See video on how this course is organised in Youtube

Self-study guide

Week 1

Keywords: Introduction, Finite Element Workflow, Weak Form, Strong Form, Finite Element Solution in 1D.

Homework: Problems P3, P5, P7, In addition, solve any additional two problems from P1-P8 to gain extra points.

Outline of Week 1 in Youtube

Pages: 4-20

Synopsis: During the first week, we have two learning goals. First goal is to develop understanding of finite element simulation workflow. Particularly, which steps are taken, e.g., to simulate the temperature inside a transistor. Second goal is understand how finite element method solves PDEs. This is illustrated by solving simple one dimensional Poisson's equation.

Week 2 - Computer exercise week with contact teaching.

Keywords: Error, evaluation of error, uniform refinement, error plot, second order finite element space.

Homework: Problems P10, P13, P14, In addition, solve any additional two problems from P10-P16 to gain extra points.

Pages: 21-31

Synopsis: During the second week we learn to evaluate finite element error and modify the finite element solver to use second order basis functions. Understanding error behaviour is important to guarantee that FE-solutions can be used to make correct design decisions. Second order basis functions produce smaller error and are often used in FE-simulation.

Week 3 - Self study week with youtube video material

Keywords: Weak derivative, Sobolev space, Poincare-Friedrichs inequality, Existence and uniqueness proof, Lax-Milgram Theorem, Variational problem, Ellipticity, Energy minimisation.

Homework: Problems P17, P19, P22, In addition, solve any additional two prob-

Outline of Week 3 in Youtube

lems from P17-P25 to gain extra points.

Pages: 31-42

Synopsis: During the third week, we first introduce sufficient mathematical framework required to prove the existence of a unique solution to weak formulation of the one dimensional Poisson's equation. This is, we define Sobolev spaces, their norms, and inner products. To properly define Sobolev spaces, we discuss weak derivatives. We apply the Lax-Milgram lemma that is an existence and uniqueness proof for abstract variational problems to study our model problem. We illustrate the importance of assumptions made in it's formulation by giving a simplified existence proof. Finally, we discuss formulation of the variational problem as an energy minimisation problem.

Week 4 - Self study week with youtube video material

Keywords: Error analysis, Cea's Lemma, Nodal interpolation, Interpolation error, Scaling argument, L^2-error estimate.

Homework: Problems P28, P30, P31, In addition, solve any additional two problems from P26-P31 to gain extra points.

Outline of Week 4 in Youtube

Pages: 42-52

Synopsis: During the fourth week, we derive error estimates for FE-solution in H^1, L^2, and energy norms. Error estimate in H^1 and energy norm follows by establishing a relation between the FE-solution and the best approximation of u form the FE-space. This relation allows us to compare the FE-solution with nodal interpolant of u. FE-error estimate is obtained by analysing the interpolation error using scaling argument. The L^2-error is analysed using duality argument and it exhibits faster convergence rate.

Week 5 - Youtube video material and Contact teaching

Keywords: FEM in 2D, Weak derivative and Sobolev spaces in dimension d, Trace, Triangular mesh, piecewise linear FE-space, Affine mapping, Integration over triangle.

Homework: Problems P32, P33, P34. In addition, solve any additional two problems from P32-P38 to gain extra points.

Outline of Week 5 in Youtube

Pages: 52-65

Synopsis: During the fifth week, we discuss application of finite element method to

solve a two dimensional model problem. First, Sobolev spaces in dimension d are discussed. Then we derive the weak form of the model problem that FEM solves approximately by posing it in finite dimensional function space. We use the space of piece-wise linear functions defined on a conforming triangular partition. The main task is to assembly a linear system related to this finite dimesional problem. This requires evaluation of integrals over triangles. These integrals are evaluated by making a change of variables to reference element and using numerical integration method.

In computer exercises, we give assistance to solving problems P33 and P34. To get most out of them, please, read Section 9.4 beforehand.

Week 6 - Youtube video material and Contact teaching

Keywords: FEM in 2D, Assembly, Reference basis functions, Implementation

Homework: Problems P40, P43, P44. In addition, solve any additional two problems from P39-P44 to gain extra points.

Outline of Week 6 in Youtube

Pages: 65-70

Synopsis: During week six, we implement finite element solver in two dimensions. The implementation is based on the one dimensional solver on p.17. The main task is to assembly the matrix A and vector $\boldsymbol{b}$. This is discussed in section 9.6 along with method for evaluating hat basis functions using affine mapping and a reference basis. The gradients of hat basis functions, required in assembly of A, are obtained by multiplying the gradient of reference basis functions with appropriate term.

In computer exercises, we give assistance to solving problems P43 and P44. To get most out of them, please, read Sections 9.6 and 9.8 beforehand.

Week 7 - Youtube video material

Keywords: FEM in 2D, Error analysis

Homework: Problems P45, P48, P49. In addition, solve any additional two problems from P45-P49 to gain extra points.

Outline of Week 7 in Youtube

Pages: 70-81

Synopsis: During week seven, we discuss FE-error analysis for our two-dimensional model problem. The analysis is almost identical with one dimensional case: We use Cea's Lemma to bound FE-error by interpolation error that is analysed using

reference interpolation error estimate and the scaling argument. The main difference to one dimensional case are requirements that we make on the shape of the triangles in the applied meshes.

Week 8 (The End)- Youtube video material

Keywords: Error analysis, Regularity

Homework: Problems P51, P52. In addition, solve problem from P50 to gain extra points.

Outline of Week 8 in Youtube

Pages: 81-84

Synopsis: During the last lecture week, we briefly discuss the $H^2(\Omega)$-regularity assumption of the solution u. The effect of non-convex corner point of Ω to convergence rate is investigated by numerical examples.

1 Introduction

Many physical quantities, such as temperature and magnetic field, are functions of spacial coordinates. Often, these unknown functions satisfy mathematical models that are expressed as Partial Differential Equations (PDEs), this is, equations involving the unknown function, it's partial derivatives, and input data. Predictions on such physical quantities are made by solving the associated function from the PDE. Explicit analytical expression for the solution can be found in very few special cases, e.g., if the PDE is posed in simple domain and has constant coefficients. Instead, the solution is *approximated* using numerical methods, such as the Finite Element Method (FEM) that is the topic of these lecture notes.

Example on modelling in Youtube

Finite Element Method is a tool for computing approximate solutions to various PDEs. It can be used on complex domains and with difficult material behaviour. It has a solid theoretical background and the properties of FE-solutions can be mathematically analysed. For example, the effect of method parameters to the accuracy of computed approximate solutions is well understood. There exists dedicated commercial and open source software for conducting finite element analysis, i.e. solving physical quantities from PDEs. All these factors make FEM a widely used tool in scientific computing.

Finite element workflow begins with the definition of the geometric shape for the object under study. Simple geometric models are created using finite element software, whereas more complex ones are imported from dedicated CAD-programs. Next, the governing PDE or PDEs and their constitutive models are defined for each part of the design. Then a *mesh* is generated for the object, this

FE-workflow in Youtube

is, it is divided into triangular, tetrahedral, quadrilateral, pyramidal, hexahedral, or prismatic subdomains called as elements. Finally, the finite element software computes an approximate solution to the PDE. The obtained FE-solution can be visualised to obtain intuition on the behaviour of the physical quantity. In addition, it can be used to compute values for design parameters, such as maximal temperature or total heat flux.

In this note, we explain the mathematical principles behind the solution step in the FE-workflow. These principles are best described by applying FEM to a simple model problem. For this purpose, we use the Poisson equation posed on the interval $(0, 1)$ as discussed in Section 2. The solution steps for complex real-world engineering problems and our model problem are nearly identical.

FEM finds an approximate solution to a PDE from a *Finite Element space* (FE-space), a finite dimensional set of functions related to the mesh. In the simplest case, the FE-space contains all continuous functions that are linear on each element (i.e., continuous piecewise linear functions). Inside finite element solver, the FE-space is defined using a basis that allows the solver to associate functions with vectors in $\mathbb{R}^n$. Basis and sets of functions are critical for understanding FEM and discussed in Section A.

FE-solver computes an approximate solution to a linear PDE by using a basis of the finite element space to transform it to a linear system. Coefficients of this linear system are computed in an *assembly step*, discussed in Section 3, by evaluating integrals related to basis functions and problem data. The resulting linear system is solved and it's solution is returned to the user. The solution is then visualised and used to compute values for design parameters in *post-processing step*.

Example 1.1 in Youtube

Example 1.1. *As an example of FE-solution process, we conduct thermal analysis of the power transistor depicted in Fig. 1. The aim in such analysis could be, e.g., to study whether the maximal temperature of the transistor stays below the limit recommended by the manufacturer or exceeds it. All FE-simulation is done in Comsol Multiphysics.*

We begin by describing the mathematical model used in thermal analysis. Let $\Omega \subset \mathbb{R}^3$ be an open set consisting of the power transistor. The stationary temperature distribution inside the transistor is a function $u : \Omega \mapsto \mathbb{R}$. Naturally, the internal temperature depends on the outside world. We use a simplified model and take the effect of the outside world into account by imposing **boundary conditions***, i.e., conditions on the behaviour of u at the outer surface of the transistor, or boundary $\partial\Omega$.*

To specify boundary conditions, the boundary $\partial\Omega$ is split into two open sets Γ_{leads} and Γ_{body}, see Fig. 2 These sets describe the part of the transistor soldered to the circuit board and the part cooled by natural convection, respectively. Denote the ambient temperature by $T_{amb} = 25$. We assume that the leads of the transistor soldered to a circuit board stay at temperature T_{amb}, and impose the condition

$$u = T_{amb} \quad on \quad \Gamma_{leads}. \tag{1}$$

Parameter	Value
Thermal conductivity for copper	385
Thermal conductivity for silica	20
Parameter h on Γ_{body}	5
Source term for semiconductor part corresponding to 2Watt power dissipation	125e6

Table 1: Parameter values used in Example 1.1. The source term is obtained by dividing the total dissipated power of two watts by the volume of the semiconductor part.

On rest of the boundary, we impose the boundary condition,

$$\frac{\partial u}{\partial n} = h(u - T_{amb}) \quad on \quad \Gamma_{body}. \tag{2}$$

This boundary condition states that the heat flux is relative to the difference between ambient and surface temperatures. Eq. (2) *is a simplified model for natural convective cooling. The stationary temperature distribution inside the transistor satisfies the PDE*

$$-\nabla \cdot (k\nabla u) = f \quad in \quad \Omega, \tag{3}$$

where function $f : \Omega \mapsto \mathbb{R}$ models heat sources and parameter $k : \Omega \mapsto \mathbb{R}$ is material dependent thermal conductivity. The heat source f corresponds to two watt power dissipation in the semiconductor region and is zero otherwise. The parameter values used in our example are given in Table 1.

To summarise, the temperature distribution $u : \Omega \to \mathbb{R}$ is solved from the system

$$\begin{aligned} -\nabla \cdot (k\nabla u) &= f && in \quad \Omega \\ \frac{\partial u}{\partial n} &= h(u - T_{amb}) && on \quad \Gamma_{body}. \\ u &= T_{amb} && on \quad \Gamma_{leads} \end{aligned} \tag{4}$$

We proceed to compute an approximate solution to (4) *using FEM. First, a geometric model of the power transistor is created. The model is constructed from geometric components conforming with the four parts of the transistor and the distinct boundary components. This way, the material parameters and boundary conditions are the same for each geometric component, and they can be easily specified to the FE-solver.*

Next, we specify the PDE to be solved and the related material parameters. Then, a mesh is generated for the geometric model. We use a mesh consisting of tetrahedral elements or sub-domains, see Fig. 2. Finally, we compute the FE-solution. The obtained approximate solution is visualised in Fig. 3.

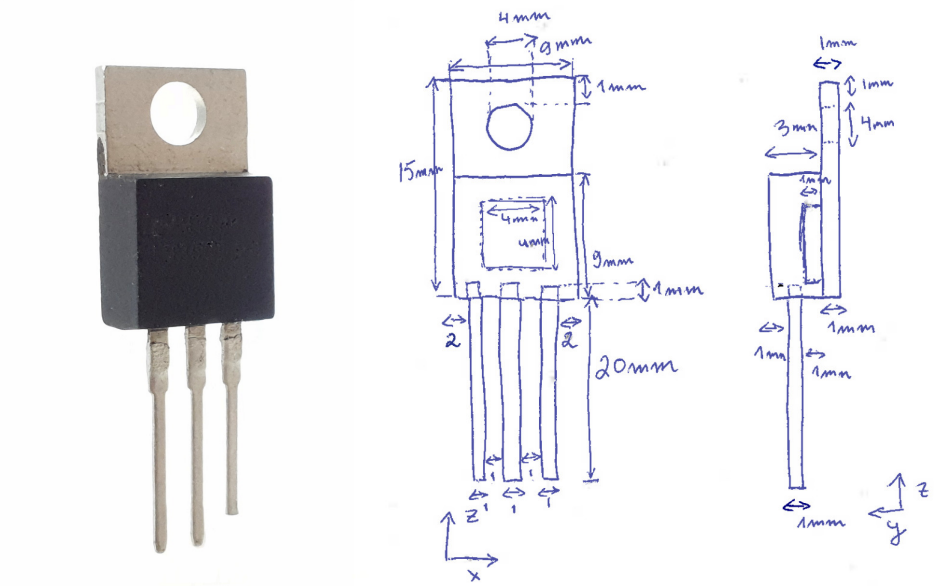

Figure 1: The power transistor studied in Example 1.1. The transistor consists of four parts: copper back plate, silicone chip, ceramic body, and copper legs. Dimensions of these parts in mm are show in the sketch on the right hand side.

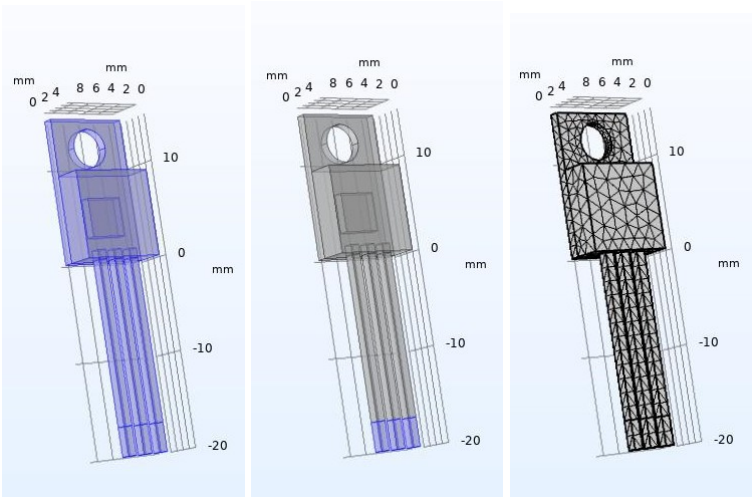

Figure 2: From left: geometric model, boundary component Γ_{leads}, and the surface mesh of the power transistor studied in Example 1.1.

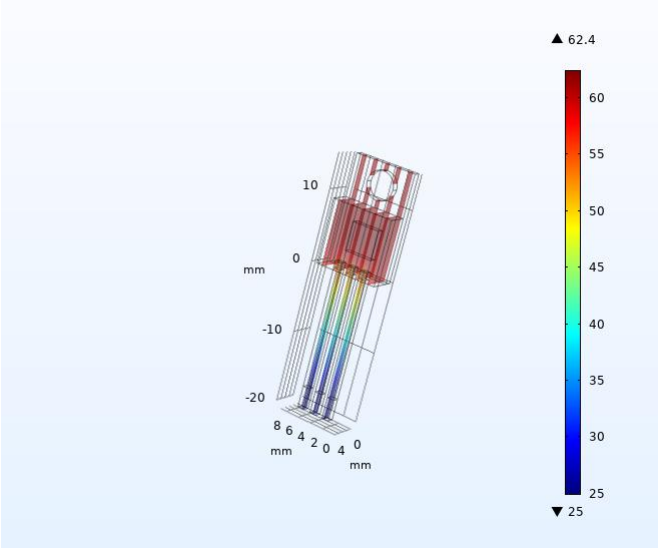

Figure 3: A slice plot of the temperature distribution insider the power transistor studied in Example 1.1. The total power dissipated in the semiconductor is two watts. Maximal temperature is 62.4 Celsius.

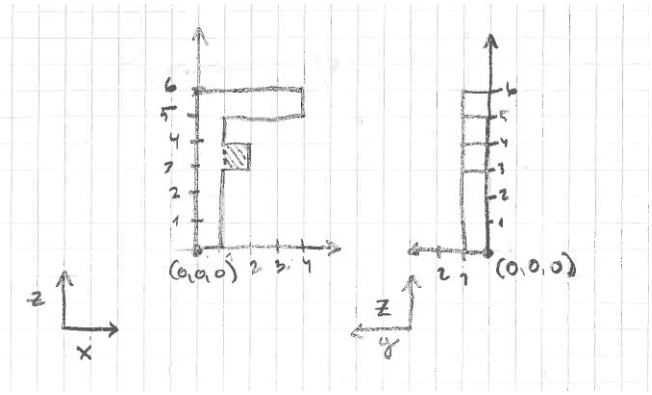

Figure 4: Domain $\Omega \subset \mathbb{R}^3$ for Problem P1. The source term is defined to have value one in the dashed region and to be zero otherwise.

1.1 Problems

P1. (1p) Conduct thermal analysis for letter F using Comsol.

(a) Draw a three dimensional model of the letter F according to the sketch given in Fig. 4. Create a separate geometric part of the subdomain marked with lines.

(b) Let $\Omega \subset \mathbb{R}^3$ be the domain drawn in (a). Solve the PDE

$$-\Delta u = f \text{ in } \Omega \quad \text{and} \quad u = 0 \text{ on } \partial\Omega.$$

The source term f has value one in the dashed subdomain is zero otherwise.

(c) What is the maximal temperature ?

2 Poisson equation in 1D

In this section, we discuss one dimensional Poisson's equation that is solved using FEM in Section 3. The Poisson's equation is: Find $u \in C^2(0,1) \cap C([0,1])$ such that

Poisson's Eq. in Youtube

$$\begin{aligned} -\frac{d^2u(x)}{dx^2} &= f \quad \text{in } (0,1) \\ u(0) = u(1) &= 0. \end{aligned} \tag{5}$$

The source function $f : (0,1) \mapsto \mathbb{R}$ is the given input and $u = 0$ imposes a *zero Dirichlet boundary condition* to function u. Variants of the Poisson's equation in dimensions $d = 1, 2, 3$ arise in several fields of science. It is used, for example, to model static electric fields or stationary temperature distributions as in Example 1.1.

The problem (5) is called as the *strong* form of the Poisson's equation. The problem is well defined only if there exists a solution that has two derivatives in $(0,1)$ and is continuous on the closed interval $[0,1]$. For such functions, we write $u \in C^2(0,1) \cap C([0,1])$, for more details see Section A.2. Unfortunately, the requirement on existence of a solution u having two derivatives is very strong, and problem (5) is not well defined for all relevant source functions f.

Example 2.1. *Consider solving problem* (5) *for*

$$f(x) = \begin{cases} -1 & x \in \left(0, \frac{1}{2}\right] \\ 1 & x \in \left(\frac{1}{2}, 1\right) \end{cases}.$$

Thus, on $(0, 1/2)$ *it holds that*

$$\frac{d^2 u_1}{dx^2} = 1$$

Integrating twice gives

$$u_1(x) = \frac{1}{2}x^2 + C_1 x + C_0$$

where C_1 *and* C_0 *are unknown constants. Imposing the boundary condition* $u(0) = 0$ *yields* $C_0 = 0$*. Using similar process on* $(\frac{1}{2}, 1)$ *gives*

$$u_2(x) = -\frac{1}{2}(x-1)^2 + D_1(x-1),$$

where D_1 *is an unknown constant. The solution u to* (5) *has to be continuous on* $[0,1]$ *and twice differentiable on* $(0,1)$*. This is the case, if we manage to choose* C_1 *and* D_1 *so that*

$$\begin{aligned} u_1\left(\frac{1}{2}\right) &= u_2\left(\frac{1}{2}\right) \\ \frac{du_1}{dx}\left(\frac{1}{2}\right) &= \frac{du_2}{dx}\left(\frac{1}{2}\right) \\ \frac{d^2u_1}{dx^2}\left(\frac{1}{2}\right) &= \frac{d^2u_2}{dx^2}\left(\frac{1}{2}\right) \end{aligned} \tag{6}$$

are satisfied. The first and the second condition give the solution candidate

$$u(x) = \begin{cases} \frac{1}{2}x^2 - \frac{1}{4}x & x \in (0, \frac{1}{2}] \\ -\frac{1}{2}(x-1)^2 - \frac{1}{4}(x-1) & x \in (\frac{1}{2}, 1) \end{cases}.$$

This solution candidate does not satisfy third condition in (6)*. This is, we could not find a solution with two derivatives to* (5) *with our example loading. The candidate function and it's derivatives are visualised in Figure 5.*

The first step in solving (5) is to relax the requirements on the solution u by deriving an alternative formulation of (5).

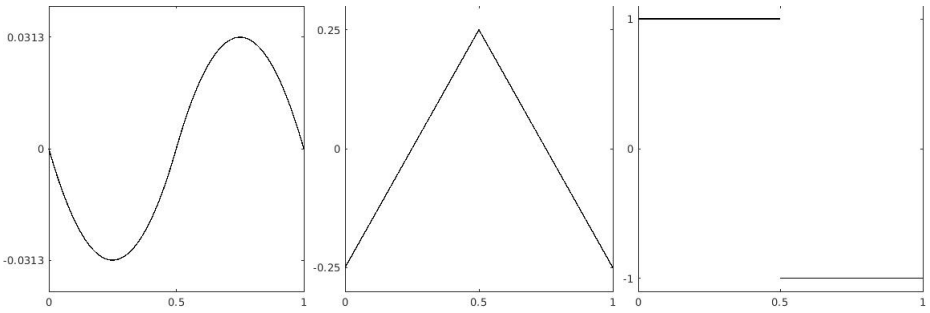

Figure 5: From left: functions $u, \frac{du}{dx}$, and $\frac{d^2u}{dx^2}$ defined in Example 2.1.

Weak form in Youtube

Weak form To obtain a well posed problem for a sufficiently large class of functions f, the strong form of the given PDE is transformed into it's weak form. The weak form is obtained by multiplying the strong problem with a test function $v \in V$, where V is an appropriate function space and integrating the right hand side by parts. For (5) this gives: find $u \in V$ such that

$$\int_0^1 \frac{du}{dx}\frac{dv}{dx}\,dx = \int_0^1 fv\,dx \quad \text{for all } v \in V \tag{7}$$

An appropriate choice for the space V is the Sobolev space $H_0^1(0,1)$, defined as

$$H_0^1(0,1) := \left\{ v \in L^2(0,1) \mid \int_0^1 \left(\frac{dv}{dx}\right)^2 dx < \infty \text{ and } v(0) = v(1) = 0 \right\}.$$

The validity of the weak problem (7) follows from the fact that one can prove existence of a unique solution $u \in V$ for a sufficiently large class of source functions f. In addition, any solution to (5) is also a solution of (7), thus, the weak problem is generalisation of the strong one.

Example 2.2. *Let*

$$f(x) = \begin{cases} -1 & x \in \left(0, \frac{1}{2}\right] \\ 1 & x \in \left(\frac{1}{2}, 1\right) \end{cases},$$

and consider solving the problem (7)*. In Example 2.1, we constructed the solution candidate*

$$u(x) = \begin{cases} \frac{1}{2}x^2 - \frac{1}{4}x & x \in (0, \frac{1}{2}] \\ -\frac{1}{2}(x-1)^2 - \frac{1}{4}(x-1) & x \in (\frac{1}{2}, 1) \end{cases}$$

that does not have two derivatives at $x = \frac{1}{2}$. Next, we investigate if u is the weak solution to (5)*. For any smooth function v satisfying $v(0) = v(1) = 0$ (or $v \in C_0^\infty(0,1)$) it holds that*

$$\int_0^1 \frac{du}{dx}\frac{dv}{dx}\,dx = \int_0^{1/2} \frac{du_1}{dx}\frac{dv}{dx}\,dx + \int_{1/2}^1 \frac{du_2}{dx}\frac{dv}{dx}\,dx$$

Integrating by parts gives

$$\int_0^{1/2} \frac{du_1}{dx}\frac{dv}{dx}\,dx = -\int_0^{1/2} \frac{d^2u_1}{dx^2} v\,dx + \frac{du_1}{dx}\left(\frac{1}{2}\right) v\left(\frac{1}{2}\right) - \frac{du_1}{dx}(0)v(0).$$

and

$$\int_{1/2}^1 \frac{du_2}{dx}\frac{dv}{dx}\,dx = -\int_{1/2}^1 \frac{d^2u_2}{dx^2} v\,dx + \frac{du_2}{dx}(1)\,v\,(1) - \frac{du_2}{dx}\left(\frac{1}{2}\right) v\left(\frac{1}{2}\right).$$

Recall, that function v is continuous and has zero boundary values, this is $v(0) = v(1) = 0$. Using these properties and the definition of u gives

$$\frac{du_1}{dx}\left(\frac{1}{2}\right) v\left(\frac{1}{2}\right) - \frac{du_2}{dx}\left(\frac{1}{2}\right) v\left(\frac{1}{2}\right) = 0$$

and further

$$\int_0^1 \frac{du}{dx}\frac{dv}{dx}\,dx = \int_0^1 fv\,dx.$$

This property extends to all test functions $v \in H_0^1(0,1)$ by **density argument**, *hence, u is a weak solution to* (5)*. Detailed discussion on the density argument is out of our scope and thus omitted.*

The motivation for posing Eq. (7) in the space $H_0^1(0,1)$ is simply by always having finite integrals on both sides of the Eq. (7). Note, that the homogeneous Dirichlet boundary condition is included into the definition of the space $H_0^1(\Omega)$.

Intuitively speaking, weak form relaxes the requirement $u \in C^2(0,1)$ or u having two derivatives in two ways: first, integration by parts reduces the number of derivatives taken from u to one. Second, the integral is blind to behaviour of a solution u in a single point. For example, the function

$$\psi(x) = \begin{cases} 2x & x < \frac{1}{2} \\ 2 - 2x & x \geq \frac{1}{2} \end{cases}$$

is not differentiable at point $x = \frac{1}{2}$, but the integral

$$\int_0^1 \left(\frac{d\psi}{dx}\right)^2 dx$$

can be evaluated as

$$\int_0^1 \left(\frac{d\psi}{dx}\right)^2 dx = \int_0^{\frac{1}{2}} \left(\frac{d\psi}{dx}\right)^2 dx + \int_{\frac{1}{2}}^1 \left(\frac{d\psi}{dx}\right)^2 dx = 4,$$

thus, it has a well defined value. The formal difficulty associated to taking a derivative of a non-differentiable function is remedied by the concept of weak derivative.

2.1 Problems

P2. (1p) Find the strong solution to the two following PDEs

(a) Find $u \in C^2(0,1) \cap C([0,1])$ satisfying

$$-\frac{d^2u}{dx^2} = 1 \quad \text{and} \quad u(0) = u(1) = 0.$$

(b) Find $u \in C^2(0,1) \cap C([0,1])$ satisfying

$$-\frac{d^2u}{dx^2} = 1, \quad u(0) = 1, \quad \text{and} \quad \left(\frac{du}{dx}\right)(1) = 0.$$

P3. (2p) Find the weak solution to the PDE

$$-\frac{d^2u}{dx^2} = f \quad \text{for} \quad f := \begin{cases} 1 & x \in (0, \frac{1}{2}] \\ 0 & x \in (\frac{1}{2}, 1) \end{cases} \quad \text{and} \quad u(0) = u(1) = 0.$$

Proceed as in Examples 2.1 and 2.2:

(a) Find a solution candidate u using integration on intervals $(0, \frac{1}{2}]$ and $(\frac{1}{2}, 1)$.

(b) Show that you candidate satisfies

$$\int_0^1 \frac{du}{dx}\frac{dv}{dx}\,dx = \int_0^1 fv\,dx$$

for any $v \in C_0^\infty$ and thus is the weak solution.

(c) Plot functions u and $\frac{du}{dx}$.

3 Finite element method in 1D

In this Section, we discuss the steps taken in the solution of one dimensional Poisson's equation in (5) using FEM. Identical steps are used to solve real engineering problems.

Limit to subspace in Youtube

Limit to a subspace In finite element method, the solution to the weak problem (7) is approximated by limiting into a finite dimensional subspace $V_h \subset V$. This is, the problem: find $u_h \in V_h$ such that

$$\int_0^1 \frac{du_h}{dx}\frac{dv}{dx}\,dx = \int_0^1 fv\,dx \quad \text{for all } v \in V_h \tag{8}$$

is solved instead of (7). Function u_h satisfying (8) is called as the Ritz-Galerkin approximation of u.

To solve (8) with computer, we need to specify a basis for the space V_h. At this point, the basis is arbitrary but later we use a specific finite element basis. Let $\{\varphi_1, \varphi_2, \ldots, \varphi_n\}$ be some basis of V_h, this is, every $v_h \in V_h$ can be written as

$$v_h = \sum_{j=1}^{n} \alpha_j \varphi_j \tag{9}$$

for some unique coordinate vector $\boldsymbol{\alpha} \in \mathbb{R}^n$. Using expansion $u_h = \sum_{j=1}^{n} \beta_j \varphi_j$ for the unknown solution u_h, equation (8) becomes: find $\boldsymbol{\beta} \in \mathbb{R}^n$ such that

$$\sum_{j=1}^{n} \beta_j \int_0^1 \frac{d\varphi_j}{dx} \frac{dv}{dx}\, dx = \int_0^1 f v\, dx \quad \forall v \in V_h. \tag{10}$$

Because the space V_h is finite dimensional, the above equation is equivalent to: find $\boldsymbol{\beta} \in \mathbb{R}^n$ such that

$$\sum_{j=1}^{n} \beta_j \int_0^1 \frac{d\varphi_j}{dx} \frac{d\varphi_i}{dx}\, dx = \int_0^1 f \varphi_i\, dx. \tag{11}$$

for each $i = 1, \ldots, n$. Observe that (11) is equivalent to the linear system: find $\boldsymbol{\beta} \in \mathbb{R}^n$ such that

$$A\boldsymbol{\beta} = \boldsymbol{b},$$

where $A \in \mathbb{R}^{n \times n}$ and $\boldsymbol{b} \in \mathbb{R}^n$ have entries

$$A_{ij} = \int_0^1 \frac{d\varphi_j}{dx} \frac{d\varphi_i}{dx}\, dx \quad \text{and} \quad b_i = \int_0^1 f \varphi_i\, dx.$$

FE-space in Youtube

Finite element space The finite element method is a systematic way to construct suitable space V_h and to evaluate the entries of the matrix A and the vector $\boldsymbol{b}$. We begin by describing the construction of the finite element space. In one spatial dimension, the interval $(0, 1)$ is first divided into sub-intervals as follows: Let $\{x_i\}_{i=1}^{N} \subset \mathbb{R}$ be a partition of the intervall $[0, 1]$, i.e.,

$$0 = x_1 < x_2 < \ldots < x_N = 1.$$

The associated sub-intervals are defined as $I_i = (x_i, x_{i+1}) \subset \mathbb{R}$ for $i = 1, \ldots, (N-1)$. These sub-intervals are called as *elements*.

The simplest example of a finite element space V_h is the space of continuous piece-wise linear functions over the partition $\{x_i\}_{i=1}^{N}$ with zero boundary conditions. Example of such function is given in Fig. 6. We formally define

$$V_h := \{\, u \in C[0,1] \mid u(0) = u(1) = 0,\ u|_{I_i} \in P^1(I_i) \text{ for } i = 1, \ldots, (N-1) \,\}.$$

In the above definition, the constraint $u \in C[0, 1]$ forces function u to be continuous. The notation $u|_{I_i}$, stands for the *restriction* of function u to the interval I_i, i.e.,

$$u|_{I_i} : I_i \mapsto \mathbb{R} \quad \text{and} \quad u|_{I_i}(x) = u(x) \quad \text{for all } x \in I_i.$$

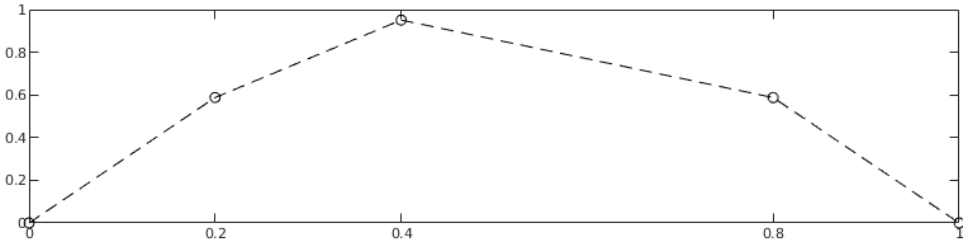

Figure 6: Example of piecewise linear function over partition $0, 0.2, 0.4, 0.8, 1$

The condition $u|_{I_i} \in P^1(I_i)$ states that u is a linear function over interval I_i, and $u(0) = u(1) = 0$ imposes the zero boundary condition at both end points. Each function in V_h is uniquely defined, when its value at points $\{x_i\}_{i=2}^{N-1}$ is known. The dimension of the space V_h is denoted by n and is $n = N - 2$.

The boundary conditions imposed in the space V_h complicate the assembly of the matrix A and the vector $\boldsymbol{b}$. Namely, the first and the last element in the partition have to be treated differently from other elements. To simplify the assembly process, it is a common practice to consider the larger space

$$\widehat{V}_h := \{\ u \in C[0,1] \mid u|_{I_i} \in P^1(I_i) \text{ for } i = 1, \ldots, (n-1)\ \},$$

without any imposed boundary conditions, and to construct matrix $\widehat{A} \in \mathbb{R}^{\widehat{n}\times\widehat{n}}$ and vector $\widehat{\boldsymbol{b}} \in \mathbb{R}^{\widehat{n}}$ related to $\widehat{V}_h$. Here $\widehat{n} = N$ is the dimension of the space $\widehat{V}_h$. The space $\widehat{V}_h$ has infinitely many possible basis. In FEM, the hat basis functions $\{\hat{\varphi}_j\}_{j=1}^{\widehat{n}}$, defined as

$$\hat{\varphi}_j \in \widehat{V}_h \quad \text{and} \quad \hat{\varphi}_j(x_p) = \begin{cases} 1 & \text{when } j = p \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

for $j \in 1, \ldots, \widehat{n}$, are used. Examples of these basisfunctions are given in Fig. 7. A basis for the space V_h is obtained as

$$V_h = span\{\hat{\varphi}_2, \ldots, \hat{\varphi}_{\widehat{n}-1}\},$$

this is, V_h is spanned by hat basis functions related to interior nodes of the partition, see Fig. 7. The hat basis functions related to the interior nodes with indeces $j \in \{2, \ldots, \widehat{n} - 1\}$ have the expression

$$\hat{\varphi}_j(x) = \begin{cases} \frac{x - x_{j-1}}{x_j - x_{j-1}} & x \in (x_{j-1}, x_j] \\ \frac{x_{j+1} - x}{x_{j+1} - x_j} & x \in (x_j, x_{j+1}) \\ 0 & \text{otherwise} \end{cases}$$

similar relations hold for boundary nodes x_1 and x_N.

The hat basis functions are preferred as they are *local*: every hat basis function is nonzero over one or two elements. Only basisfunctions $\hat{\varphi}_i$ and $\hat{\varphi}_{i+1}$ have nonzero values on element $I_i = (x_i, x_{i+1})$.

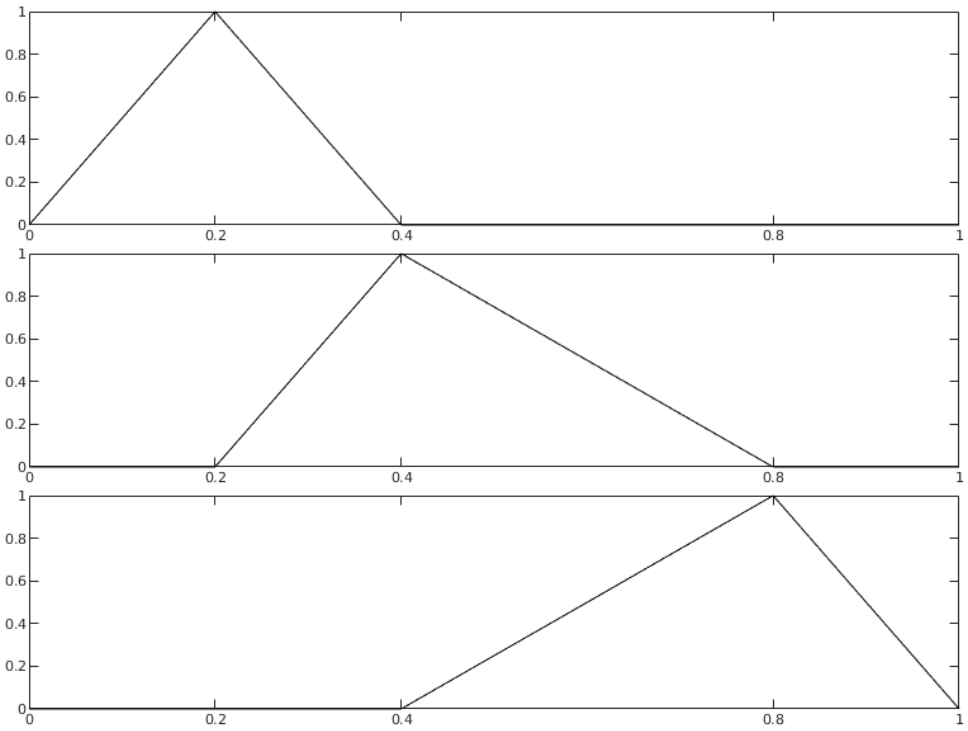

Figure 7: From top to bottom: basis functions $\hat{\varphi}_2$,$\hat{\varphi}_3$, and $\hat{\varphi}_4$ associated to the partition $\{0, 0.2, 0.4, 0.8, 1\}$. These functions are a basis for the space V_h.

Assembly of the linear system Computing entries of the matrix $\widehat{A}$ and the vector $\hat{\boldsymbol{b}}$,

$$\widehat{A}_{ij} = \int_0^1 \frac{d\hat{\varphi}_j}{dx}\frac{d\hat{\varphi}_i}{dx}\,dx \quad \text{and} \quad \hat{b}_i = \int_0^1 f\hat{\varphi}_i\,dx.$$

is called as *assembly*. entries are evaluated in a specific way: Finite element solver features a loop over elements in the mesh or partition. On each element, the solver computes all integrals related to nonzero basis functions over it. The obtained values are added to the appropriate entries of the matrix $\widehat{A}$ and vector $\hat{\boldsymbol{b}}$. This is, the entries are computed by decomposing

Assembly of A in Youtube

$$\widehat{A}_{ij} = \int_0^1 \frac{d\hat{\varphi}_j}{dx}\frac{d\hat{\varphi}_i}{dx}\,dx = \sum_{k=1}^{N-1}\int_{I_k} \frac{d\hat{\varphi}_j}{dx}\frac{d\hat{\varphi}_i}{dx}\,dx$$

and computing the terms $\int_{I_k} \frac{d\hat{\varphi}_j}{dx}\frac{d\hat{\varphi}_i}{dx}\,dx$ for each interval (observe, that most of these terms have value zero). The space $\widehat{V}_h$ is used so that each interval can be treated identically.

We consider first assembly of matrix $\widehat{A}$. On interval I_j, only basisfunctions $\hat{\varphi}_j$

and $\hat{\varphi}_{j+1}$ have non-zero values. Thus, on I_j, the solver evaluates the integrals

$$\int_{I_j} \frac{d\hat{\varphi}_j}{dx}\frac{d\hat{\varphi}_j}{dx}\,dx, \quad \int_{I_j} \frac{d\hat{\varphi}_{j+1}}{dx}\frac{d\hat{\varphi}_j}{dx}\,dx, \quad \text{and} \quad \int_{I_j} \frac{d\hat{\varphi}_{j+1}}{dx}\frac{d\hat{\varphi}_{j+1}}{dx}\,dx. \tag{13}$$

The resulting values are added to entries (j,j), $(j+1,j)$,$(j,j+1)$,$(j+1,j+1)$ of $\widehat{A}$ as

$$\widehat{A}_{lk} = \widehat{A}_{lk} + \int_{I_j} \frac{d\hat{\varphi}_l}{dx}\frac{d\hat{\varphi}_k}{dx}\,dx \quad \text{for} \quad l,k \in \{j, j+1\}.$$

The restrictions of basisfunctions $\hat{\varphi}_j$ and $\hat{\varphi}_{j+1}$ to interval I_j are

$$\hat{\varphi}_j|_{I_j} = \frac{x_{j+1}-x}{x_{j+1}-x_j} \quad \text{and} \quad \hat{\varphi}_{j+1}|_{I_j} = \frac{x_j - x}{x_j - x_{j+1}}.$$

Observe, that the above formulas are valid also on first and last element. The basisfunctions are picewise linear over the partition, hence, their derivatives are constants over interval I_j. This makes the integrals in (13) easy to evaluate.

Next, consider assembly of $\hat{\boldsymbol{b}}$. On element I_j, the integrals

$$\int_{I_j} f\hat{\varphi}_j\,dx, \quad \text{and} \quad \int_{I_j} f\hat{\varphi}_{j+1}\,dx \tag{14}$$

are evaluated. The resulting values are added to entries $\hat{b}_j$ and $\hat{b}_{j+1}$ as

Assembly of $\boldsymbol{b}$ in Youtube

$$\hat{b}_l = \hat{b}_l + \int_{I_j} f\hat{\varphi}_l\,dx \quad \text{for} \quad l \in \{j, j+1\}.$$

Integrals in (14) involve a product of linear function and the source function f. These integrals are computed approximately by using *numerical quadrature rules*. We use the midpoint rule, and approximate

$$\int_{I_j} f\hat{\varphi}_l\,dx \approx f(x_{mp})\varphi_l(x_{mp})(x_{j+1}-x_j) \quad \text{where } x_{mp} = \tfrac{1}{2}(x_j + x_{j+1}).$$

After assembly, the matrix A and vector $\boldsymbol{b}$ are extracted from $\widehat{A}$ and $\hat{\boldsymbol{b}}$. Due to the relation between the basis of V_h and $\widehat{V}_h$ it holds that

$$A_{ij} = \widehat{A}_{i+1,j+1} \quad \text{and} \quad b_i = \hat{b}_{i+1}.$$

for any $i,j = 1,\ldots,n$. In Matlab notation, $A = \widehat{A}(2:(N-1), 2:(N-1))$ and $b = \hat{b}(2:(N-1))$, where N is the number of nodes in the partition.

Implementation in Youtube

Example implementation of the assembly process described above is given below.

```
% Create uniform partition with N nodes for (0,1)
N = 1000;
```

```
x = linspace(0,1,N);

% define the load function.
f = @(x)(ones(size(x)));

% Initialise the matrix Ahat and vector bhat.
Ahat = sparse(N,N);
bhat = zeros(N,1);

% loop over the intervals
for k = 1:(length(x)-1)

    % extract endpoints of the interval
    x1 = x(k);
    x2 = x(k+1);

    % evaluate length of interval k.
    len = x2-x1;

    % evaluate derivatives of basisfunctions on interval k.
    dphi(1) = 1/(x1-x2);
    dphi(2) = 1/(x2-x1);

    % midpoint quadrature points
    t = (x1+x2)/2; w = x2-x1;

    % evaluate values of basisfunctions
    % source term at integration points
    phi(:,1) = (t-x2)./(x1-x2);
    phi(:,2) = (t-x1)./(x2-x1);

    fval = f(t);

    % enumerate the basisfunctions on interval k.
    enum([1 2]) = [k k+1];
    for i=1:2
        % evaluate intergrals related to b.
        bhat(enum(i) ) = bhat(enum(i) ) + dot(fval.*phi(:,i),w);

        for j=1:2
            % evaluate integral related to A
            Ahat(enum(i),enum(j)) = Ahat(enum(i),enum(j)) + dphi(i)*dphi(j)*len;
        end
    end
end

% remove basisfunction 1 and N+1 from the system
A = Ahat(2:(N-1), 2:(N-1) );
b = bhat(2:(N-1),1);

u(1,1) = 0;
u(N,1) = 0;
u(2:(N-1),1) = A\b;
```

```
% plot the solution
plot(x,u);
```

The linear system The linear system $A\boldsymbol{\beta} = \boldsymbol{b}$ defined in the previous step can be very large. For example, obtaining an accurate solution to (5) posed in a complicated three dimensional domain can lead to matrices of the dimension $n = 1 \cdot 10^6$. Although the matrices resulting from finite element discretisation are large, they have very few non-zero entries and are symmetric as well as positive definite.

Matrices with a large number of non-zero entries are called as *sparse*. A considerable amount of memory can be saved by storing only the non-zero entries and their indices of sparse matrices. In Matlab $n \times m$, sparse matrix data type using such storage strategy is initialised by command `A = sparse(n,m)`. Linear systems with sparse, symmetric and positive definite coefficient matrix are solved using process similar to Gaussian elimination with the difference that symmetry is utilised and the process tries to preserve all matrices as sparse in each intermediate step. Solving the linear system is typically the most costly part in FE-analysis.

3.1 Problems

P4. (1p) Let $V_h \subset H_0^1(0,1)$ be a finite dimensional subspace with basis $\{\varphi_1, \ldots, \varphi_n\}$. Show that problem: Find $u_h \in V_h$ satisfying

$$\int_0^1 \frac{du_h}{dx}\frac{dv}{dx}\,dx = \int_0^1 fv\,dx \quad \text{for all } v \in V_h \tag{15}$$

and the problem : Find $w_h \in V_h$ satisfying

$$\int_0^1 \frac{dw_h}{dx}\frac{d\varphi_j}{dx}\,dx = \int_0^1 f\varphi_j\,dx \quad \text{for all } j \in \{1, \ldots, n\}. \tag{16}$$

are equivalent. Proceed as follows:

(a) Show that any solution u_h to (15) satisfies (16).

(b) Show that any solution w_h to (16) satisfies (15).

P5. (2p) Define $V_h \subset H_0^1(\Omega)$ as $V_h := span\{x(1-x), x^2(1-x)\}$. Find $u_h \in V_h$ satisfying

$$\int_0^1 \frac{du_h}{dx}\frac{dv}{dx}\,dx = \int_0^1 \sin(\pi x)v\,dx \quad \text{for all } v \in V_h. \tag{17}$$

Proceed as follows:

(a) Write down the entries of the coefficient matrix $A \in \mathbb{R}^{2\times 2}$ and source vector $\boldsymbol{b} \in \mathbb{R}^2$ of the linear system $A\boldsymbol{\beta} = \boldsymbol{b}$ corresponding to (17)

(b) Compute the entries of A and $\boldsymbol{b}$ by hand.

(c) Use Matlab to solve $\boldsymbol{\beta}$ and plot the approximate solution u_h over $(0,1)$.

P6. (0.5p) Consider the partition $\{0, \frac{1}{2}, 1\}$ and the associated continuous piecewise linear FE-space $\widehat{V}_h$ without imposed boundary conditions.

(a) Draw the hat basisfunctions $\hat{\varphi}_1$,$\hat{\varphi}_2$, and $\hat{\varphi}_3$ of $\widehat{V}_h$.

(b) Represent the function x on $(0,1)$ as a sum of the hat basisfunctions given in (a).

(c) Represent the function $x-1$ on $(0,1)$ as a sum of the hat basisfunctions given in (a).

P7. (2p) Consider the partition $\{0, \frac{1}{2}, 1\}$ and the associated FE-space $\widehat{V}_h$. Denote the second element by $I_2 := (\frac{1}{2}, 1)$.

(a) Give formulas for functions $\hat{\varphi}_2|_{I_2}$, and $\hat{\varphi}_3|_{I_2}$.

(b) Compute by hand the entries $(2,3)$ and $(3,3)$ of matrix $\widehat{A}$ corresponding to $\widehat{V}_h$.

(c) Verify your answer of (b) by using the example FE-code.

P8. (2p) Find exact solution to the strong problem: Find $u \in C^2(0,1) \cap C([0,1])$ satisfying

$$-\frac{d^2u}{dx^2} = \sin(\pi x) \quad \text{and} \quad u(0) = u(1) = 0. \tag{18}$$

Study the effect of partition $\{x_i\}$ to the accuracy of piecewise linear FE-solution. Proceed as follows:

(a) Modify the given example FE-solver to compute piecewise linear FE-approximation to u.

(b) Compute the exact solution to (18).

(c) Compute the FE-solution using partitions having $10, 100, 1000$ uniformly spaced points. Plot FE-solutions and compare them visually to exact solution computed in (b)

(d) Generate partitions $\{x_i\}_{i=1}^N$ as follows:

$$x_i = \frac{i-1}{N-1} + \frac{i-1}{10(N-1)}\left(1 - \frac{i-1}{(N-1)}\right) \quad \text{for} \quad i \in \{1, \ldots, N\}$$

Try different number of points. For each partition, plot FE-solution and compare it visually to the exact solution computed in (b)

In (c), FEM solution has exact nodal values. This *superconvergence* is a special feature of one dimensional finite element solution of the Poisson problem on uniform partition. The partition in (d) is not uniform, FE-solution does not have exact nodal values, and the superconvergence phenomenon is not observed.

4 Error

The finite element solution u_h is an approximation of the exact solution u. In several engineering disciplines, the FE-solution is used investigate if a design satisfies given specifications. For instance, in Example 1.1 we studied if the maximal temperature inside a power transistor stays below the allowed operation temperature specified by the manufacturer. To guarantee that such decisions are correct, it is important to understand what is the error in the computed temperature or the accuracy of the finite element solution.

In this section, we discuss the accuracy of the FE-solution to the one dimensional Poisson's equation (5). We are interested in the error function

$$e = u - u_h.$$

As the basis functions are uniquely defined by the partition, the finite element approximation u_h depends on the partition $\{x_i\}$ and the problem parameters. The source function f is the only parameter of the Poisson's equation posed over interval $(0, 1)$, hence, $e = e(\{x_i\}, f)$.

In the rest of this section, we study the dependency of the error e from the partition and the source term *empirically*. We consider uniform partitions: For $N \in \mathbb{N}, N > 2$, define

$$P_N := \left\{ \frac{i-1}{N-1} \right\}_{i=1}^{N}$$

Observe, that each interval in P_N has identical length. To each uniform partition, we attach the *mesh size* $h = \frac{1}{N-1}$ and characterise the error as a function of h.

The error is studied empirically as follow: First, we choose the such source function f that the solution u can be explicitly found. Then we compute a FE-solution and the corresponding error for several uniform partitions. Finally, the dependency of FE-error on the mesh size h is studied by plotting the error as a function of h. This kind of error study is not complete; by considering particular source terms, we obtain examples on the accuracy of the finite element solution. After developing sufficient tools, the effect of f is analysed mathematically.

Example 4.1. *Let*

$$u = (1 - x)x. \tag{19}$$

By direct computation, it holds that

$$-\frac{d^2u}{dx^2} = 2 \quad \textit{on} \quad (0, 1) \quad \textit{and} \quad u(0) = u(1) = 0.$$

Hence, u in (19) *is a solution to* (5) *with source term $f = 2$. Error functions corresponding to piecewise linear FE-approximation on uniform partitions with $N = 5, 10, 20$ are depicted in Fig. 8. The FE-solution is piecewise linear over the partition, hence, its values between nodes can be evaluated simply by using the Matlab-function* `interp1`*. Example code is given below.*

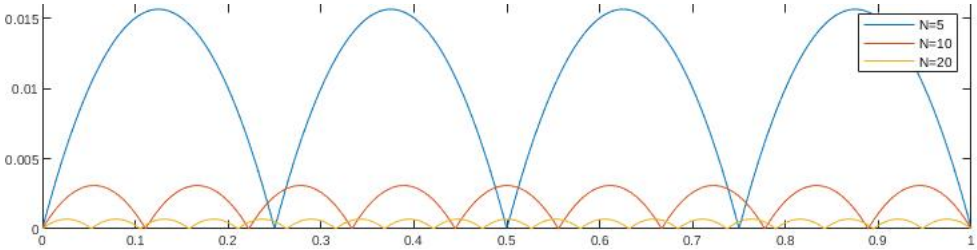

Figure 8: The error functions computed in Example 4.1

```
% Create uniform partition with N nodes for (0,1)
N_list = [5 10 20];

% define the source term.
f = @(x)(2+0*x);

% define plotting grid.
t=linspace(0,1,1000);

figure;
for i = 1:length(N_list)

    % Define the partition
    x = linspace(0,1, N_list(i) );

    % FE-solve.
    u = solver1D(x,f);

    % evaluate FE-function in plotting grid.
    uh_t = interp1(x,u,t);

    % exact solution at plotting grid.
    u_t = t.*(1-t);

    plot(t,u_t-uh_t);hold on;

end

legend('N=5','N=10','N=20');
```

It is difficult to quantify whether the errors are large or small based on Fig.8. Instead, errors are typically quantified by measuring the size of the error function

in energy-, H^1-, or L^2-norm. For our model problem these norms are:

$$\text{energy norm:} \quad \|v\|_E := \left(\int_0^1 \left(\frac{dv}{dx} \right)^2 \, dx \right)^{1/2} \tag{20}$$

$$H^1(0,1)\text{-norm:} \quad \|v\|_{H^1(0,1)} := \left(\int_0^1 \left(\frac{dv}{dx} \right)^2 + v^2 \, dx \right)^{1/2} \tag{21}$$

$$L^2(0,1)\text{-norm:} \quad \|v\|_{L^2(0,1)} := \left(\int_0^1 v^2 \, dx \right)^{1/2}. \tag{22}$$

The energy-norm is related to the left-hand side of the weak form (8) and is different for each PDE.

4.1 Evaluation of error norms

In this section, we discuss how the $L^2(0,1)$-norm of the error $e = u - u_h$ is evaluated in FE-code. Generalisation to the energy- and $H^1(0,1)$-norms is left as an exercise, see P10.

Consider partition $\{x_i\}$, and let $\{\hat{\varphi}_i\}$ be the corresponding basis of the space $\widehat{V}_h$. Recall, that the FE-solution $u_h \in V_h$. Evaluating the $L^2(0,1)$-error in V_h requires special treatment of the first and the last element. To avoid this, we write u_h in the basis of $\widehat{V}_h$ as

$$u_h := \sum_{i=1}^{\hat{n}} \hat{\beta}_i \hat{\varphi}_i. \tag{23}$$

and evaluate the error in $\widehat{V}_h$. The coordinate vector $\hat{\boldsymbol{\beta}} \in \mathbb{R}^{\hat{n}}$ is obtained from coordinate vector $\boldsymbol{\beta} \in \mathbb{R}^n$ of u_h in V_h as

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 0 \\ \boldsymbol{\beta} \\ 0 \end{bmatrix}. \tag{24}$$

To evaluate $\|e\|_{L^2(0,1)}$, the integral over $(0,1)$ is split to a sum of integrals over the elements $I_j := (x_j, x_{j+1})$,

$$\int_0^1 e^2 \, dx = \sum_{j=1}^{N-1} \int_{I_j} [e|_{I_j}]^2 \, dx.$$

Numerical integration is used to approximately evaluate the integrals over I_j. Let $(\boldsymbol{t}, \boldsymbol{w}) \in \mathbb{R}^M \times \mathbb{R}^M$ be a numerical integration rule over I_j and approximate

$$\sum_{j=1}^{N-1} \int_{I_j} e|_{I_j}^2 \, dx \approx \sum_{k=1} [e(t_k)]^2 \, w_k.$$

We proceed to evaluate the values $e(t_k) = u(t_k) - u_h(t_k)$. The restriction of the finite element solution $u_h|_{I_j}$ satisfies

$$u_h|_{I_j} = \hat{\beta}_j \hat{\varphi}_j|_{I_j} + \hat{\beta}_{j+1}\hat{\varphi}_{j+1}|_{I_j}.$$

Hence, $u_h|_{I_j}(t_k) = \hat{\beta}_j \hat{\varphi}_j|_{I_j}(t_k) + \hat{\beta}_{j+1}\hat{\varphi}_{j+1}|_{I_j}(t_k)$, where the restrictions of basisfunctions are

$$\hat{\varphi}_j|_{I_j} = \frac{x_{j+1} - x}{x_{j+1} - x_j} \quad \text{and} \quad \hat{\varphi}_{j+1}|_{I_j} = \frac{x_j - x}{x_j - x_{j+1}}.$$

Example implementation is given in the code below. Numerical integration is done using function `gaussint.m` that can be downloaded from `gaussint.m`

```
%
% x is the partition,
% uh is the coefficient vector of the solution,
% ufun is a function handle to exact solution.
%

function L2error = fem1D_error(x,u,ufun)
    val = 0;

    for i=1:(length(x)-1)

        % Gaussian quadrature rule.
        [t,w] = gaussint(2,x(i),x(i+1));

        % evaluate the finite element solution at points t.

        uh_tk = zeros(1,length(t));
        u_tk = zeros(1,length(t));

        for k=1:length(t)
            uh_tk(k) = u(i)*( x(i+1)-t(k))/(x(i+1)-x(i));
            uh_tk(k) = uh_tk(k) + u(i+1)*( x(i)-t(k))/(x(i)-x(i+1));
            u_tk(k) = ufun(t(k));
        end

        % evaluate the integral
        val = val + (uh_tk - u_tk).^2*w(:);
    end

    L2error = sqrt(val);
end
```

Plotting the error Our aim is to empirically study the dependency of FE-error measured in energy-, $H^1(0,1)$-, or $L^2(0,1)$-norm from the mesh size h and source function f. After the error has been evaluated for several uniform partitions, it is plotted. Different parameter dependencies are revealed by using `loglog` or `semilogy`-plots: Let $\{(y_i, x_i)\}_{i=1}^N \subset \mathbb{R} \times \mathbb{R}$ be the given data points.

- **The `loglog` - plot:** the set of data points is transformed as $(\log y_i, \log x_i)$. The transformed values are plotted in $\mathbb{R}^2$. A line on `loglog`-plot satisfies

$$\log y = k \log x + b \quad \text{so that} \quad y = 10^{k \log x + b} = C x^k.$$

- **The `semilogy` - plot:** the set of data points is transformed as $(\log y_i, x_i)$. The transformed values are plotted in $\mathbb{R}^2$. A line on `semilogy`-plot satisfies

$$\log y = kx + b \quad \text{so that} \quad y = 10^{kx+b} = C\rho^x.$$

For instance, the `loglog` plot suggest that the $L^2(0,1)$ error in Fig. 9 depends on the mesh size as h^2.

Example 4.2. *Next, we compute the errors corresponding to source function $f = 2$ corresponding to the exact solution $u = x(1-x)$. The $L^2(0,1)$-error is evaluated on uniform partitions with $N \in \{10, 20, 40, 80, 160, 320\}$ using the code given below. The obtained errors are depicted in Fig. 9. The computed points lie on a line. The slope is visually determined by comparing it to plots of functions h and h^2.*

```
% Create uniform partition with N nodes for (0,1)

N_list = [10 20 40 80 160 320];

% define the source term.
f = @(x)(2+0*x);

for i = 1:length(N_list)

    % Define the partition
    x = linspace(0,1, N_list(i) );

    % FE-solve.
    u = solver1D(x,f);

    % Rvaluate the error
    L2error(i) = fem1D_error(x,u, @(x)( x.*(1-x)) );

end

figure;loglog(1./(N_list-1),L2error,'k:o')
```

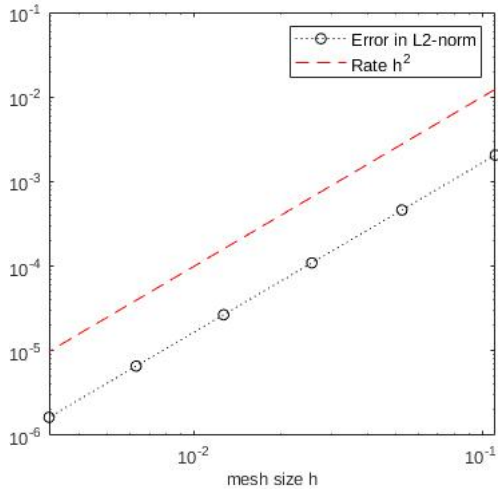

Figure 9: The $L^2(0,1)$-norm of the error computed in Example 4.2 Based on this the Figure, we the $L^2(0,1)$-error seems to behave as Ch^2 for some constant C.

4.2 Problems

P9. (1p) Consider the following approximations :

$$\int_0^1 x^2 \, dx \approx \frac{1}{N+1} \sum_{i=0}^{N} \left(\frac{i}{N+1} \right)^2 .$$

and

$$\frac{1}{1-r} \approx \sum_{i=0}^{N} r^i, r = \frac{1}{2}.$$

(a) In both cases, use Matlab to compute and to plot the approximation error as a function of N using commands `plot`, `semilogy` and `loglog`.

(b) Which graph is the most informative ?

(c) Use the graphs to determine the relation between the error and the parameter N for both approximations.

P10. (2p)

(a) Write a matlab function for evaluating the energy norm of the error.

(b) Write a Matlab function for evaluating the $H^1(0,1)$-norm of the error. Hint: observe that $\| \cdot \|_{H^1(0,1)} = \left(\| \cdot \|_E^2 + \| \cdot \|_{L^2(0,1)}^2 \right)^{1/2}$.

(c) Repeat the computation done in Example 4.2 and plot the energy- and $H^1(0,1)$-norms of the error as a function of the mesh size h. Use `loglog`-plot. How does the error depend on h ?

P11. (1p) Let

$$f(x) = \begin{cases} -1 & x \in \left(0, \frac{1}{2}\right] \\ 1 & x \in \left(\frac{1}{2}, 1\right) \end{cases} \tag{25}$$

Study the FE-error using a a sequencece of uniform partitions with $N = 2^k + 1$ nodes. Proceed as follows:

(a) Verify that the discontinuity of f matches with the nodes in the mesh.

(b) Compute the $L^2(0,1)$, $H^1(0,1)$ and energy norm errors.

(c) Plot the $L^2(0,1)$, $H^1(0,1)$ and energy norm errors. Use the plot to determine, how the error depends on h.

P12. (2p) In assembly of 1D-finite element matrices, one has to compute integral over interval $(a,b) \subset \mathbb{R}$, where $a < b$. Such integrals are approximated using numerical quarature rules as

$$\int_a^b f(x) \, dx \approx \sum_{i=1}^{m} f(x_i) w_i,$$

where $x_1, \ldots, x_m \in \mathbb{R}$ and $w_1, \ldots, w_m \in \mathbb{R}$ are called quadrature points and weights. Consider the midpoint rule,

$$m = 1, \quad x_1 = \frac{a+b}{2}, \quad \text{and} \quad w_1 = (b-a).$$

In the following, limit to interval $(0, h)$, $h > 0$.

(i) Study the accuracy of the midpoint rule as a function of interval lenght h numerically visualising the error

$$err(h) := \left| \int_0^h x^2 \, dx - x_1^2 w_1 \right|$$

for different values of h. Use logarithmic-plot. For reference, plot functions h, h^2, h^3 and guess how the error behaves.

(ii) Using integration by parts show that

$$f(x) = f\left(\frac{h}{2}\right) + f'\left(\frac{h}{2}\right)\left(x - \frac{h}{2}\right) - \int_{h/2}^x f''(t)\,(t-x)\,dt.$$

Hint: start by writing

$$f(x) = f\left(\frac{h}{2}\right) + \int_{h/2}^x f'(t)\,dt$$

(iii) Using (ii) show that

$$\left| \int_0^h f(t)\,dt - f\left(\frac{h}{2}\right) h \right| \leq \sup_{x \in (0,h)} |f''(x)| \frac{1}{24} h^3.$$

How does this estimate correspond to (i) ?

5 Second-order basisfunctions

Accuracy of finite element solution can be improved by modifying the partition or the applied FE-space V_h. In this section, we consider *second-order* FE-spaces,

$$V_h^2 := \{\, u \in C[0,1] \mid u(0) = u(1) = 0,\ u|_{I_i} \in P^2(I_i) \text{ for } i = 1, \ldots, (N-1) \,\}.$$

Similarly to Section 3, the finite element implementation is simplified by using the space

$$\widehat{V}_h^2 := \{\, u \in C[0,1] \mid u|_{I_i} \in P^2(I_i) \text{ for } i = 1, \ldots, (N-1) \,\}.$$

without imposed boundary conditions.

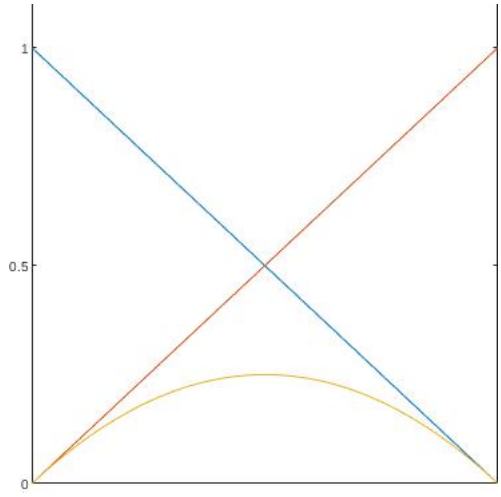

Figure 10: The second order basisfunctions φ_i, φ_{j+1}, and $\varphi^{(2)}_{I_j}$.

We proceed to define a basis for the space $\widehat{V}_h^2$. There are several possible choices for the basis, that mostly affect the implementation. We use a simple *hierarchial basis* for $\widehat{V}_h^2$ that is obtained by adding the bubble functions

$$\varphi^{(2)}_{I_j} = \begin{cases} (x - x_j)(x_{j+1} - x) & \text{for } x \in I_j \\ 0 & \text{otherwise} \end{cases}$$

for every $j \in \{1, \ldots, N-1\}$ to the hat-basis of $\widehat{V}_h$. The bubble functions are second order polynomials, and have value zero at both endpoints, see Fig. 10. The inclusion of these new basisfunctions requires few small modification of our example FE-solver:

1. **Indexing:** To keep track on which basisfunction is related to which element the basisfunctions $\hat{\varphi}^{(p)}_{I_j}$ are given global indices as

$$\hat{\varphi}_{N+j} := \hat{\varphi}^{(2)}_{I_j}$$

2. **Integration:** When using second order basisfunctions, their derivatives are linear functions. Thus, the entries of the matrix $\widehat{A}$ have to be evaluated using numerical integration. Download and use the function `gaussint.m`

3. **Elimination of boundary basisfunctions**. The matrix A is now extracted form $\widehat{A}$ by removing the rows and the columns as follows:

```
idof = setdiff(1:(2*N-1),[1 N]); A = Ahat(idof,idof);
```

same process is used for $\boldsymbol{b}$. After solution, it is best to store the solution as coordinate of $\widehat{V}_h$, e.g., as

```
u(1) = 0; u(N) = 0; u(idof) = A/b;
```

4. **Plotting:** To plot the solution, we need to evaluate the finite element solution on a set of points used for plotting. This is done similar to evaluation of $L^2(0,1)$-norm of error, see problem 13.

5.1 Problems

P13. (2p). Write a Matlab-function for plotting $v \in \widehat{V}_h^2$ and it's derivative. Proceed as follows:

(a) Loop over the elements. On element $(x(k), x(k+1))$, create a finer plotting partition, for example, as
`t = linspace(x(k),x(k+1),10);`.

(b) Evaluate the basisfunctions and their derivatives at the nodes of the plotting partition. Then, compute the value of v at the plotting partition using the formula

$$v|_{I_k} = \beta_k \hat{\varphi}_k|_{I_k} + \beta_{k+1}\hat{\varphi}_{k+1}|_{I_k} + \beta_{N+k}\hat{\varphi}^{(2)}_{I_k}.$$

Use the same expansion to evaluate the first derivative of v.

(c) Test your implementation on partition $\{0, 0.3, 1\}$ and plot $v \in \widehat{V}_h^2$ corresponding to

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 0 & 0.0525 & 0 & 0.25 & 0.25 \end{bmatrix}.$$

P14. (2p). Modify the example FE-solver to use second order polynomial basisfunctions. You have to do the modifications indicated in the list on p.29.

P15. (2p). Study the accuracy of the second order FE-solver. Proceed as follows:

(a) Write a function for evaluating the energy-, $L^2(0,1)$-, and $H^1(0,1)$-errors of second order FE-solution.

(b) Let $f = \sin \pi x$, and find the exact solution to (5) by integration.

(c) Compute the energy-, $L^2(0,1)$-, and $H^1(0,1)$-errors using uniform partitions with 10, 20 , 40, 80, 160, and 320 nodes.

(d) What do you observe ? Compare to results obtained with the first order finite element method.

P16. (2p). Let $\varphi_i = x(1-x)x^{i-1}$ and the space $V^p = \text{span}\{\varphi_i\}_{i=1}^n$. Consider the problem: find $u \in V^p$ such that

$$\int_0^1 u'(x)v'(x)\,\mathrm{d}x = \int_0^1 f(x)v(x) \quad \forall v \in V^p,$$

(i) Write a Matlab program that assembles the matrix $A \in \mathbb{R}^{n\times n}$, $A_{ij} = a(\varphi_j, \varphi_i)$, and the vector $\boldsymbol{b} \in \mathbb{R}^n$, $\boldsymbol{b}_i = L(\varphi_i)$. You can fill in the missing parts in the `ex1_p2.m` -function, see for file `ex1_p2.zip`.

(ii) Let $f(x) = \sin \pi x$ and solve the problem for different values of n. Study the accuracy of the solution by plotting the difference between the exact solution $-u_e'' = f$ and the approximate one.

6 Sobolev Spaces

In Section 3, we claim that it is reasonable to look for the weak solution u to (7) from the Sobolev space $H_0^1(0,1)$. In this Section, we discuss Sobolev spaces, a family of function spaces defined by giving conditions for the integrability of derivatives, in more detail. We restrict ourselves to Sobolev spaces based on the L^2 function space. All definitions are given for interval $I = (a, b)$.

We proceed as follows: first, we generalise the derivative to functions less regular than $C^1(I)$. Then we define the spaces $H^m(I)$ and $H_0^1(I)$, i.e. subspace of $H^1(I)$ with imposed boundary condition $u(a) = u(b) = 0$. Finally, we specify norm and inner product for these spaces and prove the Poincare-inequality that is an essential component in the existence proof.

Weak derivative in Youtube

Weak derivative Recall that the value of an integral does not depend on the behavior of the integrated function at a finite set of points. This property allows us to extend the definition of the derivative to functions that are not differentiable in the classical sense. The extension is based on the integration by parts formula

$$\int_I u'\varphi = -\int_I u\varphi' \tag{26}$$

valid for all $u \in C^\infty(I)$ and $\varphi \in C_0^\infty(I)$. Observe, that the mapping

$$\varphi \mapsto -\int_I u\varphi'$$

is well defined for any $u \in L_{loc}^1(I)$ and $\varphi \in C_0^\infty$. The notation $L_{loc}^1(I)$ denotes the space of functions that are Lebesque integrable over every compact subset of I. The space $L_{loc}^1(I)$ is larger than the space $L^1(I)$. For example, the function $\frac{1}{x}$ is integrable over every compact subset of $(0,1)$, but not over the whole interval $(0,1)$.

Weak derivative is defined as:

Definition 6.1. *The function $u \in L_{loc}^1(I)$ is* weakly differentiable *if there exists $v \in L_{loc}^1(I)$ satisfying*

$$\int_I v\varphi = -\int_I u\varphi' \quad \textit{for all} \quad \varphi \in C_0^\infty(I). \tag{27}$$

We call v as the weak derivative of u.

Intuitively speaking, the weak derivative is defined *by solving it* form the integration by parts formula (26). As all functions in (27) are under integral sign, weak derivative is defined *almost everywhere* in I or up to finite set or pointwise values. The weak derivative v is sought from $L^1_{loc}(I)$ as this is the largest possible space guaranteeing that $\int_I v\varphi$ is well defined. Observe, that there are functions $u \in L^1_{loc}$ that do not have a weak derivative, this is there does not exists $v \in L^1_{loc}(I)$ satisfying (27).

Example 6.1. *Consider computing the weak derivative of*

$$u = \begin{cases} x & \text{for } x \in (0,1) \\ (2-x) & \text{for } x \in [1,2). \end{cases}$$

There holds that

$$\int_0^2 u\varphi' = \int_0^1 x\varphi' + \int_1^2 (2-x)\varphi'.$$

Using Integration by parts, continuity of u, and continuity of φ gives

$$\int_0^2 u\varphi' = -\int_0^1 \varphi + \int_1^2 \varphi = -\int_0^2 v\varphi \quad \text{for} \quad v = \begin{cases} 1 & \text{for } x \in (0,1) \\ -1 & \text{for } x \in (1,2). \end{cases}$$

for all $\varphi \in C_0^\infty(I)$. Thus, v is the weak derivative of u. Observe, that v is not uniquely defined at $x = 1$.

Example 6.2. *Consider computing the weak derivative of*

$$u = \begin{cases} x & \text{for } x \in (0,1) \\ (3-x) & \text{for } x \in [1,2). \end{cases}$$

There holds that

$$\int_0^2 u\varphi' = \int_0^1 x\varphi' + \int_1^2 (3-x)\varphi'.$$

Using integration by parts, definition of u, and continuity of φ gives

$$\int_0^2 u\varphi' = \int_0^2 v\varphi + \varphi(1) \quad \text{for} \quad v = \begin{cases} -1 & \text{for } x \in (0,1) \\ 1 & \text{for } x \in (1,2). \end{cases}$$

for all $\varphi \in C_0^\infty(I)$. We next investigate if the RHS can be written as $\int_0^2 f\varphi$. Assume that $f \in L^1_{loc}(I)$ satisfying

$$\int_0^2 f\varphi = \int_0^2 v\varphi + \varphi(1). \tag{28}$$

exists. Now, choose a sequence $\{\psi_m\}$ such that $(f - v)\psi_m \to 0$ in $L^1_{loc}(I)$ and $\psi_m(1) = 1$. Showing the existence of such sequence $\{\psi_m\}$ is non-trivial, and

requires the use of theorems related to Lebesque integration. As this is not the topic of this lecture note, all details are omitted. There holds,

$$1 = \psi_m(1) = \int_0^2 (f - v)\psi_m \to 0,$$

which is a contradiction. Thus there does not exists f satisfying (28)*, and u is not weakly differentiable.*

The function u is said to be n-times weakly differentiable, if it's $(n-1)$th weak derivative is weakly differentiable.

Sobolev spaces We have the following definition

Sobolev spaces in Youtube

Definition 6.2. *Let $m \in \mathbb{N}$, and define the space $H^m(I) \subset L^2(I)$ as the set of those functions that are m-times weakly differentiable with all weak derivatives up to order m in $L^2(I)$.*

We proceed to specify inner product and the norm for there spaces. Norms are used to measure how large functions are. They are necessary, for instance, to study the dependency of the solution u on the source function f or to quantify the FE-error.

Definition 6.3. *Let u, v in $L^2(I)$. The inner product and the induced norm of the space $L^2(I)$ are*

$$(u, v)_{L^2(I)} = \int_I uv\, dx \quad \textit{and} \quad \|u\|_{L^2(I)} = (u, u)^{1/2}_{L^2(I)} = \left(\int_I u^2\, dx \right)^{1/2}, \quad (29)$$

respectively.

Definition 6.4. *Let $m \in \mathbb{N}$ and $u, v \in H^m(I)$. The inner product and the induced norm of $H^m(I)$ are*

$$(u, v)_{H^m(I)} = \sum_{\alpha=0}^{m} \left(\frac{d^{(\alpha)}u}{dx^{(\alpha)}}, \frac{d^{(\alpha)}v}{dx^{(\alpha)}} \right)_{L^2(I)} \quad (30)$$

and

$$\|u\|^2_{H^m(I)} = (u, u)_{H^m(I)} = \sum_{\alpha=0}^{m} \left\| \frac{d^{(\alpha)}u}{dx^{(\alpha)}} \right\|^2_{L^2(I)}, \quad (31)$$

respectively. Here we have use notation $\frac{d^{(0)}u}{dx^{(0)}} = u$.

A function v belongs to space $H^m(I)$ if it is m-times weak differentiable with derivatives in $L^2(I)$, hence, the inner product and induced norm for $H^m(I)$ in Definition 6.4 are well defined.

All inner products and the induced norms satisfy the Cauchy-Schwartz inequality. Particularly, there holds that

$$(u, v)_{L^2(I)} \leq \|u\|_{L^2(I)} \|v\|_{L^2(I)} \tag{32}$$

for all $u, v \in L^2(I)$. The above inequality is frequently used in the following.

It is often useful to define the semi-norms:

Definition 6.5. *Let $m \in \mathbb{N}$ and $u \in H^m(I)$. Define the $H^m(I)$-semi-norm as*

$$|u|_{H^m(I)} := \left\| \frac{d^{(m)}u}{dx^{(m)}} \right\|_{L^2(I)}. \tag{33}$$

This definition makes more sence in dimensions $d = 2, 3$, where it involves sum of all derivatives of order m.

The Sobolev space $H^m(I)$ is complete with respect to the norm $\|\cdot\|_m$, and thus a Hilbert space. Let $I \subset \mathbb{R}^n$ be an open and bounded set. Then $C^\infty(I)$ is dense in $H^m(I)$, and $H^m(I)$ is a completion of $C^\infty(I)$ with respect to norm $\|\cdot\|_m$. This property allows us to give most proofs for $C^\infty(I)$-functions and extend them to $H^m(I)$ by density.

6.1 Poincare inequality and space $H_0^1(I)$

$H_0^1(I)$ and P-F in Youtube

We have not specified what we mean by the boundary condition

$$u(0) = u(1) = 0.$$

This a delicate question, because the functions $u, v \in L^2(I)$ are equivalent in $L^2(I)$ if

$$\|u - v\|_{L^2(I)} = 0 \quad \text{or} \quad \int_I (u - v)^2 \, dx = 0.$$

As the value of the integral is independent of pointwise behavior of the integrated function at finite set of points, u and v are identical even if they have different values at some points. Thus, it is not meaningful to impose pointwise constrains to $L^2(I)$-functions.

In one spatial dimension, functions in the space $H^1(I)$ are *continuous*, this is $H^1(I) \subset C(\overline{I})$. Thus, the boundary values of $H^1(I)$ - functions are well defined. This is not the case in dimensions two or three, where we need to specify boundary conditions in different manner. Let $I = (a, b)$. Then the space $H_0^1(I)$ is defined as

$$H_0^1(I) := \{\, u \in H^1(I) \mid u(a) = u(b) = 0 \,\}.$$

Functions in the space $H_0^1(I)$ satisfy the Poincaré-Friedrichs inequality that relates their $L^2(I)$- and $H^1(I)$-semi-norms.

Theorem 6.1 (Poincaré-Friedrichs Inequality)**.** *Let $I = (a, b)$ and denote $s = b - a$. Then*

$$\|v\|_{L^2(I)} \leq s\|v'\|_{L^2(I)} \quad \forall v \in H_0^1(I).$$

Remark 6.1. *In the above theorem it is essential that v vanishes on the boundary. There is also variant of the theorem that does not require the function to vanish, but instead requires the mean value to vanish, that is, the above holds also for $v \in H^1(I)$ such that $\int_I v\, dx = 0$, see P19.*

Proof. Without loss of generality, we assume that $I = (0, s)$. Let $v \in C_0^\infty(I)$. Since the boundary values vanish, i.e. $v(0) = 0$, we have

$$v(x) = v(0) + \int_0^x v'(z)\, dz = \int_0^x v'(z)\, dz.$$

Using Cauchy-Schwarz we get

$$\begin{aligned}
|v(x)|^2 &= \left| \int_0^x v'(z)\, dz \right|^2 \\
&\leq \left| \left(\int_0^x 1^2\, dz \right)^{1/2} \left(\int_0^x v'(z)^2\, dz \right)^{1/2} \right|^2 \\
&\leq \left[\left(\int_0^x |1|^2\, dz \right)^{1/2} \left(\int_0^x |v'(z)|^2\, dz \right)^{1/2} \right]^2 \\
&= \int_0^x |1|^2\, dz \int_0^x |v'(z)|^2\, dz \\
&\leq \int_0^s |1|^2\, dz \int_0^s |v'(z)|^2\, dz \\
&= s\|v'\|_{L^2(I)}^2.
\end{aligned}$$

Next we integrate over the I to obtain

$$\|v\|_0^2 = \int_0^s |v(x)|^2\, dx \leq \int_0^s s|v|_1^2\, dx = s^2|v|_1^2.$$

The proof is completed by using density argument. □

This proof is easily extended to higher dimensions by using the above construction for each dimension separately.

6.2 Problems

P17. (2p) Let $\alpha \in \mathbb{R}$ and $u : (-1, 1) \mapsto \mathbb{R}$ be defined as $u(x) := (|x|^\alpha - 1)$.

(a) Plot u for different values of α.

(b) For which α is u in $L^2(-1, 1)$?

(c) For which α is u weakly differentiable ?

(d) For which α is u in $H^1(-1,1)$?

P18. (2p) Let $I = (0, s)$. The Poincaré-Friedrichs Inequality states that $\|u\|_{L^2(I)} \leq C(s)\|u'\|_{L^2(I)}$ for any $u \in H_0^1(I)$ and some $C(s) > 0$ independent of u but dependent on s. Study computationally how the constant $C(s)$ depends on s. Proceed as follows:

(a) Show that smallest possible C for interval I is characterized as

$$C^{-2} = \min_{u \in H_0^1(I)} G(u) \quad \text{where} \quad G(u) = \frac{(u', u')_{L^2(I)}}{(u, u)_{L^2(I)}}.$$

(b) Show that C^{-2} is the smallest eigenvalue λ_i of the problem: find $(\lambda_i, v_i) \in (\mathbb{R}, H_0^1(I) \setminus \{0\})$ such that

$$(v_i', \varphi') = \lambda_i (v_i, \varphi_i) \quad \forall v \in H_0^1(I).$$

Hint : the minimum is located at the critical point u of G that can be characterized as $\frac{d}{dt} G(u + tv)_{|_{t=0}} = 0 \quad \forall v \in H_0^1(I)$. Also, note the the each eigenvalue satisfies $\lambda_i = \frac{(v_i', v_i')}{(v_i, v_i)}$, in which v_i is the eigenvector corresponding to λ_i.

(c) Modify the example FE-solver to compute an approximation to the constant $C(s)$. Plot the constant as a function of s. How good is the value given in Theorem 6.1 ?

Hint: The eigenvalue problem that you need to solve is $A\boldsymbol{x} = \lambda M \boldsymbol{x}$, in which $A_{ij} = (\varphi_j', \varphi_i')$ and $M_{ij} = (\varphi_j, \varphi_i)$. In Matlab, the smallest eiganvalue of such evp. can be solved with the command `eigs(A,M,1,'SM')`.

P19. (2p) Let $s > 0$ and

$$W_0 := \{\, u \in C^\infty(0, s) \mid \int_0^s u(t)\, dt = 0 \,\}.$$

(i) Show that for any $u \in W_0$ there exists some $\xi \in (0, s)$ such that $u(\xi) = 0$.

(ii) Show that there exists a positive constant $C(s)$ independent of u such that

$$\|u\|_{L^2(0,s)} \leq C(s)\|u'\|_{L^2(0,s)} \quad \forall u \in W_0.$$

By density argument, this inequality also holds in the space $V_0 := \{\, u \in H^1(\Omega) \mid \int_0^s u = 0 \,\}$.

7 Existence of unique solution

Abstract setting in Youtube

In this section, we show that the problem: find $u \in H_0^1(0,1)$ satisfying

$$\int_0^1 \frac{du}{dx}\frac{dv}{dx}\,dx = \int_0^1 fv \quad \text{for all} \quad v \in H_0^1(0,1),$$

has a unique solution. We rely on the Lax-Milgram theorem, an existence theorem formulated for an abstract variational problem. It can also be applied to show the existence of a unique solution to other divergence-form PDEs.

Let V be a Hilbert space with inner product $\langle\cdot,\cdot\rangle_V$ and the induced norm $\|\cdot\|_V$. In addition, let $a : V \times V \to \mathbb{R}$ be a bilinear form and $L : V \to \mathbb{R}$ a linear functional.

Definition 7.1. *Mapping $a : V \times V \mapsto \mathbb{R}$ is called as bilinear, if it is linear in both of it's arguments. This is, for $u, v, w \in V$ and $\alpha \in \mathbb{R}$, we have*

$$\begin{aligned} a(u+v,w) &= a(u,w) + a(v,w),\\ a(u,v+w) &= a(u,v) + a(u,w),\\ a(\alpha u,v) &= \alpha a(u,v),\\ a(u,\alpha v) &= \alpha a(u,v). \end{aligned}$$

Definition 7.2. *Mapping $L : V \mapsto \mathbb{R}$ is linear, if it holds that*

$$L(u+v) = L(u) + L(v) \quad \textit{and} \quad L(\alpha v) = \alpha L(v).$$

The Lax-Milgram theorem concerns the *variational problem*: find $u \in V$ satisfying

$$a(u,v) = L(v) \quad \forall v \in V. \tag{34}$$

If we set $V = H_0^1(0,1)$, $a(u,v) = \int_0^1 u'v'\,dx$, and $L(v) = \int_0^1 fv\,dx$, problem (34) corresponds to the weak problem (7). Studying the above, more abstract problem allows us to develop tools that can be used to tackle other problems, e.g., the two dimensional Poisson problem.

Following assumptions are made on a and L in (34):

Assumption 7.1. *Let V be a Hilbert space with norm $\|\cdot\|_V$, $a : V \times V \mapsto \mathbb{R}$ a bilinear form, and $L : V \mapsto \mathbb{R}$ a linear functional. Assume that a and L satisfy:*

$$|a(u,v)| \le C\|u\|_V\|v\|_V \qquad \textit{(Continuity)} \tag{35}$$

$$a(u,u) \ge \alpha\|u\|_V^2 \qquad \textit{(Ellipticity)} \tag{36}$$

$$|L(v)| \le C_L\|v\|_V \qquad \textit{(Boundedness)} \tag{37}$$

for all $u, v \in V$ and constants $C, C_L\alpha > 0$ indepedent on u and v.

Assumptions (35), (37) are very mild and guarantee that the definitions of $a(\cdot,\cdot)$ and $L(\cdot)$ are reasonable. The assumption (36) is very strong and has central role in the existence proof. It is also the most difficult assumption to satisfy.

Theorem 7.1 (Lax-Milgram Theorem)**.** *Let $a : V \times V \to \mathbb{R}$ be an elliptic and continuous bilinear form and $L : V \to \mathbb{R}$ a bounded linear functional, then there exists a unique $u \in V$ such that*

$$a(u, v) = L(v) \quad \forall v \in V. \tag{38}$$

Example 7.1. *We proceed to show that the problem: find $u \in H_0^1(0,1)$ satisfying* Example 7.1 in Youtube

$$\int_0^1 \frac{du}{dx}\frac{dv}{dx}\, dx = \int_0^1 fv \quad \textit{for all} \quad v \in H_0^1(0,1),$$

has a unique solution for any $f \in L^2(I)$. Let $V = H_0^1(0,1)$, $a(u,v) = \int_0^1 u'v'\, dx$, and $L(v) = \int_0^1 fv$. Clearly, $a(u,v)$ is bilinear and $L(v)$ is linear. To apply Lax-Milgram Theorem, we have to verify that $a(\cdot,\cdot)$ and $L(\cdot)$ satisfy Assumptions 7.1.

Continuous *First, we show that $a(\cdot,\cdot)$ satisfies* (35)*. By Cauchy–Schwarz inequality Eq.* (32)*, it holds that*

$$|a(u,v)| = \left| \int_0^1 u'v'\, dx \right| = |(u', v')_{L^2(I)}| \leq \|u'\|_{L^2(I)} \|v'\|_{L^2(I)} \tag{39}$$

As $x \mapsto \sqrt{x}$ is monotonously increasing and $\|w\|_{L^2(I)} \geq 0$, it holds that

$$\|w'\|_{L^2(I)} = \left(\|w'\|_{L^2(I)}^2\right)^{1/2} \leq \left(\|w\|_{L^2(I)}^2 + \|w'\|_{L^2(I)}^2\right)^{1/2} = \|w\|_{H^1(I)}$$

for any $w \in H^1(\Omega)$. It follows from (39) *that $a(\cdot,\cdot)$ is continuous.*

Elliptic *Next, we show that $a(\cdot,\cdot)$ is elliptic. For this purpose, we use the Poincaré-Friedrichs Inequality given in Theorem 6.1: $\|w\|_{L^2(I)} \leq C\|w'\|_{L^2(I)}$ for any $w \in H_0^1(I)$ and some $C > 0$ dependent on I. There holds that*

$$a(u,u) = \int_0^1 \left(u'\right)^2\, dx = \|u'\|_{L^2(I)}^2.$$

Split $\|u'\|_{L^2(I)}^2 = \frac{1}{2}\|u'\|_{L^2(I)}^2 + \frac{1}{2}\|u'\|_{L^2(I)}^2$. Applying the Poincaré-Friedrichs inequality to the latter term yields

$$\|u'\|_{L^2(I)}^2 \geq \frac{1}{2}\|u'\|_{L^2(I)}^2 + \frac{1}{2C}\|u\|_{L^2(I)}^2 \geq \min\left\{\frac{1}{2}, \frac{1}{2C}\right\}\left(\|u'\|_{L^2(I)}^2 + \|u\|_{L^2(I)}^2\right).$$

Hence, $a(u,u) \geq \alpha\|u\|_{H^1(I)}^2$ for all $u \in H_0^1(I)$ and $\alpha = \min\left\{\frac{1}{2}, \frac{1}{2C}\right\}$.

Bounded *Finally, we show that $L(v)$ is bounded. By C-S inequality in Eq.* (32)*,*

$$|L(v)| = \left| \int_I fv \, dx \right| = \left| (f, v)_{L^2(I)} \right| \leq \|f\|_{L^2(I)} \|v\|_{L^2(I)},$$

for any $v \in H_0^1(I)$. As $\|v\|_{L^2(I)} \leq \|v\|_{H^1(I)}$, Assumptions 7.1 are satisfied, and the Lax-Milgram Theorem guarantees existence of unique solution to (7)*.*

Simplified proof of LM in Youtube

There are several alternative ways to prove the Lax-Milgram Theorem 7.1. We give a simplified proof under the assumption that V is a finite dimensional space with the aim to build intuition on the the importance of Assumptions 7.1. Proof in infinite dimensions utilises similar strategy, but is more technical. In the proof we work both with functions in V and their coordinate vectors $\boldsymbol{\beta} \in \mathbb{R}^n$ in basis $\{\varphi_1, \ldots, \varphi_n\}$ of V. We write

$$\boldsymbol{\beta} \sim u \quad \text{in } V \quad \text{if} \quad u = \sum_{j=1}^{n} \beta_j \varphi_j.$$

proof of Lax-Milgram theorem in finite dimension. Assume that V is finite dimensional. This is V has a basis $\{\varphi_j\}_{j=1}^n$ where $n = dim\, V$.

Formulate as linear system Similar to Section 3, we use the basis of V to reformulate (34) as a linear system. As V is finite dimensional, (34) is equivalent to: find $u \in V$ satisfying

$$a(u, \varphi_i) = L(\varphi_i) \quad \text{for each} \quad i \in \{1, \ldots, n\}.$$

Expanding $u = \sum_{j=1}^n \beta_j \varphi_j$ gives the linear system: find $\boldsymbol{\beta} \in \mathbb{R}^n$ satisfying

$$A\boldsymbol{\beta} = \boldsymbol{b} \quad \text{for} \quad A_{ij} = a(\varphi_j, \varphi_i) \quad \text{and} \quad b_i = L(\varphi_i)$$

for all $i, j \in \{1, \ldots, n\}$.

Study existence of unique solution to the linear system The linear system $A\boldsymbol{\beta} = \boldsymbol{b}$ has a unique solution if it has a trivial null-space this is, $N(A) = \{\boldsymbol{x} \in \mathbb{R}^n \mid A\boldsymbol{x} = 0\} = \{0\}$. This is,

$$A\boldsymbol{x} = 0 \Rightarrow \boldsymbol{x} = 0.$$

Let $A\boldsymbol{x} = 0$. Clearly, $\boldsymbol{x}^T A\boldsymbol{x} = 0$. By problem P20, it holds that $\boldsymbol{x}^T A\boldsymbol{x} = a(u, u)$ for $\boldsymbol{x} \sim u$ in V. Using ellipticity gives $\boldsymbol{x}^T A\boldsymbol{x} = a(u, u) \geq \alpha \|u\|_V^2$, and further

$$\|u\|_V^2 \leq 0.$$

As $\|\cdot\|_V$ is norm of V it follows that $\|u\|_V = 0$, $u = 0$, $\boldsymbol{x} = 0$, and $N(A) = \{0\}$. This concludes the proof. □

When V is infinite dimensional space, one has to formulate the variational problem as *an operator equation* in V using the Riesz representation theorem. The ellipticity assumption (36) is used to show that the operator equation has a unique solution. Understanding the proof requires tools from functional analysis, and thus, it is omitted.

7.0.1 Problems

P20. (1p) Let $V = span\{\varphi_1, \ldots, \varphi_n\}$, $a : V \times V \to \mathbb{R}$ be symmetric and bilinear, and $L : V \to \mathbb{R}$ be linear. In addition, let $A \in \mathbb{R}^{n\times n}, \boldsymbol{b} \in \mathbb{R}^n$ be defined as $A_{ij} = a(\varphi_j, \varphi_i)$ and $b_i = L(\varphi_i)$ for $i, j \in \{1, \ldots, n\}$. Show that

(i) $(A\boldsymbol{x})_l = a(u, \varphi_l)$ for $1 \leq l \leq k$ and $\boldsymbol{x} \sim u$ in V.

(ii) $a(u, v) = \boldsymbol{y}^T A\boldsymbol{x}$ for $\boldsymbol{y} \sim v$ and $\boldsymbol{x} \sim u$ in V

(iii) $L(v) = \boldsymbol{y}^T \boldsymbol{b}$ for $\boldsymbol{y} \sim v$ in V.

P21. (0.5p)

(i) Let $V = \mathbb{R}$ and fix $a(x, y) = 2xy$ and $L(x) = 10x$. Check that L is linear and that a is bilinear and symmetric.

(ii) From now on, let V be a finite dimensional Hilbert space with basis $\{\varphi_i\}_{i=1}^n$, $a : V \times V \to \mathbb{R}$ be symmetric and bilinear, and $L : V \to \mathbb{R}$ be linear L (general ones, not the ones given in (i)). Show that the problem: find $u \in V$ such that

$$a(u, v) = L(v), \quad \forall v \in V,$$

is equivalent to: find $u \in V$ such that

$$a(u, \varphi_j) = L(\varphi_j), \quad j = 1, \ldots, n.$$

Expand $u = \sum_{j=1}^n \beta_j \varphi_j$ and show that $\boldsymbol{\beta}$ is solution to: Find $\boldsymbol{\beta} \in \mathbb{R}^n$ such that

$$A\boldsymbol{\beta} = \boldsymbol{b},$$

where $A \in \mathbb{R}^{n\times n}, \boldsymbol{b} \in \mathbb{R}^n$ are such that $A_{ij} = a(\varphi_j, \varphi_i)$ and $b_i = L(\varphi_i)$ for $i, j \in \{1, \ldots, n\}$.

P22. (2p) Consider the problem: Find $u \in V$ such that

$$a(u, v) = L(v).$$

Show that there exists a unique solution when

(i) $V = \mathbb{R}^2$, $\|v\|_V = \left(v_1^2 + v_2^2\right)^{1/2}$, $a(u, v) = 2u_1v_1 + u_2v_1 + u_1v_2 + 2u_2v_2$ and $L(v) = v_1 + v_2$.

(ii) $V = H_0^1(0, 1)$, $\|v\|_V = \left(\|v'\|^2_{L^2(0,1)} + \|v\|^2_{L^2(0,1)}\right)^{1/2}$, $a(u, v) = \int_0^1 \sigma(t)u(t)'v(t)'\ dt$ and $L(v) = \int_0^1 v(t)'dt + \int_0^1 v(t)\ dt$. The coefficient function $\sigma \in L^\infty(0, 1)$ is bounded and positive a.e. in $(0, 1)$.

P23. (2p) Let V_0 and W_0 be as in P19, $f \in V_0$, and consider the weak problem: Find $u \in V_0$ such that

$$\int_0^s u'v'\ dt = \int_0^s fv \quad \forall v \in V_0. \tag{40}$$

(i) Show that there exists a unique solution to problem (40). Hint: use the inequality from P19 (ii).

(ii) Let $f \in W_0$, $u \in C^2(0,s) \cap C^1([0,s])$, $\int_0^s u(t)dt = 0$ satisfy

$$-u'' = f \quad \text{in } (0,s) \quad \text{and} \quad u'(0) = u'(s) = 0.$$

Show that u is a solution to problem (1).

7.1 Equivalence to minimization problem

Energy min. problem in Youtube

In this Section, we show that the variational problem (34) can be formulated as an energy minimisation problem if Assumptions 7.1 hold and $a(\cdot,\cdot)$ is in addition symmetric, i.e.

$$a(u,v) = a(v,u)$$

for all $u, v \in V$.

$$\|v\|_E^2 = a(v,v) \tag{41}$$

and an inner product for functions in V. This norm is called the energy norm of the variational problem, and it is equivalent to the norm of the Hilbert space V, i.e.,

$$c\|v\|_V \leq \|v\|_E \leq C\|v\|_V$$

for every $v \in V$ and some $c, C > 0$ independent of v. The close connection between energy norm and weak problem make it a good choise in proofs.

The solution to the variational problem (34) is equivalent to the following minimization problem: find $u \in V$ such that

$$J(u) = \min_{v \in V} J(v), \tag{42}$$

in which *the energy functions* $J : V \mapsto \mathbb{R}$ satisfies $J(v) = \frac{1}{2}a(v,v) - L(v)$. To show that (34) and (42) are equivalent, we look for the critical points of the energy functional. If $u \in V$ is the minimum, then for every $v \in V$

$$\frac{d}{d\epsilon} J(u + \epsilon v) \tag{43}$$

must vanish at $\epsilon = 0$. This yields

$$\begin{aligned}
0 =& \frac{d}{d\epsilon} J(u+\epsilon v)_{|\epsilon=0} \\
&= \Big[\frac{d}{d\epsilon}\Big(\frac{1}{2}a(u+\epsilon v, u+\epsilon v) - L(u+\epsilon v)\Big)\Big]_{|\epsilon=0} \\
&= \Big[\frac{1}{2}a(v, u+\epsilon v) + \frac{1}{2}a(u+\epsilon v, v) - L(v)\Big]_{|\epsilon=0} \\
&= \frac{1}{2}a(v,u) + \frac{1}{2}a(u,v) - L(v) \\
&= a(u,v) - L(v).
\end{aligned}$$

The converse is easy to see, if u solves the variational problem, then $J(u+v) > J(u)$ for any $v \in V$, $v \neq 0$.

7.2 Problems

P24. (1p) Let V be a Hilbert space, $a : V \times V \to \mathbb{R}$ be a symmetric, elliptic and continuous bilinear form and $L : V \to \mathbb{R}$ a continuous linear functional. In addition, let $J : V \to \mathbb{R}$ be such that

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

Show that J is bounded from below. Hint: at some point it is useful to denote $t = \|u\|_V$ and study a polynomial of t.

P25. (1p) Use he same notation and make the same assumptions as in P24. Let u be a solution to Problem $a(u, v) = L(v)$ and $\|\cdot\|_E^2 = a(\cdot, \cdot)$. Show that

(a) $J(u + v) > J(u)$ for any $v \in V$, $v \neq 0$.

(b) $\|u - v\|_E^2 = 2(J(v) - J(u))$ for any $v \in V$.

8 Error Analysis

In this section, we derive an upper bound for the $H^1(I)$-, $L^2(I)$-, and energy-norms of FE-error, $e = u - u_h$. For model problem (7), the error depends on the the source term f and the partition $\{x_j\}$, this is, $e = e(\{x_j\}, f)$. We are interested on the behavior of error norms when the mesh size h,

$$h := \max_{j \in \{1,\ldots,N-1\}} (x_{j+1} - x_j),$$

tends to zero for fixed f. Deriving such bounds is called as *convergence analysis*. We prove the error estimate:

$$\|u - u_h\|_{H^1(I)} \leq Ch\|u''\|_{L^2(I)} \leq Ch\|f\|_{L^2(I)} \tag{44}$$

for model problem (7). The error estimate (44) is valid when the exact solution $u \in H^2(I)$.

Overview The error estimate in (44) is derived by taking the following steps:

1. *Relate error to the approximation properties of the finite dimensional space*:

   Finite element method is a complicated process, and difficult to analyse directly. However, as a subspace method, FE- error is related to error of best possible approximation of the exact solution form FE-space. Hence, the error can be related to another process, that is easier to analyse. Particularly, we relate the error to the approximation of the exact solution from the finite element space by *nodal interpolation*.

2. *Study the approximation properties for the finite dimensional space*:

   We derive an upper bound for the FE-error in H^1-norm by studying how well the exact solution can be approximated by the FE-space. Deriving suitable results requires us to assume that $u \in H^2(I)$, or has two derivatives.

Higher regularity Before proceeding, we show that $u \in H_0^1(I)$ satisfying

$$\int_I u'v' \, dt = \int_I fv \, dt \quad \text{for } f \in L^2(I) \text{ and all } v \in H_0^1(I) \tag{45}$$

is in the space $H^2(I)$, i.e., $u \in H^2(I)$. Recall, that the function $u' \in L^2(I)$ is weakly differentiable with weak derivative w, if there exists $w \in L^1_{loc}(I)$ satisfying

$$\int_I wv \, dt = -\int_I u'v' \, dt \quad \text{for all } v \in C_0^\infty(I).$$

Using (45) and rearranging the terms gives

$$\int_I (w + f)v \, dt = 0 \quad \text{for all } v \in C_0^\infty(I).$$

Hence, $w = -f \in L^2(I)$ is the second weak derivative of u. Naturally, $\|u''\|_{L^2(I)} = \|f\|_{L^2(I)}$. This higher regularity of u is an important property in bounding the error.

8.1 Cea's Lemma

Let V be a Hilbert space with norm $\|\cdot\|_V$. In this section, we study the abstract variational problem: find $u \in V$ satisfying

$$a(u, v) = L(v) \quad \text{for all} \quad v \in V. \tag{46}$$

A subspace method approximates u from some finite dimensional sub-space V_h of V as: find $u_h \in V_h$ satisfying

Cea's Lemma in Youtube

$$a(u_h, v_h) = L(v_h) \quad \text{for all} \quad v_h \in V_h. \tag{47}$$

Under assumptions on a and L, the error $\|u - u_h\|_V$ satisfies Cea's Lemma:

Lemma 8.1 (Céa's Lemma)**.** *Let V be a Hilbert space with the norm $\|\cdot\|_V$, $V_h \subset V$ a subspace of V, and $u \in V$, $u_h \in V_h$ solutions to* (46) *and* (47)*, respectively. Assume that $a : V \times V \mapsto \mathbb{R}$ and $L : V \mapsto \mathbb{R}$ satisfy Assumption 7.1. Then*

$$\|u - u_h\|_V \leq \frac{C}{\alpha} \inf_{v_h \in V_h} \|u - v_h\|_V.$$

This result tells us that the error $u - u_h$ is comparable to the error of the *best approximation* of u from V_h.

Proof. Using ellipticity of a (36), gives

$$\|u - u_h\|_V^2 \leq \frac{1}{\alpha} a(u - u_h, u - u_h).$$

The proof follows from *Galerkin orthogonality*-property: As $V_h \subset V$, using (46) and (47) gives

$$a(u - u_h, v_h) = a(u, v_h) - a(u_h, v_h) = L(v_h) - L(v_h) = 0 \tag{48}$$

for any $v_h \in V_h$. Using Galerking orthogonality (48) and continuity (35) we get

$$\begin{aligned} \|u - u_h\|_V^2 &\leq \frac{1}{\alpha} a(u - u_h, u - u_h) \\ &= \frac{1}{\alpha} a(u - u_h, u - v_h) \\ &\leq \frac{C}{\alpha} \|u - u_h\|_V \|u - v_h\|_V \end{aligned}$$

for any v_h. Dividing by $\|u - u_h\|_V$ completes the proof. □

If the bilinear form a satisfies Assumptions 7.1 and is symmetric, i.e., $a(u, v) = a(v, u)$ for all $u, v \in V$, a is an inner produt on V. The induced norm $\|v\|_E := (a(v, v))^{1/2}$ is called as the energy norm. By a small modification to the proof of Cea's Lemma 8.1, it is easy to show that

Symmetry modification in Youtube

$$\|u - u_h\|_E \leq \inf_{v_h \in V_h} \|u - v_h\|_E.$$

This is, the finite element solution is the best approximation in the energy norm.

8.1.1 Problems

P26. (0.5p) Let V be a Hilbert space with norm $\|\cdot\|_V$, and bilinear form $a : V \times V \mapsto \mathbb{R}$ satisfy Assumptions 7.1. In addition assume that a is symmetric. Show that

(a) Bilinear form a is an inner product on V.

(b) $a(u - v, u + v) = \|u\|_E^2 - \|v\|_E^2$.

P27. (0.5p) Let linear functional $L : H_0^1(0, 1) \mapsto \mathbb{R}$ satisfy

$$L(v) := \int_0^1 f'v' \, dt + \int_0^1 gv \, dt \quad \text{for all } v \in H_0^1(0, 1)$$

Show that L satisfies Assumption 7.1 for $V = H_0^1(0, 1)$.

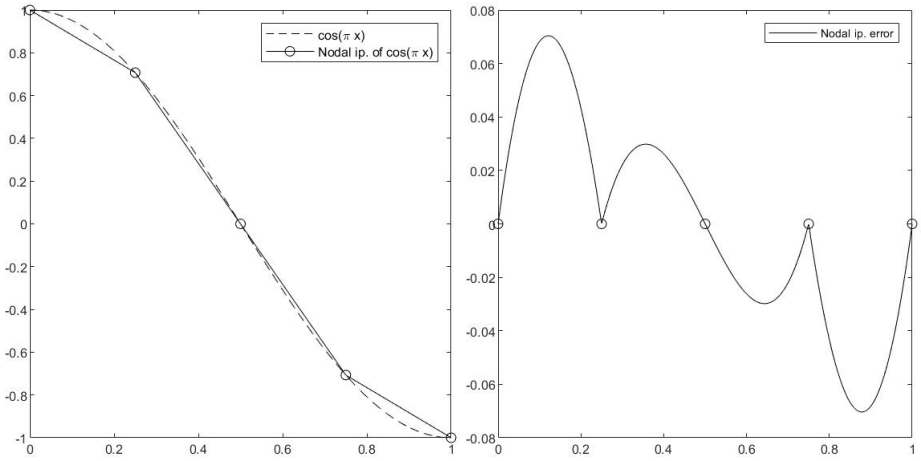

Figure 11: From left: example of nodal interpolation and the related interpolation error.

8.2 Interpolation

Let $u \in H^1(I)$. The topic of this section is to define the *nodal interpolant* $\pi u \in V_h$ of the function u and to estimate the *interpolation errors*:

Nodal Interpolation in Youtube

$$\|u - \pi u\|_{L^2(I)} \quad \text{and} \quad \|(u - \pi u)'\|_{L^2(I)}.$$

Let $\{x_i\}_{i=1}^N$ be a partition of I, $I_j = (x_j, x_{j+1})$ for $j \in \{1, \ldots, N-1\}$, and

$$\widehat{V}_h = \{v \in H^1(I) \,|\, v \in P^1(I_j) \quad \text{for } j \in \{1, \ldots, N-1\}\}. \tag{49}$$

Definition 8.1. *Let $\widehat{V}_h$ be as defined in* (49)*. The nodal interpolation operator $\pi : H^1(I) \mapsto \widehat{V}_h$ is defined as*

$$\pi v \in \widehat{V}_h \quad \textit{and} \quad (\pi v)(x_j) = v(x_j), \quad \forall j = 1, ..., N.$$

Recall, that $H^1(I) \subset C(\bar{I})$, hence, the operator π is well defined. Example of the nodal interpolant is given in Fig. 11. Choosing $V = H_0^1(I)$, $\|\cdot\|_V := \|\cdot\|_{H^1(I)}$, and combining the nodal interpolation with Céa's Lemma 8.1 gives the estimate

$$\|u - u_h\|_1 \leq \frac{C}{\alpha} \|u - \pi u\|_1 \tag{50}$$

for error between the exact and FE-solution to (7), respectively. Observe, that estimate (50) holds also for other PDEs satisfying Assumptions 7.1. Hence, by bounding $\|u - \pi u\|_1$, we obtain error estimates for several different PDEs.

We proceed to study approximation properties of the nodal interpolation operator. First, we consider single element $\hat{I} = (0, 1)$. Then, the estimates on $\hat{I}$ are extended to element (x_j, x_{j+1}) by using the *scaling argument*, a technique for extracting explicit geometry dependency of constants.

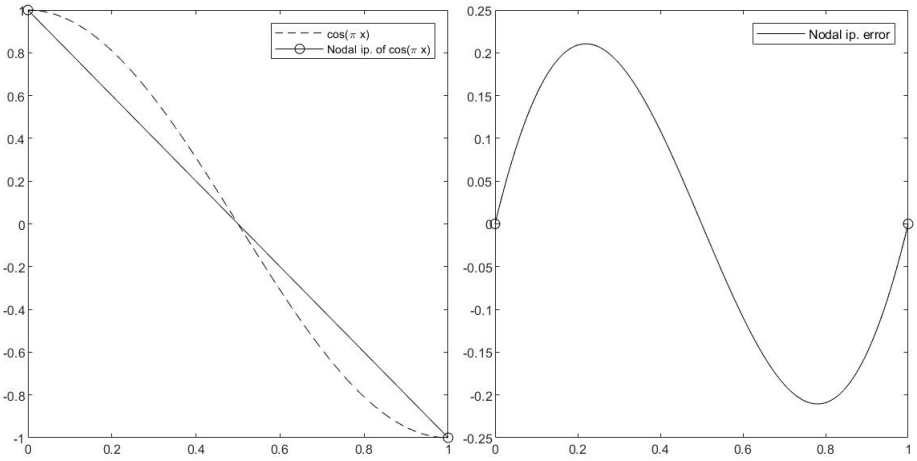

Figure 12: From left: example of nodal interpolation and the related interpolation error over the single element $(0, 1)$.

Let $\hat{u} \in H^1(0,1)$ and $\hat{\pi}\hat{u}$ satisfy

$$\hat{\pi}\hat{u} \in P^1(0,1), \quad (\hat{\pi}\hat{u})(0) = \hat{u}(0), \quad \text{and} \quad (\hat{\pi}\hat{u})(1) = \hat{u}(1). \tag{51}$$

Clearly $\hat{e}(0) = \hat{e}(1) = 0$, see Figure 12. We proceed by estimating the $L^\infty(\hat{I})$-norm of the interpolation error $\hat{e} = \hat{u} - \hat{\pi}\hat{u}$.

Theorem 8.1. *Let $\hat{u} \in H^2(0,1)$, $\hat{\pi}\hat{u} \in P^1(0,1)$ satisfy* (51)*, and $\hat{e} = \hat{u} - \hat{\pi}\hat{u}$. Then there holds that*

$$\sup_{t\in(0,1)} |\hat{e}(t)| \leq C\|\hat{u}''\|_{L^2(0,1)}$$

for some C independent of $\hat{u}$.

Proof. Let $\hat{u} \in C^\infty(\hat{I})$. As $\hat{e}(0) = \hat{e}(1) = 0$, the supremum of $|\hat{e}(t)|$ has to lie at some zero of $\hat{e}'$. Let $s \in (0,1)$ satisfy

$$\sup_{t\in(0,1)} |\hat{e}(t)| = \hat{e}(s) \quad \text{so that} \quad \hat{e}'(s) = 0.$$

By the fundamental theorem of analysis and property $\hat{e}(0) = 0$,

$$\hat{e}(s) = \int_0^s \hat{e}'(t)\, dt.$$

Integration by parts gives

$$\hat{e}(s) = s\hat{e}'(s) - \int_0^s t\hat{e}''(t)\, dt.$$

As s is extremal point of $\hat{e}$, $\hat{e}'(s) = 0$. Using the Cauchy-Schwartzh inequality gives

$$|\hat{e}(s)| \leq \left(\int_0^s t^2\, dt\right)^{1/2} \left(\int_0^s |\hat{e}''(t)|^2\, dt\right)^{1/2}.$$

Observe that $\hat{e}''(t) = \hat{u}''(t)$. Expanding the range of integration and evaluating the first term gives

$$|\hat{e}(s)| \leq \frac{1}{\sqrt{3}} \left(\int_0^1 |\hat{u}''(t)|^2 \, dt \right)^{1/2}.$$

□

The result given in Theorem 8.1 leads to estimates in the $L^2(0,1)$-norm and the $H^1(0,1)$ semi-norm.

Corollary 8.1. *Use same notation and make same assumptions as in Theorem 8.1. In addition, let $\hat{I} = (0,1)$. Then there holds that*

$$\|\hat{e}\|_{L^2(\hat{I})} \leq C_1 \|\hat{u}''\|_{L^2(\hat{I})} \tag{52}$$

$$\|\hat{e}'\|_{L^2(\hat{I})} \leq C_2 \|\hat{u}''\|_{L^2(\hat{I})}. \tag{53}$$

for some constants C_1, C_2 independent of $\hat{u}$.

Proof. Problem P28 □

Scaling argument Next, we estimate $\|(u - \pi u)'\|_{L^2(I)}$. Identical techniques are used to estimate $\|u - \pi u\|_{L^2(I)}$. First, the interpolation error is written as the sum of element-wise interpolation errors,

$$\|(u - \pi u)'\|^2_{L^2(I)} = \sum_{j=1}^{N-1} \|(u - \pi u)'|_{I_j}\|^2_{L^2(I_j)} \tag{54}$$

Each element-wise interpolation error term $\|(u - \pi u)'|_{I_j}\|^2_{L^2(I_j)}$ is estimated using Corollary 8.1 and scaling argument:

Scaling argument in Youtube

Lemma 8.2. *Let $k \in \{0, 1, \ldots\}$, $I = (a, b)$, $h_I := (b - a)$, and $r : (0,1) \mapsto (a,b)$ be defined as $r(\hat{t}) := (b-a)\hat{t} + a$. In addition, let $v \in H^k(I)$ and $\hat{v} \in H^k(0,1)$ be defined as $\hat{v}(\hat{t}) := v(r(\hat{t}))$. Then there holds that*

$$\left\| \frac{d^{(k)} v}{dt^{(k)}} \right\|_{L^2(I)} = h_I^{(1-2k)/2} \left\| \frac{d^{(k)} \hat{v}}{d\hat{t}^{(k)}} \right\|_{L^2(\hat{I})}. \tag{55}$$

Here $\frac{d^{(0)} v}{dt^{(0)}} = v$ and $\frac{d^{(0)} \hat{v}}{d\hat{t}^{(0)}} = \hat{v}$.

Proof. The proof follows by the chain rule and change of variables in integration. First, we make a change of variables $t = r(\hat{t})$.

$$\left\| \frac{d^{(k)} v}{dt^{(k)}} \right\|^2_{L^2(I)} = \int_a^b \left[\left(\frac{d^{(k)}}{dt^{(k)}} v \right)(t) \right]^2 dt = h_I \int_0^1 \left[\left(\frac{d^{(k)}}{dt^{(k)}} v \right)(r(\hat{t})) \right]^2 d\hat{t}$$

Using the chain rule gives

$$\frac{d^{(k)}}{d\hat{t}^{(k)}}\hat{v}(\hat{t}) = \frac{d^{(k)}}{d\hat{t}^{(k)}}\left[v(r(\hat{t}))\right] = h_I^k\left(\frac{d^{(k)}}{dt^{(k)}}v\right)(r(\hat{t})). \tag{56}$$

Hence,

$$h_I\int_0^1\left[\left(\frac{d^{(k)}}{dt^{(k)}}v\right)(r(\hat{t}))\right]^2 d\hat{t} = h_I^{1-2k}\int_0^1\left[\frac{d^{(k)}}{d\hat{t}^{(k)}}\hat{v}(\hat{t})\right]^2 h_I\, d\hat{t}$$

Which completes the proof. □

Theorem 8.2. *Let $u \in H^2(I)$ and $\pi : H^1(I) \mapsto \widehat{V}_h$ be as in Def. 8.1. Then the interpolation error satisfies*

Interpolation error estimate in Youtube

$$\|(u-\pi u)'\|_{L^2(I)} \le Ch\|u''\|_{L^2(I)}. \tag{57}$$

and

$$\|u-\pi u\|_{L^2(I)} \le Ch^2\|u''\|_{L^2(I)}. \tag{58}$$

Proof. We give a proof for the first claim. Error estimate in the $L^2(I)$-norm is left as homework problem. Let $r_j(\hat{t}) : (0,1) \mapsto I_j$ be defined as

$$r_j(\hat{t}) = (x_{j+1}-x_j)\hat{t} + x_j.$$

for $j \in \{1,\dots,N-1\}$. In addition, let $w_j \in H^1(I_j)$ and $\hat{w}_j \in H^1(0,1)$ be defined as $w_j = u|_{I_j} - (\pi u)|_{I_j}$ and $\hat{w}_j(\hat{t}) = w_j(r_j(\hat{t}))$. By scaling argument, Lemma 8.2, the element-wise interpolation error satisfies

$$\|(u-\pi u)'|_{I_j}\|^2_{L^2(I_j)} = \|w_j'\|^2_{L^2(I_j)} = h_{I_j}^{-1}\|\hat{w}_j'\|^2_{L^2(0,1)}.$$

The interpolation error estimate in Corollary 8.1 gives

$$\|\hat{w}_j'\|^2_{L^2(0,1)} \le C\|\hat{w}_j''\|^2_{L^2(0,1)}.$$

Using the scaling argument in Lemma 8.2 gives

$$\|\hat{w}_j''\|^2_{L^2(0,1)} \le h_{I_j}^3\|w_j''\|^2_{L^2(I_j)}.$$

Combining above equations gives

$$\|w_j'\|^2_{L^2(I_j)} = h_{I_j}^{-1}\|\hat{w}_j'\|^2_{L^2(0,1)} \le Ch_{I_j}^{-1}\|\hat{w}_j''\|^2_{L^2(0,1)} \le h_{I_j}^2\|w_j''\|^2_{L^2(I_j)}.$$

Using the Eq.(54) and estimating $h_{I_j} \le \max_{j\in\{1,\dots,N-1\}} h_{I_j} = h$ completes the proof. □

The idea of the above proof is simple: instead of using a given interpolation inequality on interval (x_j, x_{j+1}), scaling argument is used to transform the interpolation error norm to the reference element $(0, 1)$. Interpolation error estimate is derived on $(0, 1)$ and applied to the transformed error term. The last step is to map the result back to the actual element (x_j, x_{j+1}). Combining Cea's Lemma and Interpolation error theorem gives our final convergence estimate

$H^1(I)$ estimate in Youtube

Theorem 8.3. *Let $\{x_j\}_{j=1}^N$ be a partition of I, $V_h \subset H_0^1(I)$ the piecewise linear FE-space associated to $\{x_j\}$, and $u \in H_0^1(I)$, $u_h \in V_h$ solutions to* (61) *and* (62)*, respectively. Let $a : V \times V \mapsto \mathbb{R}$ and $L : V \mapsto \mathbb{R}$ satisfy Assumption 7.1. In addition, assume that $u \in H^2(I)$. Then*

$$\|u - u_h\|_{H^2(I)} \leq Ch\|u''\|_{L^2(I)}$$

where the mesh size $h = \max_{j\in\{1,\ldots,N-1\}}(x_{j+1} - x_j)$.

Proof. The proof follows by combining Cea's Lemma and interpolation error estimate. □

8.3 Problems

P28. (2p) Prove Corollary 8.1 Hint: Use Theorem 8.1. The second inequality requires you to use the integration by parts formula

$$\int_a^b e'(t)e'(t)\ dt = -\int_a^b e(t)e''(t)\ dt,$$

when $e(a) = e(b) = 0$.

P29. (2p) Show that

$$\|\hat{p}'\|_{L^2(0,1)} \leq \widehat{C}\|\hat{p}\|_{L^2(0,1)} \quad \text{for all } \hat{p} \in P^1(0, 1),$$

where constant $\widehat{C}$ is independent of $\hat{p}$. Proceed as follows:

(a) Let $\{\varphi_1, \varphi_2\}$ be a basis of $P^1(0, 1)$ and $A, M \in \mathbb{R}^{2\times 2}$ have entries $A_{ij} = \int_0^1 \varphi_i'\varphi_j'$ and $M_{ij} = \int_0^1 \varphi_i\varphi_j$ for $i, j \in \{1, 2\}$. Show that $\boldsymbol{x}^T A\boldsymbol{x} = \|p'\|^2_{L^2(0,1)}$ and $\boldsymbol{x}^T M\boldsymbol{x} = \|p\|^2_{L^2(0,1)}$ for $p = x_1\varphi_1 + x_2\varphi_2$ and further that

$$\widehat{C}^2 = \max_{\boldsymbol{x}\in\mathbb{R}^2} \frac{\boldsymbol{x}^T A\boldsymbol{x}}{\boldsymbol{x}^T M\boldsymbol{x}}$$

(b) Show that A, M are symmetric matrices. In addition, show that A is positive semi-definite and M is positive definite matrix, i.e.

$$\boldsymbol{x}^T A\boldsymbol{x} \geq 0 \quad \text{and} \quad \boldsymbol{x}^T M\boldsymbol{x} > 0 \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^2 \setminus \{0\}.$$

(c) Let $B \in \mathbb{R}^{2\times 2}$ be symmetric. Show that

$$\lambda_{min}(B)\boldsymbol{x}^T\boldsymbol{x} \leq \boldsymbol{x}^T B\boldsymbol{x} \leq \lambda_{max}(B)\boldsymbol{x}^T\boldsymbol{x}$$

for all $\boldsymbol{x} \in \mathbb{R}^2$. Here $\lambda_{min}(B)$ and $\lambda_{max}(B)$ are the smallest and largest eigenvalues of B. Further, derive the estimate

$$\widehat{C}^2 \leq \lambda_{min}(M)\lambda_{max}(A)$$

and argue that $\lambda_{min}(M), \lambda_{max}(A) > 0$.

P30. (2p) Let V_h be the first order FE-space related to partition $\{x_j\}_{j=1}^N$ of I. Assume there exists ρ independent of N and h such that

$$\min_{i\in\{1,\ldots,N-1\}}(x_{i+1} - x_i) \geq \rho h.$$

Prove *the inverse inequality*: there exists a constant C dependent on ρ but independent of v_h and h such that

$$\|v_h'\|_{L^2(I)} \leq Ch^{-1}\|v_h\|_{L^2(I)} \quad \text{for all } v_h \in V_h. \tag{59}$$

Use the scaling argument and result of Problem P29.

P31. (2p) Let $f \in L^2(0,2)$ and $\sigma : (0,2) \mapsto \mathbb{R}$ be defined as $\sigma := \begin{cases} 1 & t \in (0,1] \\ 2 & t \in (1,2) \end{cases}$.

Consider the problem: find $u \in H_0^1(0,2)$ such that

$$\int_0^2 \sigma u'v'\, dt = \int_0^2 fv\, dt \quad \text{for all } v \in H_0^1(0,2). \tag{60}$$

(a) Formulate (60) as an abstract variational problem: find $u \in V$ s.t. $a(u,v) := L(v)$ for all $v \in V$. Show that a and L satisfy Assumptions 7.1.

(b) Show that $u|_{(0,1)} \in H^2(0,1)$, $u|_{(1,2)} \in H^2(1,2)$, $\|(u|_{(0,1)})''\|_{L^2(0,1)} = \|f|_{(0,1)}\|_{L^2(0,1)}$, and $\|(u|_{(1,2)})''\|_{L^2(1,2)} = \frac{1}{2}\|f|_{(1,2)}\|_{L^2(1,2)}$

(c) Consider solving the problem (60) with first order FE-method using a partition $\{x_j\}_{j=1}^N$ having node at 1. Show that the FE-solution satisfies the error estimate $\|u - u_h\|_E \leq Ch\|f\|_{L^2(I)}$.

8.4 L^2-error estimate

So far we have derived FE-error estimate in the $H^1(I)$-norm. The same theory naturally applies also in the energy norm. In this section we derive an estimate for the FE-error in the L^2-norm using the so-called Aubin-Nitsche trick or duality argument.

$L^2(I)$ estimate in Youtube

Consider the abstract variational problem : find $u \in H_0^1(I)$ satisfying

$$a(u, v) = L(v) \quad \text{for all } v \in H_0^1(I), \tag{61}$$

and it's FE-approximation: find $u_h \in V_h$ satisfying

$$a(u_h, v_h) = L(v_h) \quad \text{for all } v_h \in V_h. \tag{62}$$

Here V_h is the first order FE-space. Make Assumptions 7.1 on a and L. In addition, assume that a is symmetric,

$$L(v) = \int_I fv \quad \text{for some } f \in L^2(I),$$

and that the solution u to (61) satisfies $u \in H^2(I)$ and $\|u''\|_{L^2(I)} \leq \|f\|_{L^2(I)}$. In other words, the $H^2(I)$ - norm of the solution is bounded by the L^2-norm of the source function $f \in L^2(I)$.

We arrive to the theorem :

Theorem 8.4. *Let $\{x_j\}_{j=1}^N$ be a partition of I, $V_h \subset H_0^1(I)$ the piecewise linear FE-space associated to $\{x_j\}$, and $u \in H_0^1(I)$, $u_h \in V_h$ solutions to* (61) *and* (62)*, respectively. Assume that $a : V \times V \mapsto \mathbb{R}$ is symmetric and satisfies Assumption 7.1, and $L : V \mapsto \mathbb{R}$ is such that $L(v) = \int_I fv$ for some $v \in L^2(I)$. In addition, assume that $u \in H^2(I)$ and $\|u''\|_{L^2(I)} \leq C\|f\|_{L^2(I)}$. Then*

$$\|u - u_h\|_{L^2(I)} \leq Ch^2\|f\|_{L^2(I)}.$$

Proof. Recall the Galerkin orthogonality property:

$$a((u - u_h), v_h) = 0 \quad \text{for all } v_h \in V_h.$$

Observe that $e := u - u_h \in H_0^1(I) \subset L^2(I)$. Hence we can set e as the load functional and study the *dual problem*: find $w \in H_0^1(\Omega)$ such that

$$a(w, v) = \int_I ev \, dt \quad \forall v \in H_0^1(\Omega). \tag{63}$$

By assumptions, $w \in H^2(I)$ and $\|w''\|_{L^2(I)} \leq C\|e\|_{L^2(I)}$. Choose $v = u - u_h \in H_0^1(\Omega)$ in (63) to get

$$\|u - u_h\|_{L^2(I)}^2 = \int_I e(u - u_h) \, dt = a(w, u - u_h).$$

Using symmetry and Galerkin orthogonality, we can insert the interpolation of w

$$\|u - u_h\|_{L^2(I)}^2 = a(w - \pi w, u - u_h).$$

Using continuity (35) yields

$$\|u - u_h\|_{L^2(I)}^2 \leq \|w - \pi w\|_{H^1(I)}\|u - u_h\|_{H^1(I)}.$$

Using the interpolation result for the first part and the usual finite element error estimate for the second part gives

$$\|u - u_h\|^2_{L^2(I)} \leq Ch^2 |w''|_{L^2(I)} \, |u''|_{L(I)}.$$

Using $|w''|_{L^2(I)} \leq C\|e\|_{L^2(I)}$ completes the proof. □

.

Above we discovered that the $L^2(I)$-error estimate has $\mathcal{O}(h^2)$ convergence. Previously we found out that the H^1-error estimate has $\mathcal{O}(h)$ convergence. If the problem is regular enough and the elements span higher order polynomials, then the rule of thumb is that the L^2-error estimate has one more power of h in convergence compared to the H^1-estimate. For example, for second order polynomials it holds that $\|u - u_h\|_1 \leq Ch^2|u|_3$ and $\|u - u_h\|_0 \leq Ch^3|u|_3$.

9 Finite element method in two dimensions

Let $\Omega \subset \mathbb{R}^2$, source term $f \in L^2(\Omega)$, and material coefficient $K \in \mathbb{R}^{2\times 2}$ be symmetric and positive definite matrix, i.e., Introduction to week 5 in Youtube

$$\lambda_{max} \boldsymbol{\xi}^T \boldsymbol{\xi} \geq \boldsymbol{\xi}^T K \boldsymbol{\xi} \geq \lambda_{min} \boldsymbol{\xi}^T \boldsymbol{\xi}$$

for all $\boldsymbol{\xi} \in \mathbb{R}^2$ and $\lambda_{max}, \lambda_{min} > 0$ independent of $\boldsymbol{\xi}$. In this section, we apply finite element method to solve the model problem: Find $u \in C^2(\Omega) \cap C(\overline{\Omega})$ satisfying

$$\begin{aligned} -\nabla \cdot K \nabla u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{64}$$

Sobolev spaces in higher dimensions Before applying finite element method, PDE (64) is reformulated as a weak problem posed in an appropriate Sobolev space. We start by defining the necessary spaces in dimension $d = 1, 2, 3$. We make the following simplifying assumptions on Ω:

Assumption 9.1. *Assume that $\Omega \subset \mathbb{R}^2$ is a simply connected polyhedral domain with finite number of boundary segments.*

This is, the boundary $\partial\Omega$ is a polygon and Ω is an open, bounded, and connected set that does not have any holes, see Fig. 13. In the following, all derivatives are interpreted in the weak sense.

Definition 9.1. *Let $u \in L^1_{loc}(\Omega)$. The function u is weakly differentiable, if there exists $w_i \in L^1_{loc}(\Omega)$ for $i \in \{1, \ldots, d\}$ such that*

$$(w_i, \varphi)_{L^2(\Omega)} = -(u, \partial_i \varphi)_{L^2(\Omega)} \quad \textit{for all } \varphi \in C_0^\infty(\Omega).$$

We call w_i as the weak ith partial derivative of u, and write $\partial_i u = w_i$.

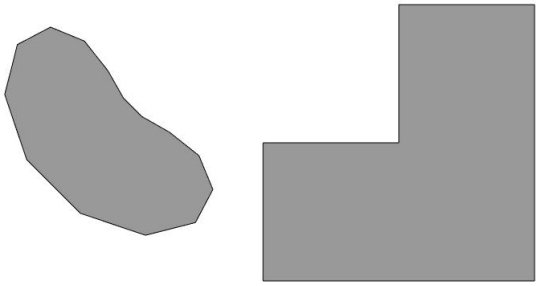

Figure 13: Examples of domains satisfying assumptions 9.1.

Higher weak partial derivatives are obtained by iterating the above definition. When defining Sobolev spaces we impose conditions on partial derivatives of order m or less that are easiest to express using a *multi-index*.

Definition 9.2. *We call vectors $\alpha \in \{0, 1, 2 \ldots, \}^d$ as multi-indices, and denote*

$$\partial_\alpha := \partial_{x_1^{\alpha_1}, \ldots, x_d^{\alpha_d}}.$$

In addition, we use the notation $|\alpha| := \alpha_1 + \ldots + \alpha_d$.

For example, we write $\partial_\alpha u \in L^2(\Omega)$ for all $|\alpha| = m$ to require that all partial derivatives of order m lie in the space $L^2(\Omega)$. The Sobolve space $H^m(\Omega)$ is defined as

Definition 9.3. *Let Ω and $u \in L^1_{loc}(\Omega)$. Then $u \in H^m(\Omega)$, if u is m-times weakly differentiable and $\partial_\alpha \in L^2(\Omega)$ for all $|\alpha| \leq m$.*

Definition 9.4. *The $H^m(\Omega)$ inner product and the induced norm are*

$$(u, v)_{H^m(\Omega)} := \sum_{|\boldsymbol{\alpha}| \leq m} (\partial_\alpha u, \partial_\alpha v)_{L^2(\Omega)}$$

and

$$\|u\|_{H^m(\Omega)} := (u, u)^{1/2}_{H^m(\Omega)} = \left(\sum_{|\alpha| \leq m} \|\partial_\alpha u\|^2_{L^2(\Omega)} \right)^{1/2}$$

Spaces $H^1(\Omega)$ and $H^2(\Omega)$ are enough for most of our purposes.

Space $H^1_0(\Omega)$ Next, we consider zero boundary conditions in more detail. The space $H^1(\Omega)$ is *not a subset of* $C(\overline{\Omega})$ for $d = 2, 3$. Further, if $d = 2, 3$, $H^1(\Omega)$-functions that have different values only on the boundary $\partial\Omega$ are considered equivalent, hence the restrictions of $H^1(\Omega)$-functions to $\partial\Omega$ has essentially no meaning.

$H^1_0(\Omega)$ space in Youtube

Zero boundary conditions are imposed on functions in $H^1(\Omega)$ using the *trace*-operator $T : H^1(\Omega) \to L^2(\partial\Omega)$. Here $L^2(\partial\Omega)$ is the space of square integrable functions over $\partial\Omega$ with the inner product and the induced norm

$$(u, v)_{L^2(\partial\Omega)} := \int_{\partial\Omega} uv \quad \text{and} \quad \|v\|_{L^2(\partial\Omega)} := (v, v)^{1/2}_{L^2(\partial\Omega)} = \left(\int_{\partial\Omega} v^2 \right)^{1/2}.$$

The trace operator is an continuous extension of the classical restriction operator, this is, operator T satisfies

$$Tv = v_{|\partial\Omega} \quad \forall v \in C^\infty(\Omega).$$

Let $v \in H^1(\Omega)$. Since the space $C^\infty(\Omega)$ is dense in the space $H^1(\Omega)$ we define Tv as the limit

$$Tv = \lim_{n\to\infty} v_{n|\partial\Omega}, \tag{65}$$

in which $\{v_n\}_{n=1}^\infty$ is a sequence in $C^\infty(\Omega)$ satisfying $v_n \to v$ in $H^1(\Omega)$. The trace operator satisfies the following trace theorem:

Theorem 9.1. *There holds that*

$$\|u\|_{L^2(\partial\Omega)} \leq C_{\partial\Omega} \|u\|_{H^1(\Omega)}$$

for any $u \in H^1(\Omega)$ and $C_{\partial\Omega}$ independent of u.

Trace theorem is important in treatment of Neumann and Robin boundary conditions. The space $H^1_0(\Omega)$ is defined follows:

$$H^1_0(\Omega) := \{\ u \in H^1(\Omega) \mid Tu = 0\ \}. \tag{66}$$

Usually, T is not explicitly written in (66), but the condition $u = 0$ is interpreted in the sense of traces, i.e., as $Tu = 0$. Functions in $H^1_0(\Omega)$ satisfy the Poincare-Friedrichs inequality.

Theorem 9.2. *There holds that*

$$\|v\|_{L^2(\Omega)} \leq C \|\nabla v\|_{L^2(\Omega)}.$$

for any $v \in H^1_0(\Omega)$ and a positive constant C independent of v.

Proof. See Theorem 6.1 for proof in $d = 1$. □

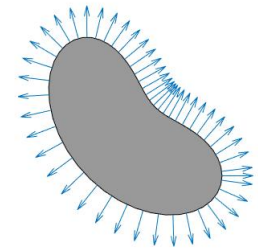

Figure 14: Example of exterior unit normal of Ω

9.1 Weak problem

Weak form in Youtube

We proceed to derive the weak form of Problem (64). First, a suitable *intergration by parts formula* is established. Let $\Omega \subset \mathbb{R}^2$ satisfy Assumptions 9.1 and $\boldsymbol{n} : \partial\Omega \mapsto \mathbb{R}^2$ be the exterior unit normal of Ω, see Fig. 14. Recall the *Gauss-Divergence* theorem:

$$\int_\Omega \nabla \cdot \boldsymbol{G} = \int_{\partial\Omega} \boldsymbol{G} \cdot \boldsymbol{n} \quad \text{for any } \boldsymbol{G} \in [C^1(\Omega)]^2. \tag{67}$$

Lemma 9.1. *Let $\boldsymbol{F} \in [C^1(\Omega)]^2$ and $\boldsymbol{n} : \partial\Omega \mapsto \mathbb{R}^2$ be the external unit normal of Ω. Then there holds that*

$$\int_\Omega \varphi \nabla \cdot \boldsymbol{F} = -\int_\Omega \boldsymbol{F} \cdot \nabla\varphi + \int_{\partial\Omega} \varphi \boldsymbol{F} \cdot \boldsymbol{n}$$

Proof. The result follows by choosing $\boldsymbol{G} = \boldsymbol{F}\varphi$ in Gauss-Divergence Theorem (67) and using the identity:

$$\nabla \cdot (\boldsymbol{F}\varphi) = \varphi \nabla \cdot \boldsymbol{F} + \boldsymbol{F} \cdot \nabla\varphi.$$

□

Multiplying first equation in (64) by test function $\varphi \in C_0^\infty(\Omega)$, integrating over Ω, using integration-by-parts formula in Lemma 9.1 with $\boldsymbol{F} = K\nabla u$, and density yields the weak-problem: find $u \in H_0^1(\Omega)$ satisfying

$$\int_\Omega K\nabla u \cdot \nabla v = \int_\Omega fv \quad \text{for all } v \in H_0^1(\Omega). \tag{68}$$

Problem (68) is posed in the space $H_0^1(\Omega)$ as this is the largest possible space where both sides of the above equation are guaranteed to have finite values. It is an instance of the abstract variational problem studied in Section 7. Hence, the existence of a unique solution to (68) follows by showing that Assumptions 7.1 are satisfied and using Lax-Milgram Lemma 7.1, see P32. In the following, we denote

$$a(u,v) := \int_\Omega K\nabla u \cdot \nabla v \quad \text{and} \quad L(v) := \int_\Omega fv,$$

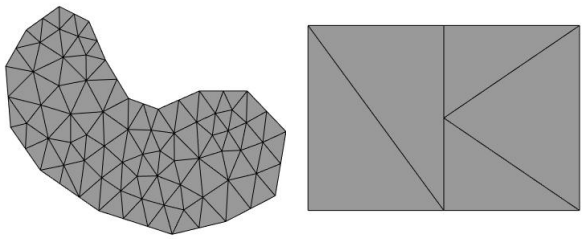

Figure 15: On left an example of a conforming triangular partition. On right, an example of a non-conforming triangular partition that has a hanging node.

so that (68) becomes: find $u \in H_0^1(\Omega)$ satisfying $a(u,v) = L(v)$ for all $v \in H_0^1(\Omega)$.

9.2 Problems

P32. Show that the weak form (68) has a unique solution. Hint: Apply Lax-Milgram Lemma 7.1 and Poincare-Friedrichs inequality in Theorem 9.2. Recall, that K is symmetric and positive definite matrix.

9.3 Piecewise linear FE-space

FEM finds an approximate solution to the weak problem (68) from a finite dimensional function space V_h. We use the space of piecewise linear continuous functions over a conforming triangular partition of Ω, see Figure 16.

FE-space in Youtube

Definition 9.5. *A conforming triangular partition of Ω is a set of closed triangular subdomains, $\mathcal{T} = \{T_1, \ldots, T_M\}$, that satisfy:*

1. $\overline{\Omega} = \cup_{i=1}^{M} T_i$
2. *The intersection $T_i \cap T_j$ for $i \neq j$ is either empty, a common vertex, or a common edge of T_i and T_j*

Examples of triangular partitions are depicted in Fig. 15. Let Ω satisfy Assumptions 9.1 and $\mathcal{T}_h$ be a conforming triangular partition of Ω. The space of piecewise-linear-continuous functions over $\mathcal{T}_h$ is defined as

$$\widehat{V}_h := \{\ v \in C(\overline{\Omega}) \mid v|_T \in P^1(T) \text{ for all } T \in \mathcal{T}_h\ \}, \tag{69}$$

and it's subspace with imposed zero Dirichlet boundary condition as

$$V_h := \{\ v \in \widehat{V}_h \mid v = 0 \text{ on } \partial\Omega\ \}. \tag{70}$$

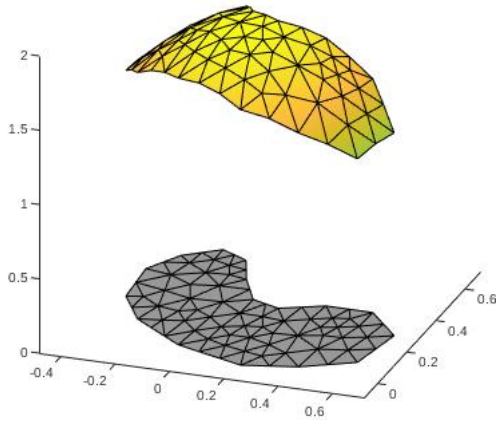

Figure 16: Example of a piecewise linear continuous function on a triangulation. The triangulation is depicted in gray.

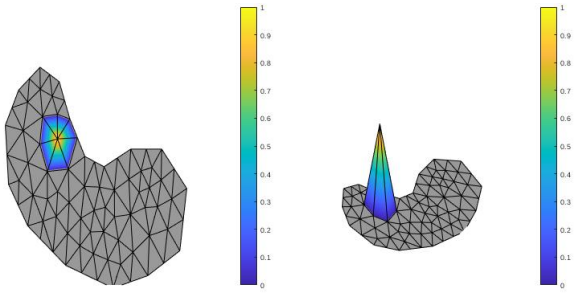

Figure 17: Example of a hat basisfunction

Hat basis-functions Let $\{\boldsymbol{n}_i\}_{i=1}^{\hat{n}} \subset \mathbb{R}^2$ be the vertices of conforming triangular partition of $\mathcal{T}_h$ and $\widehat{V}_h$ as defined in (69). The hat basis of $\widehat{V}_h$, $\{\hat{\varphi}_j\}_{j=1}^{\hat{n}}$, is defined as

$$\hat{\varphi}_j \in \widehat{V}_h \quad \text{and} \quad \hat{\varphi}_j(\boldsymbol{n}_i) = \begin{cases} 1 & j = i \\ 0 & \text{otherwise} \end{cases} \qquad \text{for } i, j \in \{1, \ldots, \hat{n}\}.$$

Examples of hat basis functions are given in Figure 17. A basis for V_h is obtained from $\{\hat{\varphi}_j\}_{j=1}^{\hat{n}}$ by omitting all basisfunctions associated to boundary vertices, i.e., those $\boldsymbol{n}_i$ that are part of $\partial\Omega$. The hat basis functions are evaluated using *reference basis functions*. This is motivated by the assembly process and discussed later.

Assembly Recall that FEM finds an approximate solution to the weak problem (68) from the subspace $V_h \subset H_0^1(\Omega)$ by solving: find $u_h \in V_h$ satisfying

$$a(u_h, v) = L(v) \quad \text{for all } v \in V_h. \tag{71}$$

As V_h is finite dimensional, Problem (71) is equivalent to: find $u_h \in V_h$ satisfying $a(u_h, \varphi_i) = L(\varphi_i)$ for all $i \in 1, \ldots, n$, and further to the linear system: find $\boldsymbol{\beta} \in \mathbb{R}^n$ satisfying

$$A\boldsymbol{\beta} = \boldsymbol{b}.$$

The matrix $A \in \mathbb{R}^{n \times n}$ and the vector $\boldsymbol{b} \in \mathbb{R}^n$ have entries $A_{ij} = a(\varphi_j, \varphi_i)$, $\boldsymbol{b}_i = L(\varphi_i)$ for $i, j \in \{1, \ldots, n\}$. The vector $\boldsymbol{\beta}$ is the coordinate vector of u_h, i.e., $u_h = \sum_{j=1}^n \beta_j \varphi_j$.

Assembling the entries of A is cumbersome to implement directly, because triangles with boundary vertices have to be treated differently from interior triangles. Hence, we assembly instead the matrix $\widehat{A} \in \mathbb{R}^{\hat{n} \times \hat{n}}$ and vector $\hat{\boldsymbol{b}} \in \mathbb{R}^{\hat{n}}$ that have the entries

$$\widehat{A}_{ij} = a(\hat{\varphi}_j, \hat{\varphi}_i), \quad \hat{\boldsymbol{b}}_i = L(\hat{\varphi}_i) \quad \text{for } i, j \in \{1, \ldots, \hat{n}\}.$$

Matrix A is obtained from $\widehat{A}$ simply by picking the entries corresponding to interior nodes. Let $I \in \mathbb{N}^n$ be an index vector satisfying

$$\varphi_k = \hat{\varphi}_{I_k} \quad \text{for } k \in \{1, \ldots, n\}.$$

Then $A_{kl} = \widehat{A}_{I_k I_l}$ for $k, l \in \{1, \ldots, n\}$. The remaining task is evaluate the entries of $\widehat{A}$ and $\hat{\boldsymbol{b}}$. The entries are evaluated by looping over triangles in the partition $\mathcal{T}_h$ and computing appropriate integrals over each T.

9.4 Integrating over a triangle

Denote a triangle T with vertices $\boldsymbol{a}$,$\boldsymbol{b}$, and $\boldsymbol{c}$ as $T \sim \begin{bmatrix} \boldsymbol{a} & \boldsymbol{b} & \boldsymbol{c} \end{bmatrix} \in \mathbb{R}^{2\times 3}$. The topic of this section is numerical evaluation of integrals

Integration over T in Youtube

$$\int_T g \, dA \quad \text{where } g \in C(T),$$

that are computed during the finite element assembly process.

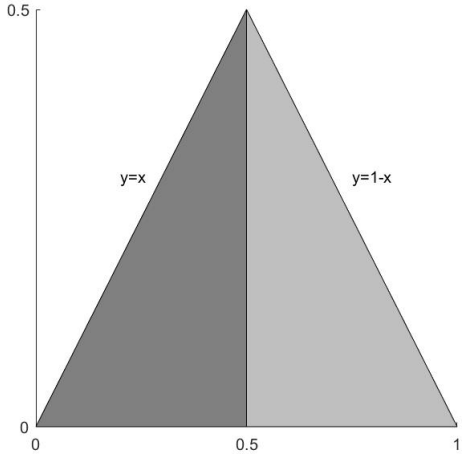

Figure 18: Triangle T related to Example 9.1. The triangles T_1 and T_2 are depicted by gray.

Example 9.1. *Let* $T \sim \begin{bmatrix} 0 & 0.5 & 1 \\ 0 & 0.5 & 0 \end{bmatrix}$, $g(\boldsymbol{x}) = xy$, *and consider the evaluation of the integral*

$$\int_T g(\boldsymbol{x})\, dA.$$

First, split

$$\int_T g(\boldsymbol{x})\, dA = \int_{T_1} g(\boldsymbol{x})\, dA + \int_{T_2} g(\boldsymbol{x})\, dA,$$

where $T_1 \sim \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0 & 0.5 & 0 \end{bmatrix}$ *and* $T_2 \sim \begin{bmatrix} 0.5 & 0.5 & 1 \\ 0 & 0.5 & 0 \end{bmatrix}$, *see Fig. 18. By the slicing principle*

$$\int_{T_1} g(\boldsymbol{x})\, dA = \int_0^{1/2} \int_0^x xy\, dydx = \frac{1}{2} \int_0^{1/2} x^3\, dx = \frac{1}{128}.$$

and

$$\int_{T_2} g(\boldsymbol{x})\, dA = \int_{1/2}^1 \int_0^{1-x} xy\, dydx = \frac{1}{2} \int_{1/2}^1 x(1-x)^2\, dx = \frac{1}{48} - \frac{1}{128}.$$

Hence, $\int_T g(\boldsymbol{x})\, dA = \frac{1}{48}$.

Integral over a triangle can be evaluated by dividing it into two parts and computing the resulting integrals using the slicing principle, see Example 9.1. This is rather cumbersome, hence, the integral is evaluated by making a change of variables form T to the reference element $\widehat{T}$,

$$\widehat{T} \sim \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

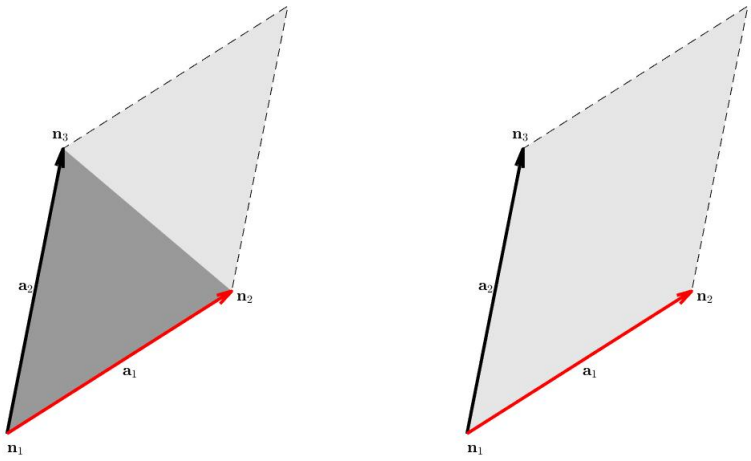

Figure 19: Vectors $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ are related to sides of triangle T and span a parallelogram.

and using numerical integration method over $\widehat{T}$. We proceed to give a mapping from a reference triangle $\widehat{T} \sim \begin{bmatrix} \hat{\boldsymbol{n}}_1 & \hat{\boldsymbol{n}}_2 & \hat{\boldsymbol{n}}_3 \end{bmatrix}$ to $T \sim \begin{bmatrix} \boldsymbol{n}_1 & \boldsymbol{n}_2 & \boldsymbol{n}_3 \end{bmatrix} \subset \mathbb{R}^2$. This mapping is used to transform integrals over T to integrals over $\widehat{T}$. Let $A_T \in \mathbb{R}^{2\times 2}$, $\boldsymbol{b}_T \in \mathbb{R}^2$, and define the affine mapping $F_T : \mathbb{R}^2 \mapsto \mathbb{R}^2$ as

$$F_T(\boldsymbol{x}) = A_T \boldsymbol{x} + \boldsymbol{b}_T.$$

The matrix A_T and vector $\boldsymbol{b}_T$ are chosen so that

$$F_T\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}\right) = \boldsymbol{n}_1, \quad F_T\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) = \boldsymbol{n}_2, \quad \text{and} \quad F_T\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right) = \boldsymbol{n}_3. \tag{72}$$

Expanding $A_T = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 \end{bmatrix}$ we obtain

$$\boldsymbol{b}_T = \boldsymbol{n}_1, \quad \boldsymbol{a}_1 = \boldsymbol{n}_2 - \boldsymbol{n}_1, \quad \text{and} \quad \boldsymbol{a}_2 = \boldsymbol{n}_3 - \boldsymbol{n}_1.$$

The vectors $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ have a geometric interpretation, see Fig. 19. Recall, that the are of the paralleogram spanned by $\boldsymbol{a}_1, \boldsymbol{a}_2$ is given by $|\det A_T|$. Hence, we have

$$|T| = \frac{1}{2}|\det A_T|, \quad \text{where } |T| \text{ denotes the area of } T.$$

The mapping F_T is one-to-one and satisfies $F_T(\widehat{T}) = T$ if $|T| > 0$, i.e., triangle T is no-degenerate.

Lemma 9.2. *Let triangle $T \sim \begin{bmatrix} \boldsymbol{n}_1 & \boldsymbol{n}_2 & \boldsymbol{n}_3 \end{bmatrix}$ satisfy $|T| > 0$, $A_T \in \mathbb{R}^{2\times 2}$, $\boldsymbol{b}_T \in \mathbb{R}^2$, and $F_T = A_T\hat{\boldsymbol{x}} + \boldsymbol{b}_T$ satisfy* (72)*. Then F_T is one-to-one and $F_T(\widehat{T}) = T$.*

Change-of-variables formula Let $F_T(\boldsymbol{x}) = A_T\boldsymbol{x} + \boldsymbol{b}_T$ satisfy (72). Then there holds that

$$\int_T f(\boldsymbol{x})\ dA = \int_{\widehat{T}} f(F_T(\hat{\boldsymbol{x}}))\ |\det A_T|\ d\widehat{A}. \tag{73}$$

Similar formula holds for differentiable mappings. Next, we give a simple justification for the change-of-variables formula (73). C-O-V in Youtube

Lemma 9.3. *Let $M \in \mathbb{R}^{2\times 2}$ be invertible, $\boldsymbol{b} \in \mathbb{R}^2$, $F(\boldsymbol{x}) = M\boldsymbol{x} + \boldsymbol{b}$, $\widehat{\Omega} = (0,1)^2$, $\Omega = F(\Omega)$, and $g : \Omega \mapsto \mathbb{R}$ satisfy*

$$|g(\boldsymbol{x}) - g(\boldsymbol{y})| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_2 \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \Omega$$

for some $L > 0$ independent of $\boldsymbol{x}, \boldsymbol{y}$. In addition, let $N \in \mathbb{N}$ and $\widehat{R}_{ij} \subset \widehat{\Omega}$, $\boldsymbol{x}_{ij} \in \widehat{\Omega}$ be defined as

$$\hat{R}_{ij} = \frac{1}{N}(i-1, i) \times (j-1, j) \text{ and } \hat{\boldsymbol{x}}_{ij} = \frac{1}{N}\begin{bmatrix} i-1 \\ j-1 \end{bmatrix} \quad i, j = 1, \ldots, N.$$

Then there holds that

$$\left| \int_\Omega g(\boldsymbol{x})\ dA - S_N \right| \leq \frac{\sqrt{2}L}{N}\|M\|_2|\Omega|.$$

and

$$\left| \int_{\widehat{\Omega}} g(F(\hat{\boldsymbol{x}}))|\det M|\ d\hat{A} - S_N \right| \leq \frac{\sqrt{2}L}{N}\|M\|_2|\Omega|.$$

where $S_N := \sum_{i,j=1}^N f(F(\hat{\boldsymbol{x}}_{ij}))\ area(F(\hat{R}_{ij}))$ and $|\Omega|$ denotes the area of Ω.

Proof. The proof is straightforward and left as an exercise problem. □

By algebraic manipulations, there holds that

$$\begin{aligned} &\int_\Omega g(\boldsymbol{x})\ dA - \int_{\widehat{\Omega}} g(F(\hat{\boldsymbol{x}}))|\det M|\ d\hat{A} \\ &\qquad = \left(\int_\Omega g(\boldsymbol{x})\ dA - S_N \right) - \left(\int_{\widehat{\Omega}} g(F(\hat{\boldsymbol{x}}))|\det M|\ d\hat{A} - S_N \right). \end{aligned}$$

Using Lemma 9.3 to take the limit $N \to \infty$ gives (73). To conclude, the factor $|\det M|$ arises from the change-of-area of the infinitesimal area element under mapping F.

Quadrature rule Integral over the reference element $\widehat{T}$ is typically evaluated using suitable numerical integration method. These methods are communicated by specifying *integration weights* $\boldsymbol{w} \in \mathbb{R}^N$ and the associated *set of integration points* $\{\boldsymbol{t}_i\}_{i=1}^N \subset \mathbb{R}^2$. The integral is then approximated as Numerical integration on $\widehat{T}$ in Youtube

$$\int_{\widehat{T}} g\ dA \approx \sum_{i=1}^N g(\boldsymbol{t}_i)w_i.$$

Simplest such rule is the midpoint rule with

$$\boldsymbol{w} = w_1 = \frac{1}{2} \quad \text{and} \quad \{\boldsymbol{t}_i\}_{i=1}^{N} = \boldsymbol{t}_1 = \begin{bmatrix} \frac{1}{3} \\ \frac{1}{3} \end{bmatrix}.$$

The midpoint rule is accurate for first order polynomials. A more accurate alternative is to use

$$\boldsymbol{w} = \begin{bmatrix} \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{bmatrix} \quad \text{and} \quad \{\boldsymbol{t}_i\}_{i=1}^{N} = \left\{ \begin{bmatrix} \frac{1}{2} \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}, \begin{bmatrix} 0 \\ \frac{1}{2} \end{bmatrix} \right\}. \tag{74}$$

The quadrature rule given in (74) is accurate for first and second order polynomials.

9.5 Problems

P33. (2p) Triangular finite element meshes can be stored in two matrices, $p \in \mathbb{R}^{2\times N_p}$ and $t \in \mathbb{N}^{3\times N_t}$. Columns of matrix p hold the nodes or vertices of the triangulation and the columns of the matrix t hold the vertex indices for each triangle in the mesh.

(a) Write a function that visualizes a given mesh. hint : `patch`, `trisurf`. Test your function with the mesh

```
N=10;
s=linspace(0,2*pi,N+1);
s=s(1:N) ;
p=[ 0 cos(s) ; 0 sin(s)];
t=[ 2:(N+1) ; [3:(N+1) 2] ; ones(1,N)];
```

(b) Let

$$\boldsymbol{n}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{n}_2 = \begin{bmatrix} 3 \\ 1 \end{bmatrix}, \quad \boldsymbol{n}_3 = \begin{bmatrix} \frac{5}{2} \\ 4 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{n}_4 = \begin{bmatrix} \frac{3}{2} \\ 3 \end{bmatrix}$$

and

$$\Omega := \{\ \boldsymbol{x} \in \mathbb{R}^2 \mid \boldsymbol{x} = \sum_{i=1}^{4} \lambda_i \boldsymbol{n}_i \text{ for } \lambda_i \in \mathbb{R}, \lambda_i > 0 \text{ and } \sum_{i=1}^{4} \lambda_i = 1\ \}.$$

Define a triangular finite element mesh for Ω by hand and by using matlab pdetoolbox (`pdetool`). Plot your mesh. Try out the `refinemesh` function.

P34. (2p) Let $f(\boldsymbol{x}) = x_2$ and Ω be as in P 33(b).

(i) Create a triangular partition $\mathcal{T}$ of Ω.

(ii) Use the $\mathcal{T}$ compute the value of the integral

$$\int_\Omega f \, dA.$$

by computing the integral element-wise as:

$$\int_\Omega f \, dA = \sum_{K \in \mathcal{T}} \int_K f \, dA.$$

Evaluate integrals over $\mathcal{T}$ by making a change of variable to the reference element and using a numerical integration method.

P35. (2p) Let

$$\hat{\varphi}_1(\hat{\boldsymbol{x}}) = 1 - \hat{x}_1 - \hat{x}_2 \quad \hat{\varphi}_2(\hat{\boldsymbol{x}}) = \hat{x}_1 \quad \text{and} \quad \hat{\varphi}_3(\hat{\boldsymbol{x}}) = \hat{x}_2.$$

Compute using Matlab entries of matrix $A \in \mathbb{R}^{3\times 3}$ such that

$$A_{ij} = \int_{\hat{K}} \hat{\nabla}\hat{\varphi}_j^T \hat{\nabla}\hat{\varphi}_i \, d\hat{x} \quad \text{for } i,j \in \{1,\ldots,3\}.$$

and the vector $\boldsymbol{b} \in \mathbb{R}^3$

$$\boldsymbol{b}_j = \int_{\hat{K}} f(\hat{x})\hat{\varphi}_j \, d\hat{x} \quad \text{for } j \in \{1,\ldots,3\},$$

where $f\hat{x} = \hat{x}_2$.

P36. (2p) Let $\{t_i\}_{i=1}^N$ and $\{w_i\}_{i=1}^N$ be the points and weights of one dimensional Gaussian-quadrature over interval $(0,1)$. These points and weights can be generated by the $gaussint.m$ - function and they integrate polynomials up to order $2N-1$ exactly over the interval $(0,1)$. In addition, define the mapping $F : \mathbb{R}^2 \to \mathbb{R}^2$ as

$$F(\hat{x}) = \begin{bmatrix} \hat{x}_1(1-\hat{x}_2) \\ \hat{x}_2 \end{bmatrix}.$$

(a) Let $\hat{R} = (0,1)\times(0,1)$ and $\hat{K} = (0,0),(1,0),(0,1)$. Use $meshgrid$ to generate grid of points to $\hat{R}$. Map these points with F to visually verify that $F(\hat{R}) = \hat{K}$ holds.

(b) Define two dimensional quadrature points and the related weights as

$$x_{ij} = \begin{bmatrix} t_i \\ t_j \end{bmatrix} \quad \text{and} \quad \hat{w}_{ij} = w_i w_j \quad i,j = 1,\ldots,N.$$

Show that these points integrate exactly functions $x_1^n x_2^m$ where $n,m = 1,\ldots,2N-1$ over $\hat{R}$.

(c) Use the change of variables formula

$$\int_{\hat{K}} f(x)\, dA = \int_{\hat{R}} f(F(\hat{x}))|\det DF(\hat{x})|\, d\hat{A}.$$

To find quadrature points and corresponding weights on $\hat{K}$. Determine how high indices n and m are integrated accurately, for functions $x_1^n x_2^m$ and fixed N.

P37. (2p) Let $F(\boldsymbol{x}) = M\boldsymbol{x} + \boldsymbol{c}$ in which $M \in \mathbb{R}^{2\times 2}, \boldsymbol{c} \in \mathbb{R}^2$. Assume that M is invertible.

(i) Show that F maps line segments to line segments. Hint : You can express line segment as $L := \{\ \boldsymbol{x} \mid \boldsymbol{x} = \boldsymbol{\alpha}t + \boldsymbol{\beta}$ for some $t \in [0,1]\ \}$ in which $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^2$ are given vectors.

(ii) Show that $|F(S)| = |\det M||S|$, in which $S = (0,a) \times (0,b)$ for some $a, b \in \mathbb{R}, a > 0, b > 0$. Hint : use (i) and map two sides of the square. Compute the area of the resulting parallelogram using determinant.

(iii) Using (ii), justify the change of variables formula

$$\int_{\hat{\Omega}} f(F(\hat{\boldsymbol{x}}))|\det M|\, d\hat{A} = \int_{\Omega} f(\boldsymbol{x})\, dA.$$

P38. (2p) Let Ω be as in P 33(b), $\hat{\Omega} = (0,1) \times (0,1)$, and define a mapping $F : \hat{\Omega} \to \Omega$ as

$$F(\boldsymbol{x}) = \sum_{i=1}^{4} \boldsymbol{n}_i \varphi_i(\boldsymbol{x}),$$

in which $\boldsymbol{n}_i$

$$\boldsymbol{n}_1 = \begin{bmatrix}1 & 0\end{bmatrix}^T, \quad \boldsymbol{n}_2 = \begin{bmatrix}3 & 1\end{bmatrix}^T, \quad \boldsymbol{n}_3 = \begin{bmatrix}\frac{5}{2} & 4\end{bmatrix}^T, \quad \text{and} \quad \boldsymbol{n}_4 = \begin{bmatrix}\frac{3}{2} & 3\end{bmatrix}^T$$

are the corner nodes of Ω and

$$\varphi_1(\boldsymbol{x}) = (1-x_1)(1-x_2), \quad \varphi_2(\boldsymbol{x}) = x_1(1-x_2),$$
$$\varphi_3(\boldsymbol{x}) = x_1 x_2, \quad \text{and} \quad \varphi_4(\boldsymbol{x}) = (1-x_1)x_2.$$

Consider $f(\boldsymbol{x}) = x_2$ and compute **by hand** the integral

$$\int_{\Omega} f\, dA. \tag{75}$$

You should take the following steps

(i) Use $meshgrid$ to generate a grid of points to $\hat{\Omega}$. Map these points with F to visually verify that $\Omega = F(\hat{\Omega})$ holds. Transform the integral in (75)

to an integral over the reference domain $\hat{\Omega}$ using the change of variables formula

$$\int_{\hat{K}} f(x)\, dA = \int_{\hat{R}} f(F(\hat{x}))|\det \mathrm{D}F(\hat{x})|\, d\hat{A},$$

where $\mathrm{D}F$ is the Jacobian of F.

(ii) Compute $\det(\mathrm{D}F)$, i.e., the determinant of the Jacobian of the mapping F, and check that it is positive in $\hat{\Omega}$.

(iii) Compute the resulting integral over the reference domain using pen and paper.

9.6 Reference basis functions

Let $\mathcal{T}_h$ be a conforming triangulation of domain Ω. Recall the definition of first order FE-space,

$$\widehat{V}_h = \{\, v \in C(\overline{\Omega}) \mid v|_K \in P^1(T) \text{ for all } T \in \mathcal{T}_h \,\}.$$

We use the hat basis $\{\hat{\varphi}_j\}_{j=1}^{\hat{n}}$ of $\widehat{V}_h$ satisfying

$$\hat{\varphi}_j \in \widehat{V}_h \quad \text{and} \quad \hat{\varphi}_j(\boldsymbol{n}_i) = \begin{cases} 1 & j = i \\ 0 & \text{otherwise} \end{cases},$$

where $\{\boldsymbol{n}_i\}_{i=1}^{\hat{n}} \subset \mathbb{R}^2$ are the vertices of $\mathcal{T}_h$. In this Section, we define the hat basis functions in a way that simplifies the evaluation of entries of the matrix $\widehat{A}$ and vector $\hat{\boldsymbol{b}}$ related to our model problem,

$$\widehat{A}_{ij} := \int_\Omega K \nabla \hat{\varphi}_j \cdot \nabla \hat{\varphi}_i \quad \text{and} \quad \hat{b}_i = \int_\Omega f \hat{\varphi}_i.$$

FE-assembly process proceeds by computing the contribution of each element $T \in \mathcal{T}_h$ to the entries of $\widehat{A}$ and $\hat{\boldsymbol{b}}$.

Indexing basis functions Let $T \in \mathcal{T}_h$ have vertices with indeces i_1, i_2, and i_3. This is,

$$T \sim \begin{bmatrix} \boldsymbol{n}_{i_1} & \boldsymbol{n}_{i_2} & \boldsymbol{n}_{i_3} \end{bmatrix}.$$

To manage the indices, we use the *index-mapping* $\sigma : \mathcal{T}_h \times \{1,2,3\} \mapsto \{1,\ldots,\hat{n}\}$ that relates the nodal indeces on each triangle to corresponding global indeces as

$$\sigma(T,1) = i_1, \quad \sigma(T,2) = i_2, \quad \text{and} \quad \sigma(T,3) = i_3.$$

In FE-implementation, σ is obtained from matrix $t \in \mathbb{R}^{3\times N_t}$ stating the nodal indeces of each triangle in the mesh. The mapping becomes more elaborate if other, e.g. second order, FE-space is used.

By definition, the hat basisfunctions $\hat{\varphi}_{\sigma(T,1)}$, $\hat{\varphi}_{\sigma(T,2)}$, $\hat{\varphi}_{\sigma(T,3)}$ have value one in some vertex of T. All other hat basisfunctions have value zero in each vertex of T. As the restriction of hat basis functions to T is linear, only $\hat{\varphi}_{\sigma(T,1)}|_T$, $\hat{\varphi}_{\sigma(T,2)}|_T$, $\hat{\varphi}_{\sigma(T,3)}|_T$ are nonzero over T. Thus the contribution of T to the entries of $\widehat{A}$ is

$$\widehat{A}_{\sigma(T,l)\sigma(T,k)} = \widehat{A}_{\sigma(T,l)\sigma(T,k)} + \int_T K \nabla \hat{\varphi}_{\sigma(T,l)} \cdot \nabla \hat{\varphi}_{\sigma(T,k)} \tag{76}$$

and to entries $\hat{\boldsymbol{b}}$

$$\hat{b}_{\sigma(T,l)} = \hat{b}_{\sigma(T,l)} + \int_T f \hat{\varphi}_{\sigma(T,l)} \tag{77}$$

for $l, k \in \{1, 2, 3\}$.

Reference basis functions Integrals appearing in (76) and (77) are evaluated using the change-of-variables formula. Recall that $\widehat{T}$ is the reference element

Assembly of $\boldsymbol{b}$ and reference basis in Youtube

$$\widehat{T} \sim \begin{bmatrix} \hat{\boldsymbol{n}}_1 & \hat{\boldsymbol{n}}_2 & \hat{\boldsymbol{n}}_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Let $A_T \in \mathbb{R}^{2\times 2}$, $\boldsymbol{b}_T \in \mathbb{R}^2$, and $F_T(\hat{\boldsymbol{x}}) = A_T \hat{\boldsymbol{x}} + \boldsymbol{b}_T$ satisfy

$$F_T(\hat{\boldsymbol{n}}_l) = \boldsymbol{n}_{\sigma(T,l)} \quad \text{for all } l \in \{1, 2, 3\}. \tag{78}$$

Then $F_T(\widehat{T}) = T$, and

$$\int_T f \hat{\varphi}_{\sigma(T,l)} = \int_{\widehat{T}} f(F_T(\hat{\boldsymbol{x}})) \hat{\varphi}_{\sigma(T,l)}(F_T(\hat{\boldsymbol{x}})) |\det A_T|.$$

Evaluating the integral requires computing values of function $\hat{\varphi}_{\sigma(T,l)}(F_T(\hat{\boldsymbol{x}}))$. We proceed by Lemma stating that the affine mapping preserves first order polynomials.

Lemma 9.4. *Let $T \in \mathcal{T}_h$, $F_T : \widehat{T} \mapsto T$ be an affine mapping satisfying* (78)*, and $p \in P^1(T)$. Then $q(\hat{\boldsymbol{x}}) := p(F_T(\hat{\boldsymbol{x}}))$ satisfies $q \in P^1(\widehat{T})$.*

Proof. P39 □

Denote $\psi_l(\hat{\boldsymbol{x}}) := \hat{\varphi}_{\sigma(T,l)}(F_T(\hat{\boldsymbol{x}}))$ for $l \in \{1, 2, 3\}$. By Lemma 9.4, $\psi_l \in P^1(\widehat{T})$. Using (78) and the definition of hat basis functions gives

$$\psi_l(\hat{\boldsymbol{n}}_k) = \begin{cases} 1 & l = k \\ 0 & \text{otherwise} \end{cases}. \tag{79}$$

Functions ψ_1, ψ_2, and ψ_3 are solved from condition (79) as

$$\psi_1(\hat{\boldsymbol{x}}) = 1 - \hat{x}_1 - \hat{x}_2, \quad \psi_2(\hat{\boldsymbol{x}}) = \hat{x}_1, \quad \psi_3(\hat{\boldsymbol{x}}) = \hat{x}_2.$$

As these function are identical for any triangle T, they are called as *reference basis functions*. The (all non-zero) restrictions of hat basis functions to triangle T are obtained from the reference basis functions as

$$\hat{\varphi}_{\sigma(T,l)}|_T(F_T(\hat{\boldsymbol{x}})) = \psi_l(\hat{\boldsymbol{x}}) \quad \text{for all } T \in \mathcal{T} \text{ and } l \in \{1,2,3\}.$$

Using reference basis functions gives the relation

$$\int_{\widehat{T}} f(F_T(\hat{\boldsymbol{x}}))\hat{\varphi}_{\sigma(T,l)}(F_T(\hat{\boldsymbol{x}}))|\det A_T| = \int_{\widehat{T}} f(F_T(\hat{\boldsymbol{x}}))\psi_l(\hat{\boldsymbol{x}})|\det A_T|.$$

The RHS is evaluated using a numerical integration method with weights $\boldsymbol{w} \in \mathbb{R}^N$ and points $\{\boldsymbol{t}_i\}_{i=1}^N$ as

$$\int_{\widehat{T}} f(F_T(\hat{\boldsymbol{x}}))\psi_l(\hat{\boldsymbol{x}})|\det A_T| \approx \sum_{i=1}^{N} f(F_T(\boldsymbol{t}_i))\psi_l(\boldsymbol{t}_i)|\det A_T|w_i$$

Observe that values of ψ_l are required *only at the integration points*. Pre-computing these values simplifies the implementation.

Entries of $\widehat{A}$ The integral

Assembly of $\widehat{A}$ in Youtube

$$\int_T K\nabla\hat{\varphi}_{\sigma(T,l)} \cdot \nabla\hat{\varphi}_{\sigma(T,k)}$$

is evaluated similar to (77). Change of variables gives

$$\int_T K\nabla\hat{\varphi}_{\sigma(T,l)}\cdot\nabla\hat{\varphi}_{\sigma(T,k)} = \int_{\widehat{T}} K(\nabla\hat{\varphi}_{\sigma(T,l)})(F_T(\hat{\boldsymbol{x}}))\cdot(\nabla\hat{\varphi}_{\sigma(T,k)})(F_T(\hat{\boldsymbol{x}}))|\det A_T|.$$

Next Lemma relates $(\nabla\hat{\varphi}_{\sigma(T,l)})(F_T(\hat{\boldsymbol{x}}))$ to gradients of the reference basis functions.

Lemma 9.5. *Let $T \in \mathcal{T}_h$, $A_T \in \mathbb{R}^{2\times 2}$, $\boldsymbol{b}_T \in \mathbb{R}^2$, and $F_T(\hat{\boldsymbol{x}}) = A_T\hat{\boldsymbol{x}} + \boldsymbol{b}_T$ be an affine mapping satisfying* (78)*. In addition, let $g \in C^1(T)$, and $\hat{g}(\hat{\boldsymbol{x}}) := g(F_T(\hat{\boldsymbol{x}}))$. Then there holds that*

$$(\nabla g)(F_T(\hat{\boldsymbol{x}})) = A_T^{-T}\hat{\nabla}\hat{g}(\hat{\boldsymbol{x}}),$$

where $\hat{\nabla}$ denotes the gradient with respect to the reference coordinates $\hat{\boldsymbol{x}}$.

Proof. P40. □

Using Lemma 9.5 gives the relation

$$(\nabla\hat{\varphi}_{\sigma(T,l)})(F_T(\hat{\boldsymbol{x}}))|_T = A_T^{-T}\hat{\nabla}\psi_l \quad \text{for all } T \in \mathcal{T}_h \text{ and } l \in \{1,2,3\}.$$

Thus,

$$\int_{\widehat{T}} K(\nabla\hat{\varphi}_{\sigma(T,l)})(F_T(\hat{\boldsymbol{x}}))\cdot(\nabla\hat{\varphi}_{\sigma(T,k)})(F_T(\hat{\boldsymbol{x}}))|\det A_T| \\ = \int_{\widehat{T}} \hat{\nabla}\psi_l(\hat{\boldsymbol{x}})^T A_T^{-1} K A_T^{-T}\hat{\nabla}\psi_k(\hat{\boldsymbol{x}})|\det A_T| \quad (80)$$

As matrices A_T, K and gradients $\hat{\nabla}\psi_1$, $\hat{\nabla}\psi_2$, $\hat{\nabla}\psi_3$ are all constant, the integral on the RHS is evaluated simply as

$$\frac{1}{2}\hat{\nabla}\psi_l^T A_T^{-1} K A_T^{-T}\hat{\nabla}\psi_k|\det A_T|.$$

Numerical integration has to be used with other FE-spaces or non-constant material coefficients.

9.7 Problems

P39. (0.5p) Prove Lemma 9.4

P40. (2p) Prove Lemma 9.5

P41. (1p) Let the finite element mesh $\mathcal{T}$ be such that

$$p = \begin{bmatrix} 0 & 1 & 2 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \quad \text{and} \quad t = \begin{bmatrix} 1 & 2 & 2 & 3 \\ 2 & 4 & 3 & 5 \\ 4 & 5 & 5 & 6 \end{bmatrix} \quad (81)$$

(a) Draw the mesh $\mathcal{T}$.

(b) Compute affine mapping from the reference element to elements 3 and 4.

(c) Consider the bilinear form $a(u,v) = (\nabla u, \nabla v)$ and assume that standard hat basis functions are used. Compute the row 3 of the system matrix

9.8 Implementation

Next we outline the modifications required in the one dimensional example FE-solver on p.17 to solve the two dimensional model problem (64).

Implementation in Youtube

1. **Mesh:** Instead of partition to intervals, triangular mesh of Ω is used. Triangular finite element meshes are often stored in two matrices, $p \in \mathbb{R}^{2\times N_p}$ and $t \in \mathbb{N}^{3\times N_t}$. Columns of matrix p hold the nodes or vertices of the triangulation and the columns of the matrix t hold the vertex indices for each triangle in the mesh. Mesh can be obtained either by specifying it manually or by using a mesh generator.

2. **Integration:** Integrals are computed with help of an affine mapping $F_T(\hat{\boldsymbol{x}}) = A_T\hat{\boldsymbol{x}} + \boldsymbol{b}_T$ satisfying (78). The contribution of element T to matrix $\widehat{A}$ is now evaluated as follows:

$$\widehat{A}_{\sigma(T,l)\sigma(T,k)} += \frac{1}{2}\hat{\nabla}\psi_l^T A_T^{-1} K A_T^{-T}\hat{\nabla}\psi_k |\det A_T|.$$

for $l, k \in \{1, 2, 3\}$. Contribution to $\boldsymbol{b}_T$ is obtained as

$$\hat{b}_{\sigma(T,l)} += \sum_{i=1}^{N} f(F_T(\boldsymbol{t}_i))\psi_l(\boldsymbol{t}_i)|\det A_T| w_i$$

for $l \in \{1, 2, 3\}$, integration points $\{\boldsymbol{t}_i\}_{i=1}^N$, and weights $\boldsymbol{w} \in \mathbb{R}^N$. Here $+=$ means add to, e.g., $x+ = 2$ means $x = x + 2$.

3. **Elimination of boundary basisfunctions**. Imposing the zero Dirichlet boundary condition requires one to find indeces of nodes that lie on the boundary of domain Ω. Mesh generator typically returns information on boundary nodes. Alternatively, boundary nodes can extracted from the t - matrix by finding all edges of mesh $\mathcal{T}_h$ that are associated only to one triangle. The problem is solved as

```
A = Ahat(idof,idof) ; b = bhat(idof)
u = zeros(ndof,1)
u(idof) = A/b;
```

where `idof` is a vector with interior node indices and `ndof` is then number of basisfunctions.

4. **Plotting:** Plotting is done using commands `patch` or `trisurf`.

9.9 Problems

P42. (2p) Let

$$\psi_1(\hat{\boldsymbol{x}}) = 1 - \hat{x}_1 - \hat{x}_2 \quad \psi_2(\hat{\boldsymbol{x}}) = \hat{x}_1 \quad \text{and} \quad \psi_3(\hat{\boldsymbol{x}}) = \hat{x}_2.$$

In addition, let $f(\boldsymbol{x}) = \sin \pi x_1 \sin \pi x_2$, $B \in \mathbb{R}^{3\times 3}$, and the vector $\boldsymbol{b} \in \mathbb{R}^3$ have entries

$$B_{ij} = \int_{\hat{T}} \hat{\nabla}\psi_j^T \hat{\nabla}\psi_i \, d\hat{x}$$

$$\boldsymbol{b}_j = \int_{\hat{T}} f(\hat{x})\psi_j \, d\hat{x}$$

for $i, j \in \{1, 2, 3\}$. Evaluate the entries of B and $\boldsymbol{b}$ using Matlab.

P43. (2p) Let $\mathcal{T}_h$ be a conforming triangular mesh, $T \in \mathcal{T}_h$, and F_T an affine mapping satisfying (78). In addition, let ψ_i be as defined in P42,

$$\varphi_i(F_T(\hat{\boldsymbol{x}})) = \psi_i(\hat{\boldsymbol{x}}), \quad \text{for } i \in \{1,2,3\},$$

$B \in \mathbb{R}^{3\times 3}$ and the vector $\boldsymbol{b} \in \mathbb{R}^3$ have entries

$$B_{ij} = \int_T \nabla\varphi_j^T \nabla\varphi_i \, dx \quad i,j = 1\ldots 3$$

$$\boldsymbol{b}_j = \int_T f(x)\varphi_j(x) \, dx \quad j = 1,\ldots,3.$$

Write a program that evaluates the entries of matrix $B \in \mathbb{R}^{3\times 3}$ and vector $\boldsymbol{b}$.

P44. (2p) Modify the FE-solver on p. 17 to solve the two dimensional Poisson's equation: find $u \in H_0^1(\Omega)$ satisfying

$$(\nabla u, \nabla v)_\Omega = (f, v)_\Omega \quad \forall v \in H_0^1(\Omega).$$

Test your implementation on domain $\Omega = (0,1)^2$ with $f = \sin\pi x \sin\pi y$. Hint: for this particular domain it is easy to find index of boundary nodes from the p-matrix. Return the code along with piture of the FE-solution u_h.

10 Error analysis in two dimensions

Outline of week 7 in Youtube

We proceed to study the accuracy of the FE-solution $u_h \in V_h$ to the two-dimensional model problem (68). Our aim is to bound the $L^2(\Omega)$, $H^1(\Omega)$, and energy norm of the error $u - u_h$. Recall that the FE-solution depends on triangulation $\mathcal{T}_h$, domain Ω, coefficient matrix K, and source function f. In one dimensional case, we derived error estimate depending on the *mesh size*, i.e., the longest interval in the applied partition. In two dimensions, we make additional assumptions on $\mathcal{T}_h$ to obtain error estimate dependent on the mesh size,

$$h = \max_{T\in\mathcal{T}_h} h_T,$$

where h_T is the diameter of the smallest sphere containing triangle T. We derive error estimates under the assumption that the exact solution $u \in H^2(\Omega)$, which in two dimensional case poses restrictions both to the load function f and the domain Ω. Particularly, Ω has to satisfy assumptions 9.1, and to be convex.

The error estimate is derived in two steps that are identical to one dimensional case discussed in Section 8:

1. *Relate error to the approximation properties of the FE-space*

The two dimensional model problem (64) is an instance of the abstract variational problem studied in Section 8.1. Particularly, we can apply Cea's Lemma 8.1 that states

$$\|u - u_h\|_E = \inf_{v \in V_h} \|u - v\|_E.$$

This is, the FE-solution is the best possible approximation of the exact solution u from the applied finite element space. The error analysis is based on *comparing* $u - u_h$ to error in the nodal interpolant of u. As $H^1(\Omega)$ functions for $\Omega \subset \mathbb{R}^d$ for $d > 1$ do not have well defined pointwise values contrary to $H^2(\Omega)$ functions, we assume that the solution u is in the space $H^2(\Omega)$. The nodal interpolant $\pi : H^2(\Omega) \mapsto \widehat{V}_h$ is defined as

$$(\pi u)(\boldsymbol{n}_i) = u(\boldsymbol{n}_i) \quad \text{for all } i \in \{1, \ldots, \hat{n}\}.$$

As u satisfies zero Dirichlet boundary condition it follows that $\pi u \in V_h$. Assume that $u \in H^2(\Omega)$. Then the FE-solution u_h satisfies

$$\|u - u_h\|_E \leq \|u - \pi u\|_E. \tag{82}$$

Similar estimate in the $H^1(\Omega)$-norm follows from the equivalence between $\|\cdot\|_E$ and $\|\cdot\|_{H^1(\Omega)}$-norms, see P46.

2. *Study approximation properties of FE-space*

   By (82), a bound for the error $\|u - u_h\|_E$ follows by bounding the interpolation error term $\|u - \pi u\|_E$. The interpolation error is bounded using identical strategy to one dimensional case: First, interpolation error estimate is obtained for the reference element $\widehat{T}$. This reference interpolation error estimate is then transferred to an arbitrary triangle by applying the scaling argument.

Bramble-Hilbert Lemma Next, we give the Bramble-Hilbert Lemma, an interpolation error estimate on the reference triangle $\widehat{T}$. The proof is rather technical, and thus omitted. Recall that the reference triangle $\widehat{T}$ satisfies

B-H lemma in Youtube

$$\widehat{T} \sim \begin{bmatrix} \hat{\boldsymbol{n}}_1 & \hat{\boldsymbol{n}}_2 & \hat{\boldsymbol{n}}_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{83}$$

Theorem 10.1 (Bramble-Hilbert Lemma)**.** *Let $\widehat{T}$ be the reference triangle satisfying* (83)*. In addition, let $\hat{\pi} : H^2(\widehat{T}) \mapsto P^1(\widehat{T})$ satisfy $(\hat{\pi}\hat{v})(\hat{\boldsymbol{n}}_i) = \hat{v}(\hat{\boldsymbol{n}}_i)$ for all $i \in \{1, 2, 3\}$ and $\hat{v} \in H^2(\widehat{T})$. Then there exist a constant $C > 0$ independent of $\hat{v}$ such that*

$$|\hat{v} - \hat{\pi}\hat{v}|_{H^1(\widehat{T})} \leq C|\hat{v}|_{H^2(\widehat{T})} \quad \textit{and} \quad \|\hat{v} - \hat{\pi}\hat{v}\|_{L^2(\widehat{T})} \leq C|\hat{v}|_{H^2(\widehat{T})}.$$

for any $\hat{v} \in H^2(\widehat{T})$.

Here $|\cdot|_{H^m(\widehat{T})}$ stands for the $H^m(\widehat{T})$ semi-norm, $|v|_{H^m(\widehat{T})} := \left(\sum_{|\alpha|=m} \|\partial_\alpha v\|^2_{L^2(\widehat{T})}\right)^{1/2}$.

Proof. See, Braess □

Scaling Argument Let $a : H_0^1(\Omega) \times H_0^1(\Omega) \mapsto \mathbb{R}$ be a bilinear form satisfying Assumptions 7.1 for $V = H_0^1(\Omega)$. Recall that the energy norm corresponding to a is defined as $\|v\|_E = (a(v,v))^{1/2}$. By continuity and Poincare inequality in Theorem 9.2, there holds that

$$\|v\|_E \leq C\|\nabla v\|_{L^2(\Omega)}$$

for any $v \in H_0^1(\Omega)$ and C independent of v, see P46. Hence, the interpolation error satisfies

$$\|u - \pi u\|_E \leq C\|\nabla(u - \pi u)\|_{L^2(\Omega)} = C\left(\sum_{T\in\mathcal{T}_h} \|\nabla(u-\pi u)\|_{L^2(T)}^2\right)^{1/2}. \quad (84)$$

The element-wise interpolation errors $\|\nabla(u - \pi u)\|_{L^2(T)}$ are estimated by using the Bramble-Hilbert Lemma and the scaling argument.

Theorem 10.2 (Scaling argument)**.** *Let triangle* $T \sim \begin{bmatrix}\boldsymbol{n}_1 & \boldsymbol{n}_2 & \boldsymbol{n}_3\end{bmatrix}$, $\widehat{T}$ *be the reference element satisfying* (83)*,* $A_T \in \mathbb{R}^{2\times 2}$, $\boldsymbol{b}_T \in \mathbb{R}^2$, *and* $F_T(\boldsymbol{x}) = A_T\boldsymbol{x} + \boldsymbol{b}_T$ *satisfy* $F_T(\hat{\boldsymbol{n}}_i) = \boldsymbol{n}_i$ *for* $i \in \{1,2,3\}$*. In addition, let* $v \in H^m(T)$ *and define* $\hat{v} \in H^m(\widehat{T})$ *as* $\hat{v}(\hat{\boldsymbol{x}}) := v(F(\hat{\boldsymbol{x}}))$ *for each* $\hat{\boldsymbol{x}} \in \widehat{T}$*. Then for any* $m \in \{0,1,2,\ldots\}$ *there exists positive constants* C_1 *and* C_2 *independent of* v *and* T *such that:*

$$|\hat{v}|_{H^m(\widehat{T})} \leq C_1\|A_T\|^m \left|\det A_T\right|^{-1/2} |v|_{H^m(T)}$$

and

$$|v|_{H^m(T)} \leq C_2\|A_T^{-1}\|^m \left|\det A_T\right|^{1/2} |\hat{v}|_{H^m(\widehat{T})}.$$

Here $H^0(T)$ *and* $H^0(\widehat{T})$ *denote* $L^2(T)$ *and* $L^2(\widehat{T})$*, respectively.*

Before proving Theorem 10.2, we give technical auxiliary results. Recall that the gradient of $\hat{v}$ is related to gradient of v as

Auxiliary results in Youtube

$$\hat{\nabla}\hat{v}(\hat{\boldsymbol{x}}) = A_T^T(\nabla v)(F(\hat{\boldsymbol{x}})). \quad (85)$$

The following Lemma is used to apply similar transformation formula to estimate the $H^2(T)$ semi-norm.

Lemma 10.1. *Let* $v \in H^2(T)$*. Denote the Hessian matrix as*

$$H = \begin{bmatrix}\partial_{x_1x_1} & \partial_{x_1x_2} \\ \partial_{x_2x_1} & \partial_{x_2x_2}\end{bmatrix}$$

Then there holds that

$$\frac{1}{2}\int_T \|(Hv)(x)\|_F^2\, dx \leq |v|_{H^2(T)}^2 \leq \int_T \|(Hv)(x)\|_F^2\, dx,$$

where $\|\cdot\|_F$ *is the Frobenius norm, defined as* $\|A\|_F^2 := \sum_{i,j=1}^n A_{ij}^2$ *for any* $A \in \mathbb{R}^{n\times n}$

Proof. By definition,

$$|v|^2_{H^2(T)} = \sum_{|\alpha|=2} \|\partial_\alpha v\|^2_{L^2(T)} \quad \text{and} \quad \int_T \|Hv\|_F^2 \, dx = \sum_{i,j=1}^{2} \|(Hv)_{ij}\|^2_{L^2(T)}.$$

The proof follows by observing that the difference between these expressions is in the cross terms; the latter one has them twice. □

Lemma 10.1 allows us to use the convenient relation between the Hessian matrix of $\hat{v}$ and v,

$$(\hat{H}\hat{v})(\hat{\boldsymbol{x}}) = A_T^T \left[(Hv)(F(\hat{\boldsymbol{x}}))\right] A_T, \tag{86}$$

in which

$$\hat{H} = \begin{bmatrix} \partial_{\hat{x}_1\hat{x}_1} & \partial_{\hat{x}_1\hat{x}_2} \\ \partial_{\hat{x}_2\hat{x}_1} & \partial_{\hat{x}_2\hat{x}_2} \end{bmatrix}.$$

Proof of scaling argument in Youtube

Proof of Theorem 10.2 (Scaling argument). We will only give a proof for the first claim and cases $m \in \{0, 1, 2\}$ that are used in the following. The second claim follows from similar arguments.

Case $m = 0$**.** Change of variables gives

$$\begin{aligned} \|v\|^2_{L^2(T)} = \int_T v(x)^2 \, dx &= \int_{\widehat{T}} v(F_T(\hat{x}))^2 |\det A_T| \, d\hat{x} \\ &= \int_{\widehat{T}} \hat{v}(\hat{x})^2 |\det A_T| \, d\hat{x} = |\det A_T| \, |\hat{v}|^2_{L^2(\widehat{T})}, \end{aligned}$$

which proves the Theorem for $m = 0$.

Case $m = 1$**.** Change of variables gives

$$|v|^2_{H^1(T)} = \int_T \|\nabla v(x)\|^2 \, dx = \int_{\hat{T}} \|(\nabla v)(F_T(\hat{x}))\|^2 |\det A_T| \, d\hat{x}.$$

Using (85) yields

$$|v|^2_{H^1(T)} = \int_{\hat{T}} \|A_T^{-T}(\hat{\nabla}\hat{v})(\hat{x})\|^2 |\det A_T| \, d\hat{x}.$$

Recall that the two-norm satisfies $\|A\boldsymbol{z}\| \le \|A\|\|\boldsymbol{z}\|$ for any $A \in \mathbb{R}^{n\times n}$ and $\boldsymbol{z} \in \mathbb{R}^n$, hence,

$$|v|^2_{H^1(T)} \le \|A_T^{-T}\|^2 \, |\det A_T| \int_{\hat{T}} \|(\hat{\nabla}\hat{v})(\hat{x})\|^2 \, d\hat{x} = \|A_T^{-T}\|^2 \, |\det A_T| \, |\hat{v}|^2_{H^1(\hat{T})},$$

which proves the claim for $m = 1$.

Case $m = 2$. Making a change of variables, using Lemma 10.1, and (86) gives

$$|v|^2_{H^2(T)} \le \int_T \|(Hv)(x)\|_F^2 \, dx = \int_{\widehat{T}} \|A_T^{-T}(\hat{H}\hat{v})(\hat{x})A_T^{-1}\|_F^2 |\det A_T| \, d\hat{x}.$$

Using properties of the Frobenius norm, $\|AB\|_F \leq \|A\|_F \|B\|_F$ and $\|A\|_F = \|A^T\|_F$, gives

$$|v|^2_{H^2(T)} \leq \|A_T^{-1}\|_F^4 |\det A_T| \int_{\widehat{T}} \|(\hat{H}\hat{v})(\hat{x})\|_F^2 \, d\hat{x} \leq 2\|A_T^{-1}\|_F^4 |\det A_T| \, |\hat{v}|^2_{H^2(\widehat{T})}.$$

Applying equivalence between two- and Frobenius-norms completes the proof for $m = 2$. □

Local interpolation error estimate Combining the scaling argument with Bramble-Hilbert Lemma gives estimate for the local interpolation error.

Local interpolation error estimate in Youtube

Theorem 10.3. *Make same assumptions as in Theorem 10.2. Then there exists a positive constant $C > 0$ independent of u and triangle T such that*

$$|u - \pi u|_{H^1(T)} \leq C\|A_T^{-1}\|\|A_T\|^2 \, |u|_{H^2(T)}$$

for any $u \in H^2(T)$.

Proof. This proof uses a standard "scaling argument technique". Let $\hat{u} \in H^2(\widehat{T})$ be defined as $\hat{u}(\hat{\boldsymbol{x}}) = u(F_T(\hat{\boldsymbol{x}}))$ for all $\hat{\boldsymbol{x}} \in \widehat{T}$. We begin by application of the scaling argument:

$$|u - \pi u|_{H^1(T)} \leq C_2\|A_T^{-1}\| \, |\det A_T|^{1/2} \, |\hat{u} - \hat{\pi}\hat{u}|_{H^1(\widehat{T})}.$$

Now, we will use the Bramble-Hilbert lemma

$$|u - \pi u|_{H^1(T)} \leq CC_2\|A_T^{-1}\| \, |\det A_T|^{1/2} \, |\hat{u}|_{H^2(\widehat{T})}.$$

Above we used the inequality on the reference element, hence the constant C above is independent of triangle T. Using the scaling argument once more completes the proof

$$|u - \pi u|_{H^1(T)} \leq C_1CC_2\|A_T^{-1}\|\|A_T\|^2|u|_{H^2(T)}.$$

□

10.1 Estimates for $\|A_T\|$ and $\|A_T^{-1}\|$

By Theorem 10.3, the local interpolation error depends on term $\|A_T^{-1}\|\|A_T\|^2$ and $H^2(T)$ semi-norm of the exact solution u. In this Section, we derive estimates for $\|A_T\|$ and $\|A_T^{-1}\|$. *Under sufficient assumptions on the triangulation $\mathcal{T}_h$*, our analysis indicates that

$$\|A_T^{-1}\|\|A_T\|^2 \leq Ch_T,$$

in which h_T is the diameter of triangle T. When combined with Theorem 10.3, (84), and (82), yields

$$\|u - u_h\|_E \leq Ch|u|_2.$$

Identical estimate holds also in the $H^1(\Omega)$ -norm, but with different constant. The terms $\|A_T\|$ and $\|A_T^{-1}\|$ are related to two geometric parameters characterising the shape of the triangle T, see Figure 20.

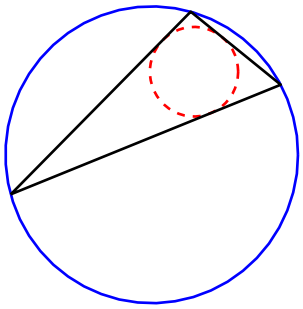

Figure 20: The smallest ball containing T and the largest ball contained in T (inscribed ball). Parameters h_T and ρ_T are the diameters of the these balls, respectively.

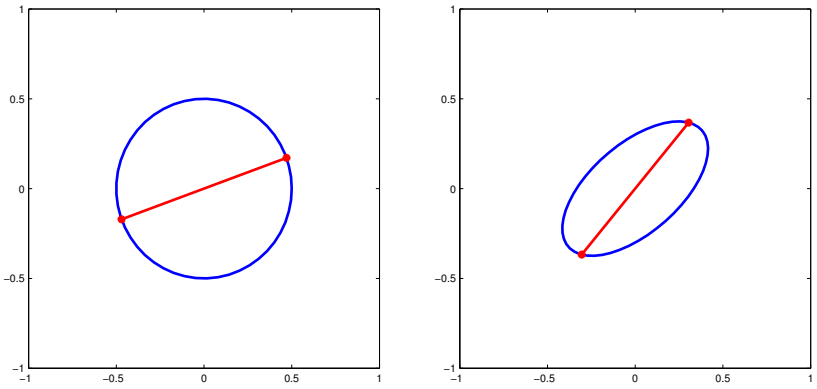

Figure 21: On the left: a circle with diameter 1. On the right: the image of this circle after a mapping with A. All vectors $\boldsymbol{x} \in \mathbb{R}^2, \|\boldsymbol{x}\| = 1$ can be obtained by connecting two points on the circle on the left. One such vector and its image are visualized in red.

Definition 10.1. *Let ρ_T be diameter of the largest ball contained in the element T.*

Definition 10.2. *Let h_T be diameter of the smallest ball containing the element T.*

Matrix two norm The two norm of $A \in \mathbb{R}^{2\times 2}$ is defined as

Matrix two-norm in Youtube

$$\|A\| := \max_{\boldsymbol{x}\in\mathbb{R}^2, \|\boldsymbol{x}\|=1} \|A\boldsymbol{x}\|, \tag{87}$$

in which $\|\boldsymbol{x}\|$ is the vector 2-norm, i.e., $\|\boldsymbol{x}\|^2 = \boldsymbol{x}^T\boldsymbol{x}$. Estimates for $\|A_T\|$ and $\|A_T^{-1}\|$ are based on a geometric interpretation of (87). The norm (87) is defined simply as follows: map any vector of length one with A and compute the length of the longest resulting vector. With this geometric interpretation in mind, we obtain

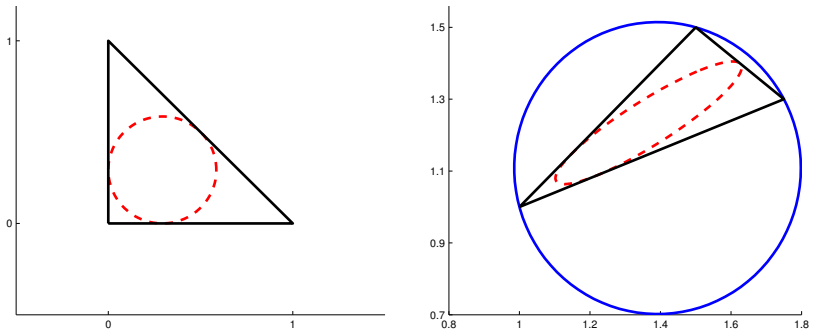

Figure 22: On the left: inscribed circle on the reference element. On the right: the image of this circle on the global element after the affine mapping and a circle containing the global element.

Theorem 10.4. *Let triangle* $T \sim \begin{bmatrix} \boldsymbol{n}_1 & \boldsymbol{n}_2 & \boldsymbol{n}_3 \end{bmatrix}$, $\widehat{T}$ *be the reference element satisfying* (83)*,* $A_T \in \mathbb{R}^{2\times 2}$, $\boldsymbol{b}_T \in \mathbb{R}^2$, *and* $F_T(\boldsymbol{x}) = A_T\boldsymbol{x} + \boldsymbol{b}_T$ *satisfy* $F_T(\hat{\boldsymbol{n}}_i) = \boldsymbol{n}_i$ *for* $i \in \{1, 2, 3\}$. *Then there holds that*

$$\|A_T\| \leq \frac{h_T}{\rho_{\widehat{T}}} \quad \textit{and} \quad \|A_T^{-1}\| \leq \frac{h_{\widehat{T}}}{\rho_T}.$$

Proof in Youtube

Proof. We will give the proof only for $\|A_T\|$. The result for $\|A_T^{-1}\|$ follows from identical arguments. By the definition of the norm (87) we have

$$\|A_T\| = \max_{\boldsymbol{x}\in\mathbb{R}^2, \|x\|=1} \|A_T\boldsymbol{x}\|.$$

To use the geometric argument, we make a change of variables $\boldsymbol{z} = \rho_{\widehat{T}}\boldsymbol{x}$. This gives

$$\|A_T\| = \frac{1}{\rho_{\widehat{T}}} \max_{\boldsymbol{z}\in\mathbb{R}^2, \|\boldsymbol{z}\|=\rho_{\widehat{T}}} \|A_T\boldsymbol{z}\|.$$

All vectors of length $\rho_{\widehat{T}}$ are generated as follows: Let C be inscribed circle of the reference triangle and $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \mathbb{R}^2$ points on opposite sides of the it's boundary, as in Fig. 21. Then any $\boldsymbol{z} \in \mathbb{R}^2$ satisfying $\|\boldsymbol{z}\| = \rho_{\widehat{T}}$ is obtained as $\boldsymbol{z} = \boldsymbol{u}_1 - \boldsymbol{u}_2$. By direct computation, there holds that

$$A_T\boldsymbol{z} = A_T\boldsymbol{u}_1 + \boldsymbol{b}_T - A_T\boldsymbol{u}_2 - \boldsymbol{b}_T = F_T(\boldsymbol{u}_1) - F_T(\boldsymbol{u}_2).$$

As $F_T(\boldsymbol{u}_1), F_T(\boldsymbol{u}_2) \in T$, it holds that $\|A_T\boldsymbol{z}\| \leq h_T$ and the result follows immediately. See Figure 22. □

Final estimate and shape regularity in Youtube

Using Theorem 10.4 we obtain more usefull form of the approximation result of Theorem 10.3.

Corollary 10.1. *Let* $T \in \mathcal{T}_h$. *Then there holds that*

$$|u - \pi u|_{H^1(T)} \leq C\frac{h_T^2}{\rho_T}|u|_{H^2(T)}.$$

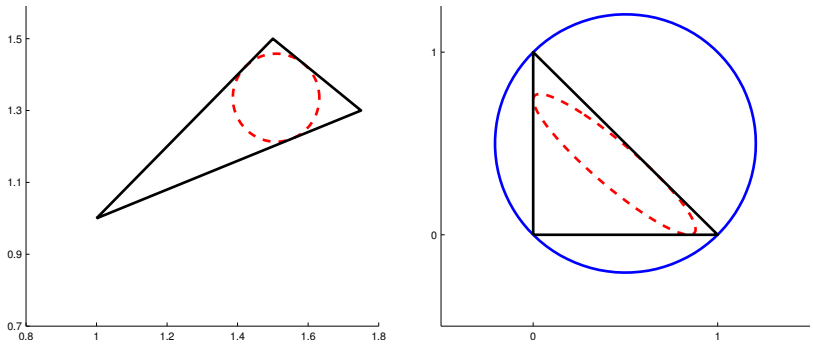

Figure 23: On the left: inscribed circle on the global element. On the right: the image of this circle on the reference element after the inverse affine mapping and a circle containing the reference element.

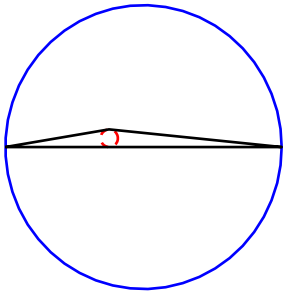

Figure 24: A badly shaped triangle. For this triangle, the ratio of $h_K \rho_K^{-1}$ is very large, which leads to large constants in the error estimates. This means that on a mesh containing such elements, error can be very large.

for any $u \in H^2(T)$ *and* C *independent of* u *and* T.

Proof. By Theorems 10.4 and 10.3

$$|u - \pi u|_{H^1(T)} \leq C \frac{h_{\hat{T}}}{\rho_{\hat{T}}^2} \frac{h_T^2}{\rho_T} |u|_{H^2(T)}.$$

Since the reference element will always remain the same, we include the term $h_{\hat{T}} \rho_{\hat{T}}^{-2}$ to the constant C. □

The upper bound stated in Corollary 10.1 depends on the shape of the triangle T. If the triangle is very flat, the term $h_T^2 \rho_T^{-1}$ can be large, regardless of the size of the triangle, see Figure 24.

Shape regularity

Usually, the convergence of finite element solution is considered using a series of refining meshes. In such a case, the interest is on the behavior of the error with

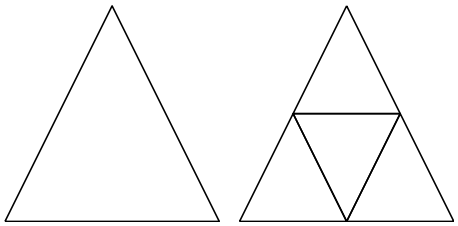

Figure 25: Division of a single triangle in the uniform mesh refinement.

respect to the mesh size h

$$h = \max_{T \in \mathcal{T}_h} h_T.$$

The series of meshes is sometimes referred as a *family of triangulations* $\{\mathcal{T}_h\}$. Above we noticed that the shape of the triangles has an effect on the error behavior, or at least on the approximation properties. Hence it is natural to place restrictions on the mesh, such as, to require that the family of triangulations is shape-regular.

Definition 10.3. *A family of triangulations* $\{\mathcal{T}_h\}$ *is called shape-regular, if there exists* $\gamma > 0$ *such that*

$$h_T \rho_T^{-1} \leq \gamma \quad \textit{for all } T \in \mathcal{T}_h \textit{ and } \mathcal{T}_h \in \{\mathcal{T}_h\}$$

The shape-regularity assumption guarantees, that the ratio $h_K \rho_K^{-1}$ stays bounded when the mesh size h varies, and the FE-error in the energy norm is characterised as

$$|u - u_h|_E \leq Ch|u|_2.$$

for C independent of h and u. Identical estimate holds in the $H^1(\Omega)$-norm but with different constant C.

Example 10.1. *The simplest example of a shape-regular family of triangulations is one generated by uniformly refining an initial triangulation. In the refinement, each triangle* K *is divided into four triangles with same shape, as in Figure 25. Since the shape of the triangles remains unchanged in the process, the family is clearly shape-regular. The mesh parameter* h *behaves as* $h_n = 2^{-n} h_0$*, where* n *is the number of refinements and* h_0 *is the mesh parameter for the initial triangulation. The first three triangulations of such a family are visualized in Figure 26.*

10.2 Problems

P45. (2p) Download a simple Matlab finite element solver util.zip. Proceed as follows:

(a) Run `demo_solver`.

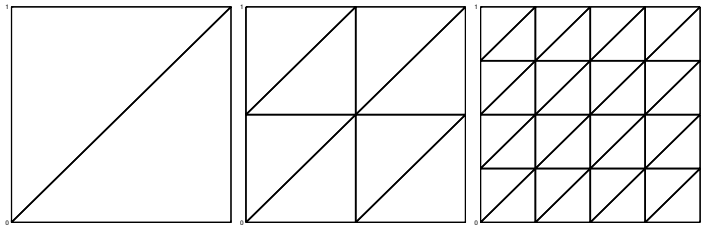

Figure 26: First three triangulations of family generated with uniform refinement.

(b) Modify the `demo_solver.m` to solve PDEs

$$-\Delta u = 1, \quad -\Delta u + u = 1, \quad \text{and} \quad -\nabla \cdot \begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix} \nabla u = 1$$

on $\Omega = (0,1)^2$ with zero Dirichlet boundary condition.

(c) Test commands `refine_tri` and `plot_2Dtri_mesh`

YOU MAY USE THIS CODE AS A BASIS IN THE FOLLOWING PROBLEMS

P46. (1p) Let $\Omega \subset \mathbb{R}^2$ and $\|\cdot\|_E$ be the energy norm associated to the bilinear form $a : H_0^1(\Omega) \times H_0^1(\Omega) \mapsto \mathbb{R}$ satisfying Assumptions 7.1 for $V = H_0^1(\Omega)$. Show that

(a) There exits $\alpha, \beta > 0$ independent of v such that

$$\alpha \|v\|_E \leq \|v\|_{H^1(\Omega)} \leq \beta \|v\|_E$$

for all $v \in H_0^1(\Omega)$.

(b) There exist C independent of v such that

$$\|v\|_E \leq C \|\nabla v\|_{L^2(\Omega)}$$

for any $v \in H_0^1(\Omega)$.

P47. (2p) Let $\Omega = (0, h_x) \times (0, h_y)$ and consider the Poincare-Friedrichs (P-F) inequality : There exists a constant $C(\Omega) > 0$ independent of u such that

$$\|u\|_{L^2(\Omega)} \leq C(\Omega) \|\nabla u\|_{L^2(\Omega)} \quad \forall u \in H_0^1(\Omega). \tag{88}$$

(a) Using the scaling argument, prove that $C(\Omega) \leq \hat{C}(\hat{\Omega}) \max\{h_x, h_y\}$, in which $\hat{C}(\hat{\Omega})$ is the constant in P-F for the domain $\hat{\Omega} = (0,1) \times (0,1)$.

(b) The smallest possible constant $C(\Omega)$ in the P-F inequality can be characterized as

$$C(\Omega)^{-2} = \min_{u \in H_0^1(\Omega)} G(u),$$

in which $G(u) = \frac{(\nabla u, \nabla u)}{(u,u)}$.

Show that $C(\Omega)^{-2}$ is the smallest eigenvalue λ_i of the problem: find $(\lambda_i, \varphi_i) \in (\mathbb{R}, H_0^1(\Omega))$ such that

$$(\nabla \varphi_i, \nabla v) = \lambda_i (\varphi_i, v) \quad \forall v \in H_0^1(\Omega).$$

Hint: the minimum is located at the critical point u that can be characterized as $\frac{d}{dt} G(u + tv)|_{t=0} = 0$ for all $v \in H_0^1(\Omega)$. Also, note that each eigenvalue satisfies $\lambda_i = \frac{(\nabla v_i, \nabla v_i)}{(v_i, v_i)}$, in which v_i is the eigenvector corresponding to λ_i.

(c) Use your finite element solver to approximate the constant $C(\Omega)$ for different values of $h_x \in (0, 1)$ and $h_y \in (0, 1)$. Plot the constant as a function of h_x and h_y. How good is your estimate?

Hint: The eigenvalue problem that you need to solve is $A\boldsymbol{x} = \lambda M \boldsymbol{x}$, in which A is the system matrix and M the mass matrix. The smallest eigenvalue can be solved with command the `eigs(A,M,1,'SM')`.

P48. (2p) Let $\Omega \subset \mathbb{R}^2$, $\mathcal{T}$ be a conforming triangular partition of Ω, and V_h the first order FE-space over $\mathcal{T}$.

(a) Let $v_h \in V_h$ and $v \in H^1(\Omega)$. Write a Matlab function for evaluating $\|\nabla(v - v_h)\|_{L^2(\Omega)}$ and $\|v - v_h\|_{L^2(\Omega)}$.

(b) Test your implementation on $\Omega = (0, 1)^2$. Use the mesh

$$p = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}, \quad t = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 4 & 4 \end{bmatrix},$$

and functions $v_h = \pi(x + 1)(y - 1)$, $v = (x + 1)(y - 1)$.

P49. (2p) Study the convergence of FE-method. Let $\Omega = (0, 1)^2$ and $\mathcal{T}_0$ be defined as

$$p = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad t = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 4 & 4 \end{bmatrix}.$$

Proceed as follows:

(a) Generate a sequence of triangular partitions $\mathcal{T}_0, \mathcal{T}_1, \ldots$ as follows: For $i \in \mathbb{N}$, the partition $\mathcal{T}_i$ is obtained from $\mathcal{T}_{i-1}$ using uniform refinement. Plot first few partitions. Hint : use util.zip

(b) Verify that the exact solution to (64) with $K = I$, and $f = 2y(1 - y) + 2x(1 - x)$ is $u = x(1 - x)y(1 - y)$.

(c) Solve (64) with $K = I$, and $f = 2y(1 - y) + 2x(1 - x)$ on $\{\mathcal{T}_1, \mathcal{T}_2, \ldots\}$ using your FE-solver. For each partition, compute the error in energy norm.

(d) Use an appropriate plot to study the dependency of FE-error in $L^2(\Omega)$ and $H^1(\Omega)$ norms on h. You may use longest edge in the mesh instead of the largest radius of the circumcircle.

10.3 Regularity of the solution

Regularity in Youtube

Consider the weak form of the Poisson's equation: find $u \in H_0^1(\Omega)$ satisfying

$$\int_\Omega \nabla u \cdot \nabla v = \int_\Omega fv \quad \text{for all } v \in H_0^1(\Omega). \tag{89}$$

The FE-error analysis relied on additional $H^2(\Omega)$-regularity of the solution u, this is, we assumed that $u \in H^2(\Omega)$, i.e.,

$$|u|_{H^2(\Omega)} < \infty.$$

In this Section we discuss how this regularity assumption depends on the domain Ω and the load function f. Detailed regularity analysis of elliptic PDEs is beyond the scope of this course, we refer interested readers to *Lawrence C. Evans: Partial Differential Equations* or *Pierre Grisvard: Singularities in boundary value problems.*

We have the following Theorem.

Theorem 10.5. *Assume that $\Omega \subset \mathbb{R}^2$ satisfies Assumptions 9.1 and is convex. Then the solution $u \in H_0^1(\Omega)$ to problem* (89) *with $f \in L^2(\Omega)$ is H^2-regular and satisfies*

$$\|u\|_{H^2(\Omega)} \leq C\|f\|_{L^2(\Omega)}$$

for some C independent of f.

The weak form of Poisson's equation (89) is an instance of the abstract problem treated in Section 8.1: Find $u \in H_0^1(\Omega)$ satisfying

$$a(u, v) = L(v) \quad \text{for all } v \in H_0^1(\Omega).$$

This problem is well defined, if the load functional $L(v) : H_0^1(\Omega) \mapsto \mathbb{R}$ satisfies Assumption 7.1, i.e., there exists C_L independent of v such that

$$|L(v)| \leq C_L\|v\|_{H^1(\Omega)} \quad \text{for all } v \in H_0^1(\Omega).$$

It is possible to construct a load functional $L(v)$ that cannot be expressed as $\int_\Omega fv$ for any $v \in H_0^1$, and the corresponding solution u is not in $H^2(\Omega)$ regardless of the domain Ω.

Assuming the load function $f \in L^2(\Omega)$, the regularity of the solution depends on the geometric properties of Ω. Let Ω satisfy Assumptions 9.1, or Ω is a

simply connected polyhedral domain without holes. If Ω is in addition convex the solution $u \in H^2(\Omega)$. For non-convex polyhedral domains, the behavior of the solution depends on the angle between two boundary segments. Let $\boldsymbol{v} \in \mathbb{R}^2$ be the intersection point of two boundary segments. Near $\boldsymbol{v}$, the solution behaves as

$$\mathcal{O}(r^\alpha), \quad \alpha = \frac{\pi}{\omega},$$

in which ω is the angle between the two segments measured inside the domain and r is the radial coordinate from the intersection point. This is, there exists $k_1, k_2 > 0$ such that

$$k_1 r^\alpha \leq u(\boldsymbol{v} + r \begin{bmatrix} \cos\varphi \\ \sin\varphi \end{bmatrix}) \leq k_2 r^\alpha$$

for sufficiently small r and such φ that $\boldsymbol{v} + r \begin{bmatrix} \cos\varphi \\ \sin\varphi \end{bmatrix} \in \Omega$.

If $\omega > \pi$, the u solution behaves as $\mathcal{O}(r^\alpha)$, $\alpha < 1$. Hence the second derivative behaves as $\mathcal{O}(r^{\alpha-2})$, $\alpha < 1$, and

$$\int_0^1 (r^{\alpha-2})^2 \, r \, dr = \int_0^1 r^{2\alpha-3} \, dr \to \infty, \quad \alpha < 1,$$

which shows that the second derivative is not square integrable, thus, $u \notin H^2(\Omega)$. To characterize such a solution, fractional Sobolev spaces H^s, $s \in \mathbb{R}$, have been defined to fill the gap between e.g. H^1 and H^2. These spaces are out of the scope of the current course. Intuitively speaking, a polynomial x^α belongs to the space $H^{s-\delta}$, $\delta > 0$, if $\alpha > s - 1$. Accordingly, the function r^α belongs to the Sobolev space $H^{1+\alpha-\delta}(\Omega)$, $\delta > 0$.

The FE-error estimate was based on local interpolation error estimate. Assume that $u \in H^2(\Omega)$. Then the local interpolation error satisies

$$|u - \pi u|_{H^1(T)} \leq C h_T |u|_{H^2(T)}$$

for every triangle $T \in \mathcal{T}_h$. Combining this estimate with Cea's Lemma led to error estimate for the FE-solution. If $u \in H^{1+s}(\Omega)$, $s \leq 1$, similar estimate holds. In particular,

$$|u - \pi u|_{H^1(T)} \leq C h_T^s |u|_{H^{1+s}(T)} \quad \forall u \in H^{1+s}(\Omega).$$

Combining this estimate with Cea's Lemma led to error estimate for the FE-solution. The convergence of the first order finite element method is of order $O(h^s)$ for functions with $H^{1+s}(\Omega)$ regularity.

Example 10.2. *A simple example of the effects of regularity to convergence of FEM is the L-shaped domain,* $(-1,1)^2 \setminus (-1,0) \times (0,1)$. *The angle between the segments at* $(0,0)$ *is* $\omega = \frac{3\pi}{2}$. *Hence, for* $f \in L^2$, *the solution behaves*

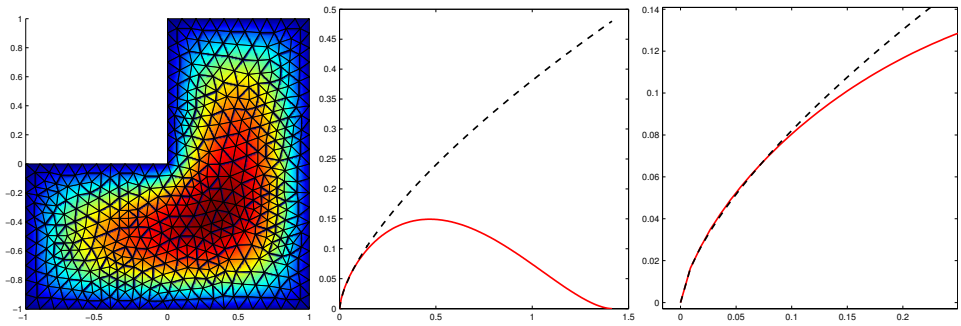

Figure 27: The solution with load $f = 1$. On left: in the whole domain. Two on the right: solution on the line $y = -x$ with red line and reference $O(r^{2/3})$ with dotted black line. Clearly, the solution behaves as $r^{2/3}$ near the corner.

as $\mathcal{O}(r^{2/3})$ close to the origin. Due to this, the solution $u \in H^{5/3-\delta}$, $\delta > 0$. Hence the first order finite element solution convergences with the rate of $h^{2/3}$. The solution with load $f = 1$ is visualized in Figure 27. In the same figure is also the behavior of this solution close to the singular point $(0, 0)$.

10.4 Problems

P50. (2p) Let $\Omega = (0, 1)^2$ and consider the problem : Find $u \in H_0^1(\Omega)$ such that

$$\int_\Omega \nabla u^T \nabla v = -\int_\Omega \left(\psi''(x)\psi(y) + \psi''(y)\psi(x)\right) v, \tag{90}$$

where $\psi(t) = t(1 - t)$. The exact solution to the above problem is $u = \psi(x)\psi(y)$. Study the effect of element shape to convergence of FE-solution. Proceed as follows :

(a) Download mesh generation functions `make_bad_LA2_mesh.m`, `make_bad_LA_mesh.m`, and `make_good_LA_mesh.m`.

(b) Implement a function for evaluating the error between exact and FE-solution to (90).

(c) Solve the problem using sequence of meshes generated using function given in (a) with parameter values $N \in \{1, 2, 3, 4, 5, 6\}$. For each mesh, evaluate the error and mesh size h (longest edge in the mesh).

(d) Plot the errors on *loglog* - scale as a function of h. What do you observe?

P51. (2p) Let the parameter $\omega \in (0, 2\pi)$ and define the domain Ω in polar coordinates as

$$\Omega := \{\ (r, \theta) \mid 0 \leq r < 1 \text{ and } 0 < \theta < \omega\ \}. \tag{91}$$

(a) Create a mesh $\mathcal{T}_0$ for the domain Ω. Hint: you can specify the domain Ω to Matlab `pdetool` as $B(0,1) \setminus T$, where $B(0,1)$ is the unit circle and T a triangle (or rectangle) with appropriately chosen corner points.

(b) Refine the mesh $\mathcal{T}_0$. If you generated the mesh using Matlab PDE-tool, use the command `refinemesh.m`. Otherwise, use function `refine_tri` as follows: after each refinement, locate all nodes on the boundary $\Gamma_C := \partial\Omega \setminus \{ (r,\theta) \mid 0 < r < 1 \text{ and } \theta \in \{0, \omega\} \}$. To obtain accurate approximation of Ω, move these nodes manually to the surface of the circle with radius 1.

(c) Test your workflow by generating meshes corresponding to $\omega \in \{\frac{\pi}{2}, \pi, \frac{3\pi}{2}, \frac{7\pi}{4}\}$. Return your code and example pictures of your meshes.

P52. (4p) Let the parameter $\omega \in (0, 2\pi)$, domain Ω be as defined in (91), and $f(\theta) := \sin \frac{\pi}{\omega}\theta$. Consider the problem: Find $u \in H^1_0(\Omega)$ such that

$$(\nabla u, \nabla v) = (f, v) \quad \text{for all } v \in H^1_0(\Omega).$$

Study the effect of ω to FE-convergence rate. Proceed as follows:

(a) Show that:

$$\|\nabla u - \nabla v\|^2_{L^2(\Omega)} = 2\left(J(v) - J(u)\right). \tag{92}$$

for any $v \in H^1_0(\Omega)$.

(b) Solve the problem using FEM on family of refining meshes. For each mesh, compute the error in the energy norm using (92) and reference energies $J(u)$ listed in Table 2.

Table 2: Reference energies $J(u)$ for P52

ω	$J(u)$
$\frac{\pi}{2}$	-0.006135923100600
π	-0.021816615503775
$\frac{3\pi}{2}$	-0.041417468112674
$\frac{7\pi}{4}$	-0.051965862140141

Plot the error as a function of the mesh size h (longest edge in the mesh) using the loglog-scale. How does the convergence rate depend on ω ?

THE END

A Mathematics toolbox

In this section, we recall the definitions of vector space, function space, basis, as well as differentiable and integrable functions. These definitions are useful for understanding the finite element method. In addition, we recall how integrals are approximately computed over one dimensional interval.

A.1 Function space and basis

Finite element solver finds an approximate solution to a PDE from a set of functions. This set is a finite dimensional function space, thus, it can defined to the solver by its basis. In this section, we recall the definitions of function space and basis. For simplicity, we limit the discussion to a simple one dimensional case.

Let $a, b \in \mathbb{R}, a < b$ and consider the interval $I = (a, b)$. In this note, we are concerned with sets of functions $f : I \mapsto \mathbb{R}$. The sum and multiplication of functions is understood point-wise: let $f, g : I \mapsto \mathbb{R}$, $\alpha \in \mathbb{R}$, and define

$$(f + g)(x) = f(x) + g(x) \quad \text{and} \quad (\alpha f)(x) = \alpha f(x)$$

for each $x \in I$.

Function space Let S be a set of functions from interval I to scalars, this is

$$S \subset \{f : I \mapsto \mathbb{R}\}.$$

Let $f, g \in S$ and $\alpha \in \mathbb{R}$. If

$$f + g \in S \quad \text{and} \quad \alpha f \in S,$$

the set S is closed under addition and multiplication by scalar. Set S satisfying these properties is called as vector space. Vector spaces are a central structure in mathematics and appear in many context. A familiar example of a vector space is $\mathbb{R}^n$.

In the following, vector spaces of functions as are called as function spaces.

Example A.1. *The set of degree n polynomials from $(0, 1) \mapsto \mathbb{R}$ is defined as*

$$P^n := \left\{ \sum_{j=0}^{n} a_j x^j \mid a_j \in \mathbb{R} \quad \textit{for} \quad j \in \{0, \ldots, n\} \right\}.$$

Let $\alpha \in \mathbb{R}$ and $f, g \in P^n$. Then there exists coefficients $a_j, b_j \in \mathbb{R}$ for $j \in \{0, \ldots, n\}$ such that

$$f(x) = \sum_{j=0}^{n} a_j x^j \quad \textit{and} \quad g(x) = \sum_{i=0}^{n} b_j x^j.$$

The sum of f and g satisfies

$$(f+g)(x) = f(x) + g(x) = \sum_{j=0}^{n} (a_j + b_j) x^j$$

for any $x \in (a,b)$, hence, $f+g \in P^n$. Similarly,

$$(\alpha f)(x) = \alpha f(x) = \sum_{j=0}^{n} \alpha a_j x^j \in P^n.$$

Thus, the set P^n is closed under addition and multiplication by a scalar and it is a function space.

Linear independence. Let the set V be a vector space, $\{v_1, \ldots, v_n\} \subset V$, and $\boldsymbol{\alpha} \in \mathbb{R}^n$. We call the set $\{v_1, \ldots, v_n\}$ as *linearly independent*, if

$$\sum_{j=1}^{n} \alpha_j v_j = 0 \Rightarrow \alpha_j = 0 \quad \text{for} \quad j \in \{1, \ldots, n\}. \tag{93}$$

This is, any v_j cannot be written as a weighted sum, or linear combination, of elements in $\{v_1, \ldots, v_n\} \setminus v_j$.

Example A.2. *Consider $\mathbb{R}^3$ and let*

$$\boldsymbol{v}_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \boldsymbol{v}_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \quad \textit{and} \quad \boldsymbol{v}_3 = \begin{bmatrix} 1 \\ 4 \\ 3 \end{bmatrix}.$$

Study if the set $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3\}$ is linearly independent. This is, investigate which $\boldsymbol{\alpha} \in \mathbb{R}^3$ satisfies the condition

$$\alpha_1 \boldsymbol{v}_1 + \alpha_2 \boldsymbol{v}_2 + \alpha_3 \boldsymbol{v}_3 = 0. \tag{94}$$

Define the matrix $A \in \mathbb{R}^{3\times 3}$ as

$$A = \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \boldsymbol{v}_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 4 \\ 0 & 1 & 3 \end{bmatrix}.$$

Using the definition of matrix-vector product, Equation (94) *is rewritten as $A\boldsymbol{\alpha} = 0$. Recall that the equation $A\boldsymbol{\alpha} = 0$ has only the zero solution $\boldsymbol{\alpha} = 0$ if $\det A \neq 0$, and infinitely many nonzero solutions if $\det A = 0$. By direct calculation*

$$\det \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 4 \\ 0 & 1 & 3 \end{bmatrix} = 0,$$

and there exists nonzero solution $\boldsymbol{\alpha}$ to (94)*. Thus, $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3\}$ are linearly dependent.*

Example A.3. *Consider the space of degree* 4 *polynomials* P^4*, and the set*

$$S = \{1 - t^2, 1 + t^3, t^2 + t^3\} \subset P^4.$$

Study, if S *is linearly independent. This is, investigate which* $\boldsymbol{\alpha} \in \mathbb{R}^3$ *satisfies*

$$\alpha_1(1 - t^2) + \alpha_2(1 + t^3) + \alpha_3(t^2 + t^3) = 0.$$

Regrouping the terms gives

$$(\alpha_1 + \alpha_2) + (\alpha_3 - \alpha_1)\, t^2 + (\alpha_2 + \alpha_3)\, t^3 = 0.$$

The above equation is satisfied when

$$\begin{cases} \alpha_1 + \alpha_2 = 0 \\ \alpha_3 - \alpha_1 = 0 \\ \alpha_2 + \alpha_3 = 0 \end{cases} \qquad \textit{or} \qquad \begin{bmatrix} 1 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \boldsymbol{\alpha} = 0. \tag{95}$$

The determinant of the coefficient matrix in (95) *is zero. Hence, there exists nonzero solutions* $\boldsymbol{\alpha}$ *to* (95)*, and we conclude that the set of functions* S *is linearly dependent.*

Basis A basis of vector space V is a linearly indepedent set $\{q_1, \ldots, q_n\} \subset V$ such that any $v \in V$ can be written as

$$v = \sum_{j=1}^{n} \alpha_j q_j \quad \text{for some coefficient vector } \boldsymbol{\alpha} \in \mathbb{R}^n.$$

The coefficient vector $\boldsymbol{\alpha} \in \mathbb{R}^n$ is called as the *coordinate* of v in the basis $\{q_1, \ldots, q_n\}$. There exists several basis for the same space. The number of elements in each possible basis is the same, and called as the *dimension* of V.

In program code, function spaces are defined using their to *basis*. We will only use *finite-dimensional* function spaces that have a finite number of basis vectors. A familiar example of a basis are the cartesian unit vectors $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$ in $\mathbb{R}^n$ satisfying $\boldsymbol{e}_i \in \mathbb{R}^n$ and

$$(\boldsymbol{e}_i)_j = \begin{cases} 1 & i = j \\ 0 & otherwise \end{cases}.$$

Example A.4. *Each of the sets*

$$\{1, t\}, \quad \{1 - t, 1 + t\}, \quad \{1, 1 - t\}, \quad \textit{and} \quad \{t, t - 1\}$$

is an example of a basis for the space of first order polynomials P^1*. Clearly, the dimension of this space is two, i.e.,* $dim(P^1) = 2$*.*

A.2 Derivative and Integral

When working with partial differential equations, it is important to have exact definition for derivative and integral as well as differentiable and integrable functions. These concepts are discussed very briefly in this section. We also define function spaces of differentiable and integrable functions that are widely used in mathematical literature dealing with PDEs. For simplicity, the discussion is limited to functions from interval $I \subset \mathbb{R}$ to real numbers.

Derivative Let $I = (a, b) \subset \mathbb{R}$ and $f : I \mapsto \mathbb{R}$. Recall, that the function f is differentiable at $x_0 \in I$ with derivative $f'(x_0)$, if the limit

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} \tag{96}$$

is a real number. The function f is said to be differentiable on the open set I, if it is differentiable at every $x_0 \in I$. If this is the case, Eq. (96) defines the derivative function $f' : I \mapsto \mathbb{R}$ by specifying it's pointwise values. One can now ask, for what kind of functions is this process well defined ?

In PDE theory, the differentiation process is used to define sets of functions whose derivatives have desired properties. Particularly, one is interested in functions that have one or several continuous derivatives over interval I. Interestingly, f' can be well defined on I but not a continuous function. If f is differentiable on I and the derivative function f' is continuous in I, we call f as continuously differentiable.

The set of continuously differentiable functions from I to real numbers is denoted by $C^1(I)$. The set $C^1(I)$ is a function space, this is, let $f, g \in C^1(I)$ and $\alpha \in \mathbb{R}$. Then

$$(f + g) \in C^1(I) \quad \text{and} \quad (\alpha f) \in C^1(I).$$

Similarly, function f is said to be n-times continuously differentiable if each $f^{(n-1)}$, is continuously differentiable. The space of n-times continously differentiable functions is denoted by $C^n(I)$. Functions that have infinitely many derivatives form the space $C^\infty(I)$. The space $C_0^\infty(I)$ denotes the function space of those functions in C^∞ that are non-zero only on some closed subset of I. Particularly, $v \in C([0, 1]) \cap C_0^\infty(0, 1)$ satisfies $v(0) = v(1) = 0$.

Integral When writing the integral

$$\int_a^b f \, dx,$$

we intuitively think about the area under the graph of function f, see Fig. 28. However, it is not straightforward to give a mathematically exact definition for

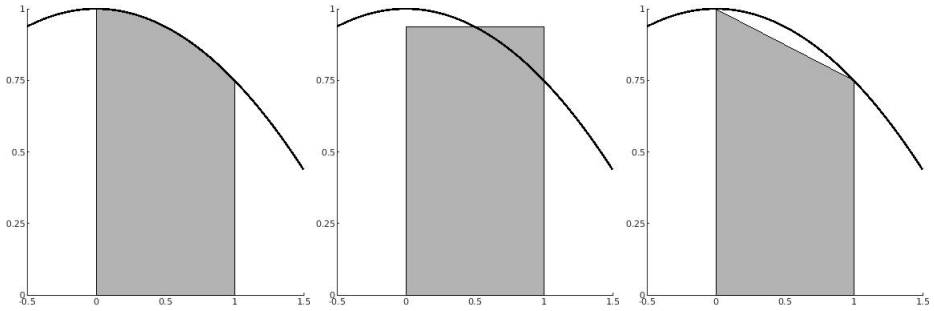

Figure 28: From left: example of definite integral, approximation by midpoint rule, approximation by trapezoidal rule

integral or design a process that computes the area. A natural idea is define a sequence of improving approximations $\{s_i\} \subset \mathbb{R}$ to $\int_a^b f \ dx$ and define the integral as a limit of s_i, when $i \to \infty$. The Riemann and Lebesque integration theories are based on such ideas. For example, the intuitive idea of Riemann integration theory is to define the sequence of approximations by bars as in Fig. 29. The challenge is to prove the convergence of the sequence $\{s_i\}$, especially for irregular functions f.

In the theory of PDEs, functions are often often approximated by a sequence of smooth functions. For example, consider function f that is approximated by pointwise converging sequence $\{f_n\} \subset C_0^\infty(0,1)$ as in Fig. 30. Many proofs in PDE theory require evaluation of limits such as

$$\lim_{n\to\infty} \int_a^b (f_n' - f')^2 \ dx. \tag{97}$$

Recall that the integration process itself is defined by taking a limit. Hence, it is not automatic that the order of limit and integral in (97) can be exchanged.

The Lebesque integration theory gives tools for investigating when the limit and integral can be exchanged. Such tools do not exist in Riemann integration theory. Due to this, PDE theory uses Lebesque integration process, however, it is not in the focus of this lecture note. We do not pay attention to limiting processes such as one in (97).

The Lebesque integral is used to define function spaces that are essential for us. Namely, we use the function space $L^2(0,1)$ consisting of all square integrable functions, i.e.

$$L^2(0,1) := \{ f : (0,1) \mapsto \mathbb{R} \mid \int_0^1 f^2 \ dx < \infty \ \}.$$

In the above expression, the integral is to be understood in the sense of Lebesque.

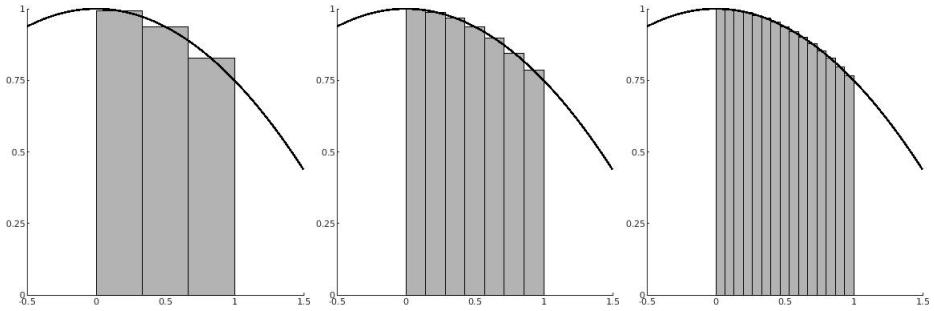

Figure 29: A sequence of improving approximations to definite integral $\int_0^1 f \; dx$. The sequence is defined by applying the midpoint rule over each interval.

Numerical Integration The finite element solver evaluates integrals approximately using numerical integration rules. The information on the applied integration rule is expressed by a pair of vectors $(\boldsymbol{w}, \boldsymbol{t}) \in \mathbb{R}^M \times \mathbb{R}^M$. Given a pair $(\boldsymbol{w}, \boldsymbol{t}) \in \mathbb{R}^M \times \mathbb{R}^M$ associated to $I = (a, b)$, we approximate

$$\int_a^b f(x) \; dx \approx \sum_{k=1}^{M} f(t_k) w_k. \tag{98}$$

Let $h := b - a$. Two simple examples of numerical integration methods are the trapezoidal rule,

$$\boldsymbol{t}_{tr} = \begin{bmatrix} a \\ b \end{bmatrix} \quad \boldsymbol{w}_{tr} = \frac{h}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{99}$$

and the midpoint rule

$$\boldsymbol{t}_{mp} = \frac{1}{2}(a + b) \quad \boldsymbol{w}_{mp} = h. \tag{100}$$

Both of these rules are based on geometric approximation of the function over (a, b), see Fig. 28. The midpoint rule is used in our example implementation of one dimensional finite element solver.

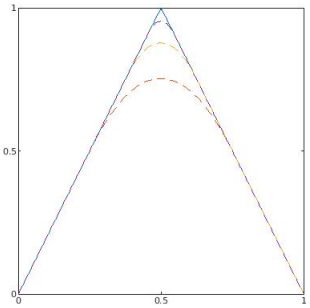

Figure 30: Dashed lines denote three first terms form a sequence of improving smooth approximations of function f indicated by solid line.