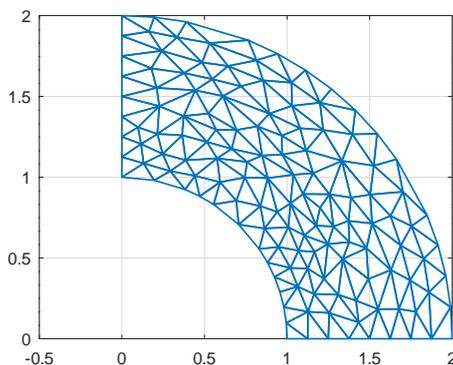
1. Use `CreateMeshTriangle()` to generate a mesh on the rectangle  $1 \leq r \leq 2$  and  $0 \leq \varphi \leq \pi/2$ .
2. With the polar coordinates use

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} r \cos \varphi \\ r \sin \varphi \end{pmatrix}$$

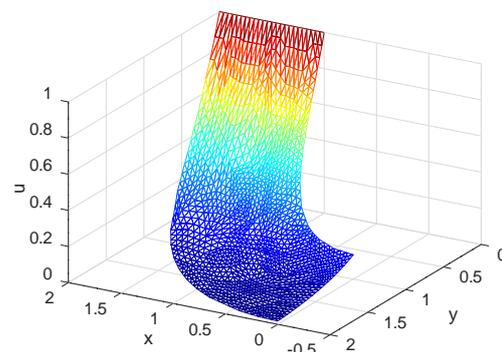
to generate the mesh on the section of a ring, visible in Figure 73(a) with the help of an appropriate function `Deform()` and the function `MeshDeform()`.

3. Then use `MeshUpgrade()` to generate a mesh with third order elements.
4. Define the coefficient functions  $a(x,y) = 1+x^2$  and the right hand side  $f(x,y) = -4(1+x^2)\exp(-2y)$  with *Octave* functions.
5. Call the function `BVP2DSym()` with appropriate arguments to calculate the approximate solution  $u(x,y)$ .
6. Use `FEMtrimesh()` to display the solution visible in Figure 73(b) and then use `FEMIntegrate()` to determine the  $L_2$ -error

$$\left( \iint_{\Omega} |u(x,y) - u_{exact}(x,y)|^2 dA \right)^{1/2}.$$



(a) the mesh



(b) the solution

Figure 73: Difference to the exact solution of a BVP

**DeformVariableCoeff.m**

```

clear *
h = 0.1
function xy_new = Deform(xy)
    xy_new = [xy(:,1).*cos(xy(:,2)), xy(:,1).*sin(xy(:,2))];
endfunction

function u = f_u_exact(xy)
    u = exp(-2*xy(:,2));
endfunction

function u = f_DDu_exact(xy)
    u = -4*(1+xy(:,1).^2).*exp(-2*xy(:,2));
endfunction

function a = f_a(xy)
    a = 1 + xy(:,1).^2;
endfunction

FEMmesh = CreateMeshTriangle('Test', [1,0,-1;2,0,-1;2,pi/2,-2;1,pi/2,-1],h^2);
FEMmesh = MeshDeform(FEMmesh,'Deform');
figure(1); FEMtrimesh(FEMmesh)
FEMmesh = MeshUpgrade(FEMmesh,'cubic');
u = BVP2Dsymb(FEMmesh,'f_a',0,'f_DDu_exact','f_u_exact',0,0);
figure(2); FEMtrimesh(FEMmesh,u)
    xlabel('x'); ylabel('y'); zlabel('u'); view([-150,30])
u_exact = f_u_exact(FEMmesh.nodes);
L2Error = sqrt(FEMIntegrate(FEMmesh,(u-u_exact).^2))
-->
L2Error = 3.3205e-06

```

**9.2 An animated wave**

With a narrow Gauss bell surface around  $(x, y) \approx (1, 0)$  as initial value and zero initial velocity observe the waves traveling away from the initial location and the different types of reflections at the boundaries. Figure 74 shows the final status.

**WaveAnimation.m**

```

if 0 %% linear elements
    FEMmesh = CreateMeshRect(linspace(0,pi,101),linspace(-pi,pi,101),-1,-2,-2,-2);
else %% quadratic elements
    FEMmesh = CreateMeshRect(linspace(0,pi,51),linspace(-pi,pi,51),-1,-2,-2,-2);
    FEMmesh = MeshUpgrade(FEMmesh);
endif
x = FEMmesh.nodes(:,1); y = FEMmesh.nodes(:,2);

m=1; alpha=0.0; a=1; b0=0; bx=0; by=0; f=0; gD=0; gN1=0; gN2=0;
t0=0; tend=3 ; steps = [150,10];

u0 = exp(-25*((x-1).^2+(y-0).^2));
v0 = zeros(length(FEMmesh.nodes),1);
[u_dyn,t] = I2BVP2D(FEMmesh,m,alpha,a,b0,bx,by,f,gD,gN1,gN2,u0,v0,t0,tend,steps);

figure(1) % show animation

```

```

for t_ii = 1:length(t)
    FEMtrimesh(FEMmesh,u_dyn(:,t_ii))
    axis([0 pi -pi pi -0.2 0.4]); xlabel('x'); ylabel('y')
    drawnow();
endfor

```

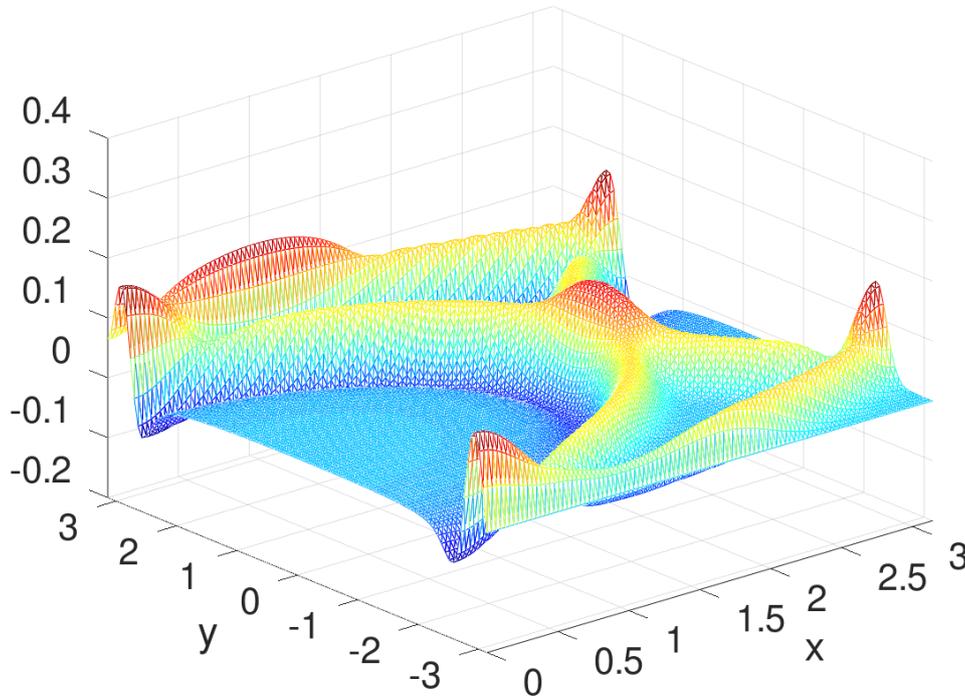


Figure 74: Traveling waves on a rectangle

### 9.3 An elliptic problem with radial symmetry, superconvergence

The Bessel function

$$u(x, y) = f(x, y) = J_0(\sqrt{x^2 + y^2})$$

is an exact solution of the BVP

$$\begin{aligned}
 -\Delta u + u &= 2f & \text{for } 0 < x, y < 1 \\
 u &= f & \text{for } (1, y) \text{ and } (x, 1) \\
 \frac{\partial u}{\partial n} &= 0 & \text{for } (0, y) \text{ and } (x, 0)
 \end{aligned}$$

A solution is shown in Figure 75. This BVP is solved by two slightly different approaches, and then the difference to the known exact solution is displayed in Figure 76. In both cases first a mesh with linear elements is generated, then upgraded to a mesh with quadratic elements, using `MeshUpgrade()`. Then a mesh with identical nodes and DOF with linear elements is generated by `MeshQuad2Linear()`.

1. Use a uniform mesh generated by `CreateMeshRect`, leading to 400 degrees of freedom. The result in Figure 76(a) shows the effect of super-convergence. Caused by the extremely regular structure of the grid points the differences are smaller than can reasonably be expected.

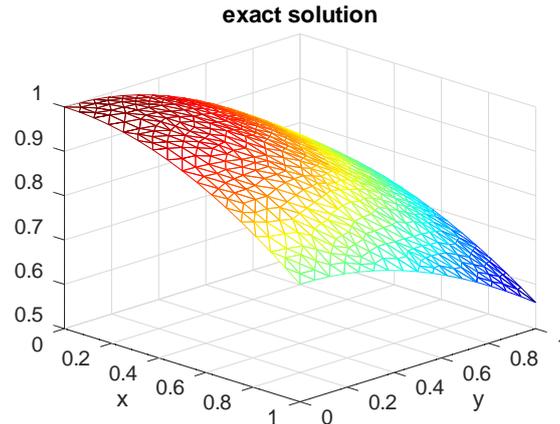


Figure 75: The radial Bessel function as solution of a BVP

2. Use a non-uniform mesh generated by `CreateMeshTriangle`, leading to 432 degrees of freedom. Thus one expects to obtain similar accuracy. The result in Figure 76(b) confirms this.

```

N = 10; Triangle = 1
if Triangle
    FEMmesh = CreateMeshTriangle('test1',[0 0 -2;1 0 -1; 1 1 -1; 0 1 -2],0.75/N^2);
    FEMmesh = MeshUpgrade(FEMmesh);
    FEMmesh1 = MeshQuad2Linear(FEMmesh);
    nDOFTri = [FEMmesh.nDOF, FEMmesh1.nDOF]
else
    FEMmesh = CreateMeshRect(linspace(0,1,N+1),linspace(0,1,N+1),-2,-1,-2,-1);
    FEMmesh = MeshUpgrade(FEMmesh,'quadratic');
    FEMmesh1 = MeshQuad2Linear(FEMmesh);
    nDOFRect = [FEMmesh.nDOF, FEMmesh1.nDOF]
endif

```

To generate Figure 77 the command `FEMgriddata()` is used to evaluate the functions on a much finer grid (not recomputing, just evaluation) and then display the difference between the approximate and exact solution. This figure illustrates that the effect of superconvergence does not provide additional accuracy one can reliably count on.

The gradient of this solution  $u$  can be determined using  $\frac{\partial}{\partial r} J_0(r) = -J_1(r)$  and

$$\begin{pmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial u}{\partial y} \end{pmatrix} = \begin{pmatrix} \cos \phi \\ \sin \phi \end{pmatrix} \frac{\partial u}{\partial r} + \begin{pmatrix} -\sin \phi \\ \cos \phi \end{pmatrix} \frac{\partial u}{\partial \phi} = - \begin{pmatrix} \cos \phi \\ \sin \phi \end{pmatrix} J_1(r).$$

Using the above FEM results compare the true partial derivative  $\frac{\partial u}{\partial x}$  with the one obtained by FEM with second order elements. Find the result in Figure 78. Observe the structure of the difference for the uniform mesh.

The above can be repeated using first order elements, leading to Figure 79. The size of the elements was set such that the same number of degrees of freedom are used. Observe that superconvergence strikes again. In this case I have a solid argument for the structural difference along the border.

Find more information on superconvergence in [Zien13, §15.2] or a short demo in [Stah08, §6.8.2].

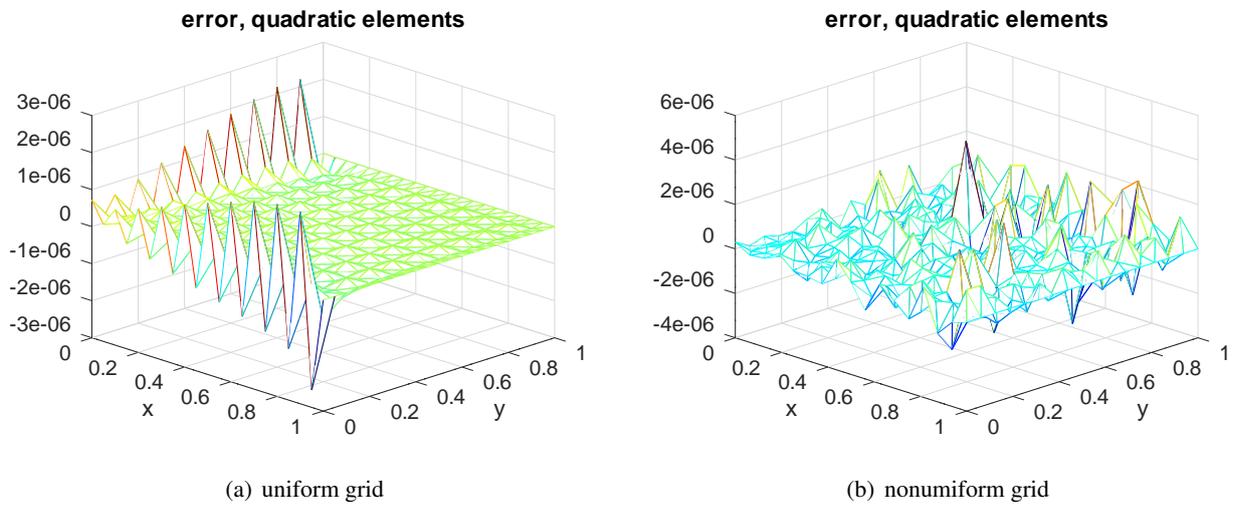


Figure 76: Difference to the exact solution of a BVP

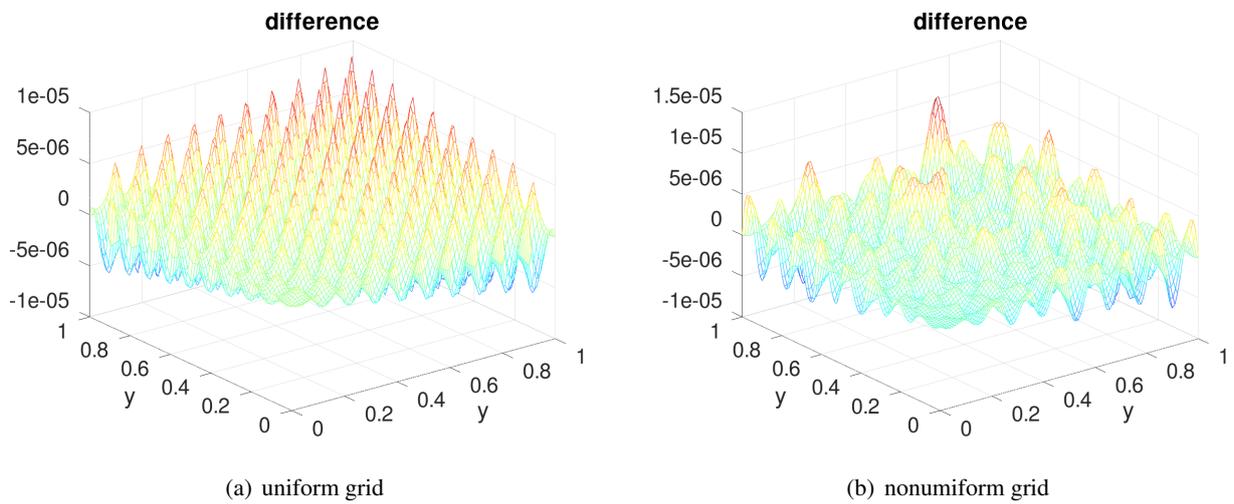


Figure 77: Difference to the exact solution of a BVP, using quadratic elements and interpolation to a finer grid.

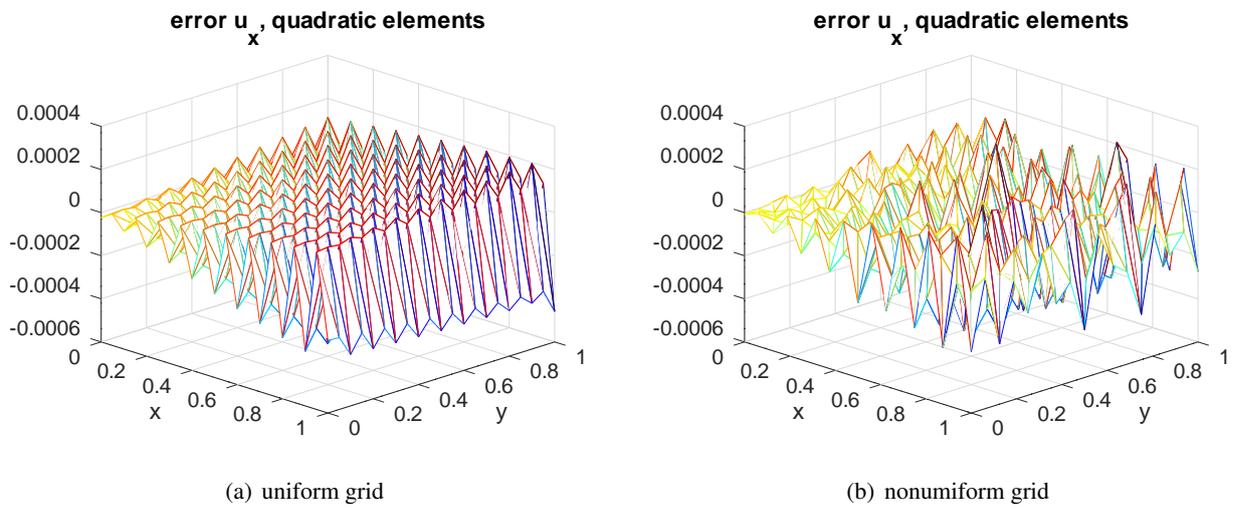


Figure 78: Difference of  $\frac{\partial u}{\partial x}$  to the exact solution, using second order elements

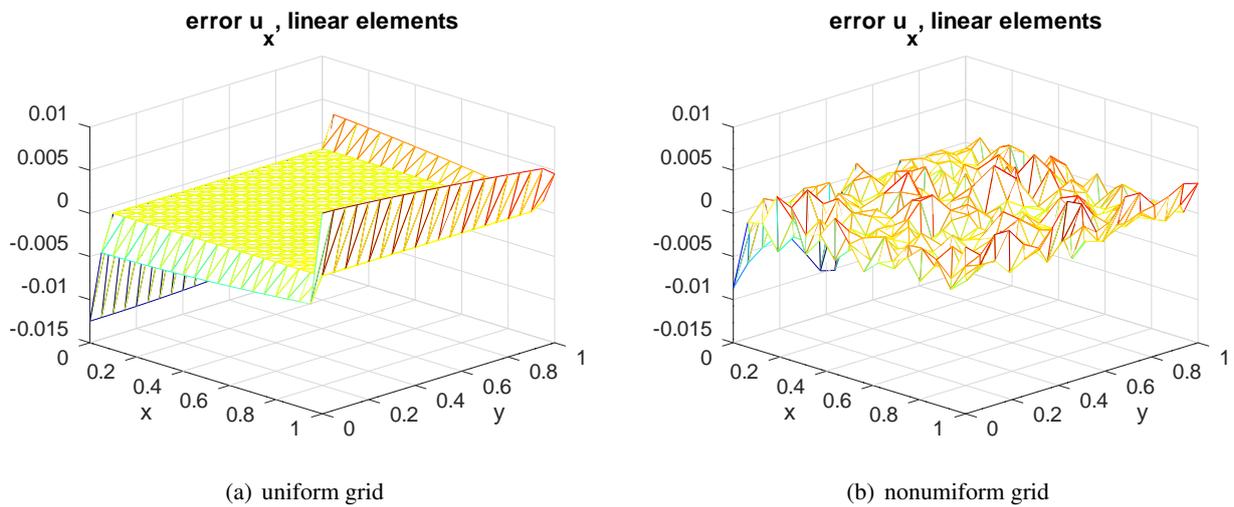


Figure 79: Difference of  $\frac{\partial u}{\partial x}$  to the exact solution, using first order elements

### 9.4 An example with limited regularity

Let  $\Omega \in \mathbb{R}^2$  be the unit square  $-1 < x, y < 1$ , with the fourth quadrant ( $x > 0, y < 0$ ) cut out. For some of the calculations identify  $(x, y) \in \mathbb{R}^2$  with  $z = x + iy \in \mathbb{C}$ . Examine the functions

$$\begin{aligned} w(z) &= z^{2/3} = (r e^{i\phi})^{2/3} = r^{2/3} e^{i\phi 2/3} = r^{2/3} (\cos(\phi 2/3) + i \sin(\phi 2/3)) \\ u(z) &= r^{2/3} \sin(\phi 2/3) \\ u(x, y) &= (x^2 + y^2)^{1/3} \sin\left(\frac{2}{3} \operatorname{atan2}(y, x)\right) \end{aligned}$$

This function satisfies  $-\Delta u = 0$  and  $u(t, 0) = u(0, -t) = 0$  for  $t > 0$ . Since  $\frac{\partial}{\partial r} u = \frac{2}{3} r^{-1/3} \sin(\frac{2}{3} \phi)$  and  $\frac{\partial}{\partial \phi} u = \frac{2}{3} r^{2/3} \cos(\frac{2}{3} \phi)$  the partial derivatives of this function have a singularity at the origin. Compute

$$\begin{aligned} \|\nabla u\|^2 &= \left|\frac{\partial u}{\partial r}\right|^2 + \left|\frac{1}{r} \frac{\partial u}{\partial \phi}\right|^2 = \frac{4}{9} r^{-2/3} + \frac{4}{9} \frac{1}{r^2} \cos^2\left(\frac{2}{3} \phi\right) \\ \iint_{\Omega} \|\nabla u\|^2 dA &= \frac{4}{9} \int_0^1 \left( \int_0^{3\pi/2} r^{-2/3} + r^{-2} \cos^2\left(\frac{2}{3} \phi\right) d\phi \right) r dr \\ &= \frac{4}{9} \int_0^1 \left( \frac{3\pi}{2} r^{-2/3} + r^{-2} \frac{3\pi}{4} \right) r dr = \frac{2\pi}{3} \int_0^1 r^{1/3} dr + \frac{\pi}{3} \int_0^1 \frac{1}{r} dr = \infty \end{aligned}$$

to observe that the gradient is not bounded in the  $L_2$  sense. Thus the standard error estimates based on  $\tilde{C}\tilde{A}\tilde{O}$ 's Lemma do not apply. Expect approximation and convergence problems close to the origin. This is confirmed by the code below and the resulting Figure 80. This example illustrates that non-convex domains with sharp corners might cause convergence problems.

#### SingularDisc.m

```
x_p = [0;1;1;-1;-1;0]; y_p = [0;0;1;1;-1;-1];

FEMmesh = CreateMeshTriangle("circle34", [x_p,y_p,-ones(size(x_p))], 0.01);
FEMmesh = MeshUpgrade(FEMmesh);

function res = gD(xy)
    phi = mod(atan2(xy(:,2),xy(:,1)),2*pi);
    res = (xy(:,1).^2+ xy(:,2).^2).^(1/3).*sin(2/3*phi);
endfunction

u = BVP2Dsym(FEMmesh,1,0,0,'gD',0,0);
figure(1); FEMtrimesh(FEMmesh,u);
    xlabel("x"); ylabel("y"); title('FEM solution'); view([30,30])

u_exact = gD(FEMmesh.nodes);
figure(2); FEMtrimesh(FEMmesh,-u+u_exact);
xlabel("x"); ylabel("y"); title('Error of FEM solution'); view([30,30])
```

The gradient in Cartesian coordinates can be determined by

$$\begin{aligned} \begin{pmatrix} \frac{\partial u}{\partial x} \\ \frac{\partial u}{\partial y} \end{pmatrix} &= \begin{pmatrix} \cos \phi \\ \sin \phi \end{pmatrix} \frac{\partial u}{\partial r} + \begin{pmatrix} -\sin \phi \\ \cos \phi \end{pmatrix} \frac{\partial u}{\partial \phi} \\ &= \begin{pmatrix} \cos \phi \\ \sin \phi \end{pmatrix} \frac{2}{3} r^{-1/3} \sin(\phi 2/3) + \begin{pmatrix} -\sin \phi \\ \cos \phi \end{pmatrix} \frac{2}{3} r^{2/3} \cos(\phi 2/3) \end{aligned}$$

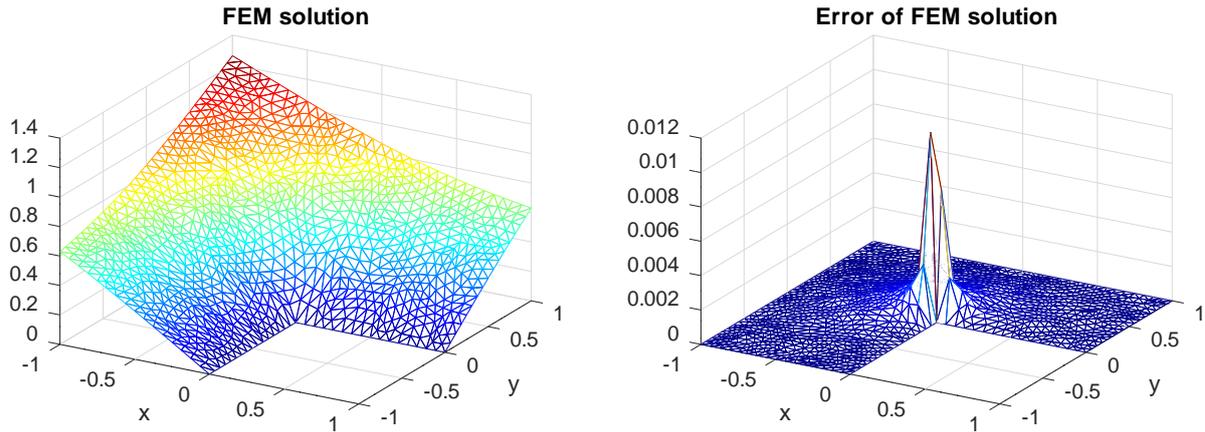


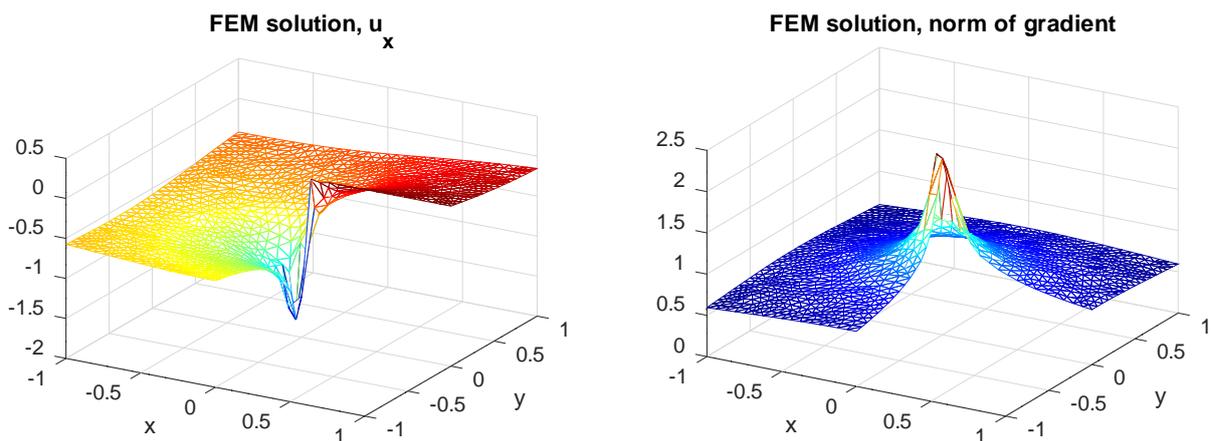
Figure 80: A solution with singular partial derivatives at the origin

and then visualized, leading to Figure 81. It is clearly visible that the FEM solution is not accurate where the gradient has a singularity.

```
[ux,uy] = FEMevaluateGradient(FEMmesh,u);
figure(3); FEMtrimesh(FEMmesh,ux);
    xlabel("x"); ylabel("y"); title('FEM solution, u_x'); view([30,30])

figure(4); FEMtrimesh(FEMmesh,uy);
    xlabel("x"); ylabel("y"); title('FEM solution, u_y'); view([30,30])

figure(5); FEMtrimesh(FEMmesh,sqrt(ux.^2+uy.^2));
    xlabel("x"); ylabel("y"); title('FEM solution, norm of gradient');
    view([30,30])
```

Figure 81: A solution with singular partial derivatives, graphs of  $\frac{\partial u}{\partial x}$  and  $\|\nabla u\|$ 

Singularities can show up in mechanical problems, e.g. the washer fastener in Section 9.30.

## 9.5 A potential flow problem

Consider a laminar flow between two plates with an obstacle between the two plates. Assume that the situation is independent on one of the spatial variables and consider a cross section only shown in Figure 82. The goal is to find the velocity field  $\vec{v}$  of the fluid.

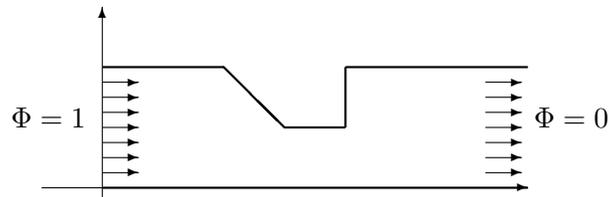


Figure 82: Fluid flow between two plates, the setup

This problem is solved by introducing a velocity potential  $\Phi(x, y)$ . The velocity vector  $\vec{v}$  is then given by

$$\vec{v} = \begin{pmatrix} v_x \\ v_y \end{pmatrix} = - \begin{pmatrix} \frac{\partial \Phi}{\partial x} \\ \frac{\partial \Phi}{\partial y} \end{pmatrix}.$$

The flow is assumed to be uniform far away from the obstacle. Thus set the potential to  $\Phi = 1$  (resp.  $\Phi = 0$ ) at the left (resp. right) end of the plates. Since the fluid can not flow through the boundaries of the plates use that the normal component of the velocity has to vanish at the upper and lower boundary. The differential equation to be satisfied by  $\Phi$  is

$$\Delta \Phi = \text{div}(\text{grad } \Phi) = 0$$

In Figure 83 the resulting flow is visualized. Observe the unrealistic velocities at the corners of the domain. The model of laminar flow is not appropriate in this situation. Selecting a finer mesh is no solution to this problem. Mathematically the effect is related to the effect illustrated in Section 9.4.

The results are generated by the code below.

### PotentialFlow.m

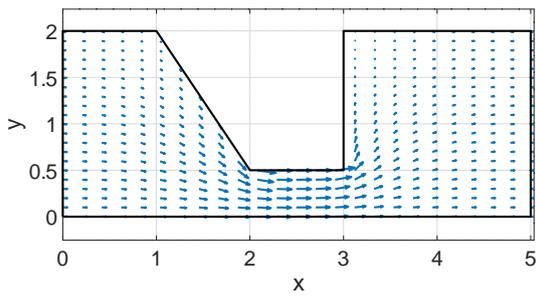
```

%% define the domain
xy = [0 0 -2; 5 0 -1; 5 2 -2; 3 2 -2; 3 0.5 -2; 2 0.5 -2; 1 2 -2; 0 2 -1];
if 1    %% linear elements
    FEMmesh = CreateMeshTriangle('PotentialFlow', xy, 0.003);
elseif 1 %% quadratic elements
    FEMmesh = CreateMeshTriangle('PotentialFlow', xy, 4*0.003);
    FEMmesh = MeshUpgrade(FEMmesh, 'quadratic');
else    %% cubic elements
    FEMmesh = CreateMeshTriangle('PotentialFlow', xy, 9*0.003);
    FEMmesh = MeshUpgrade(FEMmesh, 'cubic');
endif

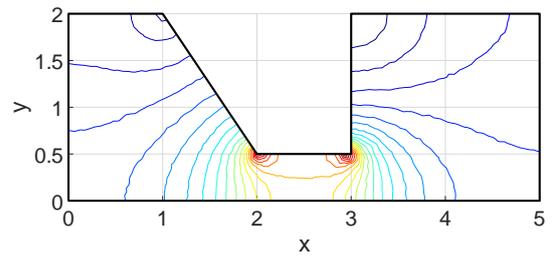
x = FEMmesh.nodes(:,1);  y = FEMmesh.nodes(:,2);
function res = gD(xy)    res = 1-xy(:,1)/5; endfunction
u = BVP2Dsym(FEMmesh, 1, 0, 0, 'gD', 0, 0);
figure(1); FEMtrimesh(FEMmesh, u)
           xlabel('x'); ylabel('y'); zlabel('potential')

[xx,yy] = meshgrid(linspace(0, 5-0.01, 25), linspace(0, 2-0.01, 21));
[u_int, ux_int, uy_int] = FEMgriddata(FEMmesh, -u, xx, yy);

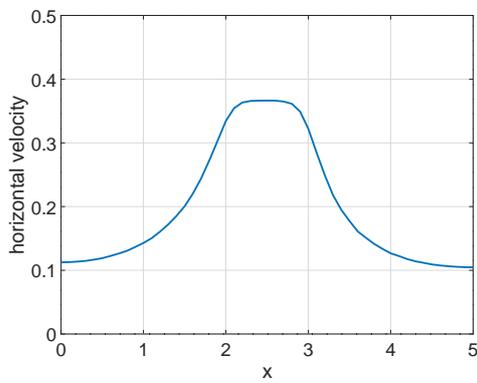
```



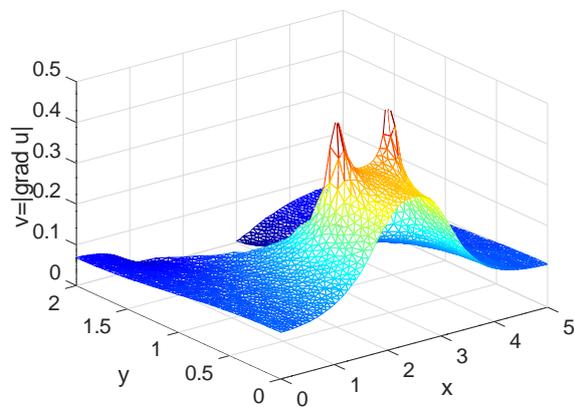
(a) field of velocity vectors



(b) velocity contours



(c) horizontal speed profile along  $y = 0.25$



(d) the velocity

Figure 83: Velocity field of an ideal fluid

```

figure(2); quiver(xx,yy,ux_int,uy_int)
           xlabel('x'); ylabel('y');
           hold on; plot([xy(:,1);0],[xy(:,2);0],'k'); hold off; axis equal

xx = linspace(0,5,101); yy = 0.25*ones(101,1);
[u_int,ux_int,uy_int] = FEMgriddata(FEMmesh,-u,xx,yy);
figure(3); plot(xx,ux_int)
           xlabel('x'); ylabel('horizontal velocity'); ylim([0 0.5])

[ux,uy] = FEMEvaluateGradient(FEMmesh,u);
figure(4); FEMtrimesh(FEMmesh,sqrt(ux.^2+ uy.^2))
           xlabel('x'); ylabel('y'); zlabel('v=|grad u|'); zlim([0 0.5])

figure(5); FEMtricontour(FEMmesh,sqrt(ux.^2+ uy.^2),21)
           xlabel('x'); ylabel('y'); zlabel(' | grad u | ');
           hold on; plot([xy(:,1);0],[xy(:,2);0],'k'); hold off
           xlim([0 5]); ylim([0 2]); axis equal

```

By integrating the horizontal velocities along vertical cuts observe the flux conservation, i.e. what's coming in on the left has to flow through the canal and leave on the right.

$$\begin{aligned}
 \text{flux at inlet } x = 0.0 &\approx 0.18337 \\
 \text{flux in middle } x = 2.5 &\approx 0.18328 \\
 \text{flux at outlet } x = 5.0 &\approx 0.18333
 \end{aligned}$$

Selecting a finer mesh or using quadratic elements will make the differences smaller.

```

yy = linspace(0,2); xx = zeros(size(yy));
vx = FEMgriddata(FEMmesh,-ux, xx, yy);           Flux_inlet_ = trapz(yy,vx)
yy = linspace(0,0.5); xx = 2.5*ones(size(yy));
vx = FEMgriddata(FEMmesh,-ux,xx,yy);           Flux_middle = trapz(yy,vx)
yy = linspace(0,2); xx = 5*ones(size(yy));
vx = FEMgriddata(FEMmesh,-ux, xx, yy);           Flux_outlet = trapz(yy,vx)

```

## 9.6 A potential flow problem in a circular pipe

An ideal liquid is flowing through a circular pipe with diminished radius in a central section. The outer radius is given by

$$R(z) = \begin{cases} 2 & \text{for } |z| \geq 1 \\ 2 - \cos^2(\frac{\pi}{2} z) & \text{for } |z| \leq 1 \end{cases} .$$

The upper half of a section is visible in Figure 84. Assuming that the solution is independent on the angle  $\theta$  the equation  $\Delta\Phi = 0$  has to be reformulated in cylindrical coordinates and simplified.

$$\begin{aligned}
 0 &= \Delta\Phi = \text{div}(\text{grad } \Phi) = \Phi_{rr} + \frac{1}{r} \Phi_r + \frac{1}{r^2} \Phi_{\theta\theta} + \Phi_{zz} \\
 0 &= r \left( \Phi_{rr} + \frac{1}{r} \Phi_r + \Phi_{zz} \right) = r \Phi_{rr} + \Phi_r + r \Phi_{zz} = \frac{\partial}{\partial r} (r \Phi_r) + \frac{\partial}{\partial z} (r \Phi_z) .
 \end{aligned}$$

Setting  $\Phi = +1$  at the left edge and  $\Phi = -1$  at the right edge, the BVP can be solved for the potential  $\Phi(z, r)$  with the help of FEMoctave. The velocity vector is again given by the gradient

$$\vec{v} = \begin{pmatrix} v_z \\ v_r \end{pmatrix} = - \begin{pmatrix} \frac{\partial \Phi}{\partial z} \\ \frac{\partial \Phi}{\partial r} \end{pmatrix}.$$

Observe that there are no singularities for the velocities, compared to the previous section 9.5, since there are no sharp corners in the domain.

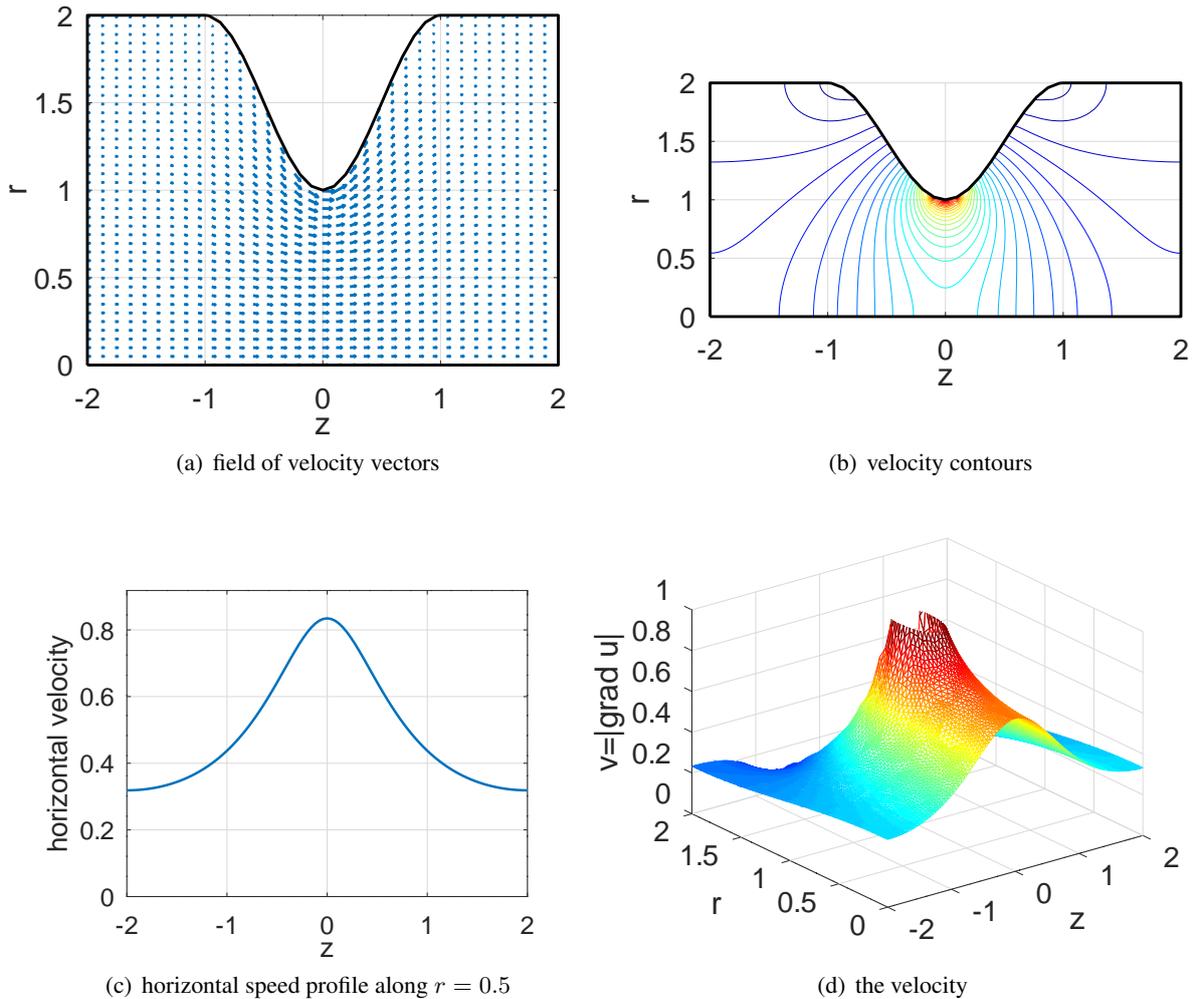


Figure 84: Velocity field of an ideal fluid in a circular pipe

#### PotentialFlowCircular.m

```
% define the domain and mesh
R = 2; R_in = 1.0; area = 0.001;
z = linspace(-1+sqrt(area), 1-sqrt(area), 21)'; r = R-R_in*cos(pi/2*z).^2;
b = -2*ones(size(z));
zr = [-2 0 -1; -2 R -2; -1 R -2; [z,r,b]; 1 R -2; 2 R -1; 2 0 -2];
if 0 %% linear elements
    FEMmesh = CreateMeshTriangle('PotentialFlow', zr, area);
elseif 0 %% quadratic elements
```

```

FEMmesh = CreateMeshTriangle('PotentialFlow', zr, 4*area);
FEMmesh = MeshUpgrade(FEMmesh, 'quadratic');
else    %% cubic elements
FEMmesh = CreateMeshTriangle('PotentialFlow', zr, 9*area);
FEMmesh = MeshUpgrade(FEMmesh, 'cubic');
endif

z = FEMmesh.nodes(:,1);  z = FEMmesh.nodes(:,2);
function res = gD(zr)    res = -zr(:,1)/2;  endfunction
function res = a_coeff(zr)  res = zr(:,2);    endfunction

u = BVP2Dsymb(FEMmesh, 'a_coeff', 0, 0, 'gD', 0, 0);

[zz, rr] = meshgrid(linspace(-2, 2-0.01, 35), linspace(0, R-0.01, 41));
[u_int, uz_int, ur_int] = FEMgriddata(FEMmesh, -u, zz, rr);

figure(1); quiver(zz, rr, uz_int, ur_int)
           xlabel('z'); ylabel('r');
           hold on; plot([zr(:,1);-2], [zr(:,2);0], 'k'); hold off
           xlim([-2, 2]); ylim([0, R]);

[uz, ur] = FEMEvaluateGradient(FEMmesh, u);
figure(2); FEMtrimesh(FEMmesh, sqrt(uz.^2+ ur.^2))
           xlabel('z'); ylabel('r'); zlabel('v=|grad u|')
           zlim([0 1]); caxis([0, 1])

zz = linspace(-2, 2, 101); rr = 0.5*ones(101, 1);
[u_int, uz_int, ur_int] = FEMgriddata(FEMmesh, -u, zz, rr);
figure(3); plot(zz, uz_int)
           xlabel('z'); ylabel('horizontal velocity');
           ylim([0 1.1*max(uz_int)])

figure(4); FEMtricontour(FEMmesh, sqrt(uz.^2+ ur.^2), 31)
           xlabel('z'); ylabel('r'); zlabel(' |grad u|')
           hold on; plot([zr(:,1);-2], [zr(:,2);0], 'k'); hold off
           xlim([-2 2]); ylim([0 R]); axis equal

```

The total flux across a vertical line  $z = \text{const}$  can be determined by the integral

$$\text{flux} = \int_0^{R(z)} v_z(r, z) 2\pi r dr = 2\pi \int_0^{R(z)} -\frac{\partial \Phi(z, r)}{\partial z} r dr.$$

```

rr = linspace(0, R); zz = -1.9*ones(size(rr)); vz = FEMgriddata(FEMmesh, -uz, zz, rr);
Flux_inlet = trapz(rr, rr.*vz)*2*pi
rr = linspace(0, R-R_in); zz = 0*ones(size(rr)); vz = FEMgriddata(FEMmesh, -uz, zz, rr);
Flux_middle = trapz(rr, rr.*vz)*2*pi
rr = linspace(0, R); zz = 1.9*ones(size(rr)); vz = FEMgriddata(FEMmesh, -uz, zz, rr);
Flux_outlet = trapz(rr, rr.*vz)*2*pi
-->
Flux_inlet = 3.3115
Flux_middle = 3.2897
Flux_outlet = 3.3115

```

The accuracy of the numerical results

$$\begin{aligned} \text{flux at inlet } z = -1.9 &\approx 3.3115 \\ \text{flux in middle } x = +0.0 &\approx 3.2897 \\ \text{flux at outlet } x = +1.9 &\approx 3.3115 \end{aligned}$$

could be improved by a finer mesh. This would verify the conservation of flux at different  $z$ -levels.

### 9.7 A minimal surface problem

Let  $u(x, y)$  be the height of a surface above the border of a 2-dimensional domain  $\Omega$  is given by a function  $g(x, y)$ . Then the function  $u$  representing the surface of minimal with has to solve a nonlinear PDE.

$$\begin{aligned} \operatorname{div}\left(\frac{1}{\sqrt{1 + |\operatorname{grad} u|^2}} \operatorname{grad} u\right) &= 0 && \text{in domain } \Omega \\ u &= g && \text{on } \Gamma = \partial\Omega \end{aligned}$$

FEMoctave is not directly capable of solving non linear problems, but a simple iteration will lead to an approxi-

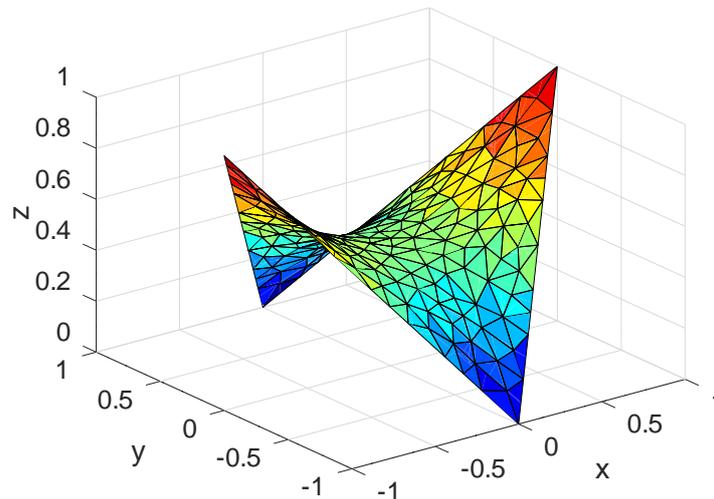


Figure 85: A minimal surface

mation of the solution.

- start with an initial solution  $u_0(x, y) = 0$
- repeat until the change in solution is small enough
  - compute the coefficient function

$$a(x, y) = \frac{1}{\sqrt{1 + |\nabla u(x, y)|^2}}$$

- Solve the boundary value problem

$$\begin{aligned} \operatorname{div}(a(x, y) \operatorname{grad} u) &= 0 && \text{in domain } \Omega \\ u &= g && \text{on } \Gamma = \partial\Omega \end{aligned}$$

The code below implements this algorithm for a square  $\Omega$  and leads to the result in Figure 85. While iterating the area of each surface is determined by integrating

$$\text{area} = \iint_{\Omega} \sqrt{1 + |\nabla u|^2} dA$$

and the average difference of subsequent solutions is computed.

#### MinimalSurface.m

```

xy = [1,0,-1;0,1,-1;-1,0,-1;0,-1,-1];
FEMmesh = CreateMeshTriangle("square",xy,0.01);
%FEMmesh = MeshUpgrade(FEMmesh,'quadratic');

x = FEMmesh.nodes(:,1); y = FEMmesh.nodes(:,2);
function res = BC(xy) res = abs(xy(:,1)); endfunction

u = BVP2Dsym(FEMmesh,1,0,0,'BC',0,0);
difference = zeros(5,1); area = difference;
for ii = 1:5
    [~,grad] = FEMEvaluateGP(FEMmesh,u);
    coeff = sqrt(1+grad(:,1).^2+ grad(:,2).^2);
    area(ii) = FEMIntegrate(FEMmesh,coeff);
    u_new = BVP2Dsym(FEMmesh,coeff,0,0,'BC',0,0);
    difference(ii) = mean(abs(u_new-u));
    u = u_new;
endfor

Area_Difference = [area,difference]
figure(1); FEMtrisurf(FEMmesh,x,y,u)
    xlabel('x'); ylabel('y'); zlabel('z')
-->
Area_Difference =     2.30454229746     0.00271116350
                    2.30609424101     0.00030136719
                    2.30586894444     0.00003705316
                    2.30589632291     0.00000508928
                    2.30589260378     0.00000078521

```

By choosing quadratic or cubic elements, or a finer mesh, one can observe that the computed minimal area will be smaller. This should not come as a surprise, the better the resolution, the smaller the minimal area.

## 9.8 Computing a capacitance

### 9.8.1 State the problem

Examine a circular plate capacitance as shown in Figure 86. Based on the radial symmetry one should be able to consider a two dimensional section only for the computations.

Consider the voltage  $u$  as unknown. On the upper conductor assume  $u = 1$  and on the lower conductor  $u = -1$ . Based on the symmetry consider a section only and use  $u = 0$  in the plane centered between the conductors. Use the Laplace operator in cylindrical coordinates. Thus the following boundary value problem has

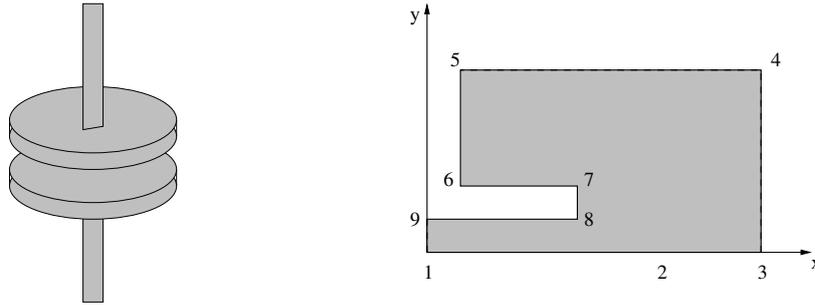


Figure 86: The capacitance and the section used for the modeling

to be solved.

$$\begin{aligned}
 \operatorname{div}(x \operatorname{grad} u(x, y)) &= 0 && \text{in domain} \\
 u(x, 0) &= 0 && \text{along edge } y = 0 \\
 u(x, y) &= 1 && \text{along edges of upper conductor} \\
 \frac{\partial}{\partial n} u(x, y) &= 0 && \text{on remaining boundary}
 \end{aligned} \tag{82}$$

Assume that the domain is embedded in the rectangle  $0 \leq x \leq R$  and  $0 \leq y \leq H$ . The lower edge of the conductor is at  $y = h$  and  $0 \leq x \leq r$ . If  $h \ll r$  expect the gradient of  $u$  to be  $1/h$  between the plates and zero away from the plates. Thus

$$\text{flux} = \iint_{\text{disk}} \vec{n} \cdot \operatorname{grad} u \, dA = 2\pi \int_0^R x \frac{\partial u}{\partial y} \, dx \approx 2\pi \int_0^r x \frac{1}{h} \, dx = \frac{\pi r^2}{h}.$$

Because the electric field will not be homogeneous around the boundaries of the disk expect deviations from the result of an idealized circular disk. With the divergence theorem and a physical argument one can verify that the flux through the midplane is proportional to the capacitance. By applying the following steps compute the capacitance by analyzing the solution of a boundary value problem.

1. Create a mesh for the domain in question.
2. Define parameters and boundary conditions.
3. Solve the partial differential equation and visualize the solution.
4. Compute the flux through the midplane as an integral to determine the capacitance.

### 9.8.2 Create the mesh and solve the BVP

According to Figure 86 create a mesh with the following data.

$h = 0.2$	distance between midplane and lower edge of capacitance
$r = 1.0$	radius of disk of the capacitance
$H = 0.5$	height of the enclosing rectangle
$R = 2.5$	radius of the enclosing rectangle

As input for the mesh generating code `triangle` (see [[www:triangle](http://www.triangle.org)]) use

- the coordinates of the corner points, numbered according to Figure 86

- a list of all the connecting edges and the type of boundary conditions to be used
- information of the desired area of the triangles to be generated

Then use two different sizes of the triangles since a finer mesh between the plates is required, expecting large variations in the solution. The file `capacitance.poly` provides this information. The numbering of the nodes is visible in Figure 86. With the above use the program `triangle` to generate a mesh.

```
triangle -pqa capacitance.poly
```

The mesh consists of 2189 nodes, forming 4036 triangles.

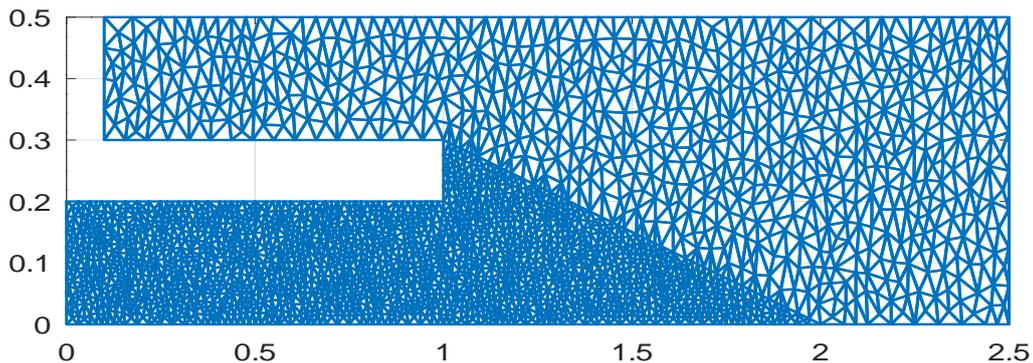


Figure 87: A mesh on the domain

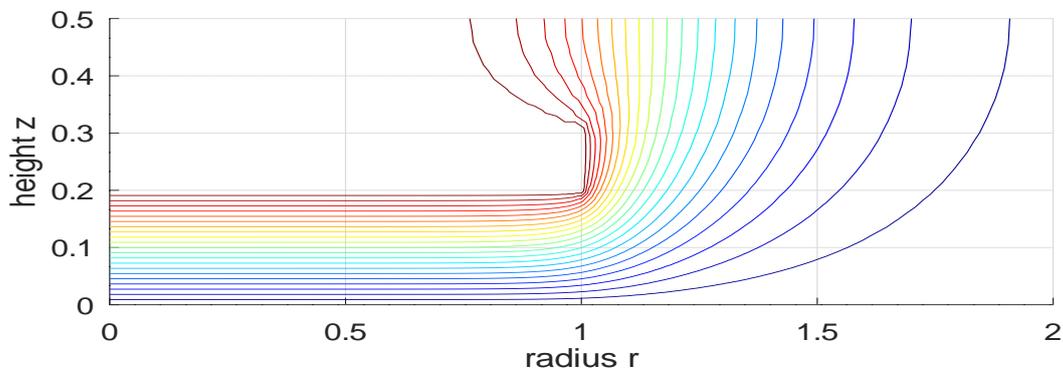


Figure 88: The contour lines of the resulting voltage

To solve the BVP (82) one needs a definition of the coefficient function and the Dirichlet boundary function. Then set up and solve the system of linear equations. This leads to a system for 1937 unknowns. Now generate a plot of the voltage  $u(x, y)$  and its level curves. Find the results in Figure 89.

### 9.8.3 Compute the capacitance

It remains to compute the flux through the midplane. For this start out by computing the gradient of the voltage  $u$  along the line  $y = 0$ . Find the plot of the normal component in Figure 89. The graph confirms that between the

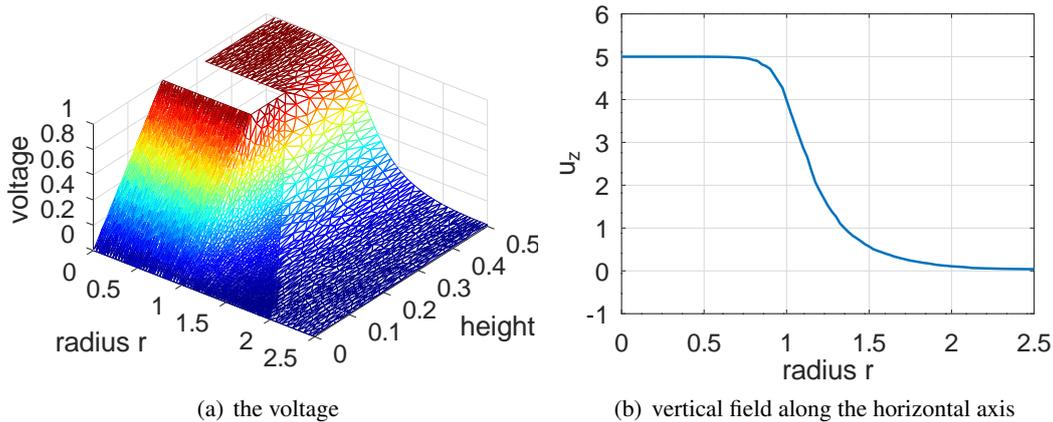


Figure 89: Voltage plot and electric field between the plates of the capacitance

plates the gradient is approximately  $1/h = 1/0.2 = 5$  and vanishes away from the plate. Then a trapezoidal rule is used to determine the flux across the midplane with the integral.

$$\text{flux} = \iint_{\text{disk}} \vec{n} \cdot \text{grad } u \, dA = 2\pi \int_0^R x \frac{\partial u}{\partial y} \, dx$$

For the selected values of  $h$ ,  $H$ ,  $r$  and  $R$  obtain a factor of 1.5 between result of the boundary value problem and the idealized approximation  $\pi r^2/h$ . Thus the simple formula is not a good approximation, the distance  $h$  is too large compared to the radius  $r$ .

#### Capacitance.m

```
FEMmesh = ReadMeshTriangle('capacitance.1');
%% FEMmesh = MeshUpgrade(FEMmesh, 'quadratic'); %% uncomment for a quadratic mesh
figure(1); FEMtrimesh(FEMmesh) %% display the generated mesh

function res = a(xy,dummy) res = xy(:,1); endfunction
function res = Volt(xy,dummy) res = xy(:,2)>0.1; endfunction

u = BVP2Dsym(FEMmesh, 'a', 0, 0, 'Volt', 0, 0);
figure(2); FEMtrimesh(FEMmesh,u);
view([38,48]); xlabel('radius r'); ylabel('height z'); zlabel('voltage')
figure(3); FEMtricontour(FEMmesh,u,21);
xlabel('radius r'); ylabel('height z');

[ux,uy] = FEMEvaluateGradient(FEMmesh,u);
xi = linspace(0,2.5,101)'; yi = zeros(101,1);
uy_i = FEMgriddata(FEMmesh,uy,xi,yi);
figure(4); plot(xi,uy_i)
xlabel('radius r'); ylabel('u_z'); ylim([-1,6])
Integral = [2*pi*trapz(xi,xi.*uy_i), pi*1^2/0.2]
-->
Integral = 23.782 15.708
```

## 9.9 Torsion of beams, Prandtl stress function

Examine the torsion of a shaft with constant cross section. Based on a few assumptions determine the deformation of the shaft under torsion. The problem is presented in [VarFEM] and more detailed in [Sout73, §12].

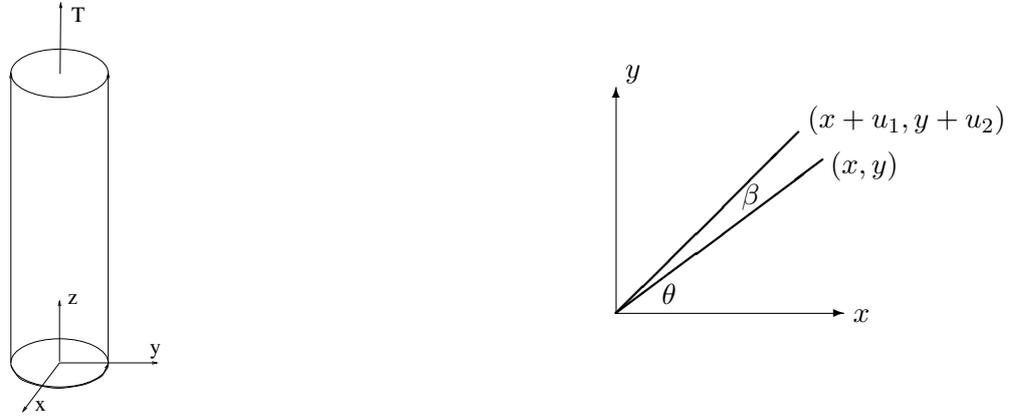


Figure 90: Torsion of a shaft

### 9.9.1 The setup with the warp function and the Prandtl stress function

Consider a vertical shaft with constant cross section. The centers of gravity of the cross section are along the  $z$  axis and the bottom of the shaft is fixed. The top surface is twisted by a total torque  $T$ . The situation of a circular cross section is shown in Figure 90. There is no exact specification of the forces and twisting moments applied to the two ends. Based on Saint-Venant principle (see [Sout73, §5.6]) assume that the stress distribution in the cross sections does not depend on  $z$ , except very close to the two ends. The twisting leads to a rotation of each cross section by an angle  $\beta$  where  $\beta = z \cdot \alpha$ . The constant  $\alpha$  is a measure of the change of angle per unit length of the shaft. Its value  $\alpha$  has to be determined, using the moment  $T$ . Based on this determine the horizontal displacements for small angles  $\beta$  by the right part of Figure 90 and a linear approximation

$$\begin{aligned} u_1(x, y) &= r \cos(\beta + \theta) - r \cos(\theta) \approx -\beta r \sin \theta = -y \beta = -y z \alpha \\ u_2(x, y) &= r \sin(\beta + \theta) - r \sin(\theta) \approx +\beta r \cos \theta = +x \beta = +x z \alpha \end{aligned}$$

It is assumed that the vertical displacement is independent of  $z$  and given by a warping function  $\phi(x, y)$ . This leads to the displacements

$$u_1 = -y z \alpha \quad , \quad u_2 = x z \alpha \quad , \quad u_3 = \alpha \phi(x, y)$$

and thus the strain components

$$\varepsilon_{xx} = \varepsilon_{yy} = \varepsilon_{zz} = \varepsilon_{xy} = 0 \quad , \quad \varepsilon_{xz} = -\frac{1}{2} \alpha y + \frac{1}{2} \alpha \frac{\partial \phi}{\partial x} \quad , \quad \varepsilon_{yz} = \frac{1}{2} \alpha x + \frac{1}{2} \alpha \frac{\partial \phi}{\partial y} .$$

Using Hooke's law find the stress components

$$\sigma_x = \sigma_y = \sigma_z = \tau_{xy} = 0 \quad , \quad \tau_{xz} = \frac{E \alpha}{2(1 + \nu)} \left( -y + \frac{\partial \phi}{\partial x} \right) \quad , \quad \tau_{yz} = \frac{E \alpha}{2(1 + \nu)} \left( x + \frac{\partial \phi}{\partial y} \right) .$$

The problem is neither plane stress ( $\tau_{xz} \neq 0, \tau_{yz} \neq 0$ ) nor plane strain ( $\phi \neq 0$ ). Using the stresses determine the horizontal forces and the torsion along a hypothetical horizontal cross section. Since the origin is the center of gravity of the cross section  $\Omega$  the first moments vanish and

$$\begin{aligned} T &= \iint_{\Omega} x \tau_{yz} - y \tau_{xz} dA = \frac{E \alpha}{2(1+\nu)} \iint_{\Omega} x \left(x + \frac{\partial \phi}{\partial y}\right) - y \left(-y + \frac{\partial \phi}{\partial x}\right) dA \\ &= \frac{E \alpha}{2(1+\nu)} \iint_{\Omega} x^2 + y^2 + x \frac{\partial \phi}{\partial y} - y \frac{\partial \phi}{\partial x} dA = \frac{E \alpha}{1+\nu} J. \end{aligned}$$

Using the **torsional rigidity**  $J$  with

$$J = \iint_{\Omega} x^2 + y^2 + x \frac{\partial \phi}{\partial y} - y \frac{\partial \phi}{\partial x} dA$$

determine the constant  $\alpha$  by

$$\alpha = \frac{2(1+\nu)}{J E} T$$

and thus for a shaft of height  $H$  the total change of angle  $\beta$  as

$$\beta = H \cdot \alpha = \frac{2(1+\nu)}{J E} H \cdot T.$$

The only difficult part is to determine the function  $\phi$ , then  $J$  is determined by an integration.

The above computations allow to compute the energy  $E$  in one cross section  $\Omega$  by

$$\begin{aligned} E &= \iint_{\Omega} \sigma_{xz} \tau_{xz} + \sigma_{yz} \tau_{yz} dA = \frac{E \alpha^2}{4(1+\nu)} \iint_{\Omega} \left(-y + \frac{\partial \phi}{\partial x}\right)^2 + \left(x + \frac{\partial \phi}{\partial y}\right)^2 dA \\ &= \frac{E \alpha^2}{4(1+\nu)} \iint_{\Omega} \left(\frac{\partial \phi}{\partial x}\right)^2 + \left(\frac{\partial \phi}{\partial y}\right)^2 - 2y \frac{\partial \phi}{\partial x} + 2x \frac{\partial \phi}{\partial y} + x^2 + y^2 dA. \end{aligned}$$

The warp function  $\phi$  has to minimize this expression. Using calculus of variations (e.g. [VarFEM]) one can show that  $\phi$  has to solve the boundary value problem

$$\begin{aligned} \operatorname{div}(\nabla \phi) = \Delta \phi &= 0 && \text{in the cross section } \Omega \\ \vec{n} \cdot \nabla \phi &= \begin{pmatrix} y \\ -x \end{pmatrix} \cdot \vec{n} && \text{on the boundary } \partial \Omega \end{aligned} \quad (83)$$

Since the stress components are given by

$$\sigma_x = \sigma_y = \sigma_z = \tau_{xy} = 0, \quad \tau_{xz} = \frac{E \alpha}{2(1+\nu)} \left(-y + \frac{\partial \phi}{\partial x}\right), \quad \tau_{yz} = \frac{E \alpha}{2(1+\nu)} \left(x + \frac{\partial \phi}{\partial y}\right)$$

the boundary condition can be written as

$$\begin{pmatrix} \tau_{xz} \\ \tau_{yz} \end{pmatrix} \cdot \vec{n} = 0.$$

This equation implies that there is no stress on the lateral surface of the shaft. This condition is consistent with the mechanical setup.

The Prandtl stress funktion  $\chi$  is characterized by

$$\frac{\partial \chi}{\partial y} = -y + \frac{\partial \phi}{\partial x} = \frac{2(1+\nu)}{E\alpha} \tau_{xz} \quad \text{and} \quad -\frac{\partial \chi}{\partial x} = x + \frac{\partial \phi}{\partial y} = \frac{2(1+\nu)}{E\alpha} \tau_{yz}.$$

By differentiating the above equations by  $y$  (resp.  $x$ ) and subtracting and using  $\frac{\partial}{\partial x} \frac{\partial \phi}{\partial y} = \frac{\partial}{\partial y} \frac{\partial \phi}{\partial x}$  find

$$\Delta \chi = \frac{\partial^2 \chi}{\partial x^2} + \frac{\partial^2 \chi}{\partial y^2} = -2.$$

To determine the boundary conditions for  $\chi$  assume that there are no external forces on the boundary.

$$\begin{pmatrix} \tau_{xz} \\ \tau_{yz} \end{pmatrix} \cdot \vec{n} = 0 \quad \implies \quad \begin{pmatrix} \frac{\partial \chi}{\partial y} \\ -\frac{\partial \chi}{\partial x} \end{pmatrix} \cdot \vec{n} = \nabla \chi \cdot \vec{t} = 0,$$

where  $\vec{t}$  is a tangential vector of the boundary curve. Assuming that there are no holes<sup>29</sup>, this implies that one can work with  $\chi = 0$  on the boundary  $\Gamma$ . Thus the Prandtl stress function is a solution of the boundary value problem

$$\begin{aligned} -\Delta \chi &= 2 & \text{in } \Omega \\ \chi &= 0 & \text{on } \Gamma \end{aligned} \quad (84)$$

The torsional rigidity is determined by

$$J = \iint_{\Omega} x^2 + y^2 + x \left( -\frac{\partial \chi}{\partial x} - x \right) - y \left( +\frac{\partial \chi}{\partial y} + y \right) dA = - \iint_{\Omega} x \frac{\partial \chi}{\partial x} + y \frac{\partial \chi}{\partial y} dA.$$

For ductile materials the von Mises stress indicates the possible fractures in the material. In this case it is given by

$$\sigma_{vM} = \sqrt{\frac{3}{2} (\tau_{xz}^2 + \tau_{yz}^2)} = \frac{E\alpha}{2(1+\nu)} \sqrt{\frac{3}{2} \left[ \left( \frac{\partial \chi}{\partial x} \right)^2 + \left( \frac{\partial \chi}{\partial y} \right)^2 \right]} = \frac{E\alpha}{2(1+\nu)} \sqrt{\frac{3}{2}} \|\nabla \chi\|.$$

### 9.9.2 On a disk with radius $R$

On a disk with radius  $R$  the solution is given by  $\chi(x, y) = \frac{1}{2} (R^2 - x^2 - y^2)$ . Thus the nonzero stresses are

$$\tau_{xz} = +\frac{E\alpha}{2(1+\nu)} \frac{\partial \chi}{\partial y} = -\frac{E\alpha}{2(1+\nu)} y \quad \text{and} \quad \tau_{yz} = -\frac{E\alpha}{2(1+\nu)} \frac{\partial \chi}{\partial x} = +\frac{E\alpha}{2(1+\nu)} x.$$

The BVP (83) for the warp function  $\phi$  is

$$\begin{aligned} \operatorname{div}(\nabla \phi) = \Delta \phi &= 0 & \text{in the cross section } \Omega \\ \vec{n} \cdot \nabla \phi &= \frac{1}{\sqrt{x^2+y^2}} \begin{pmatrix} y \\ -x \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} = 0 & \text{on the boundary } \partial \Omega \end{aligned}$$

with the unique solution  $\phi(x, y) = 0$ , i.e. no warping. The torsional rigidity is given by

$$J = \iint_{\Omega} x^2 + y^2 dA = 2\pi \int_0^R r^2 r dr = \frac{\pi}{2} R^4$$

and the von Mises stress is given by

$$\sigma_{vM} = \frac{E\alpha}{2(1+\nu)} \sqrt{\frac{3}{2} \left[ \left( \frac{\partial \chi}{\partial x} \right)^2 + \left( \frac{\partial \chi}{\partial y} \right)^2 \right]} = \frac{E\alpha}{2(1+\nu)} \sqrt{\frac{3}{2}} \sqrt{x^2 + y^2} = \frac{E\alpha}{2(1+\nu)} \sqrt{\frac{3}{2}} r.$$

<sup>29</sup>This restriction can be removed.

### 9.9.3 On a square

To examine the stiffness of a square cross section with a circular cross section examine a square with the same area as a circle with radius  $R = 1$ . Thus the length of a side is  $\sqrt{\pi} \approx 1.77$ . The code below solves the boundary value problem (84) and then computes the torsional rigidity by integrating

$$J = - \iint_{\Omega} x \frac{\partial \chi}{\partial x} + y \frac{\partial \chi}{\partial y} dA.$$

The numerical result of  $J \approx 1.39$  has to be compared to the result of  $J = \frac{\pi}{2} \approx 1.57$  for the disk with Radius 1. Thus the square cross section leads to less torsional rigidity. Then examine the von Mises stress by plotting

$$f(x, y) = \sqrt{\left(\frac{\partial \chi}{\partial x}\right)^2 + \left(\frac{\partial \chi}{\partial y}\right)^2} = \|\nabla \chi\|.$$

Find the result in Figure 91(a). The maximal value of  $\approx 1.20$  is larger than the maximal value 1 for the disk. Thus for the same twisting angle the square is exposed to a larger von Mises stress.

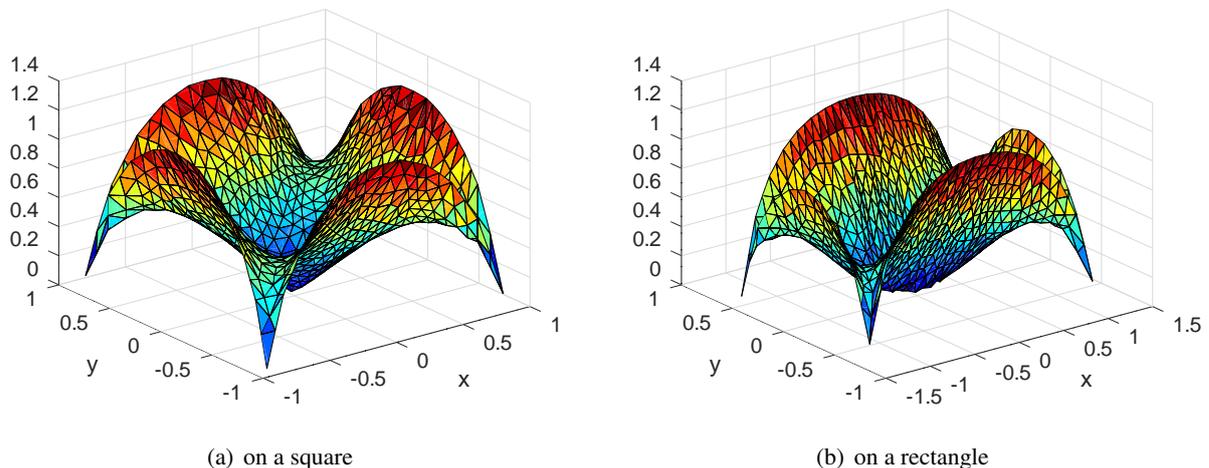


Figure 91: The von Mises stress caused by torsion of a bar with square or rectangular cross section

#### TorsionSquare.m

```

N = 10;
l = sqrt(pi)/2; al = 1; %% al = sqrt(2); % use this for the rectangle
Mesh = CreateMeshTriangle('Torsion',...
[-al*l -1/al*l -1; al*l -1/al*l -1; al*l 1/al*l -1; -al*l 1/al*l -1],pi/2/N^2);
Mesh = MeshUpgrade(Mesh,'quadratic');

chi = BVP2Dsym(Mesh,1,0,2,0,0,0);

[chiGP,gradChi] = FEMEvaluateGP(Mesh,chi);
xGP = Mesh.GP(:,1); yGP = Mesh.GP(:,2);
f = xGP.*gradChi(:,1) + yGP.*gradChi(:,2);
J = FEMIntegrate(Mesh,-f)

[chi_x,chi_y] = FEMEvaluateGradient(Mesh,chi);

```

```
figure(1); FEMtrisurf(Mesh,sqrt(chi_x.^2 + chi_y.^2))
            xlabel('x'); ylabel('y');
-->
J = 1.3873
```

### 9.9.4 On a rectangle

The above can be repeated for a rectangle with the same area but a ratio of 2 for the length of the sides. The value of  $J \approx 1.13$  indicates that the rectangle is even softer and the maximal von Mises stress of  $\approx 1.16$  is slightly smaller than for the square cross section.

### 9.10 Dynamic heat conduction problems

The dynamic heat equation with a thermal conductivity  $a(x, y)$  is of the form given in equation (4). For the simplified case with no external heating, no convection and the boundary either insulated or at a given temperature arrive at the initial boundary value problem

$$\begin{aligned} \frac{\partial}{\partial t} u - \nabla \cdot (a \nabla u) &= 0 & \text{for } (x, y, t) \in \Omega \times (0, T] \\ u &= g & \text{for } (x, y, t) \in \Gamma_1 \times (0, T] \\ \vec{n} \cdot (a \nabla u) &= 0 & \text{for } (x, y, t) \in \Gamma_2 \times (0, T] \\ u &= u_0 & \text{on } \Omega \text{ at } t = 0 \end{aligned} \quad (85)$$

In Figure 92 the upper half of the domain is shown, at the lower edge the symmetry constraint  $\frac{\partial}{\partial n} u = 0$  is used. Assume insulation on all of the boundary, except the left edge  $\Gamma_1$  at  $x = 0$ , where the temperature equals 1. As initial temperature we use  $u_0(x, y) = 0$  and observe how the domain is warming up as time advances.

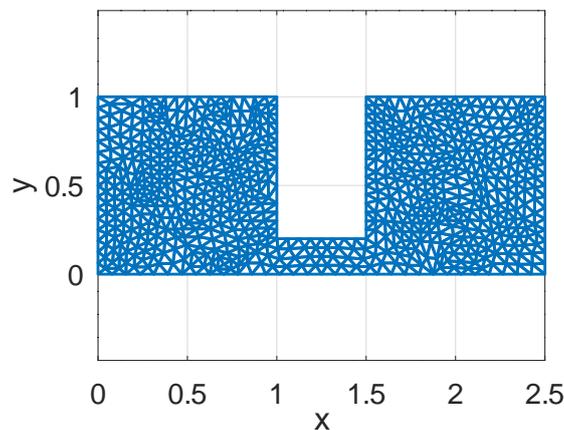


Figure 92: The mesh for a dynamic heat problem

#### 9.10.1 With a narrow section in the domain

The first case to be examined uses a narrow section between two bigger sections. The dimension of the narrow section can be changed by modifying the parameters  $h = 0.2$  and  $l = 0.5$ .

- In Figure 93 observed the delayed heating of the section on the right.

- In Figure 94 the temperature along the edge  $y = 0$  for  $0 \leq x \leq 2.5$  and  $0 \leq t \leq 10$  is shown, as surface and contour lines.
- In Figure 95 the temperature at the corner  $(x, y) = (2.5, 0)$  is shown as function of time.

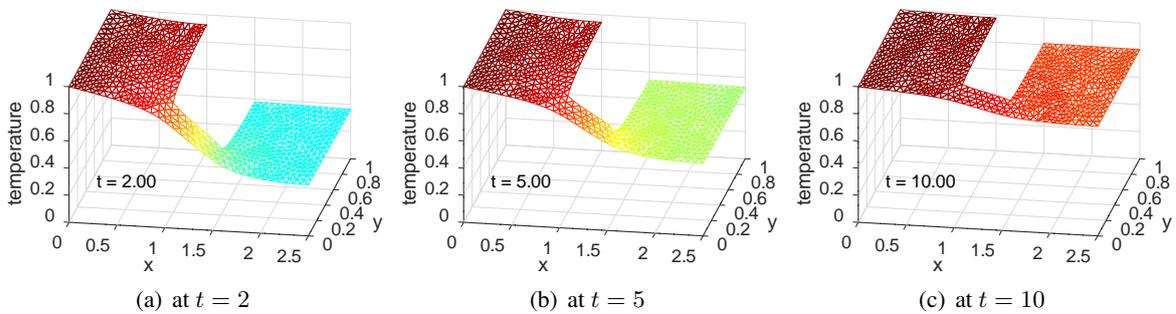
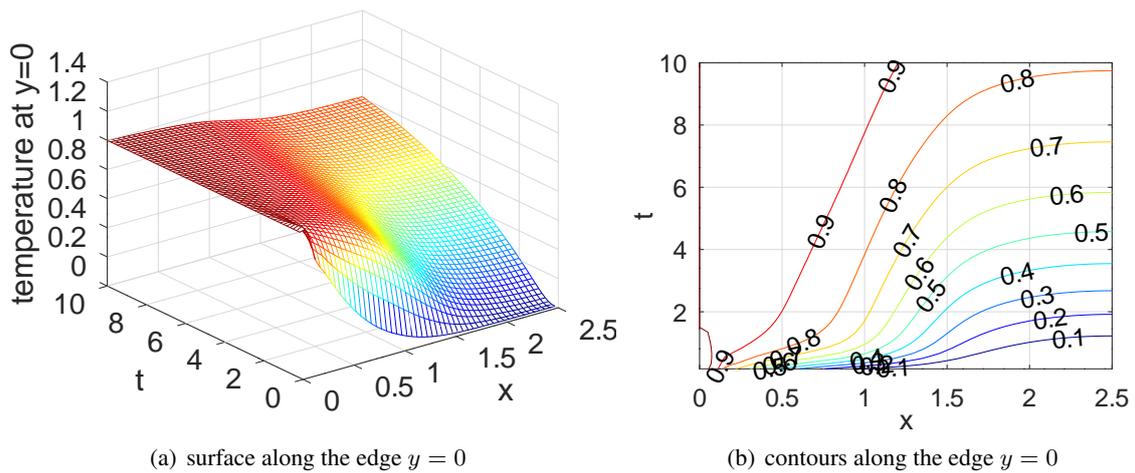


Figure 93: The evolution of the temperature surface at different times

Figure 94: The temperature surface at different times along  $y = 0$ 

### HeatDynamic.m

```

%% parameters
h = 0.2; l = 0.5; Nt = 60; %% number of time steps
FEMmesh = CreateMeshTriangle('Test',...
    [0 0,-2; 2+1 0 -2; 2+1 1,-2; 1+1 1 -2; 1+1 h -2; 1 h -2; 1 1 -2; 0 1 -1],0.01);
FEMmesh = MeshUpgrade(FEMmesh,'quadratic');

figure(1); FEMtrimesh(FEMmesh);
    axis equal; xlabel('x'); ylabel('y')

[u t] = IBVP2D(FEMmesh,1,1,0, 0, 0, 0, 0,1, 0, 0, 0, 0, 10, [Nt,10]);

figure(2); FEMtrimesh(FEMmesh,u(:,end))
    xlabel('x'); ylabel('y'); zlim([0,1]); view([10 30]); caxis([0,1]);

```

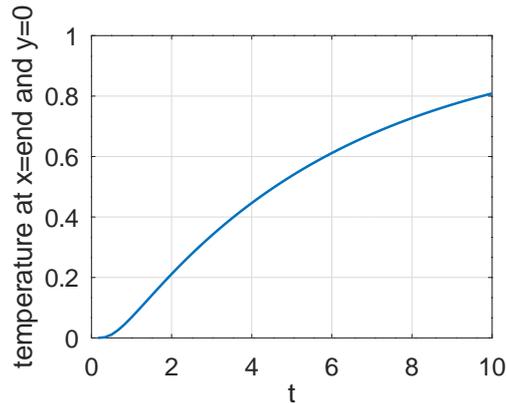


Figure 95: The temperature as function of time at the endpoint (2.5, 0)

```

text(0.2,0.2,0.2,sprintf('t = %4.2f',t(end))); zlabel('temperature')
figure(3); FEMtrimesh(FEMmesh,u(:,Nt/2+1))
xlabel('x'); ylabel('y'); zlim([0,1]); view([10 30]); caxis([0,1])
text(0.2,0.2,0.2,sprintf('t = %4.2f',t(Nt/2+1))); zlabel('temperature')
figure(4); FEMtrimesh(FEMmesh,u(:,Nt/3+1))
xlabel('x'); ylabel('y'); zlim([0,1]); view([10 30]); caxis([0,1])
text(0.2,0.2,0.2,sprintf('t = %4.2f',t(Nt/5+1))); zlabel('temperature')

x = linspace(0,2+1,51); u_int = zeros(size(t,2)-1,size(x,2));
for jj = 2:size(t,2)
    u_int(jj-1,:) = FEMgriddata(FEMmesh,u(:,jj),x,zeros(size(x)));
endfor

figure(10); mesh(x,t(2:end),u_int)
    xlabel('x'); ylabel('t'); zlabel('temperature at y=0')
figure(11); [c,h] = contour(x,t(2:end),u_int,[0:0.1:1]);
    clabel(c,h);
    xlabel('x'); ylabel('t');
figure(12); plot(t(2:end),u_int(:,end))
    xlabel('t'); ylabel('temperature at x=end and y=0')

```

### 9.10.2 With a section with lower thermal conductivity

On the modified domain visible in Figure 96 in the middle section the conductivity is considerably smaller than in the two side section, i.e.

$$a(x, y) = \begin{cases} 1 & \text{for } 0 \leq x \leq 1 \text{ and } x \geq 1.5 \\ \frac{1}{6} & \text{for } 1 < x < 1.5 \end{cases}$$

- In Figure 97 observed the delayed heating of the section on the right.
- In Figure 98 the temperature along the edge  $y = 0$  for  $0 \leq x \leq 2.5$  and  $0 \leq t \leq 10$  is shown, as surface and contour lines.
- In Figure 99 the temperature at the corner  $(x, y) = (2.5, 0)$  is shown as function of time.

Observe the similar, but not identical, behavior of the two cases examined.

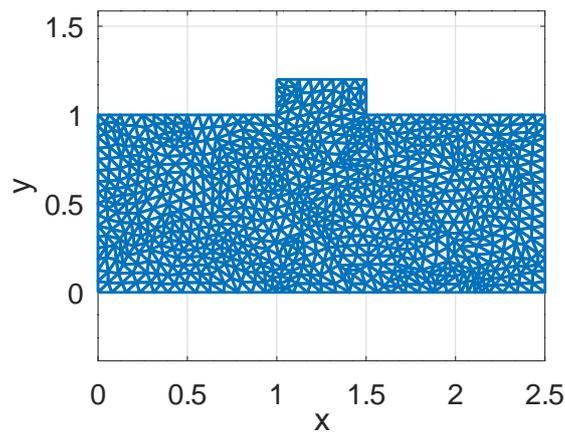


Figure 96: The mesh for a dynamic heat problem

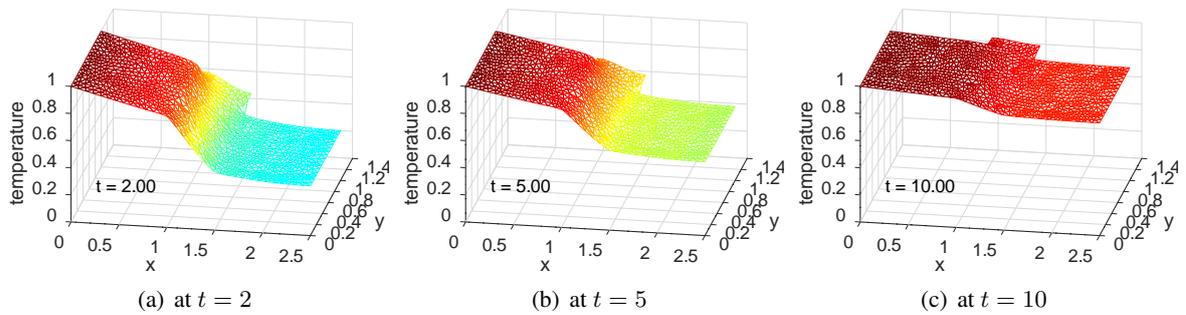


Figure 97: The evolution of the temperature surface at different times

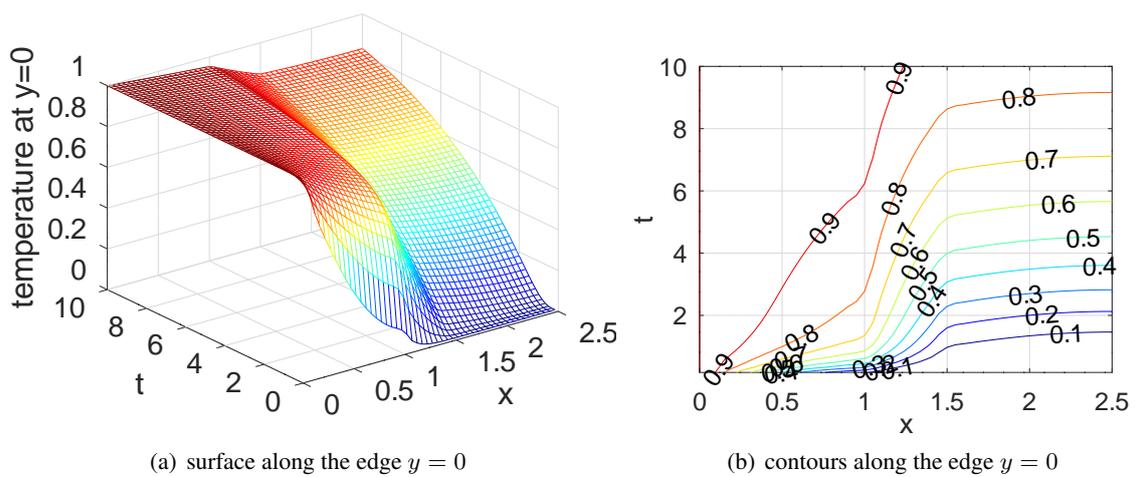
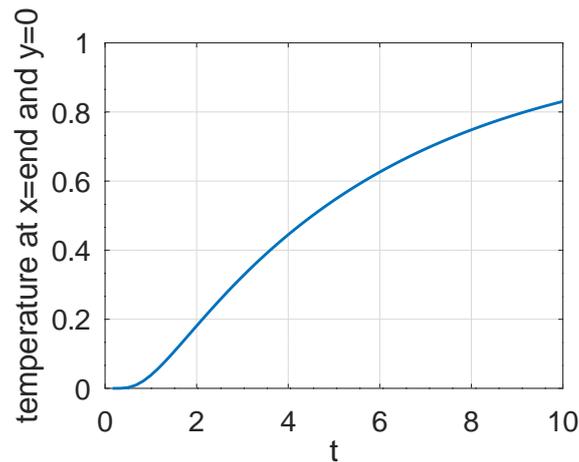


Figure 98: The temperature surface at different times along  $y = 0$

Figure 99: The temperature as function of time at the endpoint  $(2.5, 0)$ 

### HeatDynamicCoefficient.m

```

%% parameters
h = 1.2; l = 0.5; Nt = 60; %% number of time steps
FEMmesh = CreateMeshTriangle('Test', ...
    [0 0,-2; 2+1 0 -2; 2+1 1, -2; 1+1 1 -2; 1+1 h -2; 1 h -2; 1 1 -2; 0 1 -1],0.01);
FEMmesh = MeshUpgrade(FEMmesh,'quadratic');
%%FEMmesh = MeshUpgrade(FEMmesh,'cubic');

figure(1); FEMtrimesh(FEMmesh);
    axis equal; xlabel('x'); ylabel('y')

function res = a(xy,dummy)
    l = 0.5;
    res = ones(size(xy,1),1);
    res(find(abs(xy(:,1)-l-l/2)<l/2)) *= 1/6;
endfunction

[u t] = IBVP2D(FEMmesh,1,'a',0, 0, 0, 0,1, 0, 0, 0, 0, 10, [Nt,10]);

figure(2); FEMtrimesh(FEMmesh,u(:,end))
    xlabel('x'); ylabel('y'); zlim([0,1]); view([10 30]); caxis([0,1]);
    text(0.2,0.2,0.2,sprintf('t = %4.2f',t(end))); zlabel('temperature')
figure(3); FEMtrimesh(FEMmesh,u(:,Nt/2+1))
    xlabel('x'); ylabel('y'); zlim([0,1]); view([10 30]); caxis([0,1]);
    text(0.2,0.2,0.2,sprintf('t = %4.2f',t(Nt/2+1))); zlabel('temperature')
figure(4); FEMtrimesh(FEMmesh,u(:,Nt/3+1))
    xlabel('x'); ylabel('y'); zlim([0,1]); view([10 30]); caxis([0,1]);
    text(0.2,0.2,0.2,sprintf('t = %4.2f',t(Nt/5+1))); zlabel('temperature')

x = linspace(0,2+1,51); u_int = zeros(size(t,2)-1,size(x,2));
for jj = 2:size(t,2)
    u_int(jj-1,:) = FEMgriddata(FEMmesh,u(:,jj),x,zeros(size(x)));
endfor

figure(10); mesh(x,t(2:end),u_int)
    xlabel('x'); ylabel('t'); zlabel('temperature at y=0')

```

```

figure(11); [c,h] = contour(x,t(2:end),u_int,[0:0.1:1]);
            clabel(c,h);
            xlabel('x'); ylabel('t');

figure(12); plot(t(2:end),u_int(:,end))
            xlabel('t'); ylabel('temperature at x=end and y=0')

```

### 9.10.3 Cooling of a cylinder

Examine a cylinder with elliptical cross section and an initial temperature distribution  $u_0(x, y)$ , independent on  $z$ . The boundary temperature is fixed at 0. The domain and the initial temperature profile are visible in Figure 100. The selected, nonsymmetric initial temperature is

$$u_0(x, y) = \exp(-(x - 0.5)^2 - 2y^2) \cdot (4 - x^2 - y^2).$$

The initial boundary value problem is solved for times  $0 \leq t \leq 2$ . A few snapshots are visible in Figure 101. By looking at different time slices an animation can be generated.

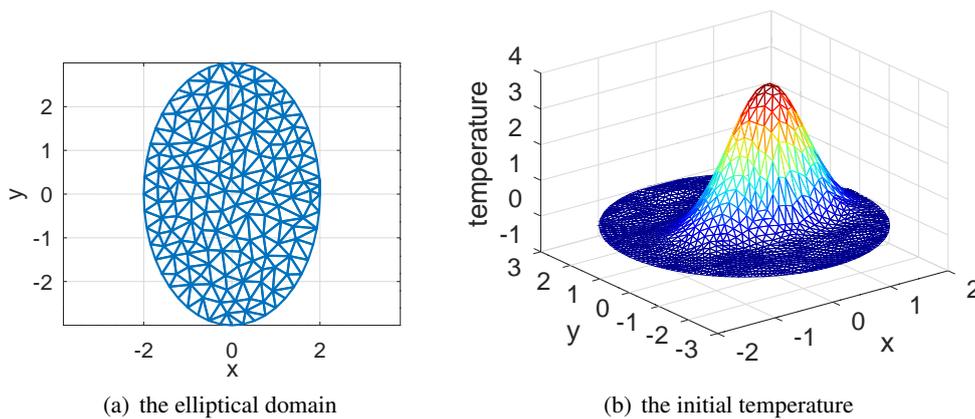


Figure 100: The domain and initial temperature

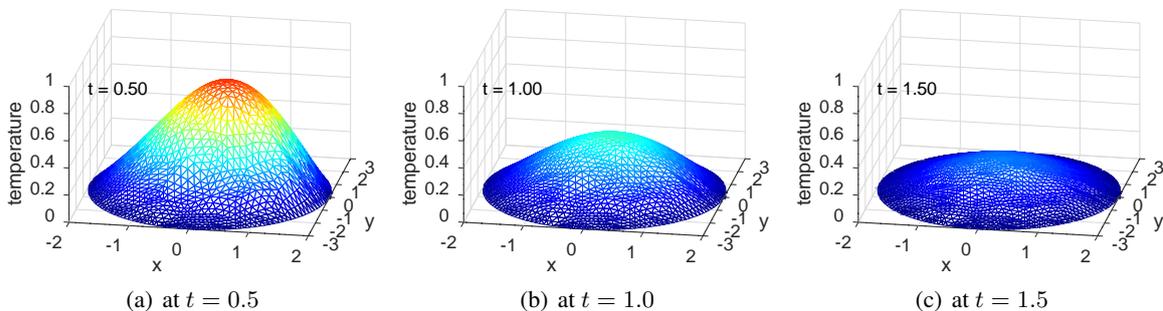


Figure 101: The temperature at different times

---

**CylinderCooling.m**


---

```

R = 2; N = 50; alpha = linspace(0,2*pi*N/(N-1),N)';
Tend = 2; Nt = 60; %% number of shown time steps
FEMmesh = CreateMeshTriangle('circle',[R*cos(alpha),1.5*R*sin(alpha),...
    -ones(size(alpha))],0.1);

figure(1); FEMtrimesh(FEMmesh)
    xlabel('x') ; ylabel('y'); axis equal
FEMmesh = MeshUpgrade(FEMmesh,'cubic');

function res = u_init(xy)
    x = xy(:,1); y = xy(:,2);
    res = exp(-(x-0.5).^2-2*y.^2) .* (2^2 -x.^2-y.^2);
endfunction

figure(2); FEMtrimesh(FEMmesh,u_init(FEMmesh.nodes))
    xlabel('x'); ylabel('y'); zlabel('temperature');

[u,t] = IBVP2Dsym(FEMmesh,1,1,0, 0, 0, 0, 0, 'u_init',0, Tend, [Nt,10]);

figure(3); FEMtrimesh(FEMmesh,u(:,Nt/4+1))
    xlabel('x'); ylabel('y'); zlim([0,1]); view([10 30]); caxis([0,1])
    text(-1.8,-2,0.9,sprintf('t = %4.2f',t(Nt/4+1))); zlabel('temperature')
figure(4); FEMtrimesh(FEMmesh,u(:,Nt/2+1))
    xlabel('x'); ylabel('y'); zlim([0,1]); view([10 30]); caxis([0,1])
    text(-1.8,-2,0.9,sprintf('t = %4.2f',t(Nt/2+1))); zlabel('temperature')
figure(5); FEMtrimesh(FEMmesh,u(:,3*Nt/4+1))
    xlabel('x'); ylabel('y'); zlim([0,1]); view([10 30]); caxis([0,1]);
    text(-1.8,-2,0.9,sprintf('t = %4.2f',t(3*Nt/4+1))); zlabel('temperature')

figure(11) %% show the animation
steps = 2;
for jj = 0:30
    FEMtrimesh(FEMmesh,u(:,jj*steps+1))
    text(-1.8,-2,0.9,sprintf('t = %4.2f',t(jj*steps+1))); zlabel('temperature')
    xlabel('x'); ylabel('y'); zlim([0,1]); view([10 30]); caxis([0,1])
    pause(0.2)
endfor

```

Obviously the temperature is decaying as time advances. To examine this behavior determine the temperatures along the center line at  $y = 0$ , as function of time. In Figure 102.

```

x = linspace(-R,R,31); u_center = zeros(length(x),length(t));
for jj = 1:length(t)
    u_center(:,jj) = FEMgriddata(FEMmesh,u(:,jj),x,zeros(size(x)));
endfor
figure(21); mesh(t,x,u_center)
    xlabel('time t'); ylabel('x'); zlabel('temperatur')
figure(22); contour(t,x,u_center,51)
    xlabel('time t'); ylabel('x');

```

The decay of the temperature at the center point  $(0,0)$  is visible in Figure 103, with linear and logarithmic scale. The exponential decay is clearly displayed in the logarithmic scale. This is consistent with the theoretical

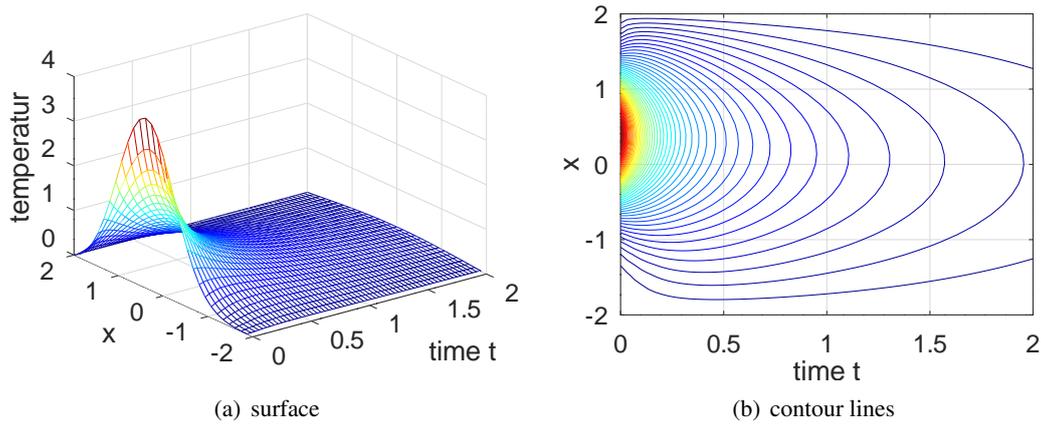


Figure 102: The temperature at different times along  $y = 0$

result

$$u(t, x, y) = \sum_{n=1}^{\infty} c_n e^{-\lambda_n t} u_n(x, y) \tag{86}$$

where  $\lambda_1 < \lambda_2 \leq \lambda_3 \leq \lambda_4 \dots$  and  $u_n(x, y)$  are the eigenvalues and eigenfunctions of the boundary value problem

$$\begin{aligned} -\nabla \cdot \nabla u_n &= \lambda_n u & \text{for } (x, y) \in \Omega \\ u &= 0 & \text{for } (x, y) \in \Gamma \end{aligned}$$

For large times  $t$  in equation (86) the first eigenvalue will dominate, i.e.

$$u(t, x, y) \approx c_1 e^{-\lambda_1 t} u_1(x, y)$$

Using the Octave command `polyfit()` with data from the right section in the logarithmic plot in Figure 103

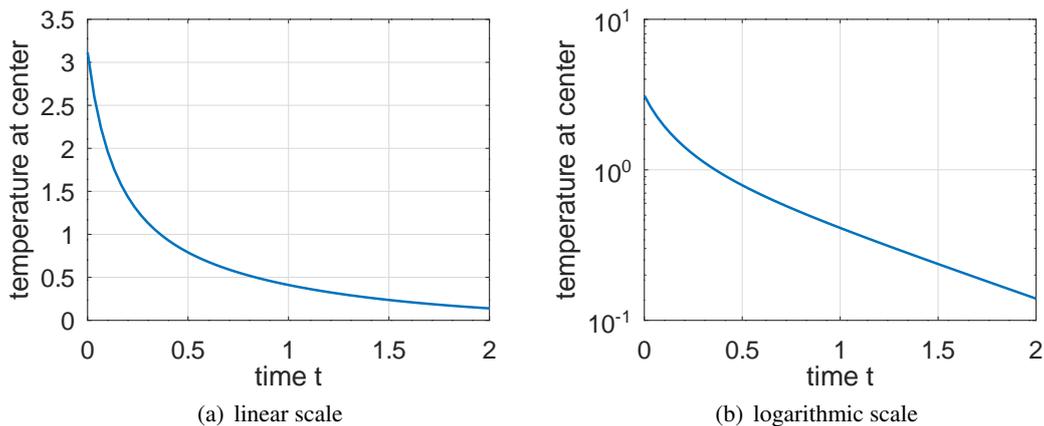


Figure 103: The temperature decay at the center  $(0, 0)$

estimate the decay by the exponential  $\exp(-1.06 t)$ . Using `BVP2Deig()` the exponent is estimated by  $\lambda_1 \approx 1.04$ , i.e. rather close to the above result by `polyfit()`, which indicates

$$\log(u(0, 0, t)) \approx 0.1556 - 1.0638 t \quad \text{or} \quad u(0, 0, t) \approx 1.1684 e^{-1.0638 t} \quad \text{for } t \text{ large.}$$

```

figure(23); plot(t,u_center(16,:))
            xlabel('time t'); ylabel('temperature at center')
figure(24); semilogy(t,u_center(16,:))
            xlabel('time t'); ylabel('temperature at center')

p = polyfit(t(40:end),log(u_center(16,40:end)),1)
EigVal = BVP2Deig(FEMmesh,1,0,1,0,3)'
-->
p      =   -1.0638    0.1556
EigVal =    1.0425    2.1314    3.1506

```

Observe that  $\lambda_2 \neq \lambda_3$ , since the domain is not circular. If the above computations are rerun on a circle of radius  $R = 2$  obtain  $\lambda_1 \approx 1.45$  and  $\lambda_2 = \lambda_3 \approx 3.68$ . The first eigenvalue  $\lambda_1 \approx 1.45$  is larger, thus the cylinder will cool down faster and the second and third eigenvalues coincide, caused by the circular symmetry of the domain.

### 9.10.4 Heat waves

In Figure 104 a domain  $\Omega \subset \mathbb{R}^2$  is visible. The heat equation (a special case of the IBVP (4)) to be solved is

$$\begin{aligned} \frac{\partial}{\partial t} u(x, y, t) - \Delta u(x, y, t) &= f(x, y, t) && \text{for } (x, y, t) \in \Omega \times (0, T] \\ \frac{\partial}{\partial n} u(x, y, t) &= 0 && \text{for } (x, y, t) \in \Gamma \times (0, T] \\ u(x, y, 0) &= 0 && \text{on } \Omega \end{aligned}$$

The function  $f(x, y, t)$  equals  $\cos(0.5 \pi t)$  for  $x \leq -0.9$  and zero otherwise. Thus there is a periodic excitation with period 5 at the very left end of the appendix for  $-1 \leq x \leq -0.9$ .

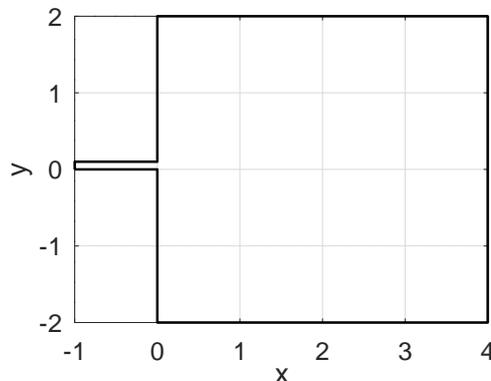


Figure 104: The domain for a heat wave propagation

The solution is generated by the command `IBVP2D()` and then evaluated along the slice at height  $y = 1$  for different values of the time  $t$ , using `FEMgriddata()`. Find the result in Figure 105.

- In Figure 105(a) the periodic behavior of the temperature is clearly visible.
- In Figure 105(b) observe the phase shift as one moves away from the heat source.

Observe that the behavior of the solution is very different from a wave equation in Section 9.11, even if the setup is comparable.

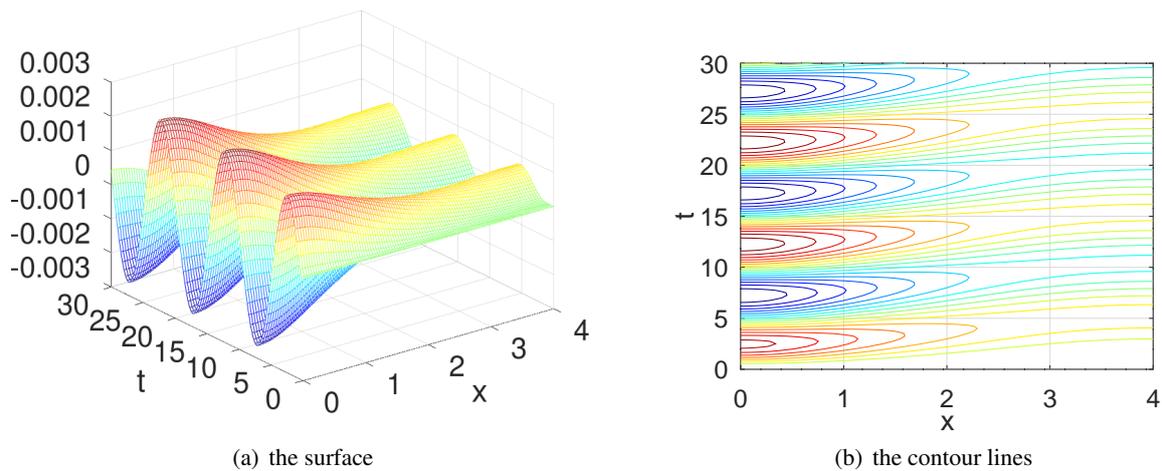


Figure 105: The propagation of a heat wave

## HeatWave.m

```

l = 1; h = 0.1; L = 4; d = 2; H = 2;
FEMmesh = CreateMeshTriangle('test',...
    [-1,0,-2;0 0 -2;0,-d,-2;L,-d,-2; L,H,-2;0,H,-2;0,h,-2;-1,h,-2],0.01);
figure(1); FEMtrimesh(FEMmesh)
    xlabel('x'); ylabel('y'); axis equal
FEMmesh = MeshUpgrade(FEMmesh,'cubic');

function res = f(xy,t)
    res = cos(0.2*pi*t)*ones(size(xy,1),1);
    res(xy(:,1)>-0.9) = 0;
endfunction

function res = u0(xy)    res = zeros(size(xy,1),1); endfunction

m = 1; a = 1; b0 = 0; bx = by = 0; f = 0; gn1 = gn2 = 0;
tic();
[u,t] = IBVP2D(FEMmesh,m,a,b0,bx,by,'f',0,gn1,gn2,'u0',0,30,[2*60,10]);
%%[u,t] = IBVP2Dsym(FEMmesh,m,a,b0,'f',0,gn1,gn2,'u0',0,30,[2*60,10]);
SolverTime = toc()

figure(2); FEMtrimesh(FEMmesh,u(:,end))
    xlabel('x'); ylabel('y'); xlim([0,L]);

umax = 0.3*max([-min(u(:)),max(u(:))]);
figure(3)
if 0 %% animation
    for jj = 1:length(t)
        FEMtrimesh(FEMmesh,u(:,jj))
        xlabel('x'); ylabel('y')
        zlim(umax*[-1 1]); caxis(0.3*umax*[-1 1]);
        text(0.8*L,0.8*H,umax,sprintf('t = %4.2f',t(jj)))
        xlim([0,L])
        pause(0.1);
    end

```

```

    endfor
else
    FEMtrimesh(FEMmesh,u(:,end))
    xlabel('x'); ylabel('y')
    zlim(umax*[-1 1]); caxis(0.3*umax*[-1 1]);
    text(0.8*L,0.8*H,umax,sprintf('t = %4.2f',t(end)))
endif

x = linspace(0,L,101); u_line = zeros(size(t,1),size(x,2));
for jj = 1:length(t)
    u_line(jj,:) = FEMgriddata(FEMmesh,u(:,jj),x,ones(size(x)));
endfor

figure(4); mesh(x,t,u_line)
    xlabel('x'); ylabel('t');
figure(5); contour(x,t,u_line,0.003*[-1:0.1:+1])
    xlabel('x'); ylabel('t');

```

### 9.10.5 Static heat equation in a ball in $\mathbb{R}^3$ , solved as a 1D problem

The dynamic heat equation in spherical coordinates  $(r, \theta, \varphi)$  is given by

$$\frac{\rho c}{k} \frac{\partial}{\partial t} u = +\Delta u + f = \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial u}{\partial r} \right) + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial u}{\partial \theta} \right) + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 u}{\partial \varphi^2} + f.$$

Examining a static problem, depending on the radius  $r$  only leads to the ordinary differential equation

$$-\frac{\partial}{\partial r} \left( r^2 \frac{\partial u}{\partial r} \right) = r^2 f(t, r).$$

At first sight FEMoctave is not set up to solve problems depending on one variable only, but one may construct solutions whose values do not depend on the second variable, i.e. transform a 1D problem artificially to a 2D problem. It has to be pointed out that this is not computationally efficient. Find the efficient 1D approach at the end of the section.

The BVP  $-\frac{\partial}{\partial r} \left( r^2 \frac{\partial u}{\partial r} \right) = 0$  with  $u(1) = 0$  and  $u(R) = 1$  is solved by<sup>30</sup>

$$u(r) = \frac{R}{R-1} \frac{r-1}{r}.$$

With FEMoctave the solution can be solved by the code below.

- Define the parameters of the system.
- Generate the mesh for  $1 \leq r \leq R = 3$  and the dummy variable  $-\frac{W}{2} \leq y \leq +\frac{W}{2}$ . At the upper and lower edge at  $y = \pm \frac{W}{2}$  apply the Neumann boundary conditions  $\frac{\partial}{\partial n} u = 0$ . This assures that the solution will depend on  $r$  only. At  $r = 1$  and  $r = R$  specify the Dirichlet conditions.
- Define the functions for the parameter  $a(r) = r^2$  and the Dirichlet conditions  $u(1) = 0$  and  $u(R) = 1$ .
- Solve the BVP with the help of `BVP2Dsym()`.
- Display the solutions. To obtain the graphs with one independent variable  $r$  use `FEMgriddata()` along the axis  $y = 0$  and  $1 \leq r \leq R = 3$ .
- In Figure 106 find the graphs of the solution  $u$ , either as function of  $r$  and  $y$ , or as function of  $r$  only.

<sup>30</sup>The ODE  $\frac{\partial}{\partial r} \left( r^2 \frac{\partial u}{\partial r} \right) = 0$  implies  $r^2 \frac{\partial u}{\partial r} = c_1$ , i.e.  $\frac{\partial}{\partial r} u = \frac{c_1}{r^2}$ . An integration leads to the general solution of the ODE  $u(r) = c_2 - \frac{c_1}{r}$ . Then use the two boundary conditions to determine  $c_1 = c_2 = \frac{R}{R-1}$  and  $u(r) = \frac{R}{R-1} \frac{r-1}{r}$ .

## HeatBall.m

```

% solve a static heat problem in a ball, using the radius r only as variable
R = 3; W = 0.1; N = 10;
FEMmesh = CreateMeshRect(linspace(1,R,N+1), [-W/2,+W/2], -2,-2,-1,-1);
%%FEMmesh = MeshUpgrade(FEMmesh,'quadratic');
%%FEMmesh = MeshUpgrade(FEMmesh,'cubic');

function res = a(r) ; res = r(:,1).^2;      endfunction
function res = gD(r); res = sign(r(:,1)-1); endfunction

u3D = BVP2Dsym(FEMmesh,'a',0,0,'gD',0,0);

figure(1); FEMtrimesh(FEMmesh,u3D)
           xlabel('r'); ylabel('dummy'); zlabel('temperature u')

r = linspace(1,R); u = FEMgriddata(FEMmesh,u3D,r,0*r);
u_exact = R/(R-1)*(r-1)./r;

figure(2); plot(r,u,r,u_exact)
           xlabel('radius r'); ylabel('temperature u')
           legend('u_{FEM}','u_{exact}','location','northwest')

figure(3); plot(r,u-u_exact)
           xlabel('radius r'); ylabel('u_{FEM}-u_{exact}')

```

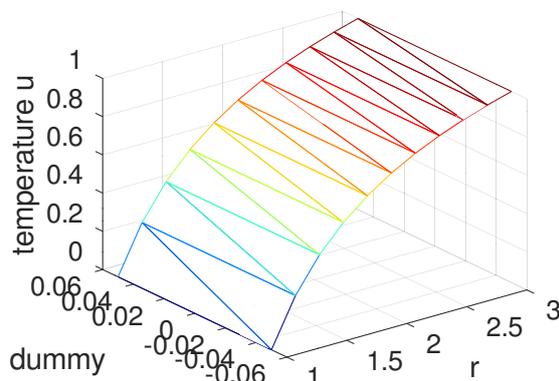
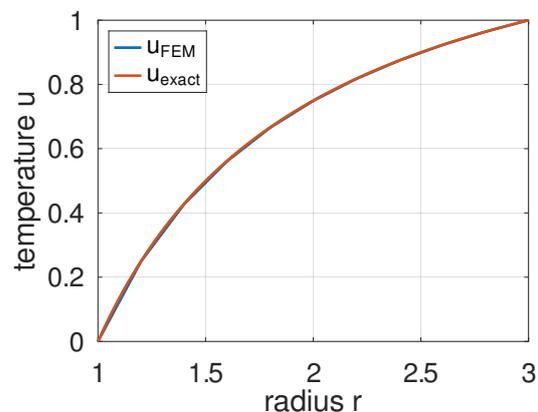
(a) as function of  $r$  and the dummy variable(b) as function of the radius  $r$  only

Figure 106: The steady state solution of a heat problem in a ball

For the results in Figure 106 linear elements are used. A simple call of `MeshUpgrade()` allows to use cubic elements. In Figure 107 find the differences to the exact solution for linear and cubic elements. Observe that the error for cubic elements is considerably smaller.

The above is a problem with one space variable only. Thus it is advisable to use `BVP1D()`. Find the results of the code below in Figure 108. Observe the effect of superconvergence

## HeatBallRadial.m

```

R1 = 1; R2 = 3; BCleft = 0; BCright = 1;
interval = linspace(R1,R2,11);

```

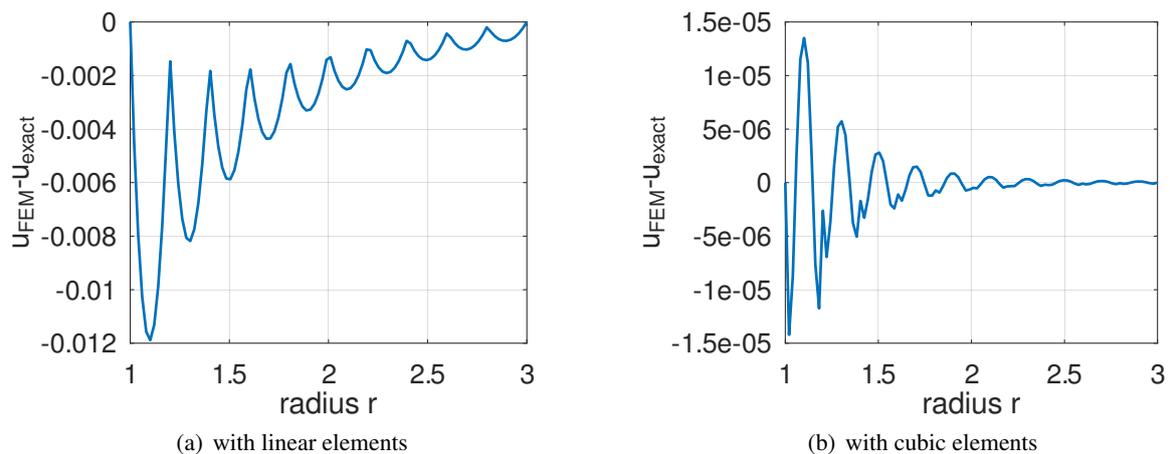


Figure 107: The errors of the steady state solution of a heat problem in a ball

```
[r,u] = BVP1D(interval,@(r)-r.^2,0,0,1,0,BCleft,BCright);
figure(4); plot(r,u,r,R2/(R2-R1)*(r-R1)./r)
    xlabel('radius r'); ylabel('temperature u')
    legend('u_{FEM}','u_{exact}','location','northwest')

figure(5); plot(r,u-R2/(R2-R1)*(r-R1)./r,'+-')
    xlabel('radius r'); ylabel('temperature u')
    legend('difference','location','southeast')

r_fine = linspace(R1,R2,501)'; u_fine = pwquadinterp(r,u,r_fine);
u_exact = R2/(R2-R1)*(r_fine-R1)./r_fine;
figure(6); plot(r_fine,u_fine-u_exact,'k',r,u-R2/(R2-R1)*(r-R1)./r,'b+')
    xlabel('radius r'); ylabel('u_{FEM}-u_{exact}')
    legend('interpolated','at nodes')
```

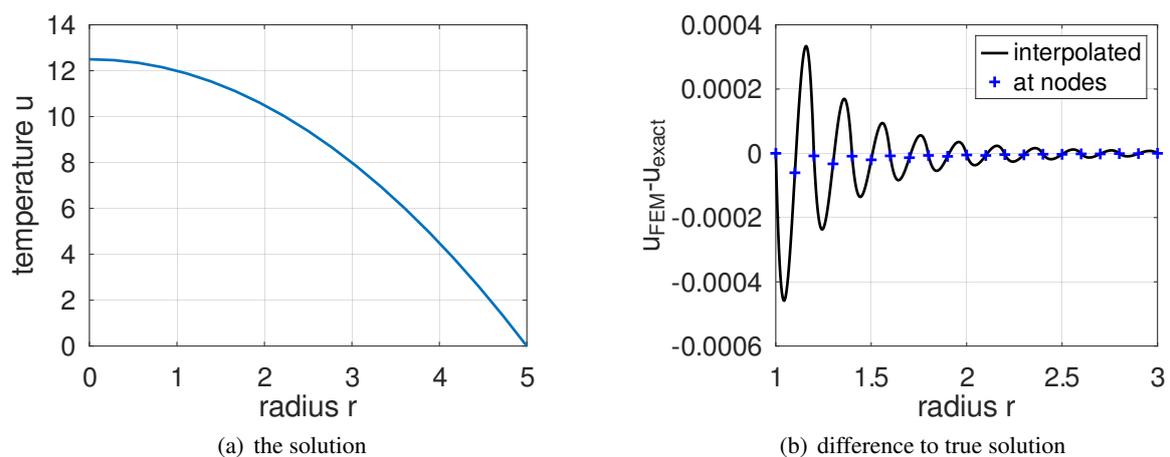


Figure 108: The steady state solution of a heat problem in a ball with second order elements, solved as true 1D problem

### 9.10.6 Dynamic heat equation in a cylinder, solved as a 1D problem

Similar to Section 9.10.5 the dynamic heat equation in cylindrical coordinates  $(r, \theta, z)$  is given by

$$\frac{\rho c}{k} \frac{\partial}{\partial t} u = +\Delta u + f = \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial u}{\partial r} \right) + \frac{1}{r^2} \frac{\partial^2 u}{\partial \theta^2} + \frac{\partial^2 u}{\partial z^2} + f.$$

Assuming that the solution depends on time  $t$  and radius  $r$  only and some rescaling leads to the dynamic equation

$$r \frac{\partial}{\partial t} u(r, t) = \frac{\partial}{\partial r} \left( r \frac{\partial u(r, t)}{\partial r} \right) + r f(r, t).$$

Examine a cylinder of radius  $R$  and apply time dependent heating in the inner section  $r < \frac{R}{2}$  of the cylinder, e.g.

$$f(r, t) = \begin{cases} \sin(t) & \text{for } 0 \leq r < \frac{R}{2} \\ 0 & \text{for } \frac{R}{2} \leq r \leq R \end{cases}.$$

#### HeatCylinder.m

```
R = 3; BCleft = [0,0]; BCright = [0,0];
f = @(r,t)sin(1*t)*(r<R/2);
u0 = 0; t0 = 0; t_end = 6*pi; steps = [121,10]; interval = linspace(0,R,21);
r = @(r)r; solver = 'RK';
[r,u,t] = IBVP1D(interval,r,r,0,0,r,f,BCleft,BCright,u0,t0,t_end,steps,'solver',solver);
figure(1); mesh(t,r,u)
    xlabel('time t'); ylabel('radius r'); zlabel('temperature u')
figure(2); contour(t,r,u,[-0.5:0.1:+0.5])
    xlabel('time t'); ylabel('radius r');
figure(3); plot(t,u(1,:),t,u(end,:))
    xlabel('time t'); ylabel('temperature u'); xlim([0,max(t)])
    legend('at center r=0','at outer edge r=R','location','south')
```

As results obtain the temperature  $u$  as function of time  $t$  and radius  $r$ , visible in Figures 109 and 110. The periodic oscillations of  $U$  are clearly visible. In Figures 110 and 111 observe the smaller amplitudes of  $u$  with respect to time at the outer edge  $r = R$  and the phase shift.

With the above code `HeatCylinder.m` it is easy to examine the effects of a different frequency of amplitude in  $k \sin(\omega t)$ .

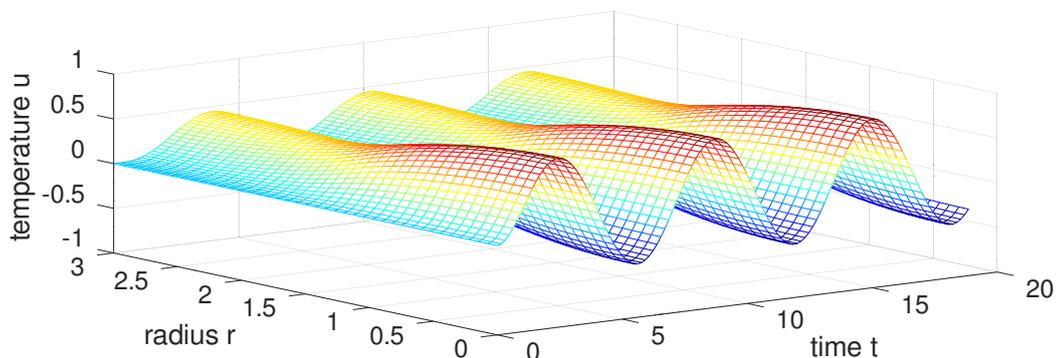


Figure 109: The temperatur  $u(r, t)$  inside the cylinder

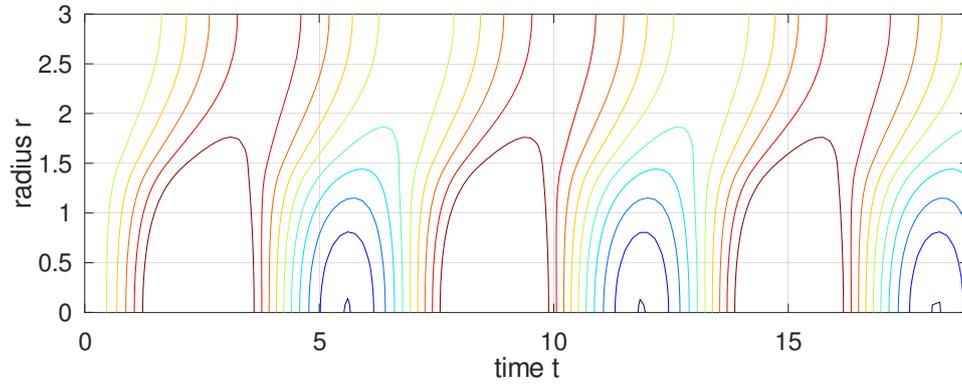


Figure 110: The contours of the temperatur  $u(r, t)$  inside the cylinder

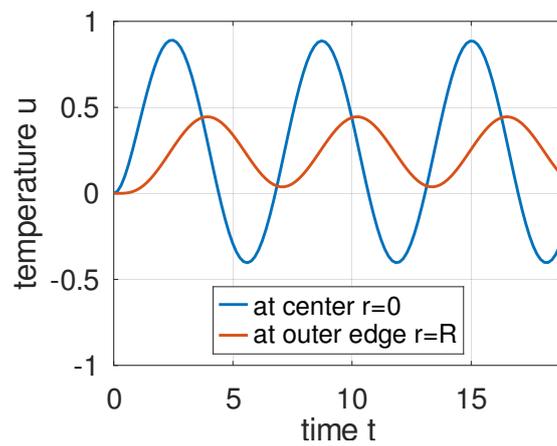


Figure 111: The temperatur  $u(r, t)$  at the inner ( $r = 0$ ) and outer ( $r = R$ ) edge

## 9.11 Wave propagation, Kirchhoff diffraction

### 9.11.1 A dynamic solution

In Figure 112 half of a domain  $\Omega \subset \mathbb{R}^2$  is visible, the lower half is generated by a reflection at the lower edge. For the computation this is taken into account by the symmetry boundary condition  $\frac{\partial u}{\partial n} = 0$ . The wave equation (a special case of the IBVP (6)) to be solved is

$$\begin{aligned} \frac{\partial^2}{\partial t^2} u(x, y, t) - \Delta u(x, y, t) &= f(x, y, t) && \text{for } (x, y, t) \in \Omega \times (0, T] \\ \frac{\partial}{\partial n} u(x, y, t) &= 0 && \text{for } (x, y, t) \in \Gamma \times (0, T] \\ u(x, y, 0) = \frac{\partial}{\partial t} u(x, y, 0) &= 0 && \text{on } \Omega \end{aligned} \quad (87)$$

The function  $f(x, y, t)$  equals  $\sin(3\pi t)$  for  $x \leq -0.9$  and zero otherwise. Thus there a periodic excitation at the very left end of the appendix for  $-1 \leq x \leq -0.9$ . The wave speed equals 1 and the appendix (more precise: the two appendices) is a source of waves.

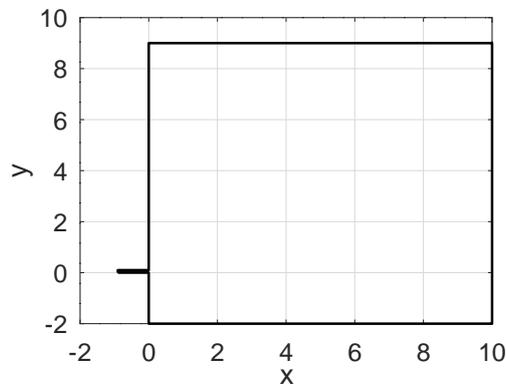


Figure 112: The domain for the wave propagation

Figure 113 shows the solution  $u(x, y, 11)$  at time  $t = 11$ .

- The wave speed equals 1, thus at  $t = 11$  the first waves are about to arrive at  $x = 10$  for  $y = 0$  and at  $y = +10$  for  $x = 0$ .
- In the top right section the unperturbed waves generated by the outlet of the appendix at  $y = 0$  are visible.
- In the top left corner the upward moving waves interfere with the waves reflected at the upper edge at  $y = 9$ .
- At the lower edge at  $y = -2$  the waves are reflected leading to interference. The result is identical to the situation of a second source at  $y = -4$ .
- In the lower part of the figure observe the result of the classical double-slit diffraction pattern by Kirchhoff, see e.g. [en.wikipedia.org/wiki/Double-slit\\_experiment](https://en.wikipedia.org/wiki/Double-slit_experiment).

Observe that the behavior of the solution is very different from a heat equation in 9.10.4, even is the setup is comparable.

In the code below you can play with the different parameters and select whether an animation is shown on the screen or the final snapshot only.

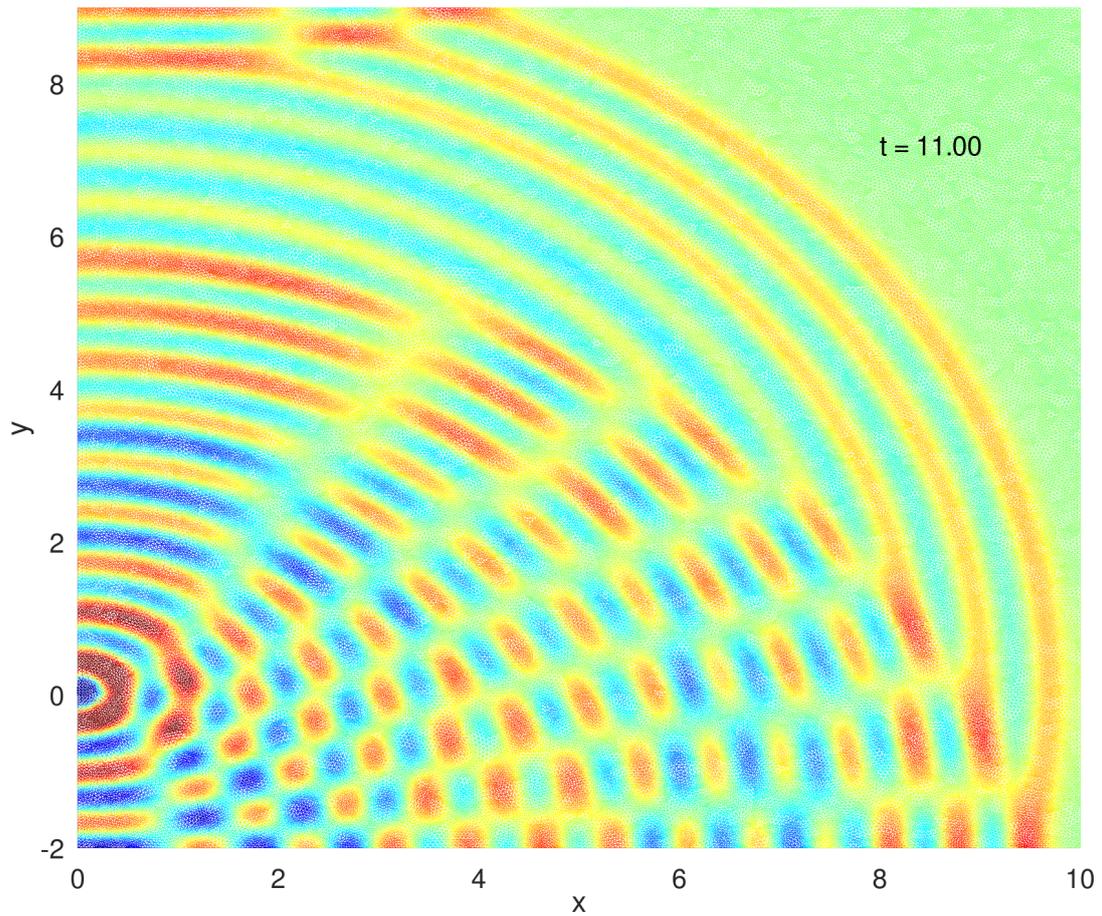


Figure 113: Wave propagation, leading to a Kirchoff diffraction pattern

## WavePropagation.m

```

l = 1; h = 0.1; L = 10; d = 2; H = 9;
FEMmesh = CreateMeshTriangle('test', ...
    [-1,0,-2;0 0 -2;0,-d,-2;L,-d,-2; L,H,-2;0,H,-2;0,h,-2;-1,h,-2],0.01);
figure(1); FEMtrimesh(FEMmesh)
    xlabel('x'); ylabel('y'); axis equal
FEMmesh = MeshUpgrade(FEMmesh,'cubic');

function res = f(x,y,t)
    res = sin(3*pi*t)*ones(size(x,y,1),1);
    res(xy(:,1)>-0.9) = 0;
endfunction
function res = v0(x,y); res = zeros(size(x,y,1),1); endfunction
function res = u0(x,y) res = zeros(size(x,y,1),1); endfunction

m = 1; a = 1; b0 = 0; bx = by = 0; f = 0; gn1 = gn2 = 0;
tic();
[u,t] = I2BVP2D(FEMmesh,m,0,a,b0,bx,by,'f',0,gn1,gn2,'u0','v0',0,11,[56,10]);
SolverTime = toc()

umax = 0.3*max([-min(u(:)),max(u(:))]);
figure(2)
if 0 %% animation
    for jj = 1:length(t)
        FEMtrimesh(FEMmesh,u(:,jj))
        xlabel('x'); ylabel('y')
        zlim(umax*[-1 1]); caxis(0.3*umax*[-1 1]);
        text(0.8*L,0.8*H,umax,sprintf('t = %4.2f',t(jj)))
        view(0,90); xlim([0,L]); ylim([-d,H]);
        pause(0.1);
    endfor
else
    FEMtrimesh(FEMmesh,u(:,end))
    xlabel('x'); ylabel('y')
    zlim(umax*[-1 1]); caxis(0.3*umax*[-1 1]);
    text(0.8*L,0.8*H,umax,sprintf('t = %4.2f',t(end)))
    view(0,90); xlim([0,L]); ylim([-d,H]);
endif

```

## 9.12 Sound waves in $\mathbb{R}^2$ and $\mathbb{R}^3$

The standard wave equation  $\frac{\partial^2}{\partial t^2} u - \Delta u = 0$  can be written in cylindrical

$$\frac{\partial^2}{\partial t^2} u(\rho, \phi, z, t) = \frac{1}{\rho} \frac{\partial}{\partial \rho} \left( \rho \frac{\partial u}{\partial \rho} \right) + \frac{1}{\rho^2} \frac{\partial^2}{\partial \phi^2} u + \frac{\partial^2}{\partial z^2} u$$

or spherical coordinates

$$\frac{\partial^2}{\partial t^2} u(r, \phi, \theta, t) = \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial u}{\partial r} \right) + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial u}{\partial \theta} \right) + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 u}{\partial \phi^2}.$$

This allows to reduce some problems to two space dimensions.

### 9.12.1 A sound wave in $\mathbb{R}^3$ with cylindrical coordinates

Assuming that the solution  $u(\rho, z, t)$  is independent on  $\phi$  and multiplying the wave equation by  $\rho$  arrive at

$$\rho \frac{\partial^2}{\partial t^2} u(\rho, z, t) - \frac{\partial}{\partial \rho} \left( \rho \frac{\partial u}{\partial \rho} \right) - \frac{\partial}{\partial z} \left( \rho \frac{\partial u}{\partial z} \right) = 0 \quad (88)$$

and thus it is in the form of the general hyperbolic equation (6) and can be solved numerically with I2BVP2D(). On a domain  $0 \leq \rho \leq R$  and  $0 \leq \theta \leq \pi$  we assume zero initial velocity  $\frac{d}{dt} u(\rho, z, 0) = 0$  and initial displacement

$$u(\rho, z, 0) = \begin{cases} 1 + \cos(10r) & \text{for } 0 \leq r \leq \frac{\pi}{10} \\ 0 & \text{for } \frac{\pi}{10} \leq r \end{cases} .$$

where we use  $r = \sqrt{\rho^2 + z^2}$ . The result of solving this initial boundary value problem will be a spherical wave moving with speed 1 and a decaying amplitude. Find the result at time  $t = 1.75$  in Figure 116. Using an energy argument the amplitude of the wave front is expected to decay like  $c \frac{1}{t}$ . Using linear regression this is confirmed in Figure 116.

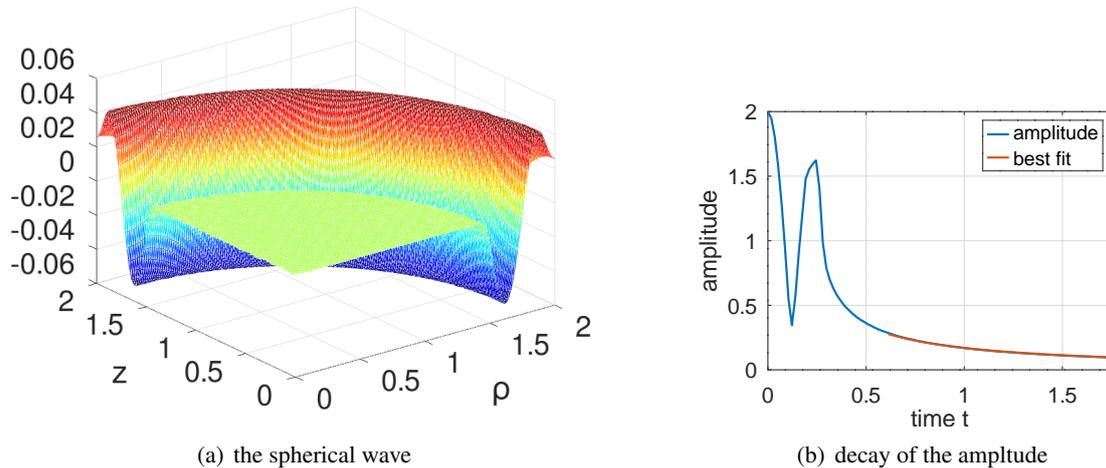


Figure 114: A spherical sound wave at time  $t = 1.75$ , and the decaying amplitude with the best fitting  $\frac{c}{t}$

#### SoundWaveSpherical.m

```
R = 2; H = 2; N = 60;
FEMmesh = CreateMeshRect(linspace(0,R,N),linspace(0,H,N),-2,-2,-2,-2);
FEMmesh = MeshUpgrade(FEMmesh,'quadratic');

function res = u_0(xy)
    r = sqrt(xy(:,1).^2+xy(:,2).^2);
    res = 1+cos(10*r);    res(r>pi/10) = 0;
endfunction
function res = rho(xy,dummy); res = xy(:,1);    endfunction;
function res = v_0(xy);    res = zeros(size(xy,1),1); endfunction

tic();
[u,t] = I2BVP2D(FEMmesh,'rho',0,'rho',0,0,0,0,0,0,0,0,'u_0','v_0',0,1.75,[100,10]);
ComputationTime = toc()
```

```

figure(1); clf
if 0 %% animation
    for jj = 1:length(t)
        FEMtrimesh(FEMmesh,u(:,jj))
        xlabel('rho'); ylabel('z'); zlim([-0.5 0.5]); caxis(0.1*[-0.5,0.5])
        pause(0.1)
    endfor
else
    FEMtrimesh(FEMmesh,u(:,end))
    xlabel('\rho'); ylabel('z')
endif

max_u = max(u) - min(u); t_start = find(t>0.6,1); t_tail = t(t_start:end)';

[p,~,~,p_var] = LinearRegression(1./t_tail,max_u(t_start:end)');
figure(2); plot(t,max_u,t_tail, p./t_tail)
                xlabel('time t'); ylabel ('amplitude'); legend('amplitude','best fit')

```

### 9.12.2 A sound wave in $\mathbb{R}^2$

In a rectangle  $0 \leq x, y \leq R$  solve the standard wave equation

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2} = 0$$

with Neumann boundary conditions  $\frac{\partial u}{\partial n} = 0$ , initial zero velocity and initial displacement

$$u(x, y, 0) = \begin{cases} 1 + \cos(10r) & \text{for } 0 \leq r \leq \frac{\pi}{10} \\ 0 & \text{for } \frac{\pi}{10} \leq r \end{cases}.$$

where we use  $r = \sqrt{x^2 + y^2}$ . The result of solving this initial boundary value problem will be a circular wave moving with speed 1 and a decaying amplitude. Find the result at time  $t = 4$  in Figure 115. Using an energy argument the amplitude of the wave front is expected to decay like  $c \frac{1}{\sqrt{t}}$ . Using linear regression this is confirmed in Figure 115.

#### SoundWave.m

```

R = 4.5; H = 4.5; N = 60;
FEMmesh = CreateMeshRect(linspace(0,R,N),linspace(0,H,N),-2,-2,-2,-2);
FEMmesh = MeshUpgrade(FEMmesh,'quadratic');

function res = u_0(xy)
    r = sqrt(xy(:,1).^2+xy(:,2).^2);
    res = 1+cos(10*r);
    res(r>pi/10) = 0;
endfunction
function res = v_0(xy) ; res = zeros(size(xy,1),1); endfunction
tic();
[u,t] = I2BVP2D(FEMmesh,1,0,1,0,0,0,0,0,0,0,0,0,0,0,'u_0','v_0',0,4,[100,10]);
ComputationTime = toc()

figure(3); clf
if 0 %% animation
    for jj = 1:length(t)

```

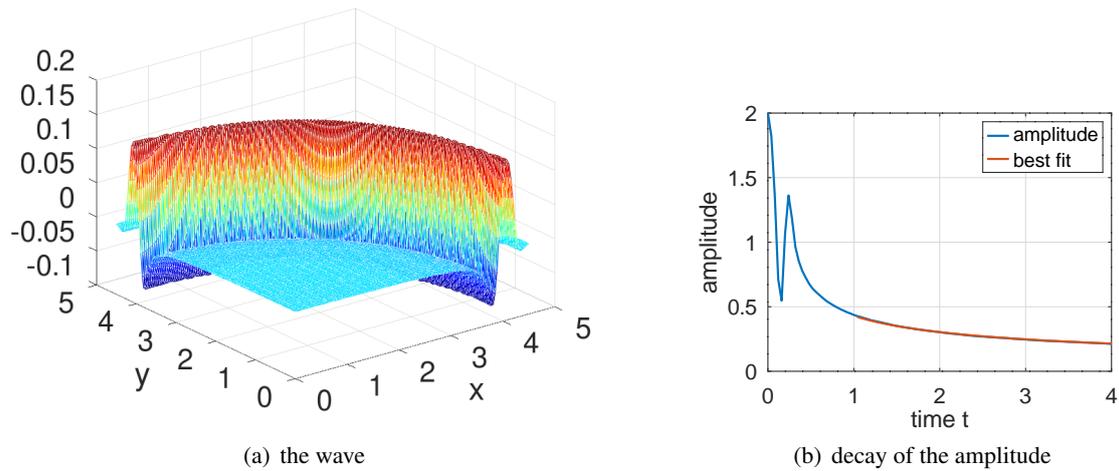


Figure 115: A circular sound wave at time  $t = 4$  and the decaying amplitude with the best fitting  $\frac{c}{\sqrt{t}}$

```

FEMtrimesh(FEMmesh,u(:,jj))
xlabel('x'); ylabel('y');
zlim(0.1*[-2 2]); caxis(0.5*[-2 2])
pause(0.1)
endfor
else
FEMtrimesh(FEMmesh,u(:,end))
xlabel('x'); ylabel('y')
endif

max_u = max(u) - min(u); t_start = find(t>1,1); t_tail = t(t_start:end)';
[p,~,~,p_var] = LinearRegression(1./sqrt(t_tail),max_u(t_start:end)');
figure(12); plot(t,max_u,t_tail, p./sqrt(t_tail))
xlabel('time t'); ylabel('amplitude'); legend('amplitude','best fit')

```

### 9.12.3 Sound waves in $\mathbb{R}^3$ and $\mathbb{R}^2$ as 1D problems

- As spherical wave:

If a solution  $u$  of the wave equation  $\frac{\partial^2}{\partial t^2} u = \Delta u$  depends on the radius  $r = \sqrt{x^2 + y^2 + z^2}$  only the IBVP is given by

$$\begin{aligned}
 r^2 \frac{\partial^2}{\partial t^2} u(r, t) &= \frac{\partial}{\partial r} \left( r^2 \frac{\partial}{\partial r} u(r, t) \right) && \text{for } 0 < r < R \text{ and } t > 0 \\
 u(r, 0) &= u_0(r) && \text{for } 0 < r < R \\
 \frac{\partial}{\partial t} u(r, 0) &= u_1(r) && \text{for } 0 < r < R
 \end{aligned}$$

For a zero initial velocity  $u_1(r) = 0$  and initial amplitude

$$u(r, 0) = u_0(r) = \begin{cases} 1 + \cos(10r) & \text{for } 0 \leq r \leq \frac{\pi}{10} \\ 0 & \text{for } \frac{\pi}{10} \leq r \end{cases}$$

and Neumann boundary conditions at  $r = 0$  and  $r = R$  the problem can be solved by `I2BVP1D()` with the code `SoundWaveSpherical1D.m` below. Find the result in Figure 116. Observe that for advanced time  $t$  the solution is equal to zero at the origin  $r = 0$ .

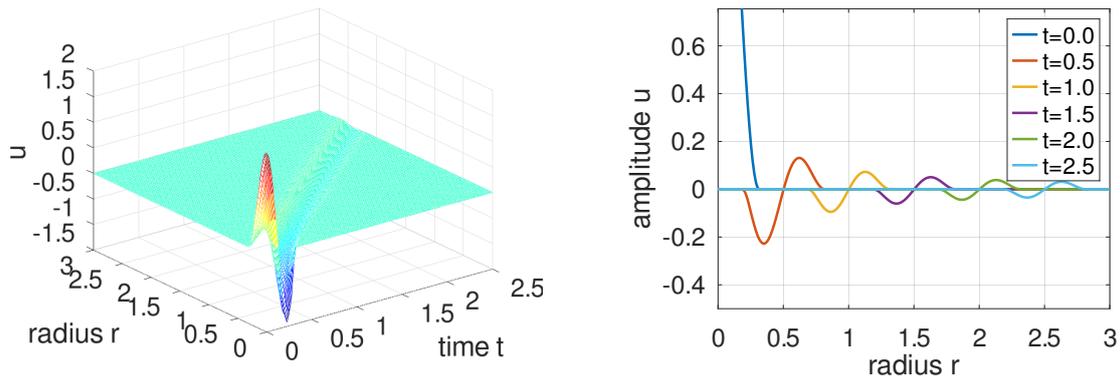


Figure 116: A spherical wave as function of time  $t$  and radius  $r = \sqrt{x^2 + y^2 + z^2}$

### SoundWaveSpherical1D.m

```

N = 2*60; R = 3; interval = linspace(0,R,N)';
f_r2 = @(r)r.^2;
u0 = @(r) (1+cos(10*r)).*(r<pi/10); u1 = 0;
w2 = f_r2; w1 = 0; a = f_r2; b = 0; c = 0; d = 1; f = 0;
BCleft = [0,0]; BCright = [0,0];
t0 = 0; tend = 2.5; steps = [100,10];
[r,u,t] = I2BVP1D(interval,w2,w1,a,b,c,d,f,BCleft,BCright,u0,u1,t0,tend,steps);

figure(1); mesh(t,r,u); xlabel('time t'); ylabel('radius r'); zlabel('u')
        xlim([min(t),max(t)]); ylim([min(r),max(r)])

[t05,t05_ind] = find(abs(t-0.5)<1e-5); [t1,t1_ind] = find(abs(t-1)<1e-5);
[t15,t15_ind] = find(abs(t-1.5)<1e-5); [t2,t2_ind] = find(abs(t-2)<1e-5);
[t25,t25_ind] = find(abs(t-2.5)<1e-5);

figure(2); plot(r,u(:,1),r,u(:,t05_ind),r,u(:,t1_ind),r,u(:,t15_ind),...
        r,u(:,t2_ind),r,u(:,t25_ind));
        ylabel('amplitude u'); ylim([-0.5,0.7]); xlabel('radius r');
        legend('t=0.0','t=0.5','t=1.0','t=1.5','t=2.0','t=2.5',...
        'location','northeast')

```

- As cylindrical wave:

If a solution  $u$  of the wave equation  $\frac{\partial^2}{\partial t^2} u = \Delta u$  depends on the radius  $r = \sqrt{x^2 + y^2}$  only the IBVP is given by

$$\begin{aligned}
 r \frac{\partial^2}{\partial t^2} u(r, t) &= \frac{\partial}{\partial r} \left( r \frac{\partial}{\partial r} u(r, t) \right) && \text{for } 0 < r < R \text{ and } t > 0 \\
 u(r, 0) &= u_0(r) && \text{for } 0 < r < R \\
 \frac{\partial}{\partial t} u(r, 0) &= u_1(r) && \text{for } 0 < r < R
 \end{aligned}$$

For a zero initial velocity  $u_1(r) = 0$  and the same initial amplitude  $u_0(r)$  as above and Neumann boundary conditions at  $r = 0$  and  $r = R$  the problem can be solved with the code `SoundWaveSpherical1D.m` below, using `I2BVP1D()`. Find the result in Figure 117. Observe that for advanced time  $t$  the solution is **not equal** to zero at the origin  $r = 0$ . The amplitudes at  $t > 0$  are larger than for the above spherical case. This can be derived analytically, using a conservation of energy argument.

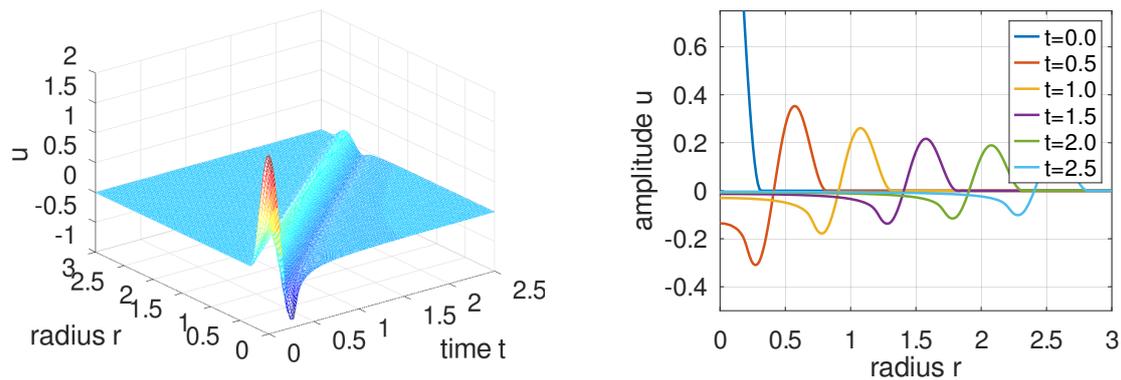


Figure 117: A cylindrical wave as function of time  $t$  and radius  $r = \sqrt{x^2 + y^2}$

#### SoundWaveCylindrical1D.m

```

N = 2*60; R = 3; interval = linspace(0,R,N)';
f_r = @(r)r;
u0 = @(r) (1+cos(10*r)).*(r<pi/10); u1 = 0;
w2 = f_r; w1 = 0; a = f_r; b = 0; c = 0; d = 1; f = 0;
BCleft = [0,0]; BCright = [0,0];
t0 = 0; tend = 2.5; steps = [100,10];
[r,u,t] = I2BVP1D(interval,w2,w1,a,b,c,d,f,BCleft,BCright,u0,u1,t0,tend,steps);

figure(1); mesh(t,r,u); xlabel('time t'); ylabel('radius r'); zlabel('u')
        xlim([min(t),max(t)]); ylim([min(r),max(r)])

[t05,t05_ind] = find(abs(t-0.5)<1e-5); [t1,t1_ind] = find(abs(t-1)<1e-5);
[t15,t15_ind] = find(abs(t-1.5)<1e-5); [t2,t2_ind] = find(abs(t-2)<1e-5);
[t25,t25_ind] = find(abs(t-2.5)<1e-5);

figure(2); plot(r,u(:,1),r,u(:,t05_ind),r,u(:,t1_ind),r,u(:,t15_ind),...
        r,u(:,t2_ind),r,u(:,t25_ind)));
        ylabel('amplitude u'); ylim([-0.5,0.75]); xlabel('radius r');
        legend('t=0.0','t=0.5','t=1.0','t=1.5','t=2.0','t=2.5',...
        'location','northeast')

```

### 9.13 Reflection and transmission by a change of impedance

A change of impedance will cause an incoming wave to be partially reflected. Examine the initial boundary value problem

$$\begin{aligned}
 \frac{\partial^2}{\partial t^2} u(x,t) &= \frac{\partial}{\partial x} \left( a(x) \frac{\partial}{\partial x} u(r,t) \right) && \text{for } 0 < x < 12 \text{ and } t > 0 \\
 \frac{\partial}{\partial x} u(0,t) = \frac{\partial}{\partial x} u(12,t) &= 0 && \text{for } t > 0 \\
 u(x,0) &= u_0(x) && \text{for } 0 < x < 12 \\
 \frac{\partial}{\partial t} u(x,0) &= u_1(x) && \text{for } 0 < x < 12
 \end{aligned}$$

with the discontinuous coefficient  $a(x)$

$$a(x) = \begin{cases} 1 & \text{for } 0 \leq x \leq 4 \\ 2 & \text{for } 4 < x \end{cases}$$

and the initial values

$$u_0(x) = \begin{cases} \sin(\pi x) & \text{for } 0 \leq x \leq 1 \\ 0 & \text{for } 1 < x \end{cases} \quad \text{and} \quad u_1(x) = \begin{cases} -\pi \cos(\pi x) & \text{for } 0 \leq x \leq 1 \\ 0 & \text{for } 1 < x \end{cases}$$

The sin-shaped pulse will travel in the positive  $x$ -direction with speed 1 and then at  $x = 4$  a part will be reflected and traveling back with speed 1 and another part will continue, but with speed  $\sqrt{2}$ . In Figure 118 find the plot of the surface and in Figure 119 a plot of the solutions at times  $t = 0, 2, 4, 6$  and  $8$ , and the contour plot.

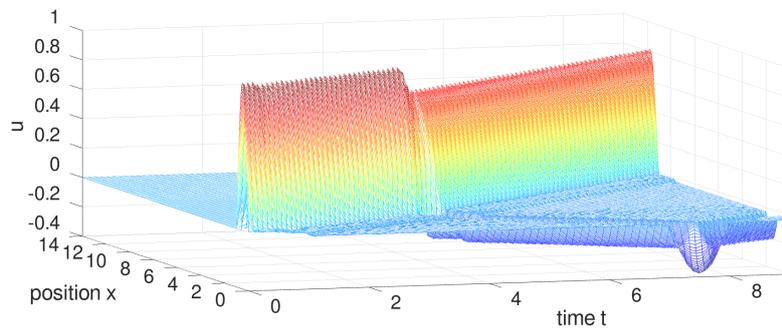


Figure 118: The amplitude  $u(x, t)$  for a reflected pulse

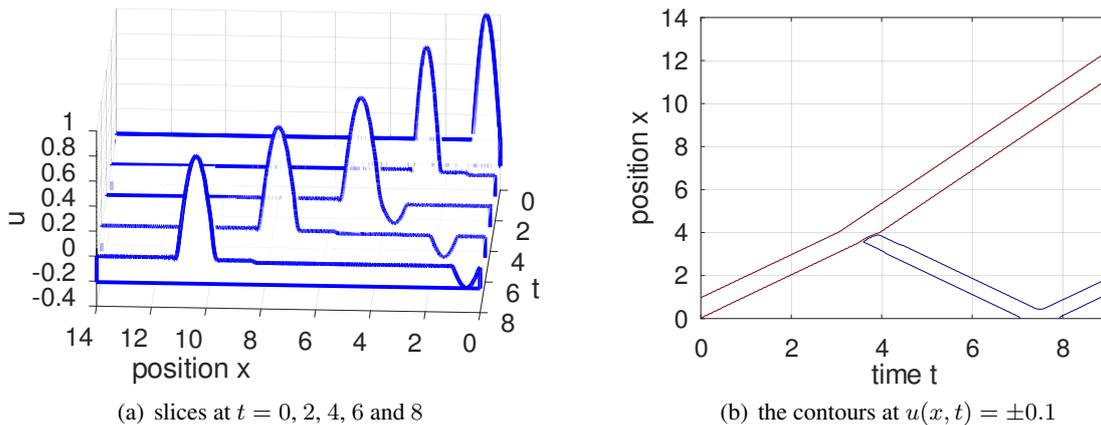


Figure 119: Waterfall plot and contour plot for the amplitude  $u(x, t)$  for a reflected pulse

**Reflection.m**

```
interval = linspace(0,14,301)';
a = @(x) 1+1*(x>4);
b = 0; c = 0; d = 1; f = 0;
w2 = 1; w1 = 0; ; BCleft = [0,0]; BCright = [0,0];
```

```

t0 = 0; tend = 9; steps = [90,100];
u0 = @(x) sin(pi*x).*(x<=1); u1 = @(x) -pi*cos(pi*x).*(x<=1);

[x,u,t] = I2BVP1D(interval,w2,w1,a,b,c,d,f,BCleft,BCright,u0,u1,t0,tend,steps);

figure(1); mesh(t,x,u); xlabel('time t'); ylabel('position x'); zlabel('u')
    xlim([min(t),max(t)]); ylim([min(x),max(x)]); view([-20,10])
t_ind = [1 21 41 61 81];
figure(2); H = waterfall(x,t(t_ind),u(:,t_ind)'); view([-177,20])
    xlabel('position x'); zlabel('u')
    set(H, 'edgecolor', [0,0.0,1]); set(H, 'linewidth', 3)
figure(3); contour(t,x,u,[-0.1,0.1]); xlabel('time t'); ylabel('position x');

```

### 9.14 Schrödinger's harmonic oscillator

The dynamic Schrödinger equation is given by<sup>31</sup>

$$i \hbar \frac{\partial}{\partial t} \Psi(x, t) = \left[ -\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} + V(x) \right] \Psi(x, t).$$

Searching for solutions of the form  $\Psi(x, t) = f(t) \psi(x)$  leads to

$$\begin{aligned}
 i \hbar \frac{d}{dt} f(t) \psi(x) &= \left[ -\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} + V(x) \right] f(t) \psi(x) \\
 E = i \hbar \frac{\frac{d}{dt} f(t)}{f(t)} &= \frac{1}{\psi(x)} \left[ -\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} + V(x) \right] \psi(x) \\
 \frac{d}{dt} f(t) &= \frac{E}{i \hbar} f(t) \implies f(t) = C \exp\left(-i \frac{E}{\hbar} t\right)
 \end{aligned}$$

With the potential  $V(x) = \frac{1}{2} m \omega^2 x^2$  for a harmonic oscillator Schrödinger's eigenvalue equation for a harmonic oscillator is given by

$$-\frac{\hbar^2}{2m} \frac{\partial^2 \psi}{\partial x^2} + \frac{1}{2} m \omega^2 x^2 \psi = E \psi \quad (89)$$

A simple rescaling with  $z = \sqrt{\frac{\hbar}{m\omega}} x$  and  $u(x) := \psi(z) = \psi\left(\sqrt{\frac{\hbar}{m\omega}} x\right)$  leads to

$$\begin{aligned}
 -\frac{\partial^2 u(x)}{\partial x^2} + x^2 u(x) &= -\frac{\partial^2 \psi(z)}{\partial z^2} \frac{\hbar}{m\omega} + \frac{m\omega}{\hbar} z^2 \psi(z) \\
 &= \frac{2}{\omega \hbar} \left( -\frac{\hbar^2}{2m} \frac{\partial^2 \psi(z)}{\partial z^2} + \frac{m\omega^2}{2} z^2 \psi(z) \right) \\
 &= \frac{2}{\omega \hbar} E \psi(z) = \lambda u(x)
 \end{aligned}$$

Thus the eigenvalue problem to be solved is

$$-\frac{\partial^2}{\partial x^2} u(x) + x^2 u(x) = \lambda u(x) \quad \text{with} \quad u(\pm\infty) = 0.$$

Then the eigenvalues of the Schrödinger equation (89) are given by  $E = \frac{\omega \hbar}{2} \lambda$ .

<sup>31</sup>Source: Wikipedia [en.wikipedia.org/wiki/Schrodinger\\_equation](https://en.wikipedia.org/wiki/Schrodinger_equation)

The code below uses the command `BVP1Deig()` to approximate the first six eigenvalues. The numerical result confirms that the eigenvalues of the Schrödinger equation (89) are given by

$$E_n = \frac{\omega \hbar}{2} \lambda_n = \frac{\omega \hbar}{2} (2n + 1) \quad \text{for } n = 0, 1, 2, 3, 4, \dots$$

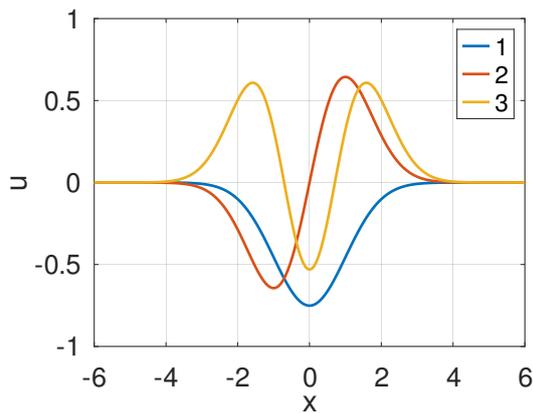
Figure 120 shows the shape of the corresponding eigenfunctions  $u_n(x)$ . The exact formulas for the eigenfunctions use the physicist's Hermite polynomials and exponential functions<sup>32</sup>.

```

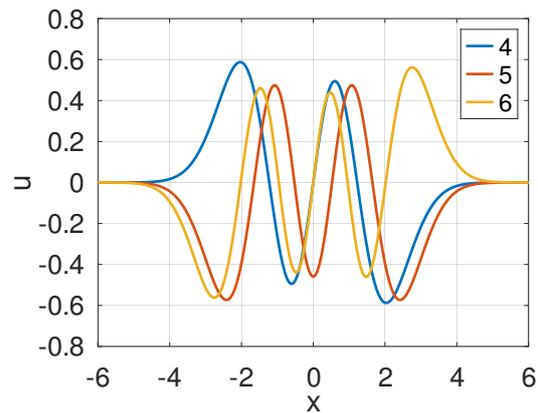
SchroedingerHarmonic.m
x_max = 6; interval = linspace(-x_max,x_max,100)';
BCleft = 0; BCright = 0;
[x,eVal, eVec] = BVP1Deig(interval,1,0,@(x)x.^2,1,BCleft,BCright,6);

Eigenvalues = eVal'
figure(1); plot(x,eVec(:,1:3)); xlabel('x'); ylabel('u'); xlim([-x_max,+x_max])
    legend('1','2','3')
figure(2); plot(x,eVec(:,4:6)); xlabel('x'); ylabel('u'); xlim([-x_max,+x_max])
    legend('4','5','6')

-->
Eigenvalues = 1.0000  3.0000  5.0000  7.0000  9.0001  11.0001
    
```



(a) the first three eigenfunctions



(b) the next three eigenfunctions

Figure 120: The first six eigenfunctions of the harmonic Schrödinger operator

### 9.15 The EIT forward problem

For a conductivity  $\sigma$  on a bounded domain  $\Omega \subset \mathbb{R}^2$  consider the PDE

$$\nabla \cdot (\sigma \nabla u) = 0 \quad \text{in } \Omega \subset \mathbb{R}^2 \tag{90}$$

- Apply a voltage  $u$  on the boundary and measure the resulting current density  $J$

$$J(z) = \sigma(z) \frac{\partial u(z)}{\partial n} \quad \text{for } z \in \partial\Omega$$

32

$$\psi_n(x) = \frac{1}{\sqrt{2^n n!}} \left( \frac{m\omega}{\pi \hbar} \right)^{1/4} e^{-\frac{m\omega x^2}{2\hbar}} H_n\left(\sqrt{\frac{m\omega}{\hbar}} x\right) \quad \text{where } H_n(z) = (-1)^n e^{z^2} \frac{d^n}{dx^n} \left( e^{-z^2} \right)$$

to obtain the **Dirichlet to Neumann** map

$$\Lambda_\sigma : u \rightarrow \sigma \frac{\partial u}{\partial n} \quad \text{on } \partial\Omega \quad (91)$$

also called voltage to current density map.

- Apply a current density  $J$  on the boundary and measure the resulting voltage  $u$ . For a static situation the total current into  $\Omega$  has to be zero, i.e.

$$\oint_{\partial\Omega} J(s) ds = \oint_{\partial\Omega} \sigma \frac{\partial u}{\partial n} ds = 0$$

to obtain the **Neumann to Dirichlet** map

$$\mathcal{R}_\sigma : \sigma \frac{\partial u}{\partial n} \rightarrow u \quad \text{on } \partial\Omega \quad (92)$$

also called current density to voltage map.

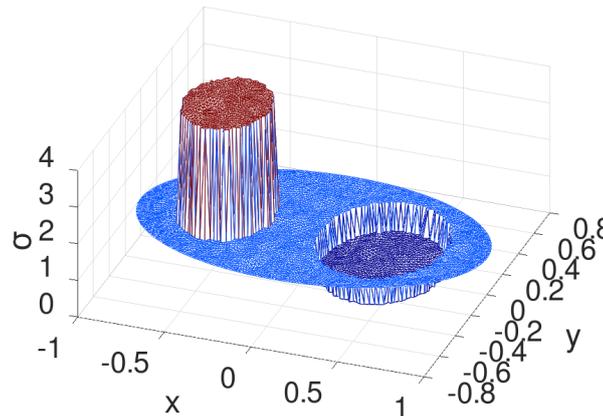


Figure 121: The conductivity with the conducting “heart” on the left and the insulating “lung” on the right

From either one of these maps it is possible to determine the conductivity  $\sigma$  in the domain. This is called Electrical Impedance Tomography, or short EIT. For a good, readable description consider the book [MuelSilt12] or the article [MuelSilt20]. The Neumann to Dirichlet map  $\mathcal{R}_\sigma$  is more reliable to measure, based on less susceptibility to noise. Using FEM examine the forward problem, i.e. apply a known current pattern and determine the resulting voltage  $u$  on the boundary. In real life this is performed by measurements. Examine the domain (a chest cross section) in Figure 121 with the graph of the conductivity  $\sigma$  shown. On the left observe a simple heart with high conductivity, caused by the blood. On the right observe a section with very low conductivity, caused by the air filled lung. Then two current patterns are examined:

1. A current input at the lower edge of the cross section in Figure 122 and a matching current outlet at an angle of approximately  $120^\circ$ . Thus the current is expected to go through the heart, mainly.
2. A similar current input at the lower edge and a matching current outlet at an angle of approximately  $60^\circ$ . Thus the current is expected to go through the lung, mainly.

The boundary  $\Gamma$  of the domain  $\Omega$  is given by

$$\begin{pmatrix} R_x \cos \alpha \\ R_y \sin \alpha \end{pmatrix} \text{ for } 0 \leq \alpha \leq 2\pi \text{ with } R_x = 1 \text{ and } R_y = 0.5,$$

with a conductivity of  $\sigma = 1$ . A simple calculation on the ellipse leads to an arc length of

$$ds = \sqrt{R_x^2 \sin^2 \alpha + R_y^2 \cos^2 \alpha} d\alpha.$$

The “heart” is given by

$$(x + 0.5)^2 + y^2 \leq 0.25^2 \text{ with conductivity } \sigma = 4$$

and the “lung” is given by

$$(x - 0.4)^2 + y^2 \leq 0.35^2 \text{ with conductivity } \sigma = \frac{1}{4}.$$

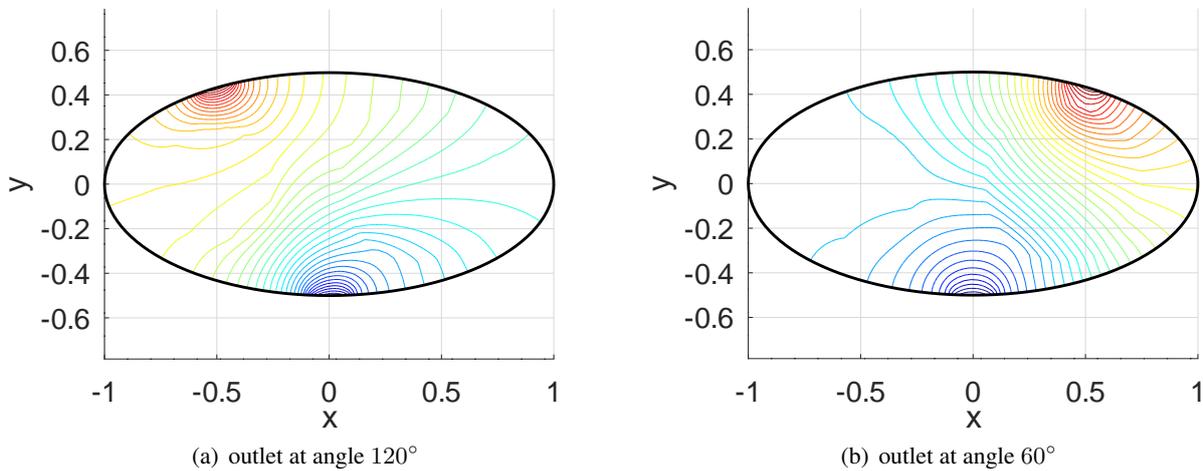


Figure 122: Contours of the voltages

FEMoctave is used twice to determine the voltage  $u$  in the domain, leading to the level curves in Figure 122. Observe that the two setups are rather similar, but not exactly symmetrical.

Since the normal derivative  $\frac{\partial u}{\partial n}$  on all of the boundary is specified, the BVP does not have a unique solution. An arbitrary constant can be added and consequently the standard FEMoctave code will fail. If the additional condition

$$u_{mean} = \frac{1}{\text{area}(\Omega)} \iint_{\Omega} u dA = 0$$

is required, the problem has a unique solution again, and there is hope to obtain a good approximation by FEM. To get around this problem use the open and free source code of FEMoctave and modify the solver in `BVP2Dsym.m`. Add an additional equation

$$\sum_{i=1}^n u_i = 0$$

by one additional line, containing `n=size(A,1); A(n+1,:)=1; b(n+1)=0;`. It is a good idea to rename the function, e.g. to `BVP2DsymMean.m`.

## BVP2DsymMean.m

```

function u = BVP2DsymMean(Mesh, a, b0, f, gD, gN1, gN2)
    if nargin ~= 7
        print_usage();
    endif
    switch Mesh.type
        case 'linear' %% first order elements
            [A,b] = FEMEquation (Mesh, a, b0, 0, 0, f, gD, gN1, gN2); % compute with compiled code
        case 'quadratic' %% second order elements
            [A,b] = FEMEquationQuad(Mesh, a, b0, 0, 0, f, gD, gN1, gN2);
        case 'cubic' %% third order elements
            [A,b] = FEMEquationCubic(Mesh, a, b0, 0, 0, f, gD, gN1, gN2); % compute with compiled code
    endswitch
    %% add the zero mean condition
    n = size(A,1); A(n+1,:) = 1; b(n+1) = 0;
    u = FEMSolve(Mesh, A, b, gD); %% solve the linear system
endfunction

```

Using the current density  $\vec{J} = -\sigma \nabla u$  the vector fields in Figure 123 can be determined. With `FEMgriddata()` determine  $\nabla u$  and then multiply by the conductivity  $\sigma$  to obtain the current density  $\vec{J}$ . Using the same starting points along  $y = -0.4$  a few streamlines are shown.

- In Figure 123(a) the current takes the path of least resistance and is attracted by the highly conducting “heart”.
- In Figure 123(b) the current tries to avoid the “lung” section with the low conductivity.

If the conductivity would be constant in all of the domain  $\Omega$ , then the two graphics in Figure 123 would be perfectly symmetric.

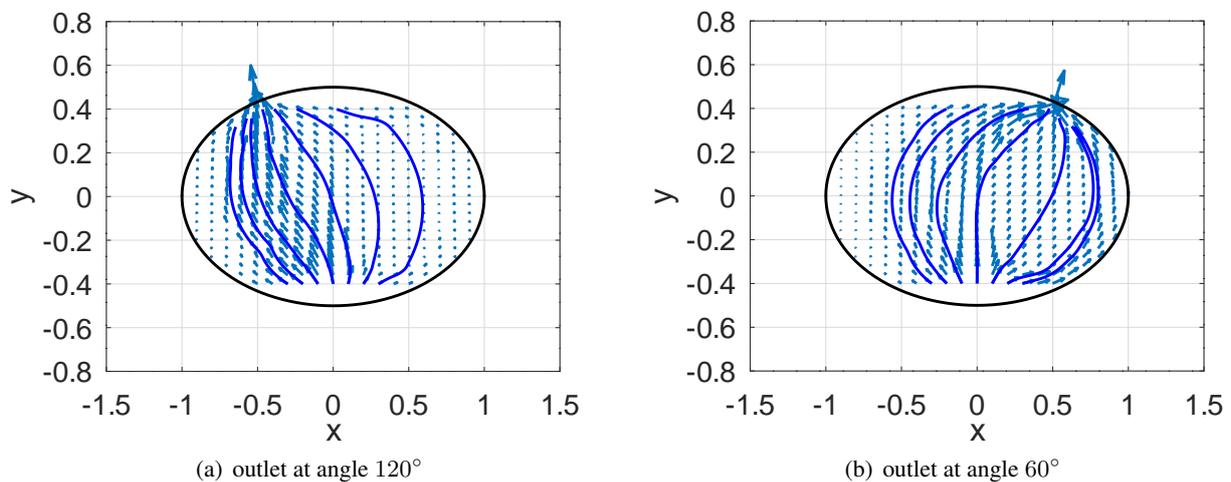


Figure 123: The vector field for the current density  $\vec{J}$  and a few streamlines

As a reference the situation of constant  $\sigma = 1$  is computed too and the resulting voltages on the boundary are shown in Figure 124. The deviations from the reference on the boundary  $\Gamma$  contain information about the conductivity inside of the domain  $\Omega$ . The deviations from the reference are shown in Figure 125. Many of those “measurements” allow to determine the Neumann to Dirichlet map, leading to the conductivity  $\sigma$  by an EIT algorithm.

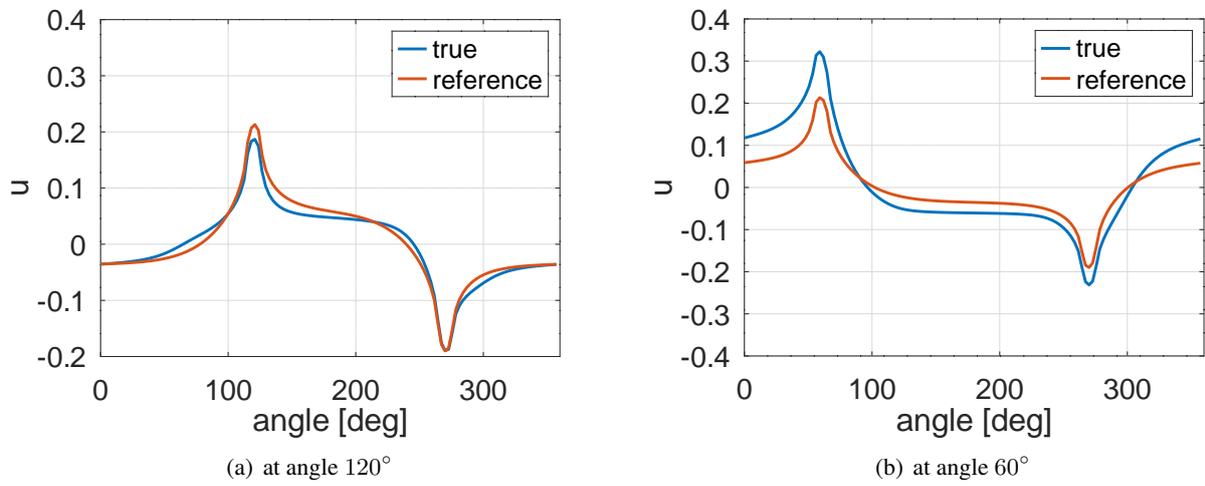


Figure 124: Voltage along the boundary

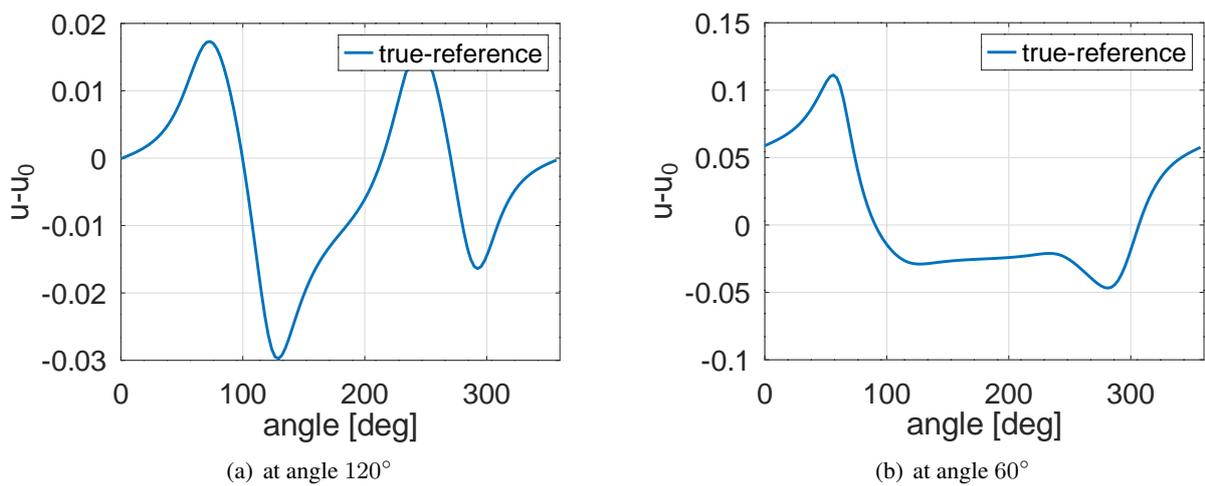


Figure 125: Differences of the voltage and the reference voltage

## EITforward.m

```

global Rx Ry dalpha my_angle
N = 2*64; %% number of angle segments
alpha = linspace(0,2*pi*(N-1)/N,N)'; Rx = 1; Ry = 0.5;
dalpha = 2*(alpha(2)-alpha(1));
x = Rx*cos(alpha); y = Ry*sin(alpha);
BC = -2*ones(size(x));
my_angle = 120 %% select the configuration, use 60 or 120

function res = sigma(xy,dummy)           %% the conductivity
    x = xy(:,1); y = xy(:,2);
    res = ones(size(x));
    res((x+0.5).^2+y.^2<=0.25^2) *= 4; %% heart on the left
    res((x-0.4).^2+y.^2<=0.35^2) *= 1/4; %% lung on the right
endfunction

FEMmesh = CreateMeshTriangle('EIT', [x,y,BC],0.003);
FEMmesh = MeshUpgrade(FEMmesh,'cubic');

figure(1); FEMtrimesh(FEMmesh,sigma(FEMmesh.nodes)); %% show the conductivity
    xlabel('x'); ylabel('y'); zlabel('\sigma'); view(20,50)

function res = flux_n(xy)               %% define the current density on the boundary
    global dalpha my_angle Rx Ry
    alpha = atan2(xy(:,2)/Ry,xy(:,1)/Rx); %% assure correct angle
    res = zeros(size(alpha));
    res(abs(alpha+pi/2) < dalpha) = -1;
    switch my_angle
        case 60
            res(abs(alpha-pi/3) < dalpha) = +1;
        case 120
            res(abs(alpha-pi*2/3) < dalpha) = +1;
    endswitch
    res = res./sqrt(Rx^2*sin(alpha).^2 + Ry^2*cos(alpha).^2); %% adjust for the arc length
endfunction

u_0 = BVP2DsymMean(FEMmesh, 1, 0,0,0,'flux_n',0); %% the reference result
u = BVP2DsymMean(FEMmesh,'sigma',0,0,0,'flux_n',0); %% the actual result

figure(2); FEMtrimesh(FEMmesh,u)           %% show the solution
    xlabel('x'); ylabel('y');

figure(3); clf; FEMtricontour(FEMmesh,u,41) %% show the contour levels
    hold on;
    plot([x;x(1)], [y;y(1)], 'k'); %% add the boundary
    hold off; xlabel('x'); ylabel('y'); axis equal

u_boundary = FEMgriddata(FEMmesh,u, x,y);
u_0_boundary = FEMgriddata(FEMmesh,u_0,x,y);

figure(4); plot(alpha*180/pi,u_boundary,alpha*180/pi,u_0_boundary)
    xlabel('angle [deg]'); ylabel('u'); xlim([0,360])
    legend('true','reference') %% show the voltages on the boundary

figure(5); plot(alpha*180/pi,u_boundary-u_0_boundary)

```

```

xlabel('angle [deg]'); ylabel('u-u_0'); xlim([0,360])
legend('true-reference') %% show the difference

%% create the vector field for the current density
[xx,yy] = meshgrid(linspace(-Rx,Rx,21),linspace(-0.8*Ry,0.8*Ry,21));
[ui,uxi,uyi] = FEMgriddata(FEMmesh,u,xx,yy);
conductivity = reshape(sigma([xx(:),yy(:)]),size(xx));
uxi = conductivity.*uxi; uyi = conductivity.*uyi;

figure(6); quiver(xx,yy,uxi,uyi,2) %% show the vector field
xlabel('x'); ylabel('y')
hold on; plot([x;x(1)], [y;y(1)], 'k'); %% add the boundary
hold off

%% create and show the streamlines
streamline(xx,yy,uxi,uyi, [-0.3 -0.2, -0.1, 0, 0.1, 0.2 0.3], -0.8*Ry*ones(1,7));

```

Since the condition

$$\oint_{\partial\Omega} J(s) ds = \oint_{\partial\Omega} \sigma \frac{\partial u}{\partial n} ds = 0$$

is critical it is a good idea to examine the numerical approximation of the flux through the boundary. For this use the normal vector

$$\vec{n} = \frac{1}{\sqrt{R_x^2 \sin^2 \alpha + R_y^2 \cos^2 \alpha}} \begin{pmatrix} R_y \cos \alpha \\ R_x \sin \alpha \end{pmatrix}$$

and then integrate over the segments where the flux is not zero

$$\int_{\text{section}} \langle \vec{n}, \nabla u \rangle ds.$$

To evaluate this numerically use `FEMgriddata()` to determine the values of the gradient  $(\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y})$  and then `trapz()` to perform a numerical integration. Observe that along the boundary the length segment is given by

$$ds = \sqrt{R_x^2 \sin^2 \alpha + R_y^2 \cos^2 \alpha} d\phi.$$

The code below leads to an inlet flux of  $\approx 0.1975$  and to outlet fluxes at either  $\approx 0.1980$  at  $60^\circ$  or  $\approx 0.1964$  at  $120^\circ$ .

#### AnalyzeBoundary.m

```

%% script to analyze the flux on the boundary
%% assumes that EITforward.m was run before
Angle = -90 % use 60, 120 or -90
Angle = deg2rad(Angle);
Section = pi/20; phi = Angle + linspace(-Section,+Section,100)';
x_b = 0.999*Rx*cos(phi); y_b = 0.999*Ry*sin(phi);

[u_boundary,ux_boundary,uy_boundary] = FEMgriddata(FEMmesh,u,x_b,y_b);

ds = sqrt(Rx^2*sin(phi).^2 + Ry^2*cos(phi).^2);
n = [Ry*cos(phi)./ds, Rx*sin(phi)./ds];

flux = (ux_boundary.*n(:,1) + uy_boundary.*n(:,2));
figure(1); plot(rad2deg(phi),flux)
xlabel('\phi [deg]'); ylabel('flux density')
TotalFlux = trapz(phi,flux.*ds)

```

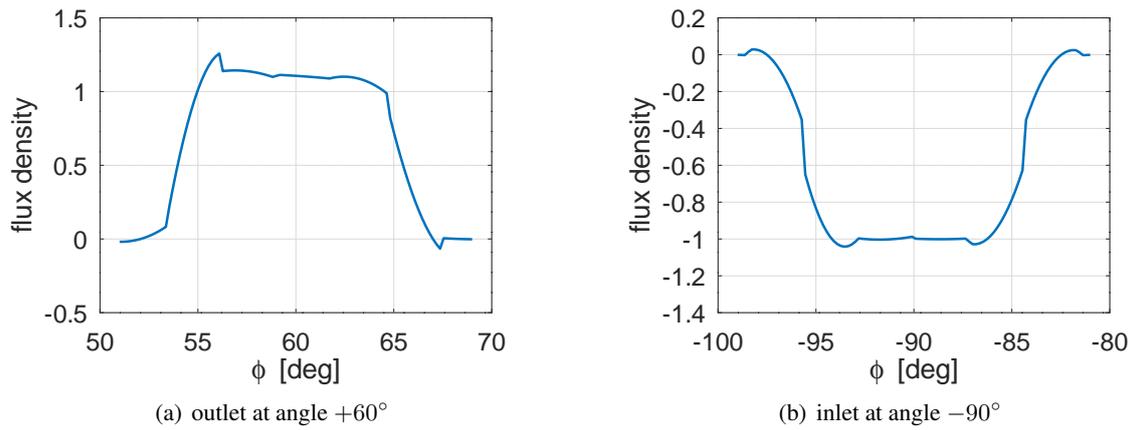


Figure 126: Flux density at inlet and outlet

### 9.16 Stretching of a beam

A beam with variable cross section  $A(x)$  is stretched by a force  $F$  applied to the right endpoint at  $x = L$ . With Young's modulus of elasticity the boundary value problem to be solved is

$$\frac{d}{dx} (E A(x) \frac{du(x)}{dx}) = F \quad \text{with} \quad u(0) = 0 \quad \text{and} \quad EA(L) \frac{du(L)}{dx} = F .$$

The solution is the displacement function  $u(x)$ . The resulting strain is given by the first derivative  $\frac{du(x)}{dx}$ . For a beam with a thinner midsection examine

$$EA(x) = \frac{1}{2} \left( 2 - \sin\left(\frac{x\pi}{L}\right) \right) .$$

In Figure 127 find the displacement  $u(x)$  and the resulting strain  $\frac{du(x)}{dx}$ , determined by 10 elements of equal length. It is clearly visible that the strain  $u'(x)$  is a piece-wise linear function. Thus just evaluating  $u'(x)$  at more grid points will not improve the appearance of the solution. Rerunning the code below with more elements (e.g.  $N=50$ ) will improve the situation.

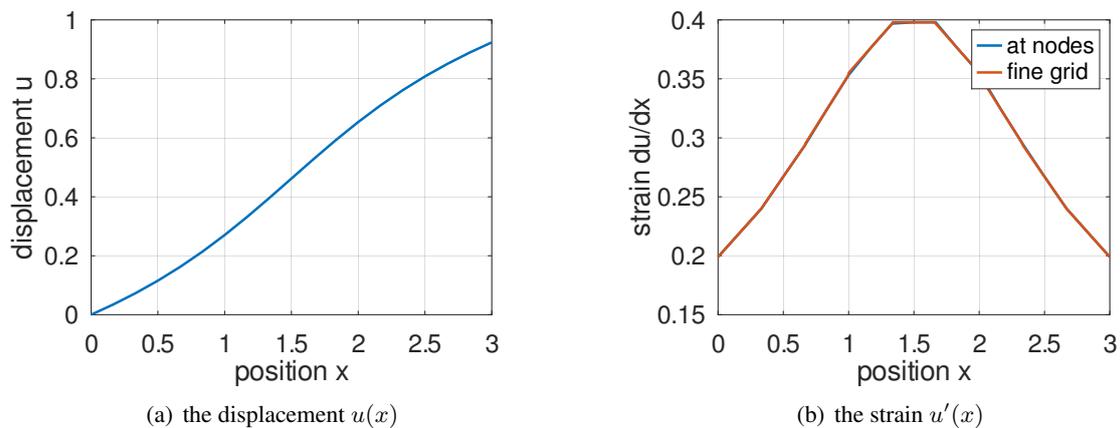


Figure 127: Displacement and strain for a beam with thinner midsection

**BeamStretch.m**

```

L = 3; F = 0.2; N = 10; EA = @(x) (2-sin(x/L*pi))/2;
[x,u] = BVP1D(linspace(0,L,N),EA,0,0,0,0,[F,0]);
figure(1); plot(x,u)
        xlabel('position x'); ylabel('displacement u')
[u2, strain] = pwquadinterp(x,u,x);           %% evaluation at nodes
x_fine = linspace(0,L,501);
[u_fine, strain_fine] = pwquadinterp(x,u,x_fine); %% interpolation to a fine grid
figure(2); plot(x, strain, x_fine, strain_fine)
        xlabel('position x'); ylabel('strain du/dx')
        legend('at nodes', 'fine grid')

```

**9.17 Keller's nonlinear boundary value problems**

In [Kell92, p. 317] the boundary value problem

$$-u''(x) = -e^{u(x)} \quad \text{with} \quad u(-1) = u(1) = 0$$

is examined. The exact solution is given by

$$u(x) = \ln \left( \frac{c^2}{1 + \cos(cx)} \right),$$

where the value of the constant  $c$  is determined as solution of the equation  $c^2 = 1 + \cos c$ . Use Newton's method to find  $c \approx 1.1765019$ . There are two possible approaches to solve this nonlinear boundary value problem, partial successive substitution and Newton's method. The command `BVP1DNL()` is based on Newton's method.

**9.17.1 Partial successive substitution**

Start with an initial guess, e.g.  $u_0(x) = 0$ , leading to  $-\exp(u_0) = -1$ . Then use the iteration

$$-u''_{n+1}(x) = -e^{u_n(x)}.$$

The coding in FEMoctave is rather straightforward, shown in the code below. Find the solution by (partial) successive substitution in Figure 128(a) and the difference to the exact solution after 5 iterations in Figure 128(b).

**Keller.m**

```

%% successive substitution
N = 51; interval = linspace(-1,1,N)';
c = 1.176501939901833; %% or use the solver
opt.TolFun = 1e-15; opt.TolX = 1e-15; c = fsolve(@(c)1+cos(c)-c^2,1,opt);

[x,u] = BVP1D(interval,1,0,0,1,-1,0,0);
u_exact = log(c^2./(1+cos(c*x)));
figure(1); plot(x,u); xlabel('x'); ylabel('u_1')
for jj = 1:4;
    [x,u] = BVP1D(interval,1,0,0,1,-exp(u),0,0);
    figure(2); plot(x,u,x,u_exact); xlabel('x'); ylabel('u')
        legend('FEM', 'exact', 'location', 'north')
    figure(3); plot(x,u-u_exact); xlabel('x'); ylabel('u')
        legend('FEM-exact')
    pause(0.2)
endfor

```

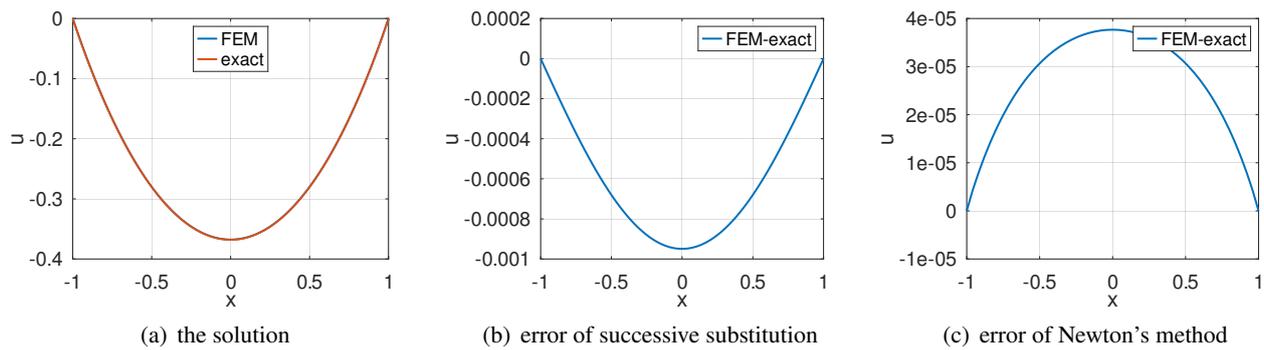


Figure 128: The solution of Keller's boundary value problem and the difference of the approximations by successive substitution and Newton's method

### 9.17.2 Newton's method

With an approximate initial guess  $u_0(x)$  (start with  $u_0(x) = 0$ ) search a sequence of new solutions of the form  $u_{n+1}(x) = u_n(x) + \phi(x)$ . Use the idea of a linear approximation to solve  $0 = f(x_n + \phi) \approx f(x_n) + f'(x_n)\phi$  to find  $\phi = -\frac{f(x_n)}{f'(x_n)}$ , leading to the usual iteration formula  $x_{n+1} = x_n + \phi = x_n - \frac{f(x_n)}{f'(x_n)}$ . Examine the linear boundary value problem for the unknown function  $\phi(x)$ . The Taylor approximation  $e^{u+\phi} \approx e^u + e^u\phi = e^u(1+\phi)$  leads to the BVP for  $\phi(x)$ .

$$\begin{aligned} -u_n''(x) - \phi''(x) &= -e^{u_n(x)+\phi(x)} \approx -e^{u_n(x)}(1+\phi(x)) \\ -\phi''(x) + e^{u_n(x)}\phi(x) &= u_n''(x) - e^{u_n(x)} \quad \text{with} \quad \phi(-1) = \phi(1) = 0 \end{aligned}$$

Use `[du, ddu] = FEM1DEvaluateDu(xn, u)` to find the values of the first and second derivatives at the nodes. The coefficient  $e^{u_n(x)}$  of the term  $\phi(x)$  depends on the previous solution  $u_n(x)$  and its values have to be known at the Gauss integration points. Find the Gauss points with the help of the function `FEM1DGaussPoints()` and use `uGauss=pwquadinterp(xn, u, xGauss)` to evaluate  $u_n(x)$  at the Gauss points. Find the solution by Newton's method in Figure 128(a) and the difference to the exact solution after 5 iterations in Figure 128(c).

#### Keller.m

```

%% Newton's method
N = 51; interval = linspace(-1,1,N)';
c = 1.176501939901833; %% or use the solver
opt.TolFun = 1e-15; opt.TolX = 1e-15; c = fsolve(@(c)1+cos(c)-c^2,1,opt);

[x,u] = BVP1D(interval,1,0,0,1,-1,0,0);
u_exact = log(c^2./(1+cos(c*x)));
figure(1); plot(x,u); xlabel('x'); ylabel('u_1')
xGauss = FEM1DGaussPoints(x);
for jj = 1:4
    [du,ddu] = FEM1DEvaluateDu(x,u);
    RHS = + ddu - exp(u);
    uGauss = pwquadinterp(x,u,xGauss); %% evaluate u at Gauss points
    u_coeff = exp(uGauss);
    [x,phi] = BVP1D(interval,1,0,u_coeff,1,RHS,0,0);
    disp(sprintf('max(abs(phi)) = %g, max(abs(RHS)) = %g',max(abs(phi)), max(abs(RHS))))
    u = u + phi;
    figure(2); plot(x,u,x,u_exact); xlabel('x'); ylabel('u')
    legend('FEM','exact','location','north')

```

```

figure(3); plot(x,u-u_exact); xlabel('x'); ylabel('u')
            legend('FEM-exact')
pause(0.2)
endfor

```

### 9.17.3 Using BVP1DNL()

Newton's method can be used with very few lines of code. Provide the partial derivative

$$\frac{\partial}{\partial u} f(x, u) = \frac{\partial}{\partial u} (-\exp(u)) = -\exp(u)$$

and start with the naive initial guess  $u_0(x) = 0$ . The graphical result and the information provided in `inform` indicate the the algorithm converged with 4 iterations and the size of the last correction was of the order  $1.4 \cdot 10^{-6}$ . The RMS of the error is approximately  $3 \cdot 10^{-8}$ , only computable since we know the exact solution. The graph shows an actual maximal error of  $\approx 2 \cdot 10^{-9}$ .

#### Keller.m

```

%% use BVP1DNL
N = 51; interval = linspace(-1,1,N)';
c = 1.176501939901833; %% or use the solver
opt.TolFun = 1e-15; opt.TolX = 1e-15; c = fsolve(@(c)1+cos(c)-c^2,1,opt);

f = {@(x,u)-exp(u) , @(x,u)-exp(u)};
[x,u,inform] = BVP1DNL(interval,1,0,0,1,f,0,0,0,'Tol',1e-5,'display','iter');
u_exact = log(c^2./(1+cos(c*x)));
figure(2); plot(x,u,x,u_exact); xlabel('x'); ylabel('u')
            legend('FEM','exact','location','north')
figure(3); plot(x,u-u_exact); xlabel('x'); ylabel('u')
            legend('FEM-exact')

inform
endswitch
RMS_difference = norm(u-u_exact)/sqrt(length(u))
-->
iteration 1, RMS(correction) = 9.323090e-02, RMS(phi) = 9.468330e-02
iteration 2, RMS(correction) = 3.300574e-04, RMS(phi) = 3.300764e-04
iteration 3, RMS(correction) = 4.332861e-09, RMS(phi) = 4.332859e-09

inform = scalar structure containing the fields:
    info      = 1
    iter      = 3
    AbsError  = 4.3329e-09

RMS_difference = 1.6647e-09

```

### 9.17.4 A very similar problem with multiple solutions

At first sight the nonlinear boundary value problem

$$-u''(x) = \frac{1}{2} e^{u(x)} \quad \text{with} \quad u(-1) = u(1) = 0$$

found in [Kell92, p. 150] is very similar to the above. But it turns out<sup>33</sup> that this problem has two solutions, one with  $u(0) \approx 0.3$  and another solution with  $u(0) \approx 3$ . By selecting appropriate initial functions  $u_0(x) = u_0(1 - x^2)$  the algorithm used in `BVP1DNL()` will determine either one of these two solutions.

```

Keller2.m
y01 = 0.3 %% more accurate y0 = 0.3290;
y02 = 3.0 %% more accurate y0 = 2.8955;
RHS = {@(x,y)0.5*exp(y), @(x,y)0.5*exp(y)};
interval = linspace(-1,1,21);
BCleft = 0; BCright = 0;
[x1,y1] = BVP1DNL(interval,1,0,0,1,RHS,BCleft,BCright,@(x)y01*(1-x.^2),
    'MaxIter',30,'Display','iter');
[x2,y2] = BVP1DNL(interval,1,0,0,1,RHS,BCleft,BCright,@(x)y02*(1-x.^2),
    'MaxIter',30,'Display','iter');
figure(1); plot(x1,y1,x2,y2); xlabel('x'); ylabel('y(x)')

```

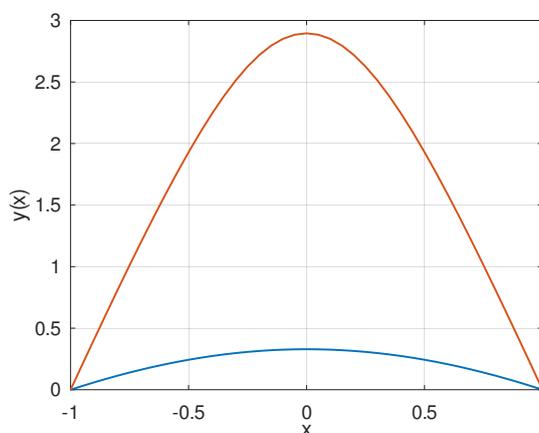


Figure 129: Two solutions of  $-u''(x) = \frac{1}{2} \exp(u(x))$  with  $u(-1) = u(1) = 0$

### 9.18 A mathematical pendulum problem

With what angular velocity does a pendulum have to start at  $u(0) = 0$  to reach the maximal angle at  $t = T$ ? What is the resulting maximal angle? The boundary value problem leading to the answers is

$$\ddot{u}(t) = -\sin(u(t)) \quad \text{with} \quad u(0) = 0 \quad \text{and} \quad \dot{u}(T) = 0. \quad (93)$$

This BVP always has the trivial solution  $u(t) = 0$ . Since  $\sin(u) \leq u$  for  $u \geq 0$  the acceleration for the mathematical pendulum is smaller than the one for the physical pendulum  $\ddot{u}(t) = -u(t)$ . The solution of the physical pendulum is  $u(t) = c \sin(t)$ . As consequence require  $T > \frac{\pi}{2}$  to obtain a nontrivial solution for the BVP (93). Conservation of energy requires<sup>34</sup>

$$\text{initial kinetic energy} = \frac{1}{2} (\dot{u}(0))^2 = (1 - \cos(u(T))) = \text{final potential energy.}$$

<sup>33</sup>Examine the initial value problem  $u''(x) = -\frac{1}{2} \exp(u(x))$  with  $u'(0) = 0$  and  $u(0) = u_0$ . Examine the value  $u(1)$  as function of  $u_0$  and use a graph or `fsolve()` to find two zeros at  $u_0 \approx 0.32895$  and  $u_0 \approx 2.8955$ .

<sup>34</sup>Conservation of energy can be used directly to determine the maximal angle as function of the initial velocity. The travel time  $T$  is then given by a singular integral.

$$\frac{1}{2} v^2 = (1 - \cos(u_{max})) - (1 - \cos(u)) = \cos(u) - \cos(u_{max})$$

The code `Pendulum.m` shown below solves the BVP (93) for  $T = 2$  and the conservation of energy is verified.

```

Pendulum.m
N = 1001; interval = linspace(0,T,N)';
T = 2.0; BCleft = 0; BCright = [0,0];
f = {@(t,u)sin(u), @(t,u)cos(u)};
[t,u] = BVP1DNL(interval,1,0,0,1,f,BCleft,BCright,@(t)t,
    'Display','off','Tol',1e-8);

figure(1); plot(t,u); xlabel('time t'); ylabel('angle u')
v = FEM1DEvaluatedu(t,u);
disp(sprintf('For T = %g: initial angular velocity v(0) = %g, maximal angle u(T) = %g',
    T,v(1),u(end)))
KineticEnergy = v(1)^2/2; Potential = 1-cos(u(end));
disp(sprintf('Kinetic energy at t=0: %g, potential energy at t=T: %g, difference: %g',
    KineticEnergy,Potential,KineticEnergy-Potential))
-->
For T = 2: initial angular velocity v(0) = 1.60481, maximal angle u(T) = 1.86263
Kinetic energy at t=0: 1.2877, potential energy at t=T: 1.2877, difference: 8.5831e-07

```

### 9.19 A BVP with multiple nonlinear contributions

In [AtkiHan09] find Exercise 5.4.1 on page 241, a nonlinear boundary value problem with multiple nonlinear contributions.

$$\begin{aligned}
 -u''(x) + u'(x)u(x) + u^3(x) &= e^x \\
 u(0) = 1 \quad , \quad u'(2) &= 2
 \end{aligned}$$

A first attempt at solving this BVP with `BVP1DNL()` directly will fail. The algorithm does not converge, caused by the initial guess  $u_0(x)$ , which has to be close to the solution. It turns out that the contribution  $u^3(x)$  leads to the blowup of the Newton algorithm. One possible way to solve the BVP is to parameterize the critical contribution. Introduce a parameter  $0 \leq \alpha \leq 1$  and examine

$$\begin{aligned}
 -u''(x) &= f_\alpha(x, u(x), u'(x)) := -u'(x)u(x) - \alpha u^3(x) + e^x \\
 f_\alpha(x, u, u') &= -u'u - \alpha u^3 + e^x \\
 \frac{\partial}{\partial u} f_\alpha(x, u, u') &= -u' - \alpha 3u^2 \\
 \frac{\partial}{\partial u'} f_\alpha(x, u, u') &= -u
 \end{aligned}$$

Then increase the value of  $\alpha$  step by step from 0 to 1. Find the result from the code `AtkinsonHan.m` below in Figure 130.

```

AtkinsonHan.m
N = 501; interval = linspace(0,1,N)'; BCleft = 1; BCright = [2,0];
u = 0;
figure(2); clf; hold on; box on
for alpha = linspace(0,1,5)

```

$$\begin{aligned}
 \frac{du}{dt} = v(u) &= \sqrt{2(\cos(u) - \cos(u_{max}))} \\
 T &= \int_0^{u_{max}} \frac{1}{v(u)} du = \int_0^{u_{max}} \frac{1}{\sqrt{2(\cos(u) - \cos(u_{max}))}} du
 \end{aligned}$$

The code in `Pendulum.m` evaluates this integral.

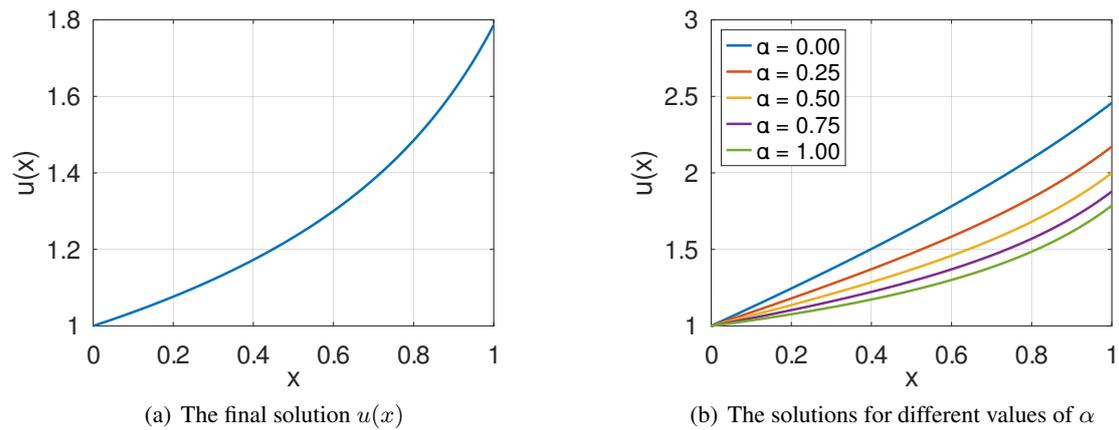


Figure 130: A BVP with multiple nonlinear contributions

```

alpha
f = {@(x,u,du)-u.*du - alpha*u.^3 + exp(x),@(x,u,du)-du -alpha*3*u.^2 ,@(x,u,du)-u};
[x,u,inform] = BVP1DNL(interval,1,0,0,1,f,BCleft,BCright,u,
    'Display','iter');
plot(x,u) ;xlabel('x'); ylabel('u(x)')
pause(0.1)
endfor
legend('\alpha = 0.00','\alpha = 0.25','\alpha = 0.50','\alpha = 0.75','\alpha = 1.00',
    'location','northwest')
figure(1); plot(x,u) ; xlabel('x'); ylabel('u(x)')

```

The output generated by the above code nicely illustrates the quadratic convergence of Newton's algorithm, i.e. the number of non-changing digits is approximately doubled at each step, after an initial search.

```

alpha = 0
iteration 1, RMS(correction) = 1.271951e+00, RMS(phi) = 9.217815e-01
iteration 2, RMS(correction) = 1.859463e-01, RMS(phi) = 1.978042e-01
iteration 3, RMS(correction) = 5.875778e-03, RMS(phi) = 5.886006e-03
iteration 4, RMS(correction) = 5.074170e-06, RMS(phi) = 5.074184e-06
alpha = 0.2500
iteration 1, RMS(correction) = 2.750162e-01, RMS(phi) = 5.594625e-01
iteration 2, RMS(correction) = 1.661185e-01, RMS(phi) = 2.075803e-01
iteration 3, RMS(correction) = 2.746584e-02, RMS(phi) = 2.827483e-02
iteration 4, RMS(correction) = 5.465197e-04, RMS(phi) = 5.468251e-04
iteration 5, RMS(correction) = 2.064430e-07, RMS(phi) = 2.064488e-07
alpha = 0.5000
iteration 1, RMS(correction) = 1.734538e-01, RMS(phi) = 4.004352e-01
iteration 2, RMS(correction) = 1.315342e-01, RMS(phi) = 1.798543e-01
iteration 3, RMS(correction) = 3.380237e-02, RMS(phi) = 3.581552e-02
iteration 4, RMS(correction) = 1.465643e-03, RMS(phi) = 1.469075e-03
iteration 5, RMS(correction) = 2.503296e-06, RMS(phi) = 2.503313e-06
alpha = 0.7500
iteration 1, RMS(correction) = 1.283952e-01, RMS(phi) = 3.183453e-01
iteration 2, RMS(correction) = 1.084098e-01, RMS(phi) = 1.567254e-01
iteration 3, RMS(correction) = 3.454831e-02, RMS(phi) = 3.743495e-02
iteration 4, RMS(correction) = 2.185036e-03, RMS(phi) = 2.195112e-03

```

```

iteration 5, RMS (correction) = 7.654415e-06, RMS (phi) = 7.654543e-06
alpha = 1
iteration 1, RMS (correction) = 1.031818e-01, RMS (phi) = 2.665817e-01
iteration 2, RMS (correction) = 9.226889e-02, RMS (phi) = 1.381142e-01
iteration 3, RMS (correction) = 3.318699e-02, RMS (phi) = 3.653564e-02
iteration 4, RMS (correction) = 2.598465e-03, RMS (phi) = 2.615903e-03
iteration 5, RMS (correction) = 1.360277e-05, RMS (phi) = 1.360325e-05
iteration 6, RMS (correction) = 3.689710e-10, RMS (phi) = 3.715615e-10

```

Another approach to find good initial values is the parametrization

$$f_{\alpha}(x, u, u') = \alpha (-u' u - u^3) + e^x .$$

With smaller steps for  $0 \leq \alpha \leq 1$  this parametrization leads to the same final result.

## 9.20 Fisher's equation

Examine Fisher's equation (named after statistician and biologist Ronald Fisher), also known as the Kolmogorov–Petrovsky–Piskunov equation, or shorter Fisher-KPP equation.

$$\frac{\partial}{\partial t} u(x, t) - \frac{\partial^2}{\partial x^2} u(x, t) = f(u(x, t)) = u(x, t) (1 - u(x, t)) \quad \text{with } u(-\infty) = 1 \text{ and } u(+\infty) = 0 . \quad (94)$$

This initial boundary value problem can exhibit traveling wave solutions that switch between the two equilibrium states given by  $f(u) = 0$ , i.e. at  $u = 0$  and  $u = 1$ . Such equations occur, e.g., in ecology, physiology, combustion, crystallization, plasma physics, and in general phase transition problems.

When searching for a traveling wave solution with speed  $c$  use  $u(x, t) = u(x - ct)$  and the above partial differential equation turns into an ordinary differential equation

$$-c u'(x) - u''(x) = u(x) (1 - u(x)) \quad \text{with } u(-\infty) = 1 \text{ and } u(+\infty) = 0 . \quad (95)$$

This ODE does not have a unique solution. Each translation  $u(x + D)$  of a solution  $u(x)$  is a solution too. Thus a naive usage of `BVP1DNL()` will not succeed. It also turns out that the condition  $u(+\infty) = 0$  leads to numerical problems. A solution of (95) was generated with the ideas and algorithm spelled out below.

- Use  $c = \frac{6}{\sqrt{5}} \approx 2.0412^{35}$  and select a value of  $M = 20 \gg 1$ . Then examine two boundary value problems:

$$\begin{aligned} -c u'_n - u''_n &= u_n (1 - u_n) \quad \text{on } -M < x < 0 \text{ and } u_n(-M) = 1 \text{ and } u_n(0) = \frac{1}{4} \\ -c u'_p - u''_p &= u_p (1 - u_p) \quad \text{on } 0 < x < +M \text{ and } u_p(0) = \frac{1}{4} \text{ and } u_p(+M) = \text{ValueAtM} \end{aligned}$$

Consider the value of  $u_p(+M)$  as parameter and determine its value such that  $u'_n(0) = u'_p(0)$ . Use a graphical hint in Figure 131(a) to select an initial interval for the Octave command `zero()` to determine a numerical solution of  $u'_n(0) = u'_p(0)$ . The code below leads to the value  $u_p(+20) \approx 8.0826 \cdot 10^{-8}$ .

- With the obtained optimal value of  $u(+M)$  solve the two boundary value problems on  $-M < x < 0$  and  $0 < x < +M$ . Since  $u_n(0) = u_p(0)$  and  $u'_n(0) = u'_p(0)$  these two solutions can be patched together at  $x = 0$  to obtain a solution of the ODE (95) on  $-M < x < M$ , leading to Figure 131(b) and the numerical confirmation of  $u'_n(0) = u'_p(0)$ .

```

first derivatives: u' (x=0) = -0.102206 and u' (x=0) = -0.102206, difference = 6.6294e-10

```

<sup>35</sup>For  $c = \frac{6}{\sqrt{5}}$  an exact solution is given by  $u(x) = (1 + \exp(\frac{-x}{\sqrt{6}}))^{-2}$ . This was also useful to find the good value for  $u_p(+M)$ .

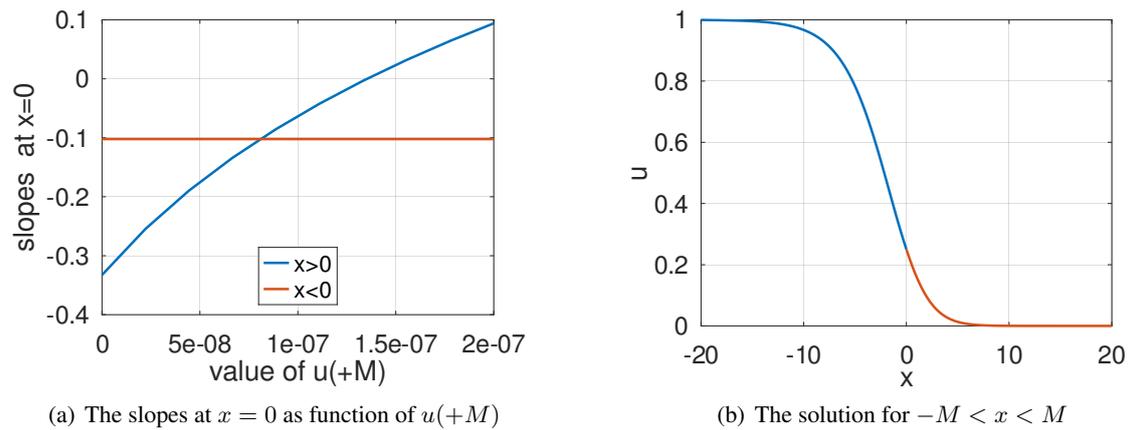


Figure 131: The solution of Fisher's equation

## Fisher.m

```
function Fisher()
c = 5/sqrt(6); M = 20; N = 50; Join = 0.25;
f = {@(x,u) (u.*(1-u)), @(x,u) (1-2*u)};
interval = linspace(0,M,N)'; intervaln = linspace(-M,0,N)';
%% find the parameter
function slopes = FindSlopes(valueM,graphs)
BCleft = Join; BCright = valueM;
[xp,u0] = BVP1D(interval,1,-c,0,1,0,BCleft,BCright);
[xp,up,inform] = BVP1DNL(interval,1,-c,0,1,f,BCleft,BCright,u0);
BCleft = 1; BCright = Join;
[xn,u0] = BVP1D(intervaln,1,-c,0,1,0,BCleft,BCright);
[xn,un,inform] = BVP1DNL(intervaln,1,-c,0,1,f,BCleft,BCright,u0);
dup = FEM1DEvaluateDu(xp,up); dun = FEM1DEvaluateDu(xn,un);
slopes = [dun(end),dup(1)];
endfunction

%%ValueMList = linspace(0,1e-3,10); %% for M = 10
ValueMList = linspace(0,2e-7,10); %% for M = 20
for jj = 1:length(ValueMList)
s = FindSlopes(ValueMList(jj)); sn(jj) = s(1); ,sp(jj) = s(2);
endfor

figure(1); plot(ValueMList,sp,ValueMList,sn);
xlabel('value of u(+M)'); ylabel('slopes at x=0');
legend('x>0','x<0','location','south')

%% find the best value at x=+M
SameSlope = @(valueM) diff(FindSlopes(valueM));
ValueAtM = fzero(SameSlope,[5,9]*1e-8) %% for M = 20

%% solve the two BVPs
BCleft = Join; BCright = ValueAtM; %% for M = 20
u0 = @(x) (xp-M).^6/M^6*Join;
[xp,up,inform] = BVP1DNL(interval,1,-c,0,1,f,BCleft,BCright,u0,
'Display','off','tol',1e-8);
```

```

figure(11); plot(xp,u0(xp),xp,up); xlabel('x'); ylabel('u');
            legend('u_0','u','location','northeast')
intervaln = linspace(-M,0,N)'; BCleft = 1; BCright = Join;
u0 = @(x) 1-(x+M).^4/M^4*(1-Join);
[xn,un,inform] = BVP1DNL(intervaln,1,-c,0,1,f,BCleft,BCright,u0,
                        'Display','off','tol',1e-8);
figure(12); plot(xn,u0(xn),xn,un); xlabel('x'); ylabel('u');
            legend('u_0','u','location','southwest')
figure(2); plot(xn,un,xp,up); xlabel('x'); ylabel('u');

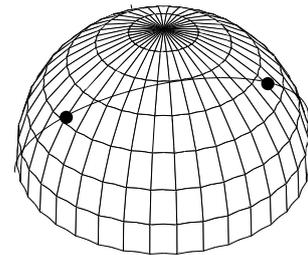
%% evaluate the derivatives at x=0
[dup, ddup] = FEM1DEvaluateDu(xp,up);
[dun, ddun] = FEM1DEvaluateDu(xn,un);
disp(sprintf("first derivatives: u'(x-0) = %g and u'(x+0) = %g, difference = %g",
            dun(end), dup(1), dun(end)-dup(1)))
endfunction

```

## 9.21 From Salt Lake City to Zürich, the shortest connection on a sphere

On the right find a graphics of the northern hemisphere (radius  $R = 6300$  km) with the two cities Zürich and Salt Lake City shown. The geographic data of the towns is given by

	latitude $\frac{\pi}{2} - \theta$	longitude $\phi$
Salt Lake City	$41^\circ$	$-112^\circ$
Zürich	$47^\circ$	$+8^\circ$



Airlines prefer to use the shortest connection possible and will fly north first, and then turn back south. Use calculus of variations to find the corresponding boundary value problem and verify that the shortest connection follows a great circle.

To examine this problem use spherical coordinates.

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = R \begin{pmatrix} \sin \theta \cos \phi \\ \sin \theta \sin \phi \\ \cos \theta \end{pmatrix}$$

The longitude equals the angle  $\phi$  and the latitude is given by  $\frac{\pi}{2} - \theta$ . Represent the connection from Salt Lake City to Zürich by writing the latitude as a function of the longitude, i.e.  $\theta = u(\phi)$ . To find the total length  $L$  express the length element  $dl$  in terms of spherical coordinates. Use basic calculus to find

$$dl = R \sqrt{\sin^2 \theta d\phi^2 + d\theta^2} = R \sqrt{\sin^2(u(\phi)) + (u'(\phi))^2} d\phi$$

$$L(u) = R \int_{\phi_0}^{\phi_1} \sqrt{\sin^2(u(\phi)) + (u'(\phi))^2} d\phi = R \int_{\phi_0}^{\phi_1} g(\phi, u(\phi), u'(\phi)) d\phi.$$

To be determined is the function  $\theta = u(\phi)$  such that the above functional  $L(u)$  is minimized, respecting the boundary conditions  $u(\phi_{SLC}) = \theta_{SLC}$  and  $u(\phi_{ZH}) = \theta_{ZH}$ . Use calculus of variations to find the Euler–Lagrange equation

$$-\frac{d}{d\phi} \left( \frac{\partial}{\partial u'} g \right) = -\frac{\partial}{\partial u} g$$

$$-\frac{d}{d\phi} \left( \frac{u'}{\sqrt{\sin^2(u) + (u')^2}} \right) = -\frac{\cos(u) \sin(u)}{\sqrt{\sin^2(u) + (u')^2}}. \quad (96)$$

This is a nonlinear boundary value problem for the function  $u(\phi)$ .

### 9.21.1 A solution based on successive substitution

With `FEMoctave` use the method of successive substitution and hope for convergence. Start with a straight line connection, i.e. the solution of  $\frac{d^2}{d\phi^2} u(\phi) = 0$ . Then hope for the iteration  $u_{n-1} \rightarrow u_n$ , where  $u_n$  solves

$$\frac{d}{d\phi} \left( \frac{u'_n}{\sqrt{\sin^2(u_{n-1}) + (u'_{n-1})^2}} \right) = \frac{\cos(u_{n-1}) \sin(u_{n-1})}{\sqrt{\sin^2(u_{n-1}) + (u'_{n-1})^2}}$$

to converge. To use the command `BVP1D()` for this iteration the following steps are performed.

- Determine the Gauss integration points by `phiG = FEM1DGaussPoints(phi)`.
- Evaluate  $u(\phi) = \theta(\phi)$  and  $u'(\phi) = \frac{d}{d\phi} \theta(\phi)$  at the Gauss points with the help of `[thetaG, DthetaG] = pwquadinterp(phi, theta, phiG)`
- Evaluate the coefficients  $a(\phi)$  and  $d(\theta)$  at the Gauss points by
  - `a = (sin(thetaG).^2 + DthetaG.^2).^(-1/2)`
  - `d = sin(thetaG).*cos(thetaG)./sqrt(sin(thetaG).^2 + DthetaG.^2)`
- Call `[phi, theta] = BVP1D(interval, a, 0, 0, d, -1, BCleft, BCright)`

To verify that the shortest connection is on a great circle use the vectors  $\vec{p}_{SLC}$  and  $\vec{p}_{ZH}$  connecting the center of the earth to the two cities. Then determine the vector  $\vec{n}$  orthogonal to the plane of the great circle connecting Salt Lake City and Zürich.

$$\vec{n} = \frac{1}{\|\vec{p}_{SLC} \times \vec{p}_{ZH}\|} \vec{p}_{SLC} \times \vec{p}_{ZH}$$

The scalar product  $\langle \vec{p}, \vec{n} \rangle$  will determine the distance of the point  $\vec{p}$  from the plane containing the great circle. In Figure 132 find the results of successive substitution performed by the code `SaltLakeCity2Zuerich.m` below. The (slow) animation also shows the length  $L$  for each iteration.

```
Iteration 1: L = 9510.058645 km
Iteration 2: L = 8532.915369 km
Iteration 3: L = 8491.805282 km
Iteration 4: L = 8485.252338 km
Iteration 5: L = 8483.738234 km
```

- Figure 132(a) shows the latitude as function of the longitude. The result of the first computations stays clearly to far south. The sequence of solutions is monotonically increasing and converging.
- Figure 132(b) shows the resulting curves in space. The length unit is the radius  $R$  of the earth.
- Figure 132(c) shows the distance from the plane of the great circle connecting Salt Lake City and Zürich. This distance converges to zero, i.e. the shortest connection is along a great circle.

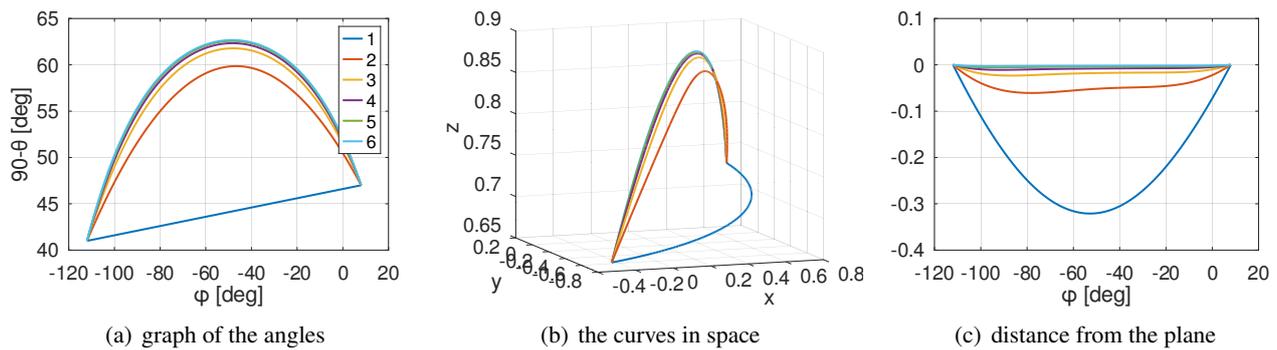


Figure 132: The connections between Salt Lake City and Zürich. The plots of latitude as function of longitude, the curves in space on the earth as sphere with radius 1 and the distances from the plane with the great circle.

### SaltLakeCity2Zuerich.m

```

N = 51; R = 6300;
Angles_ZH = [8,90-47]/180*pi; Angles_SLC = [-112,90-41]/180*pi;
Vec_ZH = R*[sin(Angles_ZH(2))*cos(Angles_ZH(1));
            sin(Angles_ZH(2))*sin(Angles_ZH(1));cos(Angles_ZH(2))];
Vec_SLC = R*[sin(Angles_SLC(2))*cos(Angles_SLC(1));
            sin(Angles_SLC(2))*sin(Angles_SLC(1));cos(Angles_SLC(2))];
n = cross(Vec_SLC,Vec_ZH); n = n/norm(n);
BCleft = Angles_SLC(2); BCright = Angles_ZH(2);
interval = linspace(Angles_SLC(1),Angles_ZH(1),N);
[phi,theta] = BVP1D(interval,1,0,0,1,0,BCleft,BCright);
figure(1); plot(phi/pi*180,90-theta/pi*180);
            xlabel('\phi [deg]'); ylabel('90-\theta [deg]'); hold on
x = sin(theta).*cos(phi); y = sin(theta).*sin(phi); z = cos(theta);
Dtheta = FEM1DEvaluateDu(phi,theta);
figure(2); plot3(x,y,z);
            xlabel('x'); ylabel('y'); zlabel('z'); view([-20,10]); hold on
disp(sprintf("Iteration 1: L = %f km",R*trapez(phi,sqrt(sin(theta).^2+Dtheta.^2)))
figure(3); plot(phi*180/pi,[x,y,z]*n);
            xlabel('\phi [deg]'); hold on;
phiG = FEM1DGaussPoints(phi);
for jj = 2:5;
    [thetaG,DthetaG] = pwquadinterp(phi,theta,phiG);
    a = (sin(thetaG).^2 + DthetaG.^2).^(-1/2);
    d = -sin(thetaG).*cos(thetaG)./sqrt(sin(thetaG).^2 + DthetaG.^2);
    [phi,theta] = BVP1D(interval,a,0,0,d,1,BCleft,BCright);
    Dtheta = FEM1DEvaluateDu(phi,theta);
    disp(sprintf("Iteration %i: L = %f km",
                jj,R*trapez(phi,sqrt(sin(theta).^2+Dtheta.^2)))
    figure(1); plot(phi/pi*180,90-theta/pi*180);
    x = sin(theta).*cos(phi); y = sin(theta).*sin(phi); z = cos(theta);
    figure(2); plot3(x,y,z);
    figure(3); plot(phi*180/pi,[x,y,z]*n);
    pause(0.2)
endfor %% jj
figure(1); legend('1','2','3','4','5'); hold off;
figure(2); hold off; figure(3); hold off

```

**9.21.2 A solution using BVP1DNL()**

When solving the BVP (96)

$$-\frac{d}{d\phi} \left( \frac{u'}{\sqrt{\sin^2(u) + (u')^2}} \right) = -\frac{\cos(u) \sin(u)}{\sqrt{\sin^2(u) + (u')^2}}.$$

with the help of BVP1DNL() the partial derivatives of the right hand side with respect to  $u$  and  $u'$  will be used.

$$\begin{aligned} f(\phi, u, u') &= -\frac{\cos(u) \sin(u)}{\sqrt{\sin^2(u) + (u')^2}} \\ f_u = \frac{\partial}{\partial u} f(\phi, u, u') &= -\frac{1 - 2 \sin^2(u)}{\sqrt{\sin^2(u) + (u')^2}} + \frac{\cos^2(u) \sin^2(u)}{\sqrt{\sin^2(u) + (u')^2}^3} \\ f_{u'} = \frac{\partial}{\partial u'} f(\phi, u, u') &= \frac{\cos(u) \sin(u) u'}{\sqrt{\sin^2(u) + (u')^2}^3} \end{aligned}$$

The the algorithm in the code below will use a combination of Newton's method and partial substitution. Thus the convergence will not be drastically faster then the pure partial substitution implemented in the above code, but there is some error control and the algorithm has not to be coded explicitly.

**SaltLakeCity2Zuerich.m**

```

N = 51; R = 6300;
Angles_ZH = [8,90-47]/180*pi; Angles_SLC = [-112,90-41]/180*pi;
BCleft = Angles_SLC(2); BCright = Angles_ZH(2);
interval = linspace(Angles_SLC(1),Angles_ZH(1),N);
a = @(x,u,du) (sin(u).^2+du.^2).^(-0.5);
f = {@(x,u,du) (-cos(u).*sin(u))./sqrt(sin(u).^2+du.^2),
    @(x,u,du) -(1-2*sin(u).^2)./sqrt(sin(u).^2+du.^2)...
    + (cos(u).^2.*sin(u).^2)./(sqrt(sin(u).^2+du.^2).^3),
    @(x,u,du) (cos(u).*sin(u).*du)./(sqrt(sin(u).^2+du.^2).^3)};
[phi2,theta2] = BVP1D(interval,1,0,0,1,0,BCleft,BCright); %% generate an intial guess
[phi2,theta2,inform] = BVP1DNL(interval,a,0,0,1,f,BCleft,BCright,theta2,
    'MaxIter',20,'Display','iter');
figure(4); plot(phi/pi*180,90-theta/pi*180);
    xlabel('\phi [deg]'); ylabel('90-\theta [deg]');
Dtheta2 = FEM1DEvaluateDu(phi2,theta2);
disp(sprintf("BVP1DNL(): L = %f km",R*trapz(phi2,sqrt(sin(theta2).^2+Dtheta2.^2)))
-->
iteration 1, RMS(correction) = 3.564728e-02, RMS(phi) = 2.463522e-02
iteration 2, RMS(correction) = 9.219981e-03, RMS(phi) = 3.524064e-03
iteration 3, RMS(correction) = 3.832637e-03, RMS(phi) = 1.109014e-03
iteration 4, RMS(correction) = 1.851140e-03, RMS(phi) = 4.995842e-04
iteration 5, RMS(correction) = 9.461144e-04, RMS(phi) = 2.493060e-04
iteration 6, RMS(correction) = 5.000805e-04, RMS(phi) = 1.302717e-04
iteration 7, RMS(correction) = 2.705884e-04, RMS(phi) = 6.998341e-05
iteration 8, RMS(correction) = 1.490106e-04, RMS(phi) = 3.832821e-05
iteration 9, RMS(correction) = 8.320151e-05, RMS(phi) = 2.129955e-05
iteration 10, RMS(correction) = 4.698144e-05, RMS(phi) = 1.197436e-05
iteration 11, RMS(correction) = 2.677860e-05, RMS(phi) = 6.796374e-06
iteration 12, RMS(correction) = 1.538521e-05, RMS(phi) = 3.888693e-06

BVP1DNL(): L = 8483.161942 km

```

## 9.22 A 1D nonlinear bending beam problem

The bending of a slender beam can be described by the angle  $\alpha(s)$  as function of the arc length  $s$ . For a beam with inertia of the cross section  $I$  and a material with Young's modulus  $E$  the curvature is given by

$$\kappa(s) = \alpha'(s) = \frac{M(s)}{EI},$$

where  $M(s)$  is the total moment at position  $s$ . Examine a beam attached at  $(x, y) = (0, 0)$  and starting out horizontally (i.e.  $\alpha(0) = 0$ ) with a vertical force  $F_2$  applied at the other endpoints  $s = L$ . The  $(x, y)$  coordinates of the beam at  $s = l$  are the given by an integral

$$\begin{pmatrix} x(s) \\ y(s) \end{pmatrix} = \int_0^s \begin{pmatrix} \cos(\alpha(\tau)) \\ \sin(\alpha(\tau)) \end{pmatrix} d\tau$$

The moment  $M(s)$  is given by

$$M(s) = F_2 (x(L) - x(s)) = -F_2 \int_s^L \cos(\alpha(\tau)) d\tau$$

Using the above equation for the bending of the beam leads to

$$\begin{aligned} \alpha'(s) &= \frac{M(s)}{EI} = \frac{F_2}{EI} \int_s^L \cos(\alpha(\tau)) d\tau \\ \frac{d}{ds} \alpha'(s) &= \frac{F_2}{EI} \frac{d}{ds} \int_s^L \cos(\alpha(\tau)) d\tau = -\frac{F_2}{EI} \cos(\alpha(s)) \\ -\alpha''(s) &= \frac{F_2}{EI} \cos(\alpha(s)) \end{aligned}$$

This ODE has to be supplemented with the boundary conditions  $\alpha(0) = \alpha'(L) = 0$  to arrive at a nonlinear BVP. For small forces the approximation  $\cos(\alpha) \approx 1$  leads to the solution

$$\alpha(s) \approx \frac{F_2}{2EI} (L^2 - (L - s)^2) = \frac{F_2}{2EI} (2Ls - s^2).$$

Use this result for code varification or as possible starting value for Newton's method.

### 9.22.1 Solving the BVP using Newton's algorithm

This can be solved using Newton's method. Use the idea of linearization to arrive at an iterative algorithm.

$$\begin{aligned} -\alpha''(s) &= \frac{F_2}{EI} \cos(\alpha(s)) \quad \text{for } 0 < s < L \quad \text{and} \quad \alpha(0) = \alpha'(L) = 0 \\ -\alpha''(s) - \phi''(s) &= \frac{F_2}{EI} \cos(\alpha(s) + \phi(s)) \approx \frac{F_2}{EI} (\cos(\alpha(s)) - \sin(\alpha(s)) \phi(s)) \\ -\phi''(s) + \frac{F_2}{EI} \sin(\alpha_n(s)) \phi(s) &= +\alpha_n''(s) + \frac{F_2}{EI} (\cos(\alpha_n(s))) \quad \text{with} \quad \phi(0) = \phi'(L) = 0 \\ \alpha_{n+1}(s) &= \alpha_n(s) + \phi(s) \end{aligned}$$

The code `BendingBeam1D.m` implements the above algorithm. For small values of  $F_2$  the results are as expected. But for larger values of  $F_2$  the resulting figure is at best surprising. In Figure 133(a) find the result for  $F_2 = 1.5$ , and this is in fact a solution of the nonlinear BVP, but not the expected solution. Newton's method works extremely well, if the initial guess is "close enough" to the desired solution. But the last point is critical and failed for this bending beam problem for large forces  $F_2$ .

**BendingBeam1D.m**

```

L = 3; EI = 1; %% as single run
F2 = 1.5; %% try values of 0.1 0.5 1.0 and 2
N = 1000; s = linspace(0,L,N);
[sn,alpha] = BVP1D(s,1,0,0,1,F2/EI,0,[0,0]);
figure(1); plot(sn,alpha); xlabel('arclength s'); ylabel('angle \alpha')
xGauss = FEM1DGaussPoints(sn);
[xGauss,Nodes2GaussU] = FEM1DGaussPoints(sn);
for jj = 1:20
    [dalpha,ddalpha] = FEM1DEvaluateDu(sn,alpha); %% evaluate derivative at nodes
    RHS = ddalpha + F2/EI*cos(alpha);
    alphaGauss = Nodes2GaussU*alpha; %% evaluate u at Gauss points
    [sn,phi] = BVP1D(s,1,0,+F2/EI*sin(alphaGauss),1,RHS,0,[0,0]);
    disp(sprintf('max(abs(phi)) = %g , max(abs(RHS)) = %g',max(abs(phi)), max(abs(RHS))))
    alpha = alpha + phi;
    figure(2); plot(sn,alpha); xlabel('x'); ylabel('angle \alpha')
    pause(0.2)
endfor
x = cumtrapz(sn,cos(alpha)); y = cumtrapz(sn,sin(alpha));
figure(3); plot(x,y); xlabel('x'); ylabel('y')

```

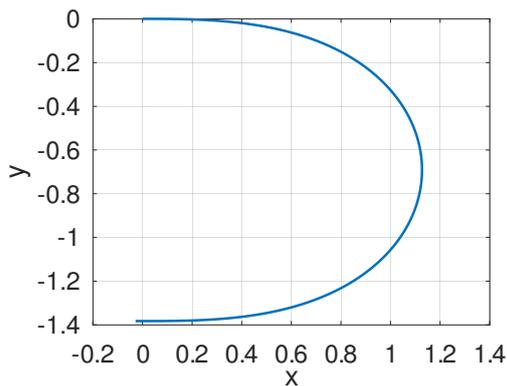
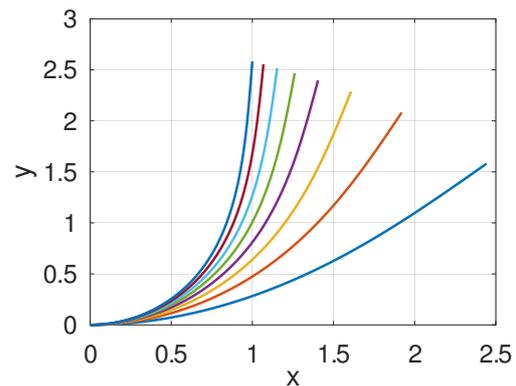
(a) the result for  $F_2 = 1.5$ (b) the parametric results for  $F_2$  up to 2

Figure 133: The solutions for a bending beam

To obtain a reliable solution for  $F_2 = 2$  it is advisable to use a parameterized approach. In the code below the value of  $F_2$  is increased from 0 to 2.0 in steps of 0.25 and at each new level of  $F_2$  the result of the previous computation is used as starting value for  $\alpha(s)$ . Find the results in Figure 133(b).

**BendingBeam1D.m**

```

L = 3; EI = 1; F2_List = [0.25:0.25:2]; %% a parametrized approach
N = 100; s = linspace(0,L,N)';
[sn,alpha] = BVP1D(s,1,0,0,1,F2_List(1)/EI,0,[0,0]);
figure(1); plot(sn,alpha); xlabel('arclength s'); ylabel('angle \alpha')
xGauss = FEM1DGaussPoints(sn);
figure(3); clf; hold off
for F2 = F2_List
    for jj = 1:10
        [dalpha,ddalpha] = FEM1DEvaluateDu(sn,alpha);
        RHS = ddalpha + F2/EI*cos(alpha);

```

```

    alphaGauss = pwquadinterp(sn,alpha,xGauss); %% evaluate u at Gauss points
    [sn,phi] = BVP1D(s,1,0,F2/EI*sin(alphaGauss),1,RHS,0,[0,0]);
    alpha = alpha + phi;
endfor % jj
x = cumtrapz(sn,cos(alpha)); y = cumtrapz(sn,sin(alpha));
figure(3); plot(x,y); xlabel('x'); ylabel('y'); hold on
endfor %% F2
hold off

```

### 9.22.2 Solving the BVP with the command BVP1DNL()

The above bending beam problem can be solved with the help of BVP1DNL(). There is less coding involved, but the problem of converging to a non desired solution remains. Observing the results of the iterations confirms that at first the algorithm is “searching for a solution”, but once it is close to a solution quadratic convergence sets in, i.e. the number of stable digits is doubled at each step. This is expected for Newton’s method.

#### BendingBeam1D.m

```

F2 = 1.5; %% try values of 0.1 0.25 0.5 1.0 and 2
f = {@(s,alpha)F2/EI*cos(alpha), @(s,alpha)-F2/EI*sin(alpha)};
N = 100; s = linspace(0,L,N)';
%% generate a good initial guess
[sn,alpha0] = BVP1D(s,1,0,0,1,F2/EI,0,[0,0]); %% as solution of linear BVP
%% alpha0 = @(s)F2/(2*EI)*(L^2-(L-s).^2); %% use the analytical solution
[sn,alpha,inform] = BVP1DNL(s,1,0,0,1,f,0,[0,0],alpha0,
    'tol',1e-8,'MaxIter',50,'Display','iter');

inform
figure(1); plot(sn,alpha); xlabel('arclength s'); ylabel('angle \alpha')
x = cumtrapz(sn,cos(alpha)); y = cumtrapz(sn,sin(alpha));
figure(3); plot(x,y); xlabel('x'); ylabel('y');

```

The output of the above code illustrates the long search for a solution by Newton’s method, and the quadratic convergence as soon as close to one of the possible solutions.

```

iteration 1, RMS(correction) = 2.148167e+00, RMS(phi) = 3.658963e+01
iteration 2, RMS(correction) = 1.534256e+00, RMS(phi) = 1.281984e+00
iteration 3, RMS(correction) = 3.233207e-01, RMS(phi) = 6.380582e-01
iteration 4, RMS(correction) = 1.715649e-01, RMS(phi) = 6.706732e-01
iteration 5, RMS(correction) = 1.442408e-01, RMS(phi) = 5.281602e-01
iteration 6, RMS(correction) = 1.380244e-01, RMS(phi) = 4.177463e-01
iteration 7, RMS(correction) = 1.388972e-01, RMS(phi) = 3.168041e-01
iteration 8, RMS(correction) = 1.277986e-01, RMS(phi) = 2.140754e-01
iteration 9, RMS(correction) = 8.810050e-02, RMS(phi) = 1.124749e-01
iteration 10, RMS(correction) = 3.198417e-02, RMS(phi) = 3.424090e-02
iteration 11, RMS(correction) = 3.261890e-03, RMS(phi) = 3.282558e-03
iteration 12, RMS(correction) = 3.022821e-05, RMS(phi) = 3.022996e-05
iteration 13, RMS(correction) = 2.563226e-09, RMS(phi) = 2.562500e-09

inform = scalar structure containing the fields:
    info      = 1
    iter      = 13
    AbsError  = 2.5632e-09

```

The parametric approach will again reliably generate the desired solution. It is using very few iterations for each value of  $F_2$ .

**BendingBeam1D.m**

```

L = 3; EI = 1;
F2_List = [0:0.25:2];
N = 100; s = linspace(0,L,N)'; sn = sort([s; s(1:end-1)+diff(s)/2]);
alpha = 0;
for F2 = F2_List
    f = {@(s,al)F2/EI*cos(al), @(s,al)-F2/EI*sin(al)};
    [sn,alpha,inform] = BVP1DNL(s,1,0,0,1,f,0,[0,0],alpha);
    inform
endfor
inform
figure(1); plot(sn,alpha*180/pi); xlabel('arclength s'); ylabel('angle \alpha [deg]')
x = cumtrapz(sn,cos(alpha)); y = cumtrapz(sn,sin(alpha));
figure(3); plot(x,y); xlabel('x'); ylabel('y');

```

**9.23 Mass transfer in a porous catalyst**

In [KubiHlav08, p. 92, Example 4.3, p. 255, Example 5.7] a nonlinear boundary value problem is examined, describing the mass transfer in a porous catalyst.

$$-y'' - \frac{a}{x} y' = -\alpha^2 y \exp\left(\frac{\gamma \beta (1-y)}{1 + \beta (1-y)}\right) \quad \text{with } y'(0) = 0 \quad \text{and } y(1) = 1$$

For  $\alpha = 0$  there is the trivial solution  $y(x) = 1$ . The nonlinear contribution and its partial derivative are given by

$$\begin{aligned}
 f(y) &= -\alpha^2 y \exp\left(\frac{\gamma \beta (1-y)}{1 + \beta (1-y)}\right) \\
 \frac{\partial}{\partial y} f(y) &= -\alpha^2 \exp\left(\frac{\gamma \beta (1-y)}{1 + \beta (1-y)}\right) + \alpha^2 y \exp\left(\frac{\gamma \beta (1-y)}{1 + \beta (1-y)}\right) \frac{-\gamma \beta (1 + \beta (1-y)) + \gamma \beta (1-y) \beta}{(1 + \beta (1-y))^2} \\
 &= -\alpha^2 \exp\left(\frac{\gamma \beta (1-y)}{1 + \beta (1-y)}\right) \left(1 + y \frac{-\gamma \beta}{(1 + \beta (1-y))^2}\right).
 \end{aligned}$$

In [KubiHlav08] the values  $a = 2$ ,  $\gamma = 20$  and  $\beta = 0.05$  are used and the parameter  $0 \leq \alpha \leq 1$  is increased from 0 to 1. BVP1DNL() can solve the problem directly for  $\alpha = 1$ , but a parametric solution is possible too. Find the solutions of the code below in Figure 134. The results coincide with the numbers in [KubiHlav08, Example 5.7].

**PorousCatalyst.m**

```

N = 10; interval = linspace(0,1,N)';
a = 2; gamma = 20; beta = 0.05;
BCleft = [0,0]; BCright = 1;
y = 1;
figure(2); clf; hold on; box on; xlabel('x'); ylabel('y')
for alpha = 0:0.2:1
    f = {@(x,y) -alpha^2*y.*exp(gamma*beta*(1-y)/(1+beta*(1-y))),
        @(x,y) -alpha^2 *exp(gamma*beta*(1-y)/(1+beta*(1-y))).*...
            (1-gamma*beta*y./((1+beta*(1-y)).^2))};
    [x,y] = BVP1DNL(interval,1,@(x)-a./x,0,1,f,BCleft,BCright,y);
    plot(x,y)
    disp(sprintf('alpha = %#3.2f, y(0) = %#g',alpha,y(1)))
endfor
legend('\alpha = 0.00', '\alpha = 0.20', '\alpha = 0.40', '\alpha = 0.60',
        '\alpha = 0.80', '\alpha = 1.00', 'location', 'southeast')

```

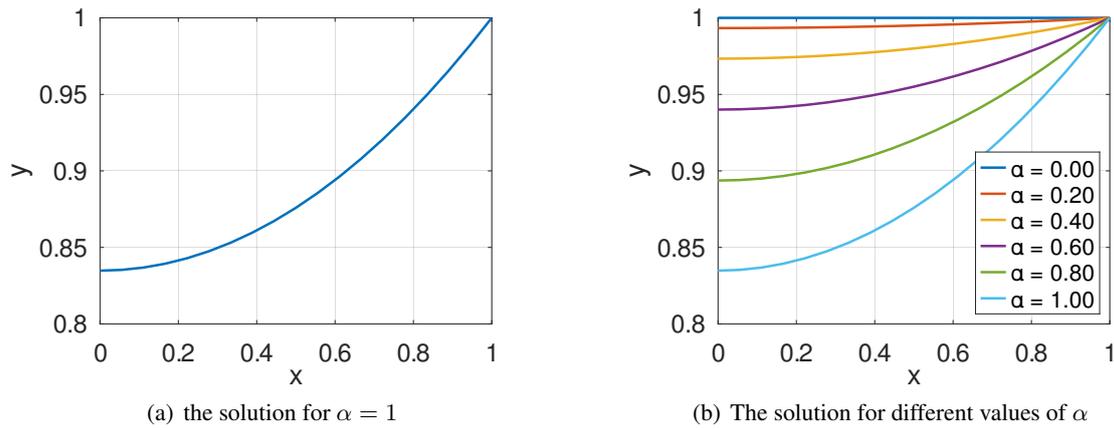


Figure 134: A BVP describing mass transfer in a porous catalyst

```
figure(1); plot(x,y); xlabel('x'); ylabel('y'); box on
-->
alpha = 0.00, y(0) = 1.00000
alpha = 0.20, y(0) = 0.993333
alpha = 0.40, y(0) = 0.973339
alpha = 0.60, y(0) = 0.940066
alpha = 0.80, y(0) = 0.893713
alpha = 1.00, y(0) = 0.834810
```

For a second set of values  $a = 0$ ,  $\gamma = 20$  and  $\beta = 0.4$  the algorithm converges nicely up to  $\alpha \approx 0.37$ . For  $\alpha > 0.4$  convergence is difficult to obtain.

In [KubiHlav08, Example 4.3] find the values for  $a = 2$ ,  $\gamma = 20$ ,  $\beta = 0.2$  and  $\alpha = 2$ . A rather sophisticated iteration scheme was used in [KubiHlav08]. By increasing the value of  $\alpha$  from 0 to 2 the results in the last row of [KubiHlav08, Table4-12] can be reproduced by `BVP1DNL()`. Find the results of the code below in Figure 135.

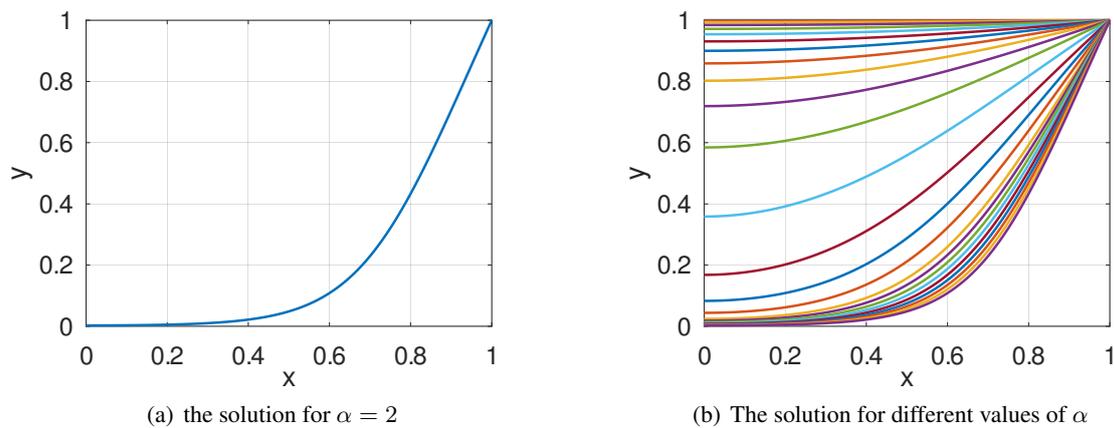


Figure 135: A BVP describing mass transfer in a porous catalyst, second setup

---

**PorousCatalyst.m**


---

```

N = 51; interval = linspace(0,1,N)';
BCleft = [0,0]; BCright = 1;
a = 2; gamma = 20; beta = 0.2;
y = 1;
figure(2); clf; hold on; box on; xlabel('x'); ylabel('y')
for alpha = [0:0.1:1.6, 1.65:0.05:2]
    f = {@(x,y) -alpha^2*y.*exp(gamma*beta*(1-y)/(1+beta*(1-y))),
        @(x,y) -alpha^2 *exp(gamma*beta*(1-y)/(1+beta*(1-y))).*...
            (1-gamma*beta*y./((1+beta*(1-y)).^2))};
    [x,y] = BVP1DNL(interval,1,@(x)-a./x,0,1,f,BCleft,BCright,y);
    plot(x,y)
endfor
figure(1); plot(x,y); xlabel('x'); ylabel('y'); box on
x_n = [0:0.2:0.8]; y_n = pwquadinterp(x,y,x_n);
disp(sprintf('y(%g)=%5.4e, y(%g)=%5.4e, y(%g)=%5.4e, y(%g)=%5.4e, y(%g)=%5.4e',
            x_n(1),y_n(1),x_n(2),y_n(2),x_n(3),y_n(3),x_n(4),y_n(4),x_n(5),y_n(5)))
-->
y(0)=2.7643e-3, y(0.2)=5.3166e-3, y(0.4)=2.1871e-2, y(0.6)=1.0853e-1, y(0.8)=4.3316e-1

```

---

## 9.24 Troesch's equation

In [IzadSuayNoei21] or [KubiHlav08, p. 244, Example 5.4] Troesch's nonlinear boundary value problem is examined.

$$y''(x) = \alpha \sinh(\alpha y(x)) \quad \text{with } y(0) = 0 \quad \text{and } y(1) = 1$$

The results for  $\alpha = 0.5$  in [IzadSuayNoei21, Table 1] can be produced by BVP1DNL() with very few lines of code.

---

**Troesch.m**


---

```

N = 11; interval = linspace(0,1,N)'; BCleft = 0; BCright = 1;
alpha = 0.5;
f = {@(x,y) -alpha*sinh(alpha*y), @(x,y)-alpha^2*cosh(alpha*y)};
[x,y,inform] = BVP1DNL(interval,1,0,0,1,f,BCleft,BCright,@(x)x);
figure(1); plot(x,y); xlabel('x'); ylabel('y(x)')
xd = [ 0.1:0.1:0.9]'; yd = pwquadinterp(x,y,xd);
Results = [xd,yd]
-->
Results =
    0.1    0.095944
    0.2    0.192129
    0.3    0.288794
    0.4    0.386185
    0.5    0.484547
    0.6    0.584133
    0.7    0.685201
    0.8    0.788017
    0.9    0.892854

```

---

To obtain the results for  $\alpha = 1$  in [IzadSuayNoei21, Table 2] just change the value of  $\alpha$  in the above code.

In [KubiHlav08, Example 5.4] a shooting method is used to determine  $\alpha$  as function of  $y'(0)$ . The results in [KubiHlav08, Table 5–6] are confirmed by the code below.

## Troesch.m

```

N = 51; interval = linspace(0,1,N)'; BCleft = 0; BCright = 1;
figure(1); clf; hold on; box on
y = 1;
for alpha = [0.792, 1.151 1.753 2.394 3.308 4.129 5.0]
f = {@(x,y) -alpha*sinh(alpha*y), @(x,y)-alpha^2*cosh(alpha*y)};
[x,y] = BVP1DNL(interval,1,0,0,1,f,BCleft,BCright,y);
figure(1); plot(x,y); xlabel('x'); ylabel('y(x)'); drawnow()
[y0,dy0] = pwquadinterp(x,y,0);
disp(sprintf("alpha = %#6.5g, y' (0)= %g",alpha,dy0))
endfor
-->
alpha = 0.79200, y' (0)= 0.900027
alpha = 1.1510, y' (0)= 0.800195
alpha = 1.7530, y' (0)= 0.600109
alpha = 2.3940, y' (0)= 0.400003
alpha = 3.3080, y' (0)= 0.200008
alpha = 4.1290, y' (0)= 0.0999462
alpha = 5.0000, y' (0)= 0.0457183

```

## 9.25 Motion of a string

Examine the motion of a string of length 5. For 2 seconds apply a force to the very left section ( $0 \leq x \leq 1$ ). Use a small damping factor, e.g. 0.25. A possible IBVP describing this setup is

$$\begin{aligned}
 \frac{\partial^2}{\partial t^2} u(x,t) + 0.25 \frac{\partial}{\partial t} u(x,t) - \frac{\partial^2}{\partial x^2} u(x,t) &= f(x,t) && \text{for } 0 < x < 5 \text{ and } 0 < t < 10 \\
 u(0,t) = u(5,t) &= 0 && \text{for } 0 \leq t \leq 10 \\
 u(x,0) = \frac{\partial}{\partial t} u(x,0) &= 0 && \text{for } 0 \leq x \leq 5
 \end{aligned}$$

with the driving function

$$f(x,t) = \begin{cases} \cos(\frac{\pi}{2} x) \sin(2\pi t) & \text{for } 0 \leq x \leq 1 \text{ and } 0 \leq t \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

This problem is examined with the help of `I2BVP1D()`. Find the result in Figure 136. Observe the two cos-shaped pulses moving with speed +1. Then they are reflected at the border at  $x = 5$  and travelling with speed -1. The amplitudes are slowly decaying, caused by the damping.

## MovingString.m

```

N = 101; x_max = 5; interval = linspace(0,x_max,N)';
omega = 2*pi;
f = @(x,t) (cos(x*pi/2) .* (x<=1)) * (sin(omega*t) * (t<=2));
u0 = 0; u1 = 0;
w2 = 1; w1 = 0.25; a = 2; b = 0; c = 0; d = 1;
BCleft = [0]; BCright = [0];
t0 = 0; tend = 10; steps = [250,10];
[x,u,t] = I2BVP1D(interval,w2,w1,a,b,c,d,f,BCleft,BCright,u0,u1,t0,tend,steps);

figure(1); mesh(t,x,u); xlabel('time t'); ylabel('position x'); zlabel('u')
xlim([min(t),max(t)]); ylim([min(x),max(x)]); view([20,20])

```

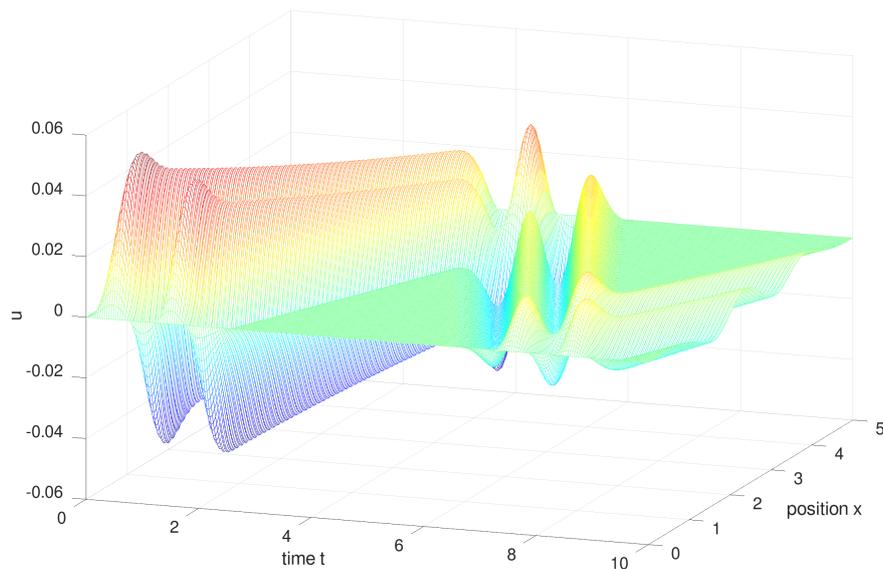


Figure 136: Motion of a string, excited by an initial pulse close to  $x = 0$

## 9.26 A pipe under pressure

Examine a pipe with a circular cross section and inner radius  $R$  and a wall with thickness  $\Delta R$ . On the inside a pressure  $P$  is applied. The pipe under pressure will expand and the wall material will stretch. For ductile materials (e.g. copper, steel) the maximal value of the von Mises stress is a good criterion to decide whether the pipe will withstand the pressure, or break. The problem can be examined by FEM as a plane strain problem or as an axially symmetric problem, or one can determine an exact solution.

As exemplary situation examine:

- a pipe with inner radius  $R = 0.1$  m and a wall thickness of  $\Delta R = 0.01$  m. A quarter of a cross section is visible in Figure 137.
- a pressure of  $P = 10$  atm =  $10^6$  Pa =  $10^6 \frac{\text{N}}{\text{m}^2}$ .
- with a copper pipe, i.e. a yield strength of  $\approx 33$  MPa or a steel pipe, i.e. a yield strength of  $\approx 350$  MPa.

### 9.26.1 As a plane strain problem

To examine this problem with FEMoctave start by defining the domain and the boundary conditions.

- Define the domain with the help of polar coordinates  $R \leq r \leq R + \Delta R$  and  $0 \leq \phi \leq \frac{\pi}{2}$ . Use `CreateMeshRect()` to create a rectangular mesh and then `MeshDeform()` to create the domain in Figure 137.
- At the lower edge at  $y = 0$  the edge is free to move in  $x$ -direction and no displacement in  $y$ -direction. This is implemented with the code `-21` in the function `CreateMeshRect()` for the boundary condition. See Table 5 on page 47 for the coding of the boundary conditions.

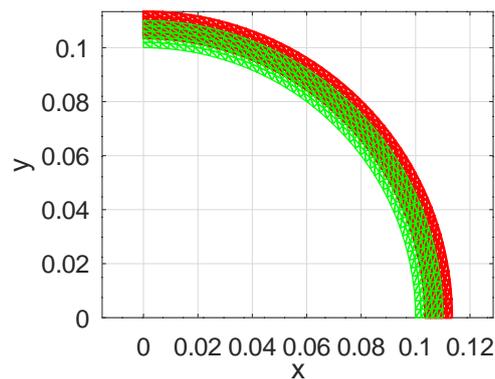


Figure 137: One quarter of a section through the pipe, the original domain (green) and the deformed domain (red)

- At the left edge at  $x = 0$  the edge is free to move in  $y$ -direction and no displacement in  $x$ -direction. This is implemented with the code `-12` for the boundary condition.
- At the inner edge at  $r = R$  pressure  $P$  is applied, leading to the code `-33` for the boundary condition.
- At the outer edge at  $r = R + \Delta R$  there is no force, leading to the code `-22` for the boundary condition.
- For good accuracy second order elements are used by calling `MeshUpgrade()`.

#### PipePressure.m

```
E = 110e9; nu = 0.35; %%% copper
%%E = 200e9; nu = 0.25; %%% steel
R = 0.1; dR = 0.01;
nR = 5; nPhi = 51; % number of layers in radial and angular direction
Estar = E/(1-nu^2); nustar = nu/(1-nu);

global P
P = 10e5; %% 10 atm pressure
FEMmesh = CreateMeshRect(linspace(R,R+dR,nR+1),linspace(0,pi/2,nPhi+1),-21,-12,-33,-22);
function new_xy = Deform(xy)
    new_xy = [xy(:,1).*cos(xy(:,2)),xy(:,1).*sin(xy(:,2))];
endfunction
FEMmesh = MeshDeform(FEMmesh,'Deform');
FEMmesh = MeshUpgrade(FEMmesh,'quadratic');
```

With the domain and the correct boundary conditions the problem can be solved.

- Start by defining the force density corresponding to the inside pressure  $P$ , i.e.

$$\text{at } \begin{pmatrix} R \cos \alpha \\ R \sin \alpha \end{pmatrix} \text{ apply the force density } \begin{pmatrix} P \cos \alpha \\ P \sin \alpha \end{pmatrix}.$$

- Assuming that the pipe will not stretch in the direction orthogonal to the cross section we end up with a plane strain problem. Thus use `PlaneStrain()` to find approximations to the displacements  $u_1$  and  $u_2$ .

## PipePressure.m

```

%% define the radial pressure to be applied on the inside
function res = gN1(xy)
    global P
    angle = atan2(xy(:,2),xy(:,1)); res = P*cos(angle);
endfunction
function res = gN2(xy)
    global P
    angle = atan2(xy(:,2),xy(:,1)); res = P*sin(angle);
endfunction

[u1,u2] = PlaneStrain(FEMmesh,E,nu,{0,0},{0,0},{'gN1','gN2'});

factor = 400;
figure(111); trimesh(FEMmesh.elem,FEMmesh.nodes(:,1)+factor*u1,...
                    FEMmesh.nodes(:,2)+factor*u2,'color','red','linewidth',2);
hold on; trimesh(FEMmesh.elem,FEMmesh.nodes(:,1),FEMmesh.nodes(:,2),...
                'color','green','linewidth',1);
hold off; axis equal; xlabel('x'); ylabel('y');

```

The last few lines in the above code generate the domain visible in Figure 137.

With the displacements determine all stresses at the nodes by using `EvaluateStress()`. Since four return arguments are asked for the plane strain setup is used. Then use `EvaluateVonMises()` to find the values of the von Mises stress, visible in Figure 138(a). The maximal value of the von Mises stress is approximated by 10 MPa, which is smaller than the yield strength 33 MPa of copper. Thus the pipe is expected to withstand the applied pressure, but the margin of error is not very large. The pipe will start cracking on the inside, where the von Mises stress is largest.

## PipePressure.m

```

[sigma_x,sigma_y,tau_xy,sigma_z] = EvaluateStress(FEMmesh,u1,u2,E,nu);
vonMises = EvaluateVonMises(sigma_x,sigma_y,tau_xy,sigma_z);

figure(2); FEMtrimesh(FEMmesh,vonMises); xlabel('x'); ylabel('y');
title('von Mises stress'); view([25,25])
vonMises_min_max = [min(vonMises),max(vonMises)]
-->
vonMises_min_max = 8.3695e+06 1.0082e+07

```

To analyze the pipe further choose an angle, e.g.  $\alpha = \frac{\pi}{4} = 90^\circ$ , and evaluate along a straight line with this angle for radii  $R \leq r \leq R + \Delta R$ .

- Start by selecting the angle  $0 \leq \alpha \leq \frac{\pi}{2}$  and define the  $x$  and  $y$  values along the arc with this angle.
- Use the above values of the stresses and `EvaluatePrincipalStress()` find the values of the principle stresses at the nodes.
- Then calls of `FEMgriddata()` will determine the values of the principle stresses along the selected arc. Use  $\sigma_3 = \nu(\sigma_1 + \sigma_2)$  to compute the third principle stress. Then a call of `plot()` will generate Figure 138(b).
- The minimal value  $-9.9770 \cdot 10^5 \approx -1$  MPa of  $\sigma_2$  shows that this is the normal stress in radial direction on the inside of the pipe, coinciding with the given pressure  $P$ .

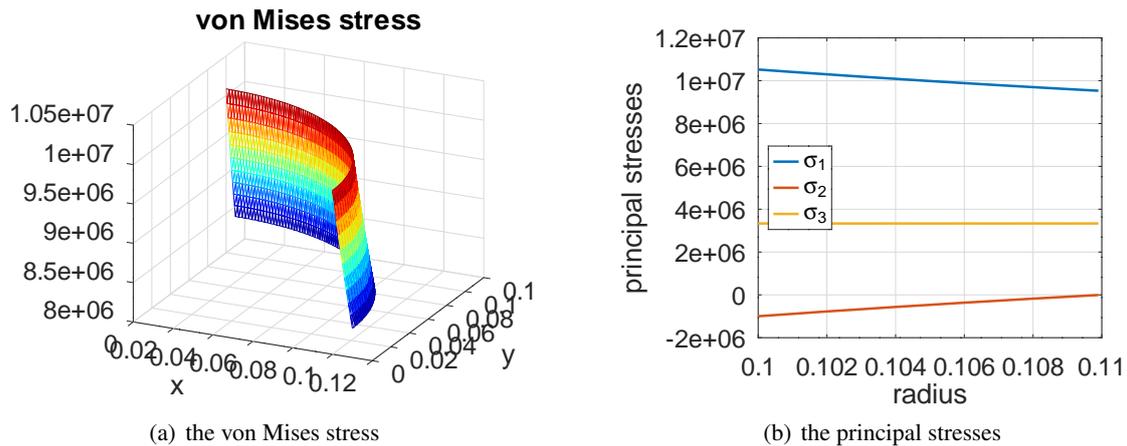


Figure 138: The von Mises stress in the cross section and the principal stresses along a radius of the pipe under pressure

- The maximal value  $-7 \cdot 10^0 \approx 0$  MPa of  $\sigma_2$  corresponds to the zero pressure on the outside.
- The values of  $\sigma_1$  are considerably larger than the values of  $\sigma_2$ . This illustrates that the wall of the pipe is severely stretched in angular direction.

#### PipePressure.m

```

%% evaluation at one angle, all radii
alpha = pi/4; Nr = 101; Nmid = (Nr+1)/2; %% use an odd number for Nr
r = linspace(R,R+dR,Nr)'; x = r*cos(alpha); y = r*sin(alpha);

[sigma_1,sigma_2] = EvaluatePrincipalStress(sigma_x,sigma_y,tau_xy);
sigma_1r = FEMgriddata(FEMmesh,sigma_1,x,y);
sigma_2r = FEMgriddata(FEMmesh,sigma_2,x,y);
sigma_3r = nu*(sigma_1r+sigma_2r);
sigma_2r_min_max = [min(sigma_2r),max(sigma_2r)]

figure(3); plot(r,[sigma_1r,sigma_2r,sigma_3r]);
            xlabel('radius'); ylabel('principal stresses')
            legend('\sigma_1','\sigma_2','\sigma_3','location','west')
-->
sigma_2r_min_max = -9.9770e+05 -7.1114e+03

```

At the midpoint in the wall of the pipe the stress matrix (tensor, to be precise) can be evaluated with the help of three calls of `FEMgriddata()`.

$$\begin{bmatrix} \sigma_x & \tau_{xy} \\ \tau_{xy} & \sigma_y \end{bmatrix} \approx \begin{bmatrix} +4.7572 & -5.2252 \\ -5.2252 & +4.7616 \end{bmatrix} \cdot 10^6$$

Then use a rotation matrix and the transformation rule for second order tensors to determine the stresses in the rotated coordinate system.

$$\begin{bmatrix} +\cos \alpha & +\sin \alpha \\ -\sin \alpha & +\cos \alpha \end{bmatrix} \begin{bmatrix} \sigma_x & \tau_{xy} \\ \tau_{xy} & \sigma_y \end{bmatrix} \begin{bmatrix} +\cos \alpha & -\sin \alpha \\ +\sin \alpha & +\cos \alpha \end{bmatrix} \approx \begin{bmatrix} -0.46582 & +0.00217 \\ +0.00217 & +9.9846 \end{bmatrix} \cdot 10^6$$

The result shows the normal, compressing pressure of  $-0.47$  MPa in radial direction and the stretching pressure of  $+10$  MPa in angular direction.

#### PipePressure.m

```

%% examine stress at middle point
x_mid = x(Nmid); y_mid = y(Nmid);

sigma_x = FEMgriddata(FEMmesh, sigma_x, x_mid, y_mid);
sigma_y = FEMgriddata(FEMmesh, sigma_y, x_mid, y_mid);
tau_xy = FEMgriddata(FEMmesh, tau_xy, x_mid, y_mid);
RotMat = [cos(alpha) -sin(alpha); +sin(alpha) cos(alpha)];
stress = [sigma_x, tau_xy; tau_xy, sigma_y]
stress_rotated = RotMat' * stress * RotMat
-->
stress          =  4.7572e+06  -5.2252e+06
                  -5.2252e+06  4.7616e+06

stress_rotated =  -4.6582e+05  2.1668e+03
                  2.1668e+03  9.9846e+06

```

With the provided code in `PipePressure.m` it is easy to modify the parameters of the above problem, e.g. change from copper to steel, examine larger radii or thinner walls.

For a pipe with a thin wall an analytical approximation is possible. Examine the section shown in Figure 137 and assume that the normal stress  $\sigma_\varphi$  in angular direction is independent on the radius. Then use a balance of force law in  $y$ -direction and an integration over the angle to conclude

$$\sigma_\varphi \Delta R = \int_0^{\pi/2} P \sin \varphi R d\varphi = P R .$$

In the above example this leads to

$$\sigma_\varphi = \frac{P R}{\Delta R} = \frac{10^6 \cdot 0.1}{0.01} = 10^7 ,$$

which is very close to the above result generated by FEMoctave. With the known angular stress and Hooke's law estimate the angular stretch, i.e.

$$\varepsilon_\varphi = \frac{\sigma_\varphi}{E} = \frac{10^7}{110 \cdot 10^9} \approx 9.09 \cdot 10^{-5} .$$

Since the angular stretching factor  $\varepsilon_\varphi$  equals the radial stretching factor  $\varepsilon_r$  estimate the change of radius by

$$R \longrightarrow R(1 + \varepsilon_r) = R + 9.09 \cdot 10^{-6} .$$

This is not too far from the FEMoctave result of  $\max\{u_1\} \approx 8.8 \cdot 10^{-6}$ . The FEM approximation allows to examine pipes with thick walls and also examines behavior within the wall.

### 9.26.2 As an axisymmetric problem

The above problem can be examined as an axially symmetric problem. The domain is given by  $R \leq r \leq R + dR$  and  $0 \leq z \leq R$ , and then rotated about the  $z$ -axis.

- At the inner edge at  $x = r = R$  the pressure  $P$  is applied in  $r$ -direction and the edge is free to move in  $z$  direction.
- The outer edge at  $x = r = R + dR$  is free to move.

- The lower and upper edge are fixed in  $z$ -direction and free to move in  $x = r$ -direction.

Start out by defining the parameters and generating the mesh. Then determine the radial displacement  $u_r$  and the  $z$ -displacement  $u_z$  by calling `AxiStress()`.

#### PipePressureAxi.m

```
R = 0.1; dR = 0.01;
if 0 %% regular mesh
    Mesh = CreateMeshRect(R+linspace(0,dR,20),linspace(0,R,10),-21,-21,-32,-22);
else %% irregular mesh
    Mesh = CreateMeshTriangle('Test',...
        [R 0 -21;R+dR 0 -22; R+dR R -21; R R -32],1e-5);
endif
Mesh = MeshUpgrade(Mesh,'quadratic');

P = 10e5; E = 110e9; nu = 0.35; f = {0,0}; gD = {0,0}; gN = {P,0};
[ur,uz] = AxiStress(Mesh,E,nu,f,gD,gN);
figure(2); FEMtrimesh(Mesh,ur);
    xlabel('r'); ylabel('z'); zlabel('u_r')
figure(3); FEMtrimesh(Mesh,uz);
    xlabel('r'); ylabel('z'); zlabel('u_z')
```

Determine all strains by using `EvaluateStrainAxi()`.

#### PipePressureAxi.m

```
[eps_xx,eps_yy,eps_zz,eps_xz] = EvaluateStrainAxi(Mesh,ur,uz);
figure(11); FEMtrimesh(Mesh,eps_xx)
    xlabel('r'); ylabel('z'); zlabel('\epsilon_{xx}')
figure(12); FEMtrimesh(Mesh,eps_yy)
    xlabel('r'); ylabel('z'); zlabel('\epsilon_{yy}')
figure(13); FEMtrimesh(Mesh,eps_zz)
    xlabel('r'); ylabel('z'); zlabel('\epsilon_{zz}')
```

Determine the normal and shearing stresses by using `EvaluateStressAxi()`.

#### PipePressureAxi.m

```
[sigma_x,sigma_y,sigma_z,tau_xz] = EvaluateStressAxi(Mesh,ur,uz,E,nu);
figure(21); FEMtrimesh(Mesh,sigma_x)
    xlabel('r'); ylabel('z'); zlabel('\sigma_x')
figure(22); FEMtrimesh(Mesh,sigma_y)
    xlabel('r'); ylabel('z'); zlabel('\sigma_y')
figure(23); FEMtrimesh(Mesh,sigma_z)
    xlabel('r'); ylabel('z'); zlabel('\sigma_z')
```

Determine the von Mises stress, the principal stresses and the Tresca stress by using the script functions `EvaluateVonMisesAxi()`, `EvaluatePrincipalStressAxi()` and `EvaluateTrescaAxi()`. The results coincide with the values from the plane strain approach in the previous section.

#### PipePressureAxi.m

```
vonMises = EvaluateVonMisesAxi(sigma_x,sigma_y,sigma_z,tau_xz);
figure(24); FEMtrimesh(Mesh,vonMises)
    xlabel('r'); ylabel('z'); zlabel('von Mises')
[sigma_1,sigma_2] = EvaluatePrincipalStressAxi(sigma_x,sigma_z,tau_xz);
r = R + linspace(0,dR,100)';
sigma_1i = FEMgriddata(Mesh,sigma_1,r,R/2*ones(size(r)));
sigma_2i = FEMgriddata(Mesh,sigma_2,r,R/2*ones(size(r)));
sigma_3i = FEMgriddata(Mesh,sigma_y,r,R/2*ones(size(r)));
```

```

figure(25); plot(r, sigma_1i, r, sigma_2i, r, sigma_3i); xlabel('r'); ylabel('z');
legend('\sigma_1', '\sigma_2', '\sigma_3', 'location', 'west')
Tresca = EvaluateTrescaAxi(sigma_x, sigma_y, sigma_z, tau_xz);
figure(26); FEMtrimesh(Mesh, Tresca); xlabel('r'); ylabel('z'); zlabel('Tresca')

```

### 9.26.3 The analytical solution

For this axisymmetric setup use that  $u_z = 0$  and  $u_r(r, z) = u_r(r)$  to determine an exact solution. The energy of the system is given by

$$\frac{U(u_r)}{2\pi} = \iint_{\Omega} \frac{rE}{2(1+\nu)(1-2\nu)} \left( (1-\nu) \left( \left( \frac{\partial u_r}{\partial r} \right)^2 + \frac{1}{r^2} u_r^2 \right) + \frac{2\nu}{r} u_r \frac{\partial u_r}{\partial r} \right) dA - RP u_r(R).$$

With the constant  $k = \frac{E}{(1+\nu)(1-2\nu)}$  and  $u(r) = u_r(r)$  the expression to be minimized is

$$\begin{aligned} U_r(u) &= \int_R^{R+\Delta R} \frac{rk}{2} \left( (1-\nu) \left( (u'(r))^2 + \frac{1}{r^2} u^2(r) \right) + \frac{2\nu}{r} u(r) u'(r) \right) dr - RP u(R) \\ &= \int_R^{R+\Delta R} \frac{k}{2} \left( (1-\nu) \left( r (u'(r))^2 + \frac{1}{r} u^2(r) \right) \right) dr + \frac{k\nu}{2} u^2(r) \Big|_{r=R}^{R+\Delta R} - RP u(R) \\ U_r(u + \phi) &= U_r(u) + \int_R^{R+\Delta R} k(1-\nu) \left( r u' \phi' + \frac{1}{r} u \phi \right) dr + k\nu u(r) \phi(r) \Big|_{r=R}^{R+\Delta R} - RP \phi(R) + O(\phi^2) \\ &= U_r(u) + \int_R^{R+\Delta R} k(1-\nu) \left( -r u' \phi + \frac{1}{r} u \phi \right) dr + \\ &\quad + k \left( (1-\nu) r u'(r) \phi(r) + \nu u(r) \phi(r) \right) \Big|_{r=R}^{R+\Delta R} - RP \phi(R) + O(\phi^2). \end{aligned}$$

Use the Euler–Lagrange equation for this problem and determine the exact solution.

$$\begin{aligned} 0 &= -r (r (u'(r)))' + u(r) \\ \text{Ansatz: } u(r) &= r^\alpha \\ 0 &= -r (r \alpha r^{\alpha-1})' + r^\alpha = -\alpha^2 r^\alpha + r^\alpha \\ 0 &= -\alpha^2 + 1 \implies \alpha = \pm 1 \\ u(r) &= c_1 r + c_2 \frac{1}{r} \end{aligned}$$

The two natural boundary conditions are

$$\begin{aligned} (1-\nu) R u'(R) + \nu u(R) &= -\frac{R}{k} P \\ (1-\nu) (R + \Delta R) u'(R + \Delta R) + \nu u(R + \Delta R) &= 0. \end{aligned}$$

Using the above solution  $u(r) = c_1 r + c_2 \frac{1}{r}$  leads to

$$\begin{aligned} R(1-\nu) \left( c_1 - \frac{1}{R^2} c_2 \right) + \nu \left( c_1 R + c_2 \frac{1}{R} \right) &= -\frac{R}{k} P \\ (R + \Delta R)(1-\nu) \left( c_1 - \frac{1}{(R + \Delta R)^2} c_2 \right) + \nu \left( c_1 (R + \Delta R) + c_2 \frac{1}{R + \Delta R} \right) &= 0 \end{aligned}$$

or as a system of linear equations

$$\begin{bmatrix} R & -\frac{1-2\nu}{R} \\ (R + \Delta R) & -\frac{1-2\nu}{R + \Delta R} \end{bmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} -\frac{R}{k} P \\ 0 \end{pmatrix}.$$

Using the above parameters the solutions are  $c_1 \approx 1.7532 \cdot 10^{-5}$  and  $c_2 \approx 7.0714 \cdot 10^{-7}$  and thus

$$u_r(r) = u(r) = c_1 r + c_2 \frac{1}{r} \approx 1.7532 \cdot 10^{-5} r + 7.0714 \cdot 10^{-7} \frac{1}{r}.$$

Using the results in Section 4.14.2 (page 83) all stresses and strain can be computed. The values coincide with the above FEM solutions.

```
k = E / ((1+nu) * (1-2*nu));
c = [R - (1-2*nu) / R; (R+dR) - (1-2*nu) / (R+dR)] \ [-R/k * P; 0];
r = R + linspace(0, dR, 100)';
u = c(1) * r + c(2) ./ r;
```

## 9.27 A crook with a weight attached

Examine the two L-shaped steel beams in Figure 139(a). Each beam has length  $L = H = 0.1$  with a square cross section of  $0.01 \times 0.01$ . The top edge is fixed and at the right end there is a force of 100 N (i.e. a weight of 10 kg) pulling the beam downwards. The corner at  $(x, y) = (0, 0)$  is slightly rounded, since the highest stresses are expected to show up in this area, see Figure 139(b). The applied force of 100 N leads to a surface force density of  $gN_2 = \frac{100 \text{ N}}{0.01^2 \text{ m}^2} = 10^6 \frac{\text{N}}{\text{m}^2}$ .

Start out by defining the domain and generating the mesh with the help of `CreateMeshTriangle()`. Since second order elements do not suffer from shear-locking use `MeshUpgrade()` to generate second order elements.

### Crook.m

```
W = 0.01; H = 0.1; Load = 1e6;;
Layers = 2*5; gap = W/5;

if 0 %% no rounding
    Domain = [-W -W -22; -W H -11; 0 H -22; 0 gap -22; ..
             0 0 -22; gap 0 -22; H 0 -23; H -W -22];
else %% with a rounded corner
    Domain = [-W -W -22; -W H -11; 0 H -22; 0 gap -22; ...
             gap*0.366 gap*0.366 -22; gap 0 -22; H 0 -23; H -W -22];
endif

FEMmesh = CreateMeshTriangle('Crook1', Domain, (W/Layers)^2);
figure(1); FEMtrimesh(FEMmesh); xlabel('x'); ylabel('y'); axis([-W 3*gap -W 3*gap])
FEMmesh = MeshUpgrade(FEMmesh, 'quadratic');
```

Then find the approximate displacements  $u_1$  and  $u_2$  by calling `PlaneStress()`. The code segment belows estimates the maximal vertical displacement by  $-8.96 \cdot 10^{-4}$  m, i.e. approximately 0.9 mm. Then examine the vertical displacement generated by the horizontal section of the crook. To verify the order of magnitude of the displacement use two arguments:

1. For a bending Euler beam with the dimensions of one arm obtain

$$u_2(L) = -\frac{4F}{EWH^3} L^3 \approx -\frac{4 \cdot 10^2}{200 \cdot 10^9 \cdot 0.01^4} 0.1^3 = -2 \cdot 10^{-4},$$

i.e. a displacement of 0.2 mm.

2. The slope of the lower arm at the left starting point is estimated by  $-6.56 \cdot 10^{-3}$  and with the length  $H = L = 0.1$  this leads to another contribution of  $\approx 0.56$  mm.

The sum of the two contributions is not too far from the result 0.9 mm by FEMoctave.

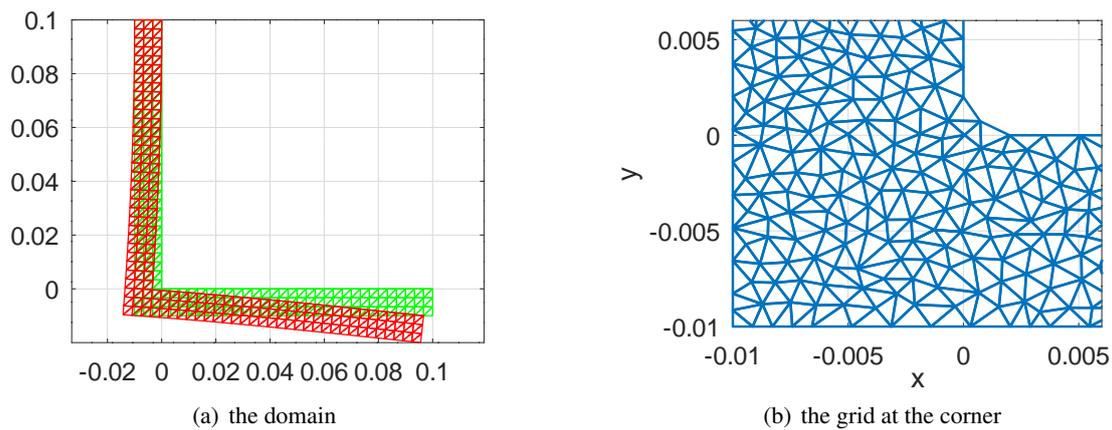


Figure 139: Original and deformed domain for the hook with attached weight at the right edge

### Crook.m

```

E = 200e9; nu = 0.25; %%% steel
[u1,u2] = PlaneStress(FEMmesh,E,nu,{0,0},{0,0},{0,-Load});

MaximalDisplacement = min(u2)
[~,slope_x,~] = FEMgriddata(FEMmesh,u2,0,-W/2)
i = linspace(-0.01,0.1)'; xi = -0.005*ones(size(yi));
uli = FEMgriddata(FEMmesh,u1,xi,yi);
figure(8); plot(yi,uli)
           xlabel('y'); ylabel('u_1')
p = polyfit(yi,uli,2); % linear regression of a polynomial of degree 2
slope = polyval([2*p(1) p(2)],-W/2) % evaluate the derivative of the polynomial
-->
MaximalDisplacement = -8.9570e-04
slope_x = -6.5645e-03
slope = 6.6130e-03

```

To determine the slope of the horizontal beam at the left starting point the result by FEMoctave was used above. One can use an analytical approximation by using the moment applied to the vertical beam, generated by the force at the right endpoint. Along the centerline of the vertical beam use

$$\frac{\partial^2 u_1(y)}{\partial y^2} = \frac{-F(H+W/2)}{EI} = \frac{-F(H+W/2)}{E \frac{1}{12} W^3 W} \approx 6.3 \cdot 10^{-2}.$$

Then use the conditions  $u_1(H) = u_1(0.1) = \frac{\partial u_1(H)}{\partial y} = 0$  at the top edge to estimate  $\frac{\partial u_1(-W/2)}{\partial y} \approx 6.62 \cdot 10^{-3}$ , which is rather close to the FEMoctave result of  $6.56 \cdot 10^{-3}$ . The horizontal displacement  $u_1$  along the centerline of the vertical beam is shown in Figure 142(a).

To generate Figure 139(a) with the scaled deformation also shown, start out by creating a coarse mesh and evaluate the displacement at those nodes. Then show the original and deformed mesh with different colors.

### Crook.m

```

CoarseMesh = CreateMeshRect([-W:W/3:H],[-W:W/3:H],[-11,-11,-11,-11]);
x = CoarseMesh.nodes(:,1); y = CoarseMesh.nodes(:,2);
uli = FEMgriddata(FEMmesh,u1,x,y); u2i = FEMgriddata(FEMmesh,u2,x,y);

```

```
x(isnan(u1)) = NaN;

figure(2); clf; factor = H/10/abs(min(u2));
trimesh(CoarseMesh.elem,x,y,'color','green','linewidth',1); hold on;
trimesh(CoarseMesh.elem,x+factor*u1,y+factor*u2i,'color','red','linewidth',1)
axis equal; hold off
```

To examine the mechanical load of the structure evaluate the stresses by calling `EvaluateStress()`. By asking for three return arguments a plane stress model is used. It is easy to generate graphs of the whole structure, but more insight might be gained by a closer look at some slices.

- At height  $y = \frac{H}{2} = 0.05$  examine the normal stress  $\sigma_y$  in  $y$ -direction. The result in Figure 140(a) show a compression in the left segment and traction on the right. This corresponds to the bending on the vertical arm. By integrating  $\sigma_y$  along this slice one should obtain the value of the force applied on the right edge of the crook, i.e.

$$W \int_{-W}^0 \sigma_y(x, 0.05) dx \approx \text{Force} = 100 .$$

For the moment with respect to the origin  $(0, 0)$  we expect

$$W \int_{-W}^0 x \sigma_y(x, 0.05) dx \approx H \cdot \text{Force} = 10 .$$

Both results are confirmed by the code below. The values of the normal stress  $\sigma_x$  are approximately zero.

- At  $x = \frac{H}{2} = 0.05$  examine the normal stress  $\sigma_x$  in  $x$ -direction along a vertical slice. The result in Figure 140(b) shows a compression in the lower segment and traction in the upper segment. This corresponds to the downward bending on the horizontal arm. For the moment with respect to the point  $(\frac{H}{2}, 0)$  we expect

$$W \int_{-W}^0 y \sigma_x(0.05, y) dy \approx \frac{1}{2} H \cdot \text{Force} = 5 .$$

The values of the normal stress  $\sigma_y$  are approximately zero. By integrating the shearing stress  $\tau_{xy}$  obtain again the applied force, i.e.

$$W \int_{-W}^0 \tau_{xy}(0.05, y) dy \approx \text{Force} = -100 .$$

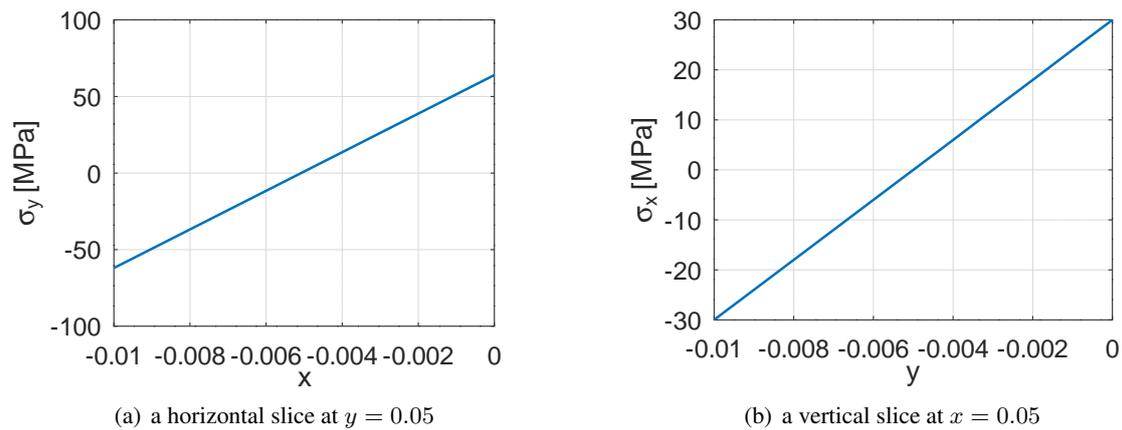
- Observe that the stress values in the vertical arm are considerably larger than in the horizontal arm. The values in Figure 140(b) are at  $x = 0.05$ . For larger values of  $x$  the strains  $\sigma_x$  will be even smaller.

### Crook.m

```
[sigma_x,sigma_y,tau_xy] = EvaluateStress(FEMmesh,u1,u2,E,nu);

dist = linspace(-W,0,100)'; HH = H/2*ones(size(dist));
sigma_y_slice_H = FEMgriddata(FEMmesh,sigma_y,dist,HH);
figure(3); plot(dist,sigma_y_slice_H/1e6);
            xlabel('x'); ylabel('\sigma_y [MPa]');xlim([-W,0])

sigma_y_slice_H(isnan(sigma_y_slice_H)) = 0;
Integral_sigma_y = W*trapz(dist,sigma_y_slice_H)
Integral_Moment = W*trapz(dist,dist.*sigma_y_slice_H)
```

(a) a horizontal slice at  $y = 0.05$ (b) a vertical slice at  $x = 0.05$ Figure 140: A horizontal slice with  $\sigma_y$  shown and a vertical slice with  $\sigma_x$  shown

```

sigma_x_slice_V = FEMgriddata(FEMmesh,sigma_x,HH,dist);
tau_xy_slice_V  = FEMgriddata(FEMmesh,tau_xy,HH,dist);
figure(4); plot(dist,sigma_x_slice_V/1e6);
                xlabel('y'); ylabel('\sigma_x [MPa]'); xlim([-W,0])
Integral_Moment_x = W*trapz(dist,dist.*sigma_x_slice_V)
Integral_tau_xy   = W*trapz(dist,tau_xy_slice_V)
-->
Integral_sigma_y = 99.769
Integral_Moment  = 10.002
Integral_Moment_x = 5.0005
Integral_tau_xy  = -99.985

```

Since steel is a ductile material one can use the von Mises stress to decide whether the crook will withstand the force of 100 N. Use `EvaluateVonMises()` to find the values of the von Mises stress at the nodes and then `FEMtrisurf()` and `FEMtricontour()` to generate Figure 141. The contour lines in Figure 141 are supplemented with the borders of the domain. The spikes of the von Mises stress at the corner  $(0,0)$  should be no surprise to mechanical engineers. One possible measure to reduce the maximal value of von Mises is the rounding visible in Figure 139(b). To obtain more insight the von Mises stress is evaluated along the straight line connecting  $(-W,-W)$  and  $(0,0)$ , using `FEMgriddata()`, leading to Figure 142(b). Since the yield strength of steel is  $\approx 330$  MPa the crook should be able to support the force of 100 N.

#### Crook.m

```

vonMises = EvaluateVonMises(sigma_x,sigma_y,tau_xy);
figure(5); FEMtrisurf(FEMmesh,vonMises/1e6);
            xlabel('x'); ylabel('y'); zlabel('von Mises [MPa]'); view(160,25)
            colorbar(); shading interp
figure(6); clf; FEMtricontour(FEMmesh,vonMises/1e6,1e1*[0:0.5:6]);
            xlabel('x'); ylabel('y'); title('von Mises [MPa]');
            caxis(1e2*[0 0.7]); axis equal; colorbar(); hold on
            plot([Domain(:,1);Domain(1,1)], [Domain(:,2);Domain(1,2)], ...
                'color','black','linewidth',1); hold off

dist = linspace(-W,gap,100)'; HH = H/2*ones(size(dist));
vonMises_slice = FEMgriddata(FEMmesh,vonMises,dist,dist);
figure(7); plot(dist,vonMises_slice*1e-6);

```

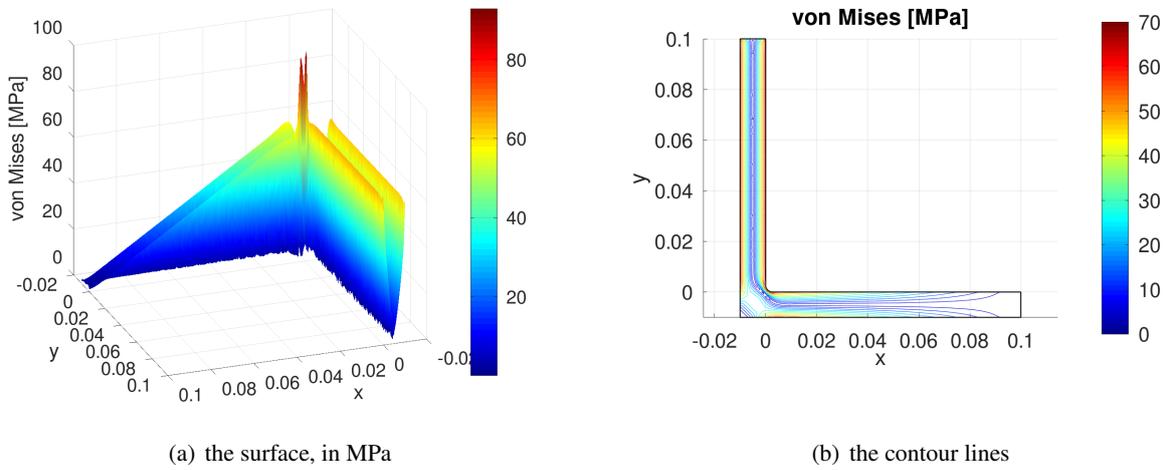


Figure 141: The von Mises stress on the crook, as surface and level curves

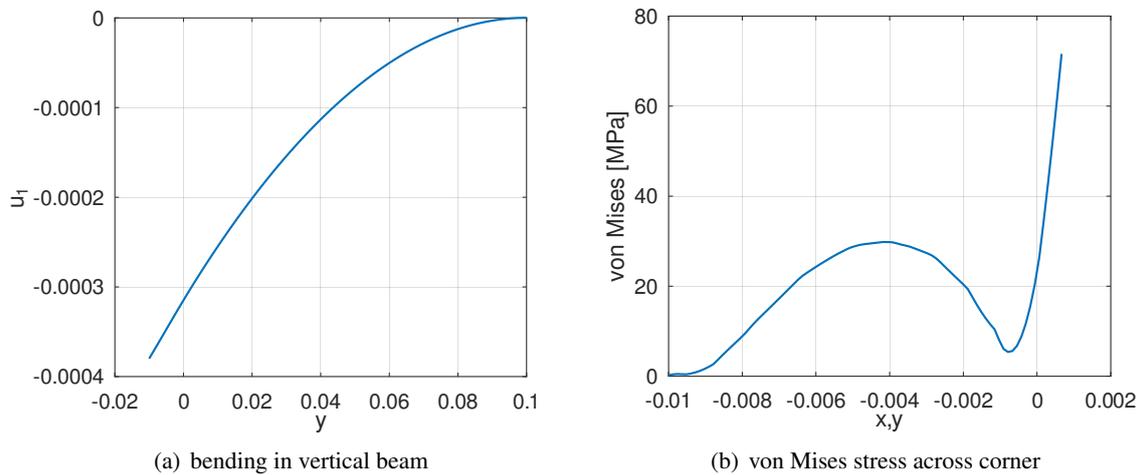


Figure 142: The bending of the centerline of the vertical beam and the von Mises stress on the 45° line through the origin

```
xlabel('x,y'); ylabel('von Mises [MPa]')
```

## 9.28 A wrench

A classical example application for mechanical FEM is a wrench. With a digital image of a typical wrench the tool `xinput()` is used in *Octave* to grab the contour data from the screen and written to the file `WrenchData.m`, see [Stah22, §3.9]. Then rescale the contour to obtain a typical length of 0.15 m of the wrench in Figure 143(a). Then setup an appropriate configuration of the wrench.

- The material is steel with the parameters  $E = 200$  GPa and  $\nu = 0.25$ .
- Most of the boundary is force free, thus with the code `-22`, according to Table 5 on page 47.
- Along the two horizontal sections on the very left the displacements are zero, modeling the screw head in the wrench. A closer look at the contour data shows that these are sections 1 and 4 of the contour, used with the code `-11` for the boundary condition.
- The applied force is 100 N over a length of 0.05 m and width 0.005 m, leading to a force density of  $\frac{100}{0.05 \cdot 0.005} = 4 \cdot 10^5 \frac{\text{N}}{\text{m}^2}$ . This load is applied on segment 17 of the contour, used with the code `-23` for no force in  $x$  direction and the given load in  $y$  direction.

With this data the mesh is generated by calling `CreateMeshTriangle()`. Then the mesh of linear elements should be upgraded to quadratic or cubic elements with the help of `MeshUpgrade()`. Then use `PlaneStress()` to solve for the displacements  $u_1$  and  $u_2$ .

### Wrench.m

```
load WrenchData.m           %% load the contour data
scale = 0.15/max(x);       %% scale the contour data
x = scale*x; y = scale*y;
Order = 3;                 %% select the order of the elements 1,2 or 3
BC = -22*ones(size(x));    %% default is a force free boundary
BC([1 4]) = -11; BC(17) = -23; %% fixed at the two horizontal section on the left
                                %% vertical force on top right segment
Load = 100/(0.05*0.005);   %% 100 N, distributed over length 0.05 and width 0.005

Mesh = CreateMeshTriangle('Wrench', [x,y,BC], 0.01^2/4); %% create the mesh
switch Order
  case 2 Mesh = MeshUpgrade(Mesh, 'quadratic');
  case 3 Mesh = MeshUpgrade(Mesh, 'cubic');
endswitch
E = 200e9; nu = 0.25; gN = {0, -Load}; %% data for steel
[u1,u2] = PlaneStress(Mesh,E,nu,{0,0},{0,0},gN); %% solve the plane stress problem
```

With the solution the original and deformed shape can be displayed with the applied load visualized by a few vectors, see Figure 143(a).

### Wrench.m

```
%%display the original and deformed wrench, with the applied force
scale = 0.001*max(y)/max(u2);
x_force = linspace(x(17),x(18),8); y_force = 0.038*ones(size(x_force))+0.02;
vec_x = zeros(size(x_force)); vec_y = -0.02*ones(size(x_force));
figure(1); clf
trimesh(Mesh.elem,Mesh.nodes(:,1),Mesh.nodes(:,2),...
        'color','green','linewidth',1); hold on
```

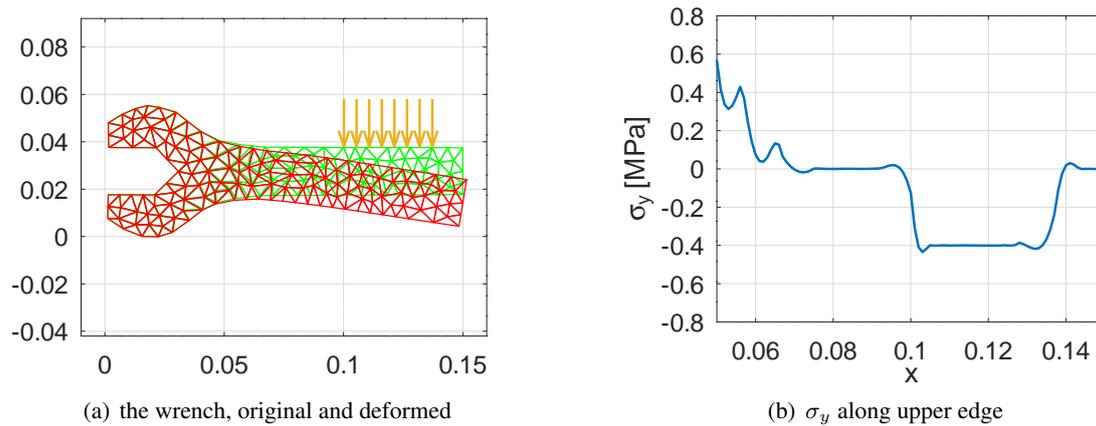


Figure 143: The deformed wrench and the stress  $\sigma_y$  along upper edge with the applied load

```

trimesh(Mesh.elem,Mesh.nodes(:,1)+scale*u1,Mesh.nodes(:,2)+scale*u2,...
        'color','red','linewidth',1)
quiver(x_force,y_force,vec_x,vec_y,0)
hold off; axis equal; xlim([-0.01, 0.16])

```

Then evaluate the stresses at the nodes, including the von Mises Stress. By asking for three return arguments the plane stress situation is used. By a piecewise linear interpolation and `FEMgriddata()` the vertical stress  $\sigma_y$  can be evaluated along the upper edge, leading to Figure 143(b). The external load of  $-0.4$  MPa is clearly visible.

#### Wrench.m

```

[sigma_x,sigma_y,tau_xy] = EvaluateStress(Mesh,u1,u2,E,nu); %% basic stress
vonMises = EvaluateVonMises(sigma_x,sigma_y,tau_xy); %% von Mises stress

xi = linspace(0.05,0.15,101)'; yi = interp1(x(14:19),y(14:19),xi);
sigma_y_interp = FEMgriddata(Mesh,sigma_y,xi,yi);

figure(2); plot(xi,sigma_y_interp/1e6)
xlabel('x'); ylabel('\sigma_y [MPa]'); xlim([0.05,0.15])

```

Since steel is a ductile metal the von Mises stress can be used to examine the effect on the wrench. In Figure 144(a) find the surface plot of the von Mises stress. The highest stress is on the boundary at the mid section, but spikes are also visible at the sharp corners on the left. Figure 144(b) shows the contour lines and the position of the highest and lowest von Mises stress. It should be no surprise that the section on the very right is almost stress free.

#### Wrench.m

```

figure(3); clf; FEMtrimesh(Mesh,vonMises/1e6)
zlabel('von Mises stress');
colorbar(); view([40 75])
xlim([0 0.15]); ylim([-0.025 0.09])
set(gca, 'XTickLabel', [], 'yTickLabel', [], 'zTickLabel', [])

MaxVonMises = max(vonMises); MinVonMises = min(vonMises);
Max_Min_vonMises_MPa = [MaxVonMises,MinVonMises]/1e6
MaxInd = find(vonMises == MaxVonMises); MaxPosition = Mesh.nodes(MaxInd,:);

```

```

MinInd = find(vonMises == MinVonMises); MinPosition = Mesh.nodes(MinInd, :);

figure(4); clf; FEMtricontour(Mesh, vonMises/1e6, 41)
    hold on; plot([x;x(1)], [y;y(1)], 'k');
    plot(MaxPosition(1), MaxPosition(2), ' *r', MinPosition(1), MinPosition(2), ' *b');
    hold off; axis equal

-->
Max_Min_vonMises_MPa = 1.2415e+01 2.4078e-03

```

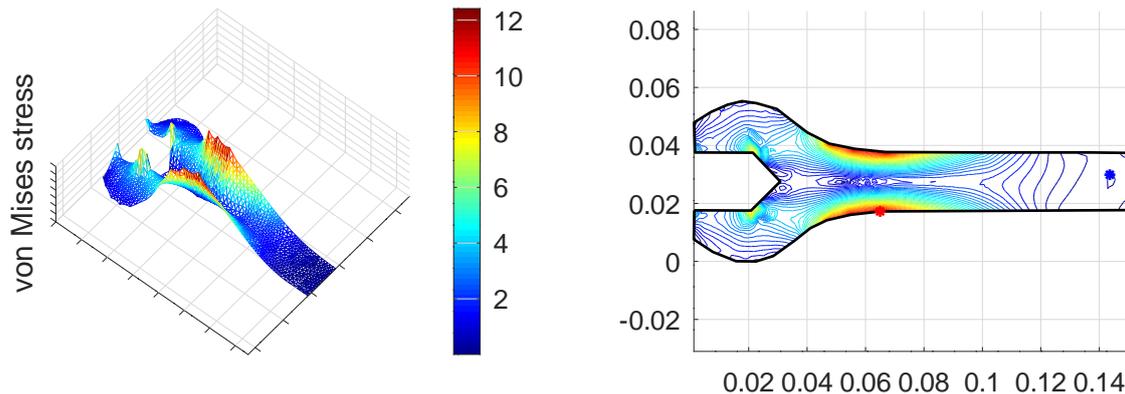


Figure 144: Surface and contour plot of the von Mises stress in [MPa]

### 9.29 A rotating rubber cylinder

A cylinder with radius  $R = 0.2$  and height  $2H = 0.2$  is rotating about the  $z$ -axis with 10 revolutions per second. The wall consist of a Silicone rubber of thickness 0.01 and the cover and bottom are 0.02 thick. The goal is to determine the resulting deformation and the von Mises stress.

Using an axially symmetric setup only a cross section in the  $y = 0$  plane for  $x = r > 0$  have to be examined. Since the setup is symmetric with respect to the plane  $z = 0$  only the upper half has to be modeled, using the zero  $z$  displacement at the lower edge.

Start by defining the parameters and generating the mesh. In this case third order elements are used. Then define the function for the centrifugal force and solve for the two displacements  $u_r$  and  $u_z$  with the help of `AxiStress()`.

#### RubberBox.m

```

rho = 1100; E = 1e6; nu = 0.47; %% Silicone rubber
H = 0.1; R = 0.2; W = 0.01;
Contour = [0 H -11; 0 H-2*W -22; R-2*W H-2*W -22; R-W H-2*W -22;
           R-W 0 -21; R 0 -22; R H-W -22; R-W H -22];
Mesh = CreateMeshTriangle('RubberBox', Contour, 3e-5);
Mesh = MeshUpgrade(Mesh, 'cubic');

function res=fr(xy, dummy)
    freq = 10; omega = freq*2*pi; rho = 1100;
    res = rho*xy(:,1)*omega^2;
endfunction

[ur, uz] = AxiStress(Mesh, E, nu, {'fr', 0}, {0, 0}, {0, 0});

```

Then display the original and the deformed domain in Figure 145 and the displacements in Figure 146.

**RubberBox.m**

```
factor = 1;
figure(10); trimesh(Mesh.elem,Mesh.nodes(:,1)+factor*ur,...
    Mesh.nodes(:,2)+factor*uz,'color','red','linewidth',2);
hold on;trimesh(Mesh.elem,Mesh.nodes(:,1),Mesh.nodes(:,2),'color','green','linewidth',1);
hold off; axis equal; xlabel('x'); ylabel('y');

figure(11); FEMtrimesh(Mesh,ur); xlabel('r'); ylabel('z'); zlabel('u_r')
figure(12); FEMtrimesh(Mesh,uz); xlabel('r'); ylabel('z'); zlabel('u_z')
```

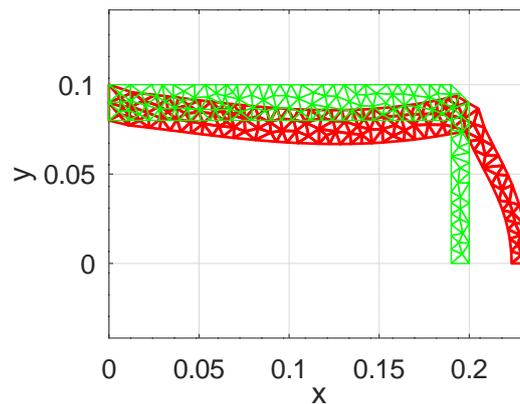


Figure 145: The upper half of the original and deformed domain for the rotating rubber box

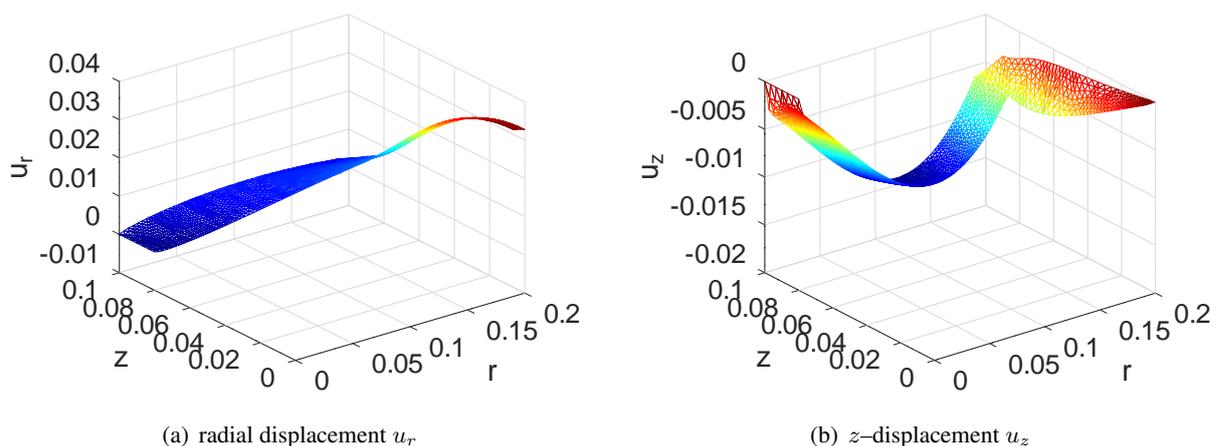


Figure 146: The displacements  $u_r$  and  $u_z$  for the rotating rubber box

As last step evaluate the stresses and then the von Mises stress, leading to the surface and contour plots in Figure 147.

**RubberBox.m**

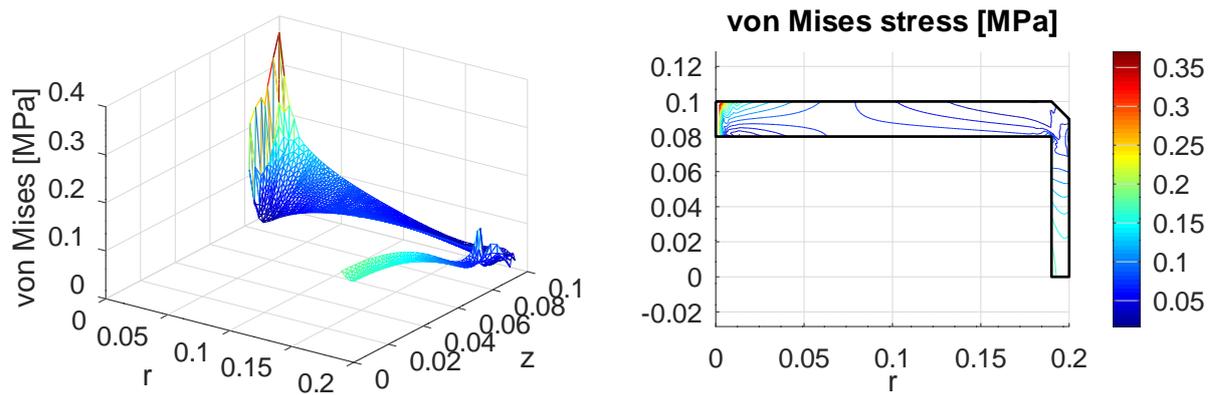


Figure 147: The von Mises stress for the rotating rubber box

```
[sigma_x,sigma_y,sigma_z,tau_xz] = EvaluateStressAxi(Mesh,ur,uz,E,nu);
vonMises = EvaluateVonMisesAxi(sigma_x,sigma_y,sigma_z,tau_xz);
figure(13); FEMtrimesh(Mesh,vonMises/1e6)
    xlabel('r'); ylabel('z'); zlabel('von Mises [MPa]'); view([35 30])
figure(14); clf; FEMtricontour(Mesh,vonMises/1e6)
    xlabel('r'); ylabel('z'); zlabel('von Mises [MPa]')
    hold on; plot([Contour(:,1);Contour(1,1)], [Contour(:,2);Contour(1,2)], 'k')
    hold off; axis equal; colorbar; %%shading interp
    title('von Mises stress [MPa]')
```

### 9.30 A washer fastener examined as spring

In this example a washer fastener design is examined. The goal is to determine the force required to deform the washer.

#### 9.30.1 The setup

- The material is aluminum, with density  $\rho = 2700 \frac{\text{kg}}{\text{m}^3}$ , Young's modulus  $E = 70 \text{ GPa}$  and Poisson ratio  $\nu = 0.33$ .
- The intersection of the washer with the plane  $y = 0$  is almost rectangular. The inner part is moved up slightly and there are two horizontal sections, one at the inner/upper location at height  $z = 0.001$  and the second at the lower/outer section at height  $z = 0$ . Find the domain in Figure 148. The corners of the domain are given by the six points

$r$ [m]	0.0020	0.0020	0.0044	0.0050	0.0050	0.0026
$z$ [m]	0.0010	0.0004	0	0	0.0006	0.0010

and connected by straight line segments. This domain is then rotated about the  $z$ -axis to obtain the washer in  $\mathbb{R}^3$ .

To determine the resulting deformation of the washer the boundary conditions have to be specified.

- The outer/lower edge at height  $z = 0$  is fixed in  $z$ -direction, but free in radial direction.

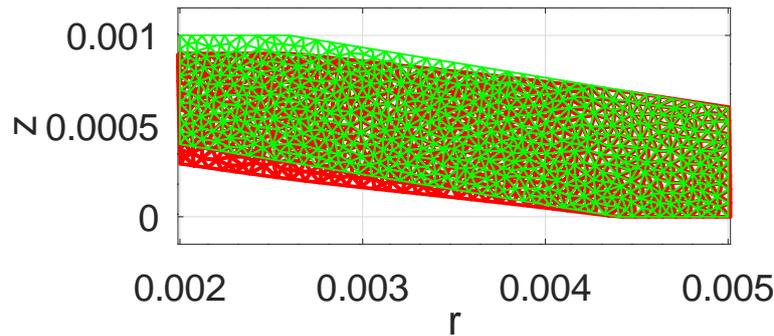


Figure 148: The original and deformed domain

- The inner/upper edge at height  $z = 0.001$  is moved downward by  $0.0001$  m and free in radial direction.
- All other edges are force free.

With this information the boundary value problem can be solved, using the command `AxiStress()`. The code contains additional configurations with different boundary conditions on the inside and outside.

#### WasherSpring.m

```
pkg load femoctave
rho = 2700; E = 70e9; nu = 0.33; %% Aluminum
H = 0.001; Ri = 0.002; Ro = 0.005; D = 0.0006; H = 0.0004;
global Offset
Offset = 1*1e-4;

if 1 %% free sides
    Contour = [Ri H+D -22; Ri H -22;Ro-D 0 -21; Ro 0 -22; Ro D -22;Ri+D H+D -21];
elseif 0 %% clamped on the outside
    Contour = [Ri H+D -22; Ri H -22;Ro-D 0 -21; Ro 0 -12; Ro D -22;Ri+D H+D -21];
else %% clamped on both sides
    Contour = [Ri H+D -12; Ri H -22;Ro-D 0 -21; Ro 0 -12; Ro D -22;Ri+D H+D -21];
endif

Mesh = CreateMeshTriangle('Washer',Contour,2.5e-9);
%%Mesh = MeshUpgrade(Mesh,'quadratic');
Mesh = MeshUpgrade(Mesh,'cubic');

function res = gDz(xy,dummy)
    global Offset
    res = -Offset*(xy(:,2)>Offset);
endfunction

[ur,uz] = AxiStress(Mesh,E,nu,{0,0},{0,'gDz'},{0,0});
factor = 1;
figure(10);clf; trimesh(Mesh.elem,Mesh.nodes(:,1)+factor*ur,...
    Mesh.nodes(:,2)+factor*uz,'color','red','linewidth',2);
hold on; trimesh(Mesh.elem,Mesh.nodes(:,1),Mesh.nodes(:,2),...
    'color','green','linewidth',1);
hold off; axis equal; xlabel('r'); ylabel('z');
xticks([2:5]/1000); yticks([0:0.5:1]/1000)
```

Display the radial displacement  $u_r$  in Figure 149 and the height displacement  $u_z$  in Figure 150.

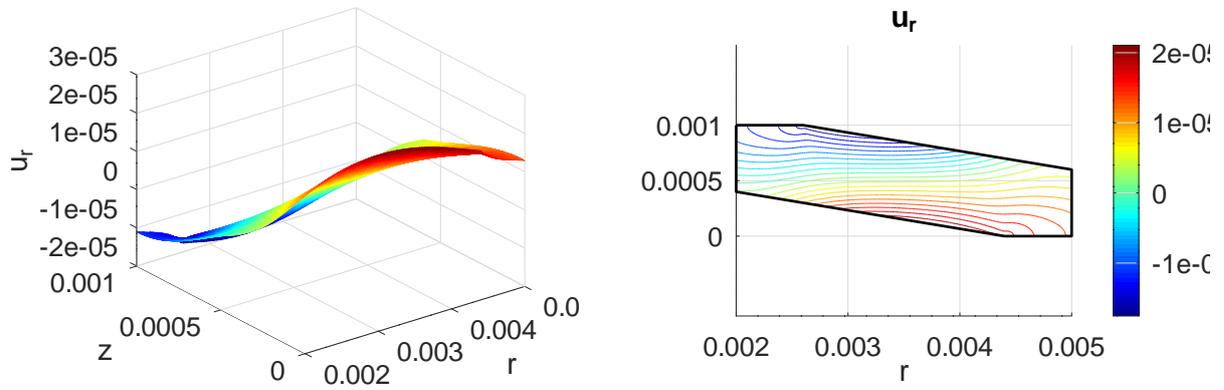


Figure 149: The radial displacement  $u_r$

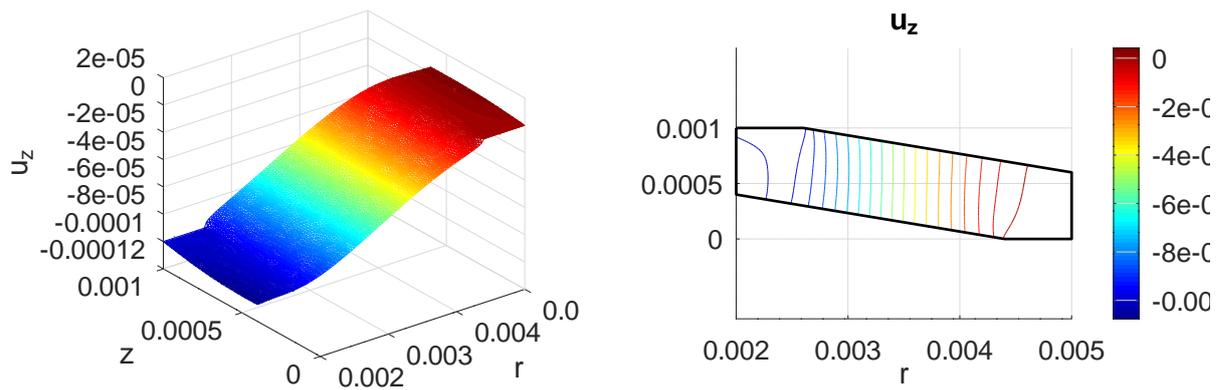


Figure 150: The height displacement  $u_z$

## WasherSpring.m

```

figure(11); FEMtrimesh(Mesh,ur)
    xlabel('r'); ylabel('z'); zlabel('u_r')
    xticks([2:5]/1000); yticks([0:0.5:1]/1000)
Cx = [Contour(:,1);Contour(1,1)]; Cy = [Contour(:,2);Contour(1,2)];
figure(21); clf; FEMtricontour(Mesh,ur)
    hold on ; plot(Cx,Cy,'k'); hold off
    xlabel('r'); ylabel('z'); title('u_r');
    axis equal; colorbar;xticks([2:5]/1000); yticks([0:0.5:1]/1000)
figure(12); FEMtrimesh(Mesh,uz)
    xlabel('r'); ylabel('z'); zlabel('u_z');
    xticks([2:5]/1000); yticks([0:0.5:1]/1000)
figure(22); clf; FEMtricontour(Mesh,uz)
    hold on ; plot(Cx,Cy,'k'); hold off
    xlabel('r'); ylabel('z'); title('u_z');
    axis equal; colorbar;xticks([2:5]/1000); yticks([0:0.5:1]/1000)

```

## 9.30.2 Evaluate the force by integrating the normal stress

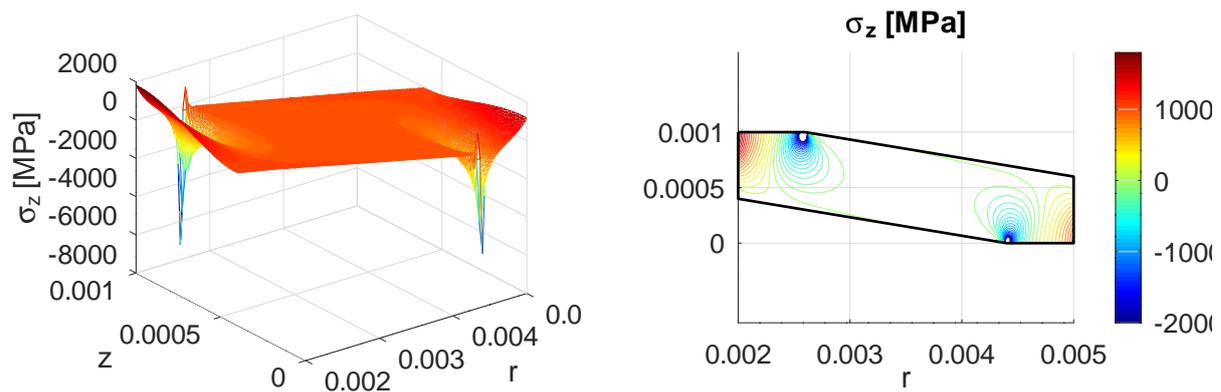
To determine the force  $F$  required to push the upper edge down by 0.1 mm use the normal stress  $\sigma_z$  in vertical direction.

## WasherSpring.m

```

[sigma_x,sigma_y,sigma_z,tau_xz] = EvaluateStressAxi(Mesh,ur,uz,E,nu);
figure(13); FEMtrimesh(Mesh,sigma_z*1e-6)
    xlabel('r'); ylabel('z'); zlabel('\sigma_z [MPa]');
    xticks([2:5]/1000); yticks([0:0.5:1]/1000)
figure(23); clf; FEMtricontour(Mesh,sigma_z/1e6,[-20:1:20]*100)
    hold on ; plot(Cx,Cy,'k'); hold off
    xlabel('r'); ylabel('z'); title('\sigma_z [MPa]');
    axis equal; colorbar;xticks([2:5]/1000); yticks([0:0.5:1]/1000)

```

Figure 151: The normal stress  $\sigma_z$ 

At any height  $0 \leq h \leq 0.001$  examine the slice  $a \leq r \leq b$  in the domain visible in Figure 148 and perform an integration to determine  $F$ .

$$F = 2\pi \int_{r=a}^b r \sigma_z(r, h) dr$$

Examine the graph of the normal stress  $\sigma_z$  in Figure 151 and observe the singularities at the corners of the edges with fixed displacement. These singularities cause serious numerical trouble when trying to integrate along the upper or lower edges.

On a mesh with elements of order 3 with 5968 free nodes obtain the numerical results<sup>36</sup>

$$\begin{aligned} F_{up} &\approx 1415.1 \text{ N} \\ F_{middle} &\approx 2401.6 \text{ N} \\ F_{low} &\approx 1085.8 \text{ N} \end{aligned}$$

and the graphs in Figure 152. By changing the element types or the size of the meshes the results at the lower and upper edges can change substantially, while the result at half height remains stable and thus is more reliable. By changing the height of the slice in the code below (modify the value of  $s$ ) one may observe that the results for heights between 20% and 80% are stable.

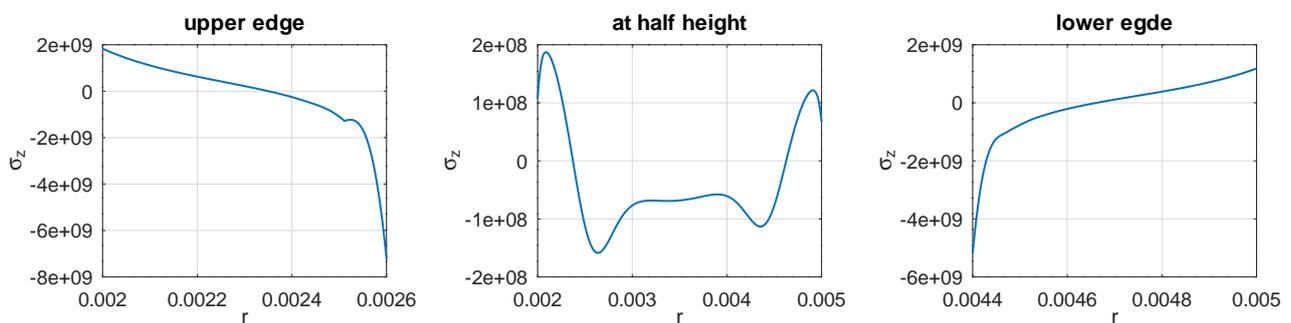


Figure 152: The normal pressures  $\sigma_z$  along upper and lower edge and at half height

#### WasherSpring.m

```
r = linspace(0,D,1000)';
sigma_up = FEMgriddata(Mesh,sigma_z,Ri+r,(H+D)*ones(size(r)));
figure(31); plot(Ri+r,sigma_up); xlabel('r'); ylabel('\sigma_z'); title('upper edge')
xlim([Ri,Ri+D]); xticks([2:0.2:2.6]/1000);
Force_up = 2*pi*trapz(Ri+r,sigma_up.*(Ri+r))

sigma_low = FEMgriddata(Mesh,sigma_z,Ro-D+r,zeros(size(r)));
figure(32); plot(Ro-D+r,sigma_low); xlabel('r'); ylabel('\sigma_z'); title('lower edge')
xlim([Ro-D, Ro]); xticks([4.4:0.2:5]/1000);
Force_low = 2*pi*trapz(Ro-D+r,sigma_low.*(Ro-D+r))

s = 0.5; %% select the height
r_mid = linspace(Ri,Ro,1000)';
sigma_mid = FEMgriddata(Mesh,sigma_z,r_mid,s*(H+D)*ones(size(r_mid)));
ind = find(isfinite(sigma_mid));
r_mid = r_mid(ind); sigma_mid = sigma_mid(ind);
figure(33); plot(r_mid,sigma_mid); xlabel('r'); ylabel('\sigma_z');
title('at half height'); xticks([2:5]/1000);
Force_mid = 2*pi*trapz(r_mid,sigma_mid.*r_mid)
```

<sup>36</sup>A computation with Comsol Multiphysics lead to a force of 2395.5 N at half height and an elastic energy of 0.12077 J. The shape of the graphs in Figure 152 is confirmed.

### 9.30.3 Evaluate the force by an energy argument

Since the above evaluation of the required force  $F$  is rather delicate, it is a good idea to examine an alternative approach. Since the problem is linear the force  $F$  depends linearly on the displacement  $d$  of the upper edge, i.e. with a displacement  $0 \leq s \leq d$  obtain  $F(s) = ks = \frac{s}{d} F(d) = \frac{s}{d} F$ . An elementary integration leads to the energy  $U$  in the system.

$$U = \int_0^d F(s) ds = \int_0^d \frac{s}{d} F ds = \frac{1}{2} d F$$

Thus find the force  $F = \frac{2U}{d}$ . Using the results in Section 8.6 (starting on page 190) the elastic energy  $U$  is given by

$$\begin{aligned} U(\vec{u}) = & 2\pi \iint_{\Omega} \frac{r E}{2(1+\nu)(1-2\nu)} \left( (1-\nu) \left( \left( \frac{\partial u_r}{\partial r} \right)^2 + \left( \frac{\partial u_z}{\partial z} \right)^2 + \frac{1}{r^2} u_r^2 \right) + \right. \\ & \left. + 2\nu \left( \left( \frac{\partial u_r}{\partial r} \right) \left( \frac{\partial u_z}{\partial z} \right) + \frac{1}{r} u_r \left( \frac{\partial u_r}{\partial r} + \frac{\partial u_z}{\partial z} \right) \right) \right) dA + \\ & + 2\pi \iint_{\Omega} \frac{r E}{1+\nu} \frac{1}{4} \left( \frac{\partial u_r}{\partial z} + \frac{\partial u_z}{\partial r} \right)^2 dA . \end{aligned}$$

Since the displacements  $u_r$  and  $u_z$  are available use the function `FEMIntegrate()`. It is best to evaluate the displacements and their partial derivatives at the Gauss points by using `FEMEvaluateGP()` and then integrate. For this example the result is  $F = 2402.1$  N, which is close to the above integration at half height of the normal stress  $\sigma_z$ .

#### WasherSpring.m

```
[urGP,ur_rz] = FEMEvaluateGP(Mesh,ur); [uzGP,uz_rz] = FEMEvaluateGP(Mesh,uz);
rGP = Mesh.GP(:,1);
w = E/(2*(1+nu)*(1-2*nu))*rGP.*(1-nu)*(ur_rz(:,1).^2+uz_rz(:,2).^2+(urGP./rGP).^2)...
    + 2*nu*(ur_rz(:,1).*uz_rz(:,2) + 1./rGP.*urGP.*(ur_rz(:,1)+uz_rz(:,2)))...
    + E/(4*(1+nu))*rGP.*(ur_rz(:,2)+uz_rz(:,1)).^2;
U_elast = 2*pi*FEMIntegrate(Mesh,w);
Force_energy = 2*U_elast/Offset
```

### 9.30.4 Comparison of linear, quadratic and cubic elements

The above results were generated with a mesh consisting of 1294 triangle and piecewise cubic functions. It is easy to recompute, using the same number of triangles, but linear or quadratic functions. Find the results in Table 16.

- The FEM algorithm is minimizing the energy  $U$  of the system amongst the functions to be used. The space of piecewise linear functions is a strict subspace of the piecewise quadratic functions. Thus the minimal energy will be smaller when using elements of order 2 than with elements of order 1. As a consequence linear element will overestimate the resulting force  $F = \frac{2U}{d}$ .
- The space of piecewise quadratic functions is a strict subspace of the piecewise cubic functions. Thus the minimal energy will be smaller when using elements of order 3 than with elements of order 1. The force  $F$  evaluated with elements of order 3 will be the smallest. This is confirmed in Table 16.
- Using finer meshes will lead to smaller minimal energies  $U$  and thus smaller forces  $F$ .
- The estimates in Section 5.4 on page 104 lead to factors of 0.5, 2 or 4.5 for the ratio of number of nodes divided by the number of triangles for linear, quadratic or cubic elements. This is confirmed in Table 16.

element type	linear	quadratic	cubic
number of nodes	696	2685	5968
elastic energy $U$	0.12726	0.12118	0.12010
force $F = \frac{2U}{d}$	2545.2	2423.3	2402.1

Table 16: Comparison of different elements for the washer fastener example

### 9.30.5 Different boundary conditions

The above setup can be modified by changing the boundary conditions at the inner or outer edge.

- In the original setup both sides are free to move in radial direction.
- The second setup prevents the outer side to move in radial direction.
- The third setup prevents both sides to move in radial direction.

This is implemented by switches at in the first section of the code. The parameter `Contour` contains the values of the flags indicating the boundary conditions on the vertical segments.

```

if 1 %% free sides
    Contour = [Ri H+D -22; Ri H -22;Ro-D 0 -21; Ro 0 -22; Ro D -22;Ri+D H+D -21];
elseif 0 %% clamped on the outside
    Contour = [Ri H+D -22; Ri H -22;Ro-D 0 -21; Ro 0 -12; Ro D -22;Ri+D H+D -21];
else %% clamped on both sides
    Contour = [Ri H+D -12; Ri H -22;Ro-D 0 -21; Ro 0 -12; Ro D -22;Ri+D H+D -21];
endif

```

The results computed by the energy argument are

$$\begin{aligned}
 F &= 2402.1 \text{ N} && \text{with both sides free} \\
 F &= 2880.4 \text{ N} && \text{with inner side free and outer side fixed} \\
 F &= 3809.1 \text{ N} && \text{with both sides fixed}
 \end{aligned}$$

As expected the additional constraints lead to a stiffer system.

## Bibliography

- [AtkiHan09] K. Atkinson and W. Han. *Theoretical Numerical Analysis*. Number 39 in Texts in Applied Mathematics. Springer, 2009.
- [AxelBark84] O. Axelsson and V. A. Barker. *Finite Element Solution of Boundary Value Problems*. Academic Press, 1984.
- [Butc03] J. Butcher. *Numerical Methods for Ordinary Differential Equations*. John Wiley & Sons, Ltd, second edition, 2003.
- [Demm97] J. W. Demmel. *Applied Numerical Linear Algebra*. SIAM, Philadelphia, 1997.
- [GoluVanLoan96] G. Golub and C. Van Loan. *Matrix Computations*. Johns Hopkins University Press, third edition, 1996.
- [IzadSuayNoei21] M. Izadi, S. Yüzbaşı, and S. Noeiaghdam. Approximating Solutions of Non-Linear Troesch's Problem via an Efficient Quasi-Linearization Bessel Approach. *Mathematics*, 9(16), 2021.
- [Kell92] H. B. Keller. *Numerical Methods for Two-Point Boundary Value Problems*. Dover, 1992.
- [KubiHlav08] M. Kubicek and V. Hlavacek. *Numerical Solution of Nonlinear Boundary Value Problems with Applications*. Dover books on engineering. Dover, 2008.
- [MuelSilt12] J. Mueller and S. Siltanen. *Linear and Nonlinear Inverse Problems with Practical Applications*. Computational Science and Engineering. Society for Industrial and Applied Mathematics, 2012.
- [MuelSilt20] J. Mueller and S. Siltanen. The D-bar method for electrical impedance tomography—demystified. *Inverse Problems*, 36:28, 2020.
- [Prze68] J. Przemieniecki. *Theory of Matrix Structural Analysis*. McGraw-Hill, 1968. Republished by Dover in 1985.
- [Sege77] L. A. Segel. *Mathematics Applied to Continuum Mechanics*. MacMillan Publishing Company, New York, 1977. republished by Dover 1987.
- [www:triangle] J. R. Shewchuk. <https://www.cs.cmu.edu/~quake/triangle.html>.
- [Sout73] R. W. Soutas-Little. *Elasticity*. Prentice-Hall, 1973.
- [VarFEM] A. Stahel. Calculus of Variations and Finite Elements. Lecture Notes used at HTA Biel, 2000.
- [Stah08] A. Stahel. Numerical Methods. lecture notes, BFH-TI, 2008.
- [Stah22] A. Stahel. *Octave and MATLAB for Engineering Applications*. Springer Fachmedien Wiesbaden, Wiesbaden, first edition, 2022.
- [TongRoss08] P. Tong and J. Rossettos. *Finite Element Method, Basic Technique and Implementation*. MIT, 1977. Republished by Dover in 2008.
- [Wein74] R. Weinstock. *Calculus of Variations*. McGraw-Hill, New York, 1962. republished by Dover.
- [Zien13] O. Zienkiewicz, R. Taylor, and J. Zhu. *The Finite Element Method: Its Basis and Fundamentals*. Butterworth-Heinemann, 7 edition, 2013.

## List of Figures

1	A semi-disk as domain in $\mathbb{R}^2$ and a solution of a BVP . . . . .	9
2	Deformation of an elastic solid . . . . .	13
3	Definition of strain: rectangle before and after deformation . . . . .	13
4	Components of stress in space . . . . .	14
5	Solution of $-\Delta u = 0.25$ on a rectangle . . . . .	21
6	Solution of the Laplace equation in cylindrical coordinates . . . . .	22
7	Solution of a diffusion problem on a L-shaped domain . . . . .	23
8	Solution of a diffusion convection problem . . . . .	23
9	The fourth eigenfunction of $\Delta u = \lambda u$ on a disc . . . . .	25
10	The first four radial eigenmodes of the Laplace operator on a disk of radius 1 . . . . .	26
11	Solution of a dynamic heat equation . . . . .	28
12	Solution of a wave equation . . . . .	30
13	The solution for a simple 1D boundary value problem and the sparsity of the matrix <b>A</b> . . . . .	31
14	The solution of a steady state heat problem and the heat flux across spheres of radius $r$ . . . . .	32
15	The solution of dynamic heating of a ball, $u$ as function of $r$ or $t$ . . . . .	33
16	The solution of dynamic heating of a ball, $u$ as function of $r$ and $t$ . . . . .	33
17	The amplitude of a vibrating string, $u$ as function of $x$ and $t$ , the surface . . . . .	34
18	The amplitude of a vibrating string, $u$ as function of $x$ and $t$ , the contour lines . . . . .	35
19	The computational domain and the two displacement functions $u_1$ and $u_2$ . . . . .	38
20	The normal strains $\varepsilon_{xx}$ , $\varepsilon_{yy}$ and the shearing strain $\varepsilon_{xy}$ . . . . .	38
21	The normal stresses $\sigma_x$ and $\sigma_y$ and the shearing stress $\tau_{xy}$ . . . . .	39
22	The von Mises and Tresca stress . . . . .	39
23	The computational domain and the two displacement functions $u_1$ and $u_2$ . . . . .	40
24	The normal strains $\varepsilon_{xx}$ , $\varepsilon_{yy}$ and the shearing strain $\varepsilon_{xy}$ . . . . .	41
25	The normal stresses $\sigma_x$ and $\sigma_y$ and the shearing stress $\tau_{xy}$ . . . . .	42
26	The von Mises and Tresca stress . . . . .	42
27	The original and deformed domain and the von Mises stress for an axially symmetric setup . . . . .	44
28	Stresses for an axially symmetric setup . . . . .	44
29	A domain with a hole and a finer mesh at the lower edge . . . . .	51
30	The deformed lever and the bending of the center line . . . . .	52
31	A domain with a two different mesh sizes . . . . .	52
32	The same mesh with linear or quadratic elements . . . . .	54
33	A mesh generated by a Delaunay triangulation and the solution of a BVP . . . . .	55
34	A function evaluated on a uniform grid . . . . .	60
35	Convergence results for linear, quadratic and cubic elements . . . . .	91
36	An linear, equilateral triangle, the Gauss integration points and the element stiffness matrix . . . . .	93
37	Uniform meshes consisting of equilateral triangles . . . . .	94
38	An equilateral, quadratic triangle, the Gauss integration points and the element stiffness matrix . . . . .	95
39	A right triangle, the Gauss integration points and the element stiffness matrix . . . . .	96
40	Uniform meshes consisting of rectangular triangles . . . . .	96
41	A right angle triangle, the Gauss integration points and the element stiffness matrix . . . . .	97
42	The number of nonzero entries in each row . . . . .	100
43	The mesh and the solution for a BVP . . . . .	101
44	Difference to the exact solution and values of $\frac{\partial u}{\partial y}$ , using a first order mesh . . . . .	102
45	Difference to the exact solution and values of $\frac{\partial u}{\partial y}$ , using a second order mesh . . . . .	102
46	Difference of the approximate values of $\frac{\partial u}{\partial y}$ to the exact values . . . . .	103

47	Difference of the approximate values of $u$ and $\frac{\partial u}{\partial y}$ to the exact values for cubic elements . . . . .	103
48	Meshes for linear, quadratic and cubic elements . . . . .	106
49	The solution and the first derivative, evaluated at the nodes and by interpolation . . . . .	108
50	The first derivative, complete graph and zoomed in at $x = 0.5$ . . . . .	109
51	Differences of the FEM solution $u_{FEM}(x)$ to the exact solution $u_{exact}(x) = \exp(x)$ . . . . .	109
52	The solution and difference to the exact solution for a 1D BVP . . . . .	110
53	The derivative of the numerical solution and difference to the exact solution . . . . .	111
54	The second derivative of the numerical solution and difference to the exact solution . . . . .	112
55	The solution generated by the implicit time stepper . . . . .	113
56	The solution generated by the Crank–Nicolson time stepper . . . . .	113
57	The solution at time $t = 2$ generated by the explicit and Runge–Kutta time steppers . . . . .	113
58	Original and deformed domain and the Gauss integration points for linear elements . . . . .	114
59	The strains $\varepsilon_{xx}$ and $\varepsilon_{xy}$ with two layers in each direction . . . . .	115
60	The original shape of the a beam and its (exaggerated) deformed shape, using two layers of elements . . . . .	119
61	Meshes for linear and quadratic elements with one layer, with the integration points . . . . .	120
62	The elastic energy density of the bending beam with one or five layers . . . . .	121
63	Classical and weak solutions, minimizers and FEM . . . . .	127
64	A few triangular elements . . . . .	129
65	Transformation of the standard triangle $\Omega$ to a general triangle $E$ . . . . .	130
66	Gauss integration of order 2 on the standard triangle, using 3 integration points . . . . .	131
67	Gauss integration of order 5 on the standard triangle, using 7 integration points . . . . .	132
68	Local and global numbering of nodes . . . . .	133
69	Basis functions for second order triangular elements . . . . .	140
70	Transformation of the cubic standard triangle $\Omega$ to a general triangle $E$ . . . . .	150
71	The 10 basis functions for third order triangular elements . . . . .	152
72	The interpolation from four nodes to three Gauss points on an interval $[-\frac{h}{2}, +\frac{h}{2}]$ . . . . .	159
73	Difference to the exact solution of a BVP . . . . .	197
74	Traveling waves on a rectangle . . . . .	199
75	The radial Bessel function as solution of a BVP . . . . .	200
76	Difference to the exact solution of a BVP . . . . .	201
77	Difference to the exact solution of a BVP, using quadratic elements and interpolation to a finer grid . . . . .	201
78	Difference of $\frac{\partial u}{\partial x}$ to the exact solution, using second order elements . . . . .	202
79	Difference of $\frac{\partial u}{\partial x}$ to the exact solution, using first order elements . . . . .	202
80	A solution with singular partial derivatives at the origin . . . . .	204
81	A solution with singular partial derivatives, graphs of $\frac{\partial u}{\partial x}$ and $\ \nabla u\ $ . . . . .	204
82	Fluid flow between two plates, the setup . . . . .	205
83	Velocity field of an ideal fluid . . . . .	206
84	Velocity field of a ideal fluid in a circular pipe . . . . .	208
85	A minimal surface . . . . .	210
86	The capacitance and the section used for the modeling . . . . .	212
87	A mesh on the domain . . . . .	213
88	The contour lines of the resulting voltage . . . . .	213
89	Voltage plot and electric field between the plates of the capacitance . . . . .	214
90	Torsion of a shaft . . . . .	215
91	The von Mises stress caused by torsion of a bar with square or rectangular cross section . . . . .	218
92	The mesh for a dynamic heat problem . . . . .	219
93	The evolution of the temperature surface at different times . . . . .	220
94	The temperature surface at different times along $y = 0$ . . . . .	220
95	The temperature as function of time at the endpoint $(2.5, 0)$ . . . . .	221

96	The mesh for a dynamic heat problem . . . . .	222
97	The evolution of the temperature surface at different times . . . . .	222
98	The temperature surface at different times along $y = 0$ . . . . .	222
99	The temperature as function of time at the endpoint $(2.5, 0)$ . . . . .	223
100	The domain and initial temperature . . . . .	224
101	The temperature at different times . . . . .	224
102	The temperature at different times along $y = 0$ . . . . .	226
103	The temperature decay at the center $(0, 0)$ . . . . .	226
104	The domain for a heat wave propagation . . . . .	227
105	The propagation of a heat wave . . . . .	228
106	The steady state solution of a heat problem in a ball . . . . .	230
107	The errors of the steady state solution of a heat problem in a ball . . . . .	231
108	The steady statesolution of a heat problem in a ball, solved as 1D problem . . . . .	231
109	The temperatur $u(r, t)$ inside the cylinder . . . . .	232
110	The contours of the temperatur $u(r, t)$ inside the cylinder . . . . .	233
111	The temperatur $u(r, t)$ at the inner ( $r = 0$ ) and outer ( $r = R$ ) edge . . . . .	233
112	The domain for the wave propagation . . . . .	234
113	Wave propagation, leading to a Kirchhoff diffraction pattern . . . . .	235
114	A spherical sound wave at time $t = 1.75$ , and the decaying amplitude with the best fitting $\frac{c}{t}$ . . . . .	237
115	A circular sound wave at time $t = 4$ and the decaying amplitude with the best fitting $\frac{c}{\sqrt{t}}$ . . . . .	239
116	A spherical wave as function of time $t$ and radius $r = \sqrt{x^2 + y^2 + z^2}$ . . . . .	240
117	A cylindrical wave as function of time $t$ and radius $r = \sqrt{x^2 + y^2}$ . . . . .	241
118	The amplitude $u(x, t)$ for a reflected pulse . . . . .	242
119	Waterfall plot and contour plot for the amplitude $u(x, t)$ for a reflected pulse . . . . .	242
120	The first six eigenfunctions of the harmonic Schrödinger operator . . . . .	244
121	The conductivity . . . . .	245
122	Contours of the voltages . . . . .	246
123	The vector field for the current density $\vec{J}$ and a few streamlines . . . . .	247
124	Voltage along the boundary . . . . .	248
125	Differences of the voltage and the reference voltage . . . . .	248
126	Flux density at inlet and outlet . . . . .	251
127	Displacement and strain for a beam with thinner midsection . . . . .	251
128	Keller's boundary value problem . . . . .	253
129	Two solutions of $-u''(x) = \frac{1}{2} \exp(u(x))$ with $u(-1) = u(+1) = 0$ . . . . .	255
130	A BVP with multiple nonlinear contributions . . . . .	257
131	The solution of Fisher's equation . . . . .	259
132	The connections between Salt Lake City and Zürich . . . . .	262
133	The solutions for a bending beam . . . . .	265
134	A BVP describing mass transfer in a porous catalyst . . . . .	268
135	A BVP describing mass transfer in a porous catalyst, second setup . . . . .	268
136	Motion of a string, excited by an initial pulse close to $x = 0$ . . . . .	271
137	One quarter of a section through the pipe . . . . .	272
138	Von Mises stress and principal stresses . . . . .	274
139	Domain for the hook with attached weight . . . . .	279
140	A horizontal slice with $\sigma_y$ shown and a vertical slice with $\sigma_x$ shown . . . . .	281
141	The von Mises stress on the crook, as surface and level curves . . . . .	282
142	Bending of vertical beam and von Mises Stress at corner . . . . .	282
143	The deformed wrench and the stress $\sigma_y$ along upper edge with the applied load . . . . .	284
144	Surface and contour plot of the von Mises stress in [MPa] . . . . .	285

145	The upper half of the original and deformed domain for the rotating rubber box . . . . .	286
146	The displacements $u_r$ and $u_z$ for the rotating rubber box . . . . .	286
147	The von Mises stress for the rotating rubber box . . . . .	287
148	The original and deformed domain . . . . .	288
149	The radial displacement $u_r$ . . . . .	289
150	The height displacement $u_z$ . . . . .	289
151	The normal stress $\sigma_z$ . . . . .	290
152	The normal pressures $\sigma_z$ along upper and lower edge and at half height . . . . .	291

## List of Tables

1	Commands to solve BVPs, IBVPs and elasticity problems . . . . .	11
2	Normal and shear strains in space . . . . .	14
3	Description of normal and tangential stress in space . . . . .	14
4	Elements of a mesh structure . . . . .	46
5	Codes for the boundary conditions . . . . .	47
6	Commands to solve and examine 1D boundary value problems . . . . .	67
7	Commands to solve and examine plane elasticity problems . . . . .	78
8	Commands to solve and examine axially symmetric elasticity problems . . . . .	83
9	Results for elements of order 1, 2 and 3 . . . . .	107
10	Elastic energy contributions for shearing . . . . .	116
11	Different values for the deformation of a bending beam, depending on the size of the grid . . . . .	122
12	Properties of triangular elements . . . . .	129
13	Coordinates of the nodes in the standard quadratic triangle . . . . .	139
14	Coordinates of the nodes in the standard cubic triangle . . . . .	150
15	Properties of the ODE solvers used in IBVP1D ( ) . . . . .	176
16	Comparison of different elements for the washer fastener example . . . . .	293

## Index

- AxiStress(), 44, 83, 277, 286, 289  
AxiStressEquationCubicM(), 89  
AxiStressEquationM(), 89, 125  
AxiStressEquationQuadM(), 89  
axisymmetric, 19
- basis function, 140, 151  
boundary condition, 173  
BVP, 21  
    boundary value problem, 9  
    eigenvalue, 10  
    elliptic, 9  
    symmetric, 10  
BVP1D(), 32, 68, 69, 108–111, 231, 252–254, 262, 265, 266  
BVP1Deig(), 26, 27, 74, 75, 245  
BVP1DNL(), 36, 37, 75, 255–257, 259, 264, 267–270  
BVP2D(), 23–25, 64, 228  
BVP2Deig(), 25, 64, 227  
BVP2Dsym(), 21–23, 63, 198, 230
- conforming, 60, 102, 108–111, 130  
constant strain triangle, 116, 183  
convection, 24  
convergence, 91  
Crank–Nicolson, 72, 113, 164  
CreateMeshRect(), 21, 22, 24, 28, 41, 47, 55, 59, 98, 117, 121, 199, 201, 238, 239, 272, 280  
CreateMeshTriangle(), 23, 25, 30, 38, 49, 51, 57, 61, 92, 93, 96, 102, 198, 201, 204, 206, 209, 212, 219, 221, 222, 225, 228, 235, 279, 284, 289  
CST, 116, 183
- Delaunay, 56  
Delaunay2Mesh(), 56  
DIRK, 178
- eigenvalue, 25, 88, 165, 167  
eigs, 88  
EIT, 246  
element stiffness matrix, 134  
EvaluatePrincipalStress(), 40, 43, 82, 274, 275  
EvaluatePrincipalStressAxi(), 85, 277  
EvaluateStrain(), 39, 42, 80, 118  
EvaluateStrainAxi(), 84, 277  
EvaluateStress(), 39, 42, 80, 274, 281, 285  
EvaluateStressAxi(), 44, 85, 277, 287, 288  
EvaluateTresca(), 40, 43, 82  
EvaluateTrescaAxi(), 86, 277
- EvaluateVonMises(), 40, 43, 81, 274, 282, 285  
EvaluateVonMisesAxi(), 44, 85, 277, 287, 288
- FEM1DEvaluateDu(), 68, 70, 108–111, 176, 254, 257, 262  
FEM1DGaussPoints(), 68, 70, 254, 262, 265, 266  
FEMEquation(), 86, 93  
FEMEquationCubic(), 87  
FEMEquationQuad(), 87  
FEMEvaluateGP(), 59, 119, 293  
FEMEvaluateGradient(), 58, 205  
FEMgriddata(), 60, 116, 117, 201, 228, 230, 248, 251, 274, 275, 279–282, 285  
FEMIntegrate(), 59, 118, 119, 198, 293  
FEMInterpolBoundaryWeight(), 88  
FEMInterpolWeight(), 87  
FEMtricontour(), 23–25, 28, 56, 57, 206, 209, 215, 282, 285, 288  
FEMtrimesh(), 21–23, 25, 28, 30, 56, 57, 102, 198, 199, 204–206, 209, 215, 228, 235, 285, 288  
FEMtrisurf(), 57, 212, 219, 282  
Fisher’s equation, 259
- Gauss integration, 59, 131–133, 138, 141, 148, 152, 161, 170, 171, 185, 187  
GenerateFEM1D(), 68, 69  
GenerateWeight1D(), 68
- heat equation, 10, 24, 28, 163–165  
Hooke’s law, 14, 16
- I2BVP1D(), 35, 68, 73, 240, 241, 243, 271  
I2BVP2D(), 30, 67, 199, 235, 238, 239  
IBVP, 21  
    hyperbolic, 11, 67, 165  
    parabolic, 10, 66, 164  
IBVP1D(), 33, 68, 71, 72, 113, 233  
IBVP2D(), 28, 66, 228  
IBVP2Dsym(), 66, 228  
interpolation, 170
- Jaccobi determinant, 131
- MeshCubic2Linear(), 54, 55  
MeshDeform(), 57, 198, 272  
MeshQuad2Linear(), 25, 54, 55, 200  
MeshUpgrade(), 25, 49, 50, 54, 198, 200, 231, 235, 273, 279, 284  
minimal surface, 211

- PlaneStrain(), 41, 78, 273  
PlaneStress(), 38, 77, 279, 284  
potential flow, 206, 208  
Prandtl stress function, 218  
principal stress, 40, 43, 82  
PStressEquationCubicM(), 89  
PStressEquationM(), 89, 125  
PStressEquationQuadM(), 89  
pwquadinterp(), 32, 68, 70, 108–111, 176, 231, 252,  
254, 262, 265, 266, 269, 270
- ReadMeshTriangle(), 49, 50, 54  
Runge–Kutta, 72, 113, 178
- scheme  
  implicit, 178
- Schrödinger equation, 244  
shear locking, 117, 121  
singular problem, 204  
smallEig(), 88
- solution  
  classical, 127, 128  
  weak, 128, 129
- stability  
  A–stability, 177  
  conditional, 72, 113, 177  
  L–stability, 72, 113, 177  
  unconditional, 72, 113, 177
- stiffness matrix  
  element, 129  
  global, 129
- strain, 14
- streamline(), 248
- stress  
  principal, 40, 43, 82, 277  
  Tresca, 40, 43, 82, 277  
  von Mises, 40, 43, 81, 218, 274, 277, 282, 285,  
  287
- successive substitution, 253, 262  
superconvergence, 91, 111, 200, 231
- tensor  
  infinitesimal strain, 14
- time step  
  Crank–Nicolson, 164, 178  
  explicit, 178  
  implicit, 178  
  Runge–Kutta, 178
- torsional rigidity, 217  
Tresca stress, 40, 43, 82, 277
- triangle, 7, 49, 50, 90, 213  
tricontour(), 57, 90
- wave equation, 28, 163, 165, 167