~

Triangulations, finite elements, finite volumes

Scientific Computing Winter 2016/2017

Part III

With material from A. Ern, J.-L. Guermond: "Theory and Practice of Finite Elements"

Jürgen Fuhrmann

juergen.fuhrmann@wias-berlin.de

made wit pandoc

Delaunay triangulations

- Given a finite point set $X \subset \mathbb{R}^d$. Then there exists simplicial a complex called *Delaunay triangulation* of this point set such that
  - X is the set of vertices of the triangulation
  - The union of all its simplices is the convex hull of X.
  - (Delaunay property): For any given d-simplex $\Sigma \subset \Omega$ belonging to the triangulation, the interior of its circumsphere does not contain any vertex $x_k \in X$.
- Assume that the points of X are in general position, i.e. no $n+2$ points lie on one sphere. Then the Delaunay triangulation is unique.

Voronoi diagram

- Given a finite point set $X \subset \mathbb{R}^d$. Then the Voronoi diagram is a partition of $\mathbb{R}^d$ into convex nonoverlapping polygonal regions defined as

$$\mathbb{R}^d = \bigcup_{k=1}^{N_x} V_k$$

$$V_k = \{x \in \mathbb{R}^d : ||x - x_k|| < ||x - x_l|| \forall x_l \in X, l \neq k\}$$

Voronoi - Delaunay duality

- Given a point set $X \subset \mathbb{R}^d$ in general position. Then its Delaunay triangulation and its Voronoi diagram are dual to each other:
  - Two Voronoi cells V_k, V_l have a common facet if and only if $\overline{x_k x_l}$ is an edge of the triangulation.

Boundary conforming Delaunay triangulations

- Domain $\Omega \subset \mathbb{R}^n$ (we will discuss only $n = 2$) with polygonal boundary $\partial\Omega$.
- Partition (triangulation) $\Omega = \bigcup_{s=1}^{N_\Sigma} \Sigma$ into non-overlapping simplices Σ_s such that this partition represents a simplicial complex. Regard the set of nodes $X = \{x_1 \dots x_{N_x}\}$.
- It induces a partition of the boundary into lower dimensional simplices: $\partial\Omega = \bigcup_{t=1}^{N_\sigma} \sigma_t$. We assume that in 3D, the set $\{\sigma_t\}_{t=1}^{N_\sigma}$ includes all edges of surface triangles as well. For any given lower ($d-1$ or $d-2$) dimensional simplex σ, its *diametrical sphere* is defined as the smallest sphere containing all its vertices.
- *Boundary conforming Delaunay property*:
  - (Delaunay property): For any given d-simplex $\Sigma_s \subset \Omega$, the interior of its circumsphere does not contain any vertex $x_k \in X$.
  - (Gabriel property) For any simplex $\sigma_t \subset \partial\Omega$, the interior of its diametrical sphere does not contain any vertex $x_k \in X$.
- Equivalent formulation in 2D:
  - For any two triangles with a common edge, the sum of their respective angles opposite to that edge is less or equal to 180°.
  - For any triangle sharing an edge with $\partial\Omega$, its angle opposite to that edge is less or equal to 90°.

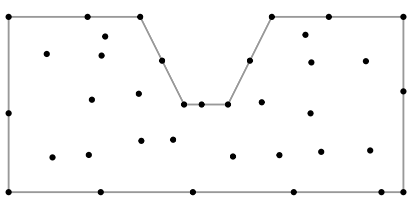
Restricted Voronoi diagram

- Given a boundary conforming Delaunay discretization of Ω, the *restricted Voronoi diagram* consists of the *restricted Voronoi cells* corresponding to the node set X defined by

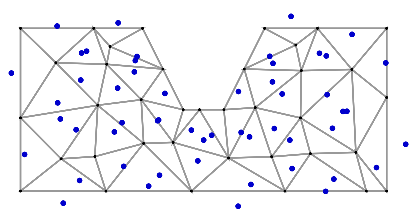
$$\omega_k = V_k \cap \Omega = \{x \in \Omega : ||x - x_k|| < ||x - x_l|| \forall x_l \in X, l \neq k\}$$

- These restricted Voronoi cells are used as control volumes in a finite volume discretization

Piecewise linear description of computational domain with given point cloud

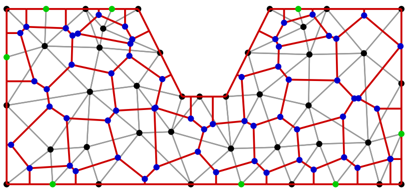


Delaunay triangulation of domain and triangle circumcenters.



- Blue: triangle circumcenters
- Some boundary triangles have larger than 90° angles opposite to the boundary $\Rightarrow$ their circumcenters are outside of the domain

Boundary conforming Delaunay triangulation



- Automatically inserted additional points at the boundary (green dots)
- Restricted Voronoi cells (red).

General approach to triangulations

- Obtain piecewise linear descriptiom of domain
- Call mesh generator (triangle, TetGen, NetGen ...) in order to obtain triangulation
- Performe finite volume or finite element discretization of the problem.

Alternative way:

- Construction "by hand" on regular structures

~

Partial Differential Equations

DIfferential operators

- Bounded domain $\Omega \subset \mathbb{R}^d$, with piecewise smooth boundary
- Scalar function $u : \Omega \to \mathbb{R}$
- Vector function $\mathbf{v} : \Omega \to \mathbb{R}^d$
- Write $\partial_i u = \frac{\partial u}{x_i}$
- For a multindex $\alpha = (\alpha_1 \dots \alpha_d)$, write $|\alpha| = \alpha_1 + \dots + \alpha_d$ and define $\partial^\alpha u = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}$
- Gradient $\mathrm{grad} = \nabla$: $u \mapsto \nabla u = \begin{pmatrix} \partial_1 u \\ \vdots \\ \partial_d u \end{pmatrix}$
- Divergence $\mathrm{div} = \nabla \cdot$: $\mathbf{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_d \end{pmatrix} \mapsto \nabla \cdot \mathbf{v} = \partial_1 v_1 + \dots + \partial_d v_d$
- Laplace operator $\Delta = \mathrm{div} \cdot \mathrm{grad} = \nabla \cdot \nabla$: $u \mapsto \Delta u = \partial_{11} u + \dots + \partial_{dd} u$

Matrices from PDE revisited

Given:

- Domain $\Omega = (0, X) \times (0, Y) \subset \mathbb{R}^2$ with boundary $\Gamma = \partial\Omega$, outer normal $\mathbf{n}$
- Right hand side $f : \Omega \to \mathbb{R}$
- "Conductivity" λ
- Boundary value $v : \Gamma \to \mathbb{R}$
- Transfer coefficient α

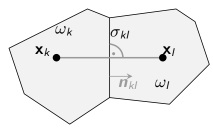
Search function $u : \Omega \to \mathbb{R}$ such that

$$\begin{aligned} -\nabla \cdot \lambda \nabla u &= f \quad \text{in}\Omega \\ \lambda \nabla u \cdot \mathbf{n} + \alpha(u - v) &= 0 \quad \text{on}\Gamma \end{aligned}$$

- Example: heat conduction:
  - u: temperature
  - f: volume heat source
  - λ: heat conduction coefficient
  - v: Ambient temperature
  - α: Heat transfer coefficient

The finite volume idea revisited

- Assume Ω is a polygon
- Subdivide the domain Ω into a finite number of **control volumes** : $\bar{\Omega} = \bigcup_{k \in \mathcal{N}} \bar{\omega}_k$ such that
  - ω_k are open (not containing their boundary) convex domains
  - $\omega_k \cap \omega_l = \emptyset$ if $\omega_k \neq \omega_l$
  - $\sigma_{kl} = \bar{\omega}_k \cap \bar{\omega}_l$ are either empty, points or straight lines
    - we will write $|\sigma_{kl}|$ for the length
    - if $|\sigma_{kl}| > 0$ we say that ω_k, ω_l are neigbours
    - neigbours of ω_k: $\mathcal{N}_k = \{l \in \mathcal{N} : |\sigma_{kl}| > 0\}$
- To each control volume ω_k assign a **collocation point**: $\mathbf{x}_k \in \bar{\omega}_k$ such that
  - **admissibility condition**: if $l \in \mathcal{N}_k$ then the line $\mathbf{x}_k \mathbf{x}_l$ is orthogonal to σ_{kl}
  - if ω_k is situated at the boundary, i.e. $\gamma_k = \partial\omega_k \cap \partial\Omega \neq \emptyset$, then $\mathbf{x}_k \in \partial\Omega$



- Now, we know how to construct this partition
  - obtain a boundary conforming Delaunay triangulation
  - construct restricted Voronoi cells

Discretization ansatz

- Given control volume ω_k, integrate equation over control volume

$$\begin{aligned} 0 &= \int_{\omega_k} (-\nabla \cdot \lambda \nabla u - f)\, d\omega \\ &= -\int_{\partial\omega_k} \lambda \nabla u \cdot \mathbf{n}_k d\gamma - \int_{\omega_k} f d\omega \qquad \text{(Gauss)} \\ &= -\sum_{L \in \mathcal{N}_k} \int_{\sigma_{kl}} \lambda \nabla u \cdot \mathbf{n}_{kl} d\gamma - \int_{\gamma_k} \lambda \nabla u \cdot \mathbf{n} d\gamma - \int_{\omega_k} f d\omega \\ &\approx \sum_{L \in \mathcal{N}_k} \frac{\sigma_{kl}}{h_{kl}} (u_k - u_l) + |\gamma_k| \alpha (u_k - v_k) - |\omega_k| f_k \end{aligned}$$

- Here,
  - $u_k = u(\mathbf{x}_k)$
  - $v_k = v(\mathbf{x}_k)$
  - $f_k = f(\mathbf{x}_k)$

Solvability of discrete problem

- $N = |\mathcal{N}|$ equations (one for each control volume)
- $N = |\mathcal{N}|$ unknowns (one in each collocation point $\equiv$ control volume)
- Graph of discretzation matrix $\equiv$ edge graph of triangulation $\Rightarrow$ matrix is irreducible
- Matrix is symmetric
- Main diagonal entries are positive, off diagonal entries are non-positive
- The matrix is diagonally dominant
- For positive heat transfer coefficients, the matrix becomes irreducibly diagonally dominant

$\Rightarrow$ the discretization matrix has the M-property.

Note on matrix M property and discretization methods

- Finite volume methods on boundary conforming Delaunay triangulations can be *practically* constructed on large classes of 2D and 3D polygonal domains using *provable* algorithms
  - Results mostly by J. Shewchuk (triangle) and H. Si (TetGen)
- Later we will discuss the finite element method. It has a significantly simpler convergence theory than the finite volume method.
  - For constant heat conduction coefficients, in 2D it yields the same discretization matrix as the finite volume method.
  - However this is *not true* in 3D.
  - Consequence: there is no provable mesh construction algorithm which leads to the M-Propertiy of the finite element discretization matrix in 3D.

Convergence theory

For an excurse into convergence theory, we need to recall a number of concepts from functional analysis.

See e.g. Appendix of the book of Ern/Guermond.

Lebesgue integral, $L^1(\Omega)$ I

- Let Ω have a boundary which can be represented by continuous, piecewise smooth functions in local coordinate systems, without cusps and other+ degeneracies (more precisely: Lipschitz domain).
  - Polygonal domains are Lipschitz.
- Let $C_c(\Omega)$ be the set of continuous functions $f : \Omega \to \mathbb{R}$ with compact support.
- For these functions, the Riemann integral $\int_\Omega f(x)dx$ is well defined, and $||f|| := \int_\Omega |f(x)|dx$ provides a norm, and induces a metric
- A Cauchy sequence is a sequence f_n of functions where the norm of the difference between two elements can be made arbitrarily small by increasing the element numbers:

$$\forall \varepsilon > 0\ \exists N \in \mathbb{N} : \forall m, n > n, ||f_n - f_m|| < \varepsilon$$

- All convergent sequences of functions are Cauchy sequences
- A metric space is *complete* if all Cauchy sequences of its element have a limit within this space

Lebesgue integral, $L^1(\Omega)$ II

- Let $L^1(\Omega)$ be the completion of $C_c(\Omega)$ with respect to the metric defined by the integral norm, i.e. "include" all limites of Cauchy sequences
- Defined via sequences, $\int_\Omega |f(x)|dx$ is defined for all functions in $L^1(\Omega)$.
- Equality of L^1 functions is elusive as they are not necessarily continuous: best what we can say is that they are equal "almost everywhere".
- Examples for Lebesgue integrable (measurable) functions:
  - Step functions
  - Bounded functions continuous except in a finite number of points

Spaces of integrable functions

- For $1 \leq p \leq \infty$, let $L^p(\Omega)$ be the space of measureable functions such that

$$\int_\Omega |f(x)|^p dx < \infty$$

equipped with the norm

$$||f||_p = \left(\int_\Omega |f(x)|^p dx\right)^{\frac{1}{p}}$$

- These spaces are *Banach spaces*, i.e. complete, normed vector spaces.
- The space $L^2(\Omega)$ is a *Hilbert space*, i.e. a Banach space equipped with a scalar product $(\cdot, \cdot)$ whose norm is induced by that scalar product, i.e. $||u|| = \sqrt{(u, u)}$. The scalar product in L^2 is

$$(f, g) = \int_\Omega f(x)g(x)dx.$$

Green's theorem

- Green's theorem for *smooth* functions: Let $u, v \in C^1(\bar{\Omega})$ (continuously differentiable). Then for $\mathbf{n} = (n_1 \dots n_d)$ being the outward normal to Ω,

$$\int_\Omega u\partial_i v dx = \int_{\partial\Omega} uvn_i ds - \int_\Omega v\partial_i u dx$$

In particular, if $v = 0$ on $\partial\Omega$ one has

$$\int_\Omega u\partial_i v dx = -\int_\Omega v\partial_i u dx$$

Weak derivative

- Let $L^1_{loc}(\Omega)$ the set of functions which are Lebesgue integrable on every compact subset $K \subset \Omega$. Let $C_0^\infty(\Omega)$ be the set of functions infinitely differentiable with zero values on the boundary.

For $u \in L^1_{loc}(\Omega)$ we *define* $\partial_i u$ by

$$\int_\Omega v\partial_i u dx = -\int_\Omega u\partial_i v dx \quad \forall v \in C_0^\infty(\Omega)$$

and $\partial^\alpha u$ by

$$\int_\Omega v\partial^\alpha u dx = (-1)^{|\alpha|}\int_\Omega u\partial_i v dx \quad \forall v \in C_0^\infty(\Omega)$$

if these integrals exist.

Sobolev spaces

- For $k \geq 0$ and $1 \leq p < \infty$, the *Sobolev space* $W^{k,p}(\Omega)$ is the space functions where all up to the k-th derivatives are in L^p:

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

with then norm

$$||u||_{W^{k,p}(\Omega)} = \left(\sum_{|\alpha| \leq k} ||\partial^\alpha u||^p_{L^p(\Omega)}\right)^{\frac{1}{p}}$$

- Alternatively, they can be defined as the completion of C^∞ in the norm $||u||_{W^{k,p}(\Omega)}$
- $W_0^{k,p}(\Omega)$ is the completion of C_0^∞ in the norm $||u||_{W^{k,p}(\Omega)}$
- The Sobolev spaces are Banach spaces.

Fractional Sobolev spaces and traces

- For $0 < s < 1$ define the *fractional Sobolev space*

$$W^{s,p}(\Omega) = \left\{ u \in L^p(\Omega) : \frac{u(x) - u(y)}{||x-y||^{s+\frac{d}{p}}} \in L^p(\Omega \times \Omega) \right\}$$

- Let $H^{\frac{1}{2}}(\Omega) = W^{\frac{1}{2},2}(\Omega)$
- A priori it is hard to say what the value of a function from L^p on the boundary is like.
- For Lipschitz domains there exists unique continuous *trace mapping* $\gamma_0 : W^{1,p}(\Omega) \to L^p(\partial\Omega)$ where $\frac{1}{p} + \frac{1}{p'} = 1$ such that
  - $\mathrm{Im}\gamma_0 = W^{\frac{1}{p'},p}(\partial\Omega)$
  - $\mathrm{Ker}\gamma_0 = W_0^{1,p}(\Omega)$

Sobolev spaces of square integrable functions

- $H^k(\Omega) = W^{k,2}(\Omega)$ with the scalar product

$$(u,v)_{H^k(\Omega)} = \sum_{|\alpha| \le k} \int_\Omega \partial^\alpha u \partial^\alpha v \, dx$$

  is a Hilbert space.
- $H^k(\Omega)_0 = W_0^{k,2}(\Omega)$ with the scalar product

$$(u,v)_{H^k(\Omega)} = \sum_{|\alpha| \le k} \int_\Omega \partial^\alpha u \partial^\alpha v \, dx$$

  is a Hilbert space as well.
- The initally most important:
  - $L^2(\Omega)$ with the scalar product $(u,v)_{L^2(\Omega)} = \int_\Omega uv \, dx$
  - $H^1(\Omega)$ with the scalar product $(u,v)_{H^1(\Omega)} = \int_\Omega (uv + \nabla u \cdot \nabla v) \, dx$
  - $H_0^1(\Omega)$ with the scalar product $(u,v)_{H_0^1(\Omega)} = \int_\Omega (\nabla u \cdot \nabla v) \, dx$

Heat conduction revisited: Derivation of weak formulation

- Sobolev space theory provides the necessary framework to formulate existence and uniqueness of solutions of PDEs.
- Heat conduction equation with homogeneous Dirichlet boundary conditions:

$$\begin{aligned} -\nabla \cdot \lambda \nabla u &= f \text{ in } \Omega \\ u &= 0 \text{ on } \partial\Omega \end{aligned}$$

Multiply and integrate with an arbitrary *test function* from $C_0^\infty(\Omega)$:

$$\begin{aligned} -\int_\Omega \nabla \cdot \lambda \nabla u v \, dx &= \int_\Omega fv \, dx \\ \int_\Omega \lambda \nabla u \nabla v \, dx &= \int_\Omega fv \, dx \end{aligned}$$

Weak formulation of homogeneous Dirichlet problem

- Search $u \in H_0^1(\Omega)$ such that

$$\int_\Omega \lambda \nabla u \nabla v \, dx = \int_\Omega fv \, dx \;\forall v \in H_0^1(\Omega)$$

- Then,

$$a(u,v) := \int_\Omega \lambda \nabla u \nabla v \, dx$$

  is a self-adjoint bilinear form defined on the Hilbert space $H_0^1(\Omega)$
- $f(v) = \int_\Omega fv \, dx$ is a linear functional on $H_0^1(\Omega)$. For Hilbert spaces V the dual space V' (the space of linear functionals) can be identified with the space itself.

The Lax-Milgram lemma

Let V be a Hilbert space. Let $a : V \times V \to \mathbb{R}$ be a self-adjoint bilinear form, and f a linear functional on V. Assume a is coercive, i.e.

$$\exists \alpha > 0 : \forall u \in V, a(u,u) \ge \alpha ||u||_V^2.$$

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

$$a(u,v) = f(v) \;\forall v \in V$$

admits one and only one solution with an a priori estimate

$$||u||_V \le \frac{1}{\alpha} ||f||_{V'}$$

Heat conduction revisited

Let $\lambda > 0$. Then the weak formulation of the heat conduction problem: search $u \in H_0^1(\Omega)$ such that

$$\int_\Omega \lambda \nabla u \nabla v \, dx = \int_\Omega fv \, dx \;\forall v \in H_0^1(\Omega)$$

has an unique solution.

Weak formulation of inhomogeneous Dirichlet problem

$$\begin{aligned} -\nabla \cdot \lambda \nabla u &= f \text{ in } \Omega \\ u &= g \text{ on } \partial\Omega \end{aligned}$$

If g is smooth enough, there exists a *lifting* $u_g \in H^1(\Omega)$ such that $u_g|_{\partial\Omega} = g$. Then, we can re-formulate:

$$\begin{aligned} -\nabla \cdot \lambda \nabla (u - u_g) &= f + \nabla \cdot \lambda \nabla u_g \text{ in } \Omega \\ u - u_g &= 0 \text{ on } \partial\Omega \end{aligned}$$

- Search $u \in H^1(\Omega)$ such that

$$\begin{aligned} u &= u_g + \phi \\ \int_\Omega \lambda \nabla \phi \nabla v \, dx &= \int_\Omega fv \, dx + \int_\Omega \lambda \nabla u_g \nabla v \;\; \forall v \in H_0^1(\Omega) \end{aligned}$$

  Here, necessarily, $\phi \in H_0^1(\Omega)$ and we can apply the theory for the homogeneous Dirichlet problem.

Weak formulation of Robin problem

$$\begin{aligned} -\nabla \cdot \lambda \nabla u &= f \text{ in } \Omega \\ \lambda \nabla u \cdot \mathbf{n} + \alpha(u - g) &= 0 \text{ on } \partial\Omega \end{aligned}$$

Multiply and integrate with an arbitrary *test function* from $C_c^\infty(\Omega)$:

$$\begin{aligned} -\int_\Omega (\nabla \cdot \lambda \nabla u) v \, dx &= \int_\Omega fv \, dx \\ \int_\Omega \lambda \nabla u \nabla v \, dx + \int_{\partial\Omega} (\lambda \nabla u \cdot \mathbf{n}) v ds &= \int_\Omega fv \, dx \\ \int_\Omega \lambda \nabla u \nabla v \, dx + \int_{\partial\Omega} \alpha uv \, ds &= \int_\Omega fv \, dx + \int_{\partial\Omega} \alpha gv \, ds \end{aligned}$$

Weak formulation of Robin problem II

- Let

$$a^R(u,v) := \int_\Omega \lambda \nabla u \nabla v \, dx + \int_{\partial\Omega} \alpha uv \, ds$$
$$f^R(v) := \int_\Omega fv \, dx + \int_{\partial\Omega} \alpha gv \, ds$$

  The integrals over $\partial\Omega$ must be understood in the sense of the trace space $H^{\frac{1}{2}}(\partial\Omega)$.
- Search $u \in H^1(\Omega)$ such that

$$a^R(u,v) = f^R(v) \; \forall v \in H^1(\Omega)$$

- If $\lambda > 0$ and $\alpha > 0$ then $a^R(u,v)$ is coercive.

Neumann boundary conditions

Homogeneous Neumann:

$$\lambda \nabla u \cdot \mathbf{n} = 0 \text{ on } \partial\Omega$$

Inhomogeneous Neumann:

$$\lambda \nabla u \cdot \mathbf{n} = g \text{ on } \partial\Omega$$

Weak formulation:

- Search $u \in H^1(\Omega)$ such that

$$\int_\omega \nabla u \nabla v dx = \int_{\partial\Omega} gvds \; \forall v \in H^1(\Omega)$$

  Not coercive due to the fact that we can add an arbitrary constant to u and $a(u,u)$ stays the same!

Further discussion on boundary conditions

- Mixed boundary conditions:
  One can have differerent boundary conditions on different parts of the boundary. In particular, if Dirichlet or Robin boundary conditions are applied on at least a part of the boundary of measure larger than zero, the binlinear form becomes coercive.
- Natural boundary conditions: Robin, Neumann
  These are imposed in a "natural" way in the weak formulation
- Essential boundary conditions: Dirichlet
  Explicitely imposed on the function space
- Coefficients $\lambda, \alpha \dots$ can be functions.

The Dirichlet penalty method

- Robin problem: search $u_\alpha \in H^1(\Omega)$ such that

$$\int_\Omega \lambda \nabla u_\alpha \nabla v \, dx + \int_{\partial\Omega} \alpha u_\alpha v \, ds = \int_\Omega fv \, dx + \int_{\partial\Omega} \alpha gv \, ds \forall v \in H^1(\Omega)$$

- Dirichlet problem: search $u \in H^1(\Omega)$ such that

$$u = u_g + \phi \quad \text{where } u_g|_{\partial\Omega} = g$$
$$\int_\Omega \lambda \nabla \phi \nabla v \, dx = \int_\Omega fv \, dx + \int_\Omega \lambda \nabla u_g \nabla v \quad \forall v \in H_0^1(\Omega)$$

- Penalty limit:

$$\lim_{\alpha \to \infty} u_\alpha = u$$

- Formally, the convergence rate is quite low
- Implementing Dirichlet boundary conditions directly leads to a number of technical problems
- Implementing the penalty method is technically much simpler
- Proper way of handling the parameter leads to exact fulfillment of Dirichlet boundary condition in the floating point precision

The Galerkin method I

- Weak formulations "live" in Hilbert spaces which essentially are infinite dimensional
- For computer representations we need finite dimensional approximations
- The finite volume method provides one possible framework which in many cases is close to physical intuition. However, its error analysis is hard.
- The Galerkin method and its modifications provide a general scheme for the derivation of finite dimensional appoximations

The Galerkin method II

- Let V be a Hilbert space. Let $a : V \times V \to \mathbb{R}$ be a self-adjoint bilinear form, and f a linear functional on V. Assume a is coercive with coercivity constant α, and continuity constant γ.
- Continuous problem: search $u \in V$ such that

$$a(u,v) = f(v) \; \forall v \in V$$

- Let $V_h \subset V$ be a finite dimensional subspace of V
- "Discrete" problem $\equiv$ Galerkin approximation:
  Search $u_h \in V_h$ such that

$$a(u_h, v_h) = f(v_h) \; \forall v_h \in V_h$$

By Lax-Milgram, this problem has a unique solution as well.

Céa's lemma

- What is the connection between u and u_h ?
- Let $v_h \in V_h$ be arbitrary. Then

$$\begin{aligned}
\alpha ||u - u_h||^2 &\le a(u - u_h, u - u_h) \quad \text{(Coercivity)} \\
&= a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h) \\
&= a(u - u_h, u - v_h) \quad \text{(Galerkin Orthogonality)} \\
&\le \gamma ||u - u_h|| \cdot ||u - v_h|| \quad \text{(Boundedness)}
\end{aligned}$$

- As a result

$$||u - u_h|| \le \frac{\gamma}{\alpha} \inf_{v_h \in V_h} ||u - v_h||$$

- Up to a constant, the error of the Galerkin approximation is the error of the best approximation of the solution in the subspace V_h.

From the Galerkin method to the matrix equation

- Let $\phi_1 \dots \phi_n$ be a set of basis functions of V_h.
- Then, we have the representation $u_h = \sum_{j=1}^n u_j \phi_j$
- In order to search $u_h \in V_h$ such that

$$a(u_h, v_h) = f(v_h) \; \forall v_h \in V_h$$

  it is actually sufficient to require

$$\begin{aligned}
a(u_h, \phi_i) &= f(\phi_i) \; (i = 1 \dots n) \\
a\left(\sum_{j=1}^n u_j \phi_j, \phi_i\right) &= f(\phi_i) \; (i = 1 \dots n) \\
\sum_{j=1}^n a(\phi_j, \phi_i) u_j &= f(\phi_i) \; (i = 1 \dots n) \\
AU &= F
\end{aligned}$$

  with $A = (a_{ij}), a_{ij} = a(\phi_i, \phi_j)$, $F = (f_i), f_i = F(\phi_i)$, $U = (u_i)$.
- Matrix dimension is $n \times n$. Matrix sparsity ?

Obtaining a finite dimensional subspace

- Let $\Omega = (a, b) \subset \mathbb{R}^1$
- Let $a(u, v) = \int_\Omega \lambda(x)\nabla u \nabla v dx$.
- Analysis I provides a finite dimensional subspace: the space of sin/cos functions up to a certain frequency $\Rightarrow$ *spectral method*
- Ansatz functions have global support $\Rightarrow$ full $n \times n$ matrix
- OTOH: rather fast convergence for smooth data
- Generalization to higher dimensions possible
- Big problem in irregular domains: we need the eigenfunction basis of some operator...
- Spectral methods are successful in cases where one has regular geometry structures and smooth/constant coefficients – e.g. “Spectral Einstein Code”

The finite element idea

- Choose basis functions with local support. In this case, the matrix becomes sparse, as only integrals of basis function pairs with overlapping support contribute to the matrix.
- Linear finite elements in $\Omega = (a, b) \subset \mathbb{R}^1$:
- Partition $a = x_1 \le x_2 \le \dots \le x_n = b$
- Basis functions (for $i = 1 \dots n$)

$$\phi_i(x) = \begin{cases} \frac{x - x_{i-1}}{x_i - x_{i-1}}, & i > 1, x \in (x_{i-1}, x_i) \\ \frac{x_{i+1} - x}{x_{i+1} - x_i}, & i < n, x \in (x_i, x_{i+1}) \\ 0, & \text{else} \end{cases}$$

- Any function $u_h \in V_h = \operatorname{span}\{\phi_1 \dots \phi_n\}$ is piecewise linear, and the coefficients in the representation $u_h = \sum_{i=1}^n u_i \phi_i$ are the values $u_h(x_i)$.
- Fortunately, we are working with a weak formulation, and weak derivatives are well defined !

1D matrix elements

($\lambda = 1$, $x_{i+1} - x_i = h$) - The integrals are nonzero for $i = j, i + 1 = j, i - 1 = j$
Let $j = i + 1$

$$\begin{aligned} a_{ij} = a(\phi_i, \phi_{i+1}) &= \int_\Omega \nabla \phi_i \nabla \phi_j dx = \int_{x_i}^{x_{i+1}} \nabla \phi_i \nabla \phi_j dx = -\int_{x_i}^{x_{i+1}} \frac{1}{h^2} dx \\ &= \frac{1}{h} dx \end{aligned}$$

Similarly, $a(\phi_i, \phi_{i-1}) = -\frac{1}{h}$

For $1 < i < N$:

$$\begin{aligned} a_{ii} = a(\phi_i, \phi_i) &= \int_\Omega \nabla \phi_i \nabla \phi_i dx = \int_{x_{i-1}}^{x_{i+1}} \nabla \phi_i \nabla \phi_i dx = \int_{x_{i-1}}^{x_{i+1}} \frac{1}{h^2} dx \\ &= \frac{2}{h} dx \end{aligned}$$

For $i = 1$ or $i = N$, $a(\phi_i, \phi_i) = \frac{1}{h}$

1D matrix elements II

Adding the boundary integrals yields

$$A = \begin{pmatrix} \alpha + \frac{1}{h} & -\frac{1}{h} & & & & & \\ -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & & & & \\ & -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & & & \\ & \ddots & \ddots & \ddots & \ddots & & \\ & & -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & & \\ & & & -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & \\ & & & & -\frac{1}{h} & \frac{1}{h} + \alpha & \end{pmatrix}$$

... the same matrix as for the finite volume method...

Definition of a Finite Element (Ciarlet)

Triplet $\{K, P, \Sigma\}$ where

- $K \subset \mathbb{R}^d$: compact, connected Lipschitz domain with non-empty interior
- P: finite dimensional vector space of functions $p : K \to \mathbb{R}^m$ (mostly, $m = 1, m = d$)
- $\Sigma = \{\sigma_1 \dots \sigma_s\} \subset \mathcal{L}(P, \mathbb{R})$: set of linear forms defined on P called *local degrees of freedom* such that the mapping

$$\begin{aligned} \Lambda_\Sigma : P &\to \mathbb{R}^s \\ p &\mapsto (\sigma_1(p) \dots \sigma_s(p)) \end{aligned}$$

is bijective, i.e. Σ is a basis of $\mathcal{L}(P, \mathbb{R})$.

Local shape functions

- Due to bijectivity of Λ_Σ, for any finite element $\{K, P, \Sigma\}$, there exists a basis $\{\theta_1 \dots \theta_s\} \subset P$ such that

$$\sigma_i(\theta_j) = \delta_{ij} \quad (1 \le i, j \le s)$$

- Elements of such a basis are called *local shape functions*

Unisolvence

- Bijectivity of Λ_Σ is equivalent to the condition

$$\forall (\alpha_1 \dots \alpha_s) \in \mathbb{R}^s \; \exists! p \in P \text{ such that } \sigma_i(p) = \alpha_i \quad (1 \le i \le s)$$

i.e. for any given tuple of values $a = (\alpha_1 \dots \alpha_s)$ there is a unique polynomial $p \in P$ such that $\Lambda_\Sigma(p) = a$.
- Equivalent to *unisolvence*:

$$\begin{cases} \dim P = |\Sigma| = s \\ \forall p \in P : \; \sigma_i(p) = 0 \, (i = 1 \dots s) \; \Rightarrow \; p = 0 \end{cases}$$

Lagrange finite elements

- A finite element $\{K, P, \Sigma\}$ is called *Lagrange* finite element (or *nodal* finite element) if there exist a set of points $\{a_1 \dots a_s\} \subset K$ such that

$$\sigma_i(p) = p(a_i) \quad 1 \le i \le s$$

- $\{a_1 \dots a_s\}$: *nodes* of the finite element
- *nodal basis: $\{\theta_1 \dots \theta_s\} \subset P$ such that

$$\theta_j(a_i) = \delta_{ij} \quad (1 \le i, j \le s)$$

Hermite finite elements

- All or a part of degrees of freedoms defined by derivatives of p in some points

Local interpolation operator

- Let $\{K, P, \Sigma\}$ be a finite element with shape function bases $\{\theta_1 \dots \theta_s$. Let $V(K)$ be a normed vector space of functions $v : K \to \mathbb{R}^m$ such that
  - $P \subset V(K)$
  - The linear forms in Σ can be extended to be defined on $V(K)$
- *local interpolation operator*

$$
\begin{aligned}
\mathcal{I}_K : V(K) &\to P \\
v &\mapsto \sum_{i=1}^{s} \sigma_i(v)\theta_i
\end{aligned}
$$

- P is invariant under the action of $\mathcal{I}_K$, i.e. $\forall p \in P, \mathcal{I}_K(p) = p$:
  - Let $p = \sum_{j=1}^{s} \alpha_j \theta_j$ Then,

$$
\begin{aligned}
\mathcal{I}_K(p) &= \sum_{i=1}^{s} \sigma_i(p)\theta_i = \sum_{i=1}^{s}\sum_{j=1}^{s} \alpha_j \sigma_i(\theta_j)\theta_i \\
&= \sum_{i=1}^{s}\sum_{j=1}^{s} \alpha_j \delta_{ij}\theta_i = \sum_{j=1}^{s} \alpha_j \theta_j
\end{aligned}
$$

Local Lagrange interpolation operator

- Let $V(K) = (\mathcal{C}^0(K))^m$

$$
\begin{aligned}
\mathcal{I}_K : V(K) &\to P \\
v &\mapsto I_K v = \sum_{i=1}^{s} v(a_i)\theta_i
\end{aligned}
$$

Simplices

- Let $\{a_0 \dots a_d\} \subset \mathbb{R}^d$ such that the d vectors $a_1 - a_0 \dots a_d - a_0$ are linearly independent. Then the convex hull K of $a_0 \dots a_d$ is called *simplex*, and $a_0 \dots a_d$ are called *vertices* of the simplex.
- *Unit simplex*: $a_0 = (0...0), a_1 = (0, 1 \dots 0) \dots a_d = (0 \dots 0, 1)$.

$$
K = \left\{ x \in \mathbb{R}^d : x_i \geq 0 \; (i = 1 \dots d) \text{ and } \sum_{i=1}^{d} x_i \leq 1 \right\}
$$

- A general simplex can be defined as an image of the unit simplex under some affine transformation
- F_i: face of K opposite to a_i
- $\mathbf{n}_i$: outward normal to F_i

Barycentric coordinates

- Let K be a simplex.
- Functions λ_i $(i = 0 \dots d)$:

$$
\begin{aligned}
\lambda_i : \mathbb{R}^d &\to \mathbb{R} \\
x &\mapsto \lambda_i(x) = 1 - \frac{(x - a_i) \cdot \mathbf{n}_i}{(a_j - a_i) \cdot \mathbf{n}_i}
\end{aligned}
$$

  where a_j is any vertex of K situated in F_i.
- For $x \in K$, one has

$$
\begin{aligned}
1 - \frac{(x - a_i) \cdot \mathbf{n}_i}{(a_j - a_i) \cdot \mathbf{n}_i} &= \frac{(a_j - a_i) \cdot \mathbf{n}_i - (x - a_i) \cdot \mathbf{n}_i}{(a_j - a_i) \cdot \mathbf{n}_i} \\
&= \frac{(a_j - x) \cdot \mathbf{n}_i}{(a_j - a_i) \cdot \mathbf{n}_i} = \frac{\mathrm{dist}(x, F_i)}{\mathrm{dist}(a_i, F_i)} \\
&= \frac{\mathrm{dist}(x, F_i)|F_i|/d}{\mathrm{dist}(a_i, F_i)|F_i|/d} \\
&= \frac{\mathrm{dist}(x, F_i)|F_i|}{|K|}
\end{aligned}
$$

  i.e. $\lambda_i(x)$ is the ratio of the volume of the simplex $K_i(x)$ made up of x and the vertices of F_i to the volume of K.

Barycentric coordinates II

- $\lambda_i(a_j) = \delta_{ij}$
- $\lambda_i(x) = 0 \; \forall x \in F_i$
- $\sum_{i=0}^{d} \lambda_i(x) = 1 \; \forall x \in \mathbb{R}^d$
  (just sum up the volumes)
- $\sum_{i=0}^{d} \lambda_i(x)(x - a_i) = 0 \; \forall x \in \mathbb{R}^d$
  (due to $\sum \lambda_i(x)x = x$ and $\sum \lambda_i a_i = x$ as the vector of linear coordinate functions)
- Unit simplex:
  - $\lambda_0(x) = 1 - \sum_{i=1}^{d} x_i$
  - $\lambda_i(x) = x_i$ for $1 \leq i \leq d$

Polynomial space $\mathbb{P}_k$

- Space of polynomials in $x_1 \dots x_d$ of total degree $\leq k$ with real coefficients $\alpha_{i_1 \dots i_d}$:

$$
\mathbb{P}_k = \left\{ p(x) = \sum_{\substack{0 \leq i_1 \dots i_d \leq k \\ i_1 + \dots + i_d \leq k}} \alpha_{i_1 \dots i_d} x_1^{i_1} \dots x_d^{i_d} \right\}
$$

- Dimension:

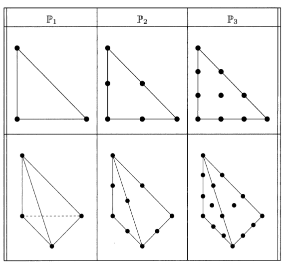
$$
\dim \mathbb{P}_k = \binom{d+k}{k} = \begin{cases} k+1, & d = 1 \\ \frac{1}{2}(k+1)(k+2), & d = 2 \\ \frac{1}{6}(k+1)(k+2)(k+3), & d = 3 \end{cases}
$$

$$
\dim \mathbb{P}_1 = d + 1
$$

$$
\dim \mathbb{P}_2 = \begin{cases} 3, & d = 1 \\ 6, & d = 2 \\ 10, & d = 3 \end{cases}
$$

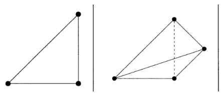
$\mathbb{P}_k$ simplex finite elements

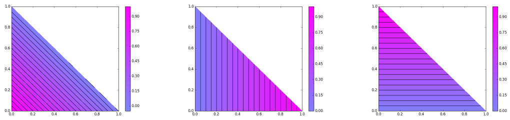
- K: simplex spanned by $a_0 \dots a_d$ in $\mathbb{R}^d$
- $P = \mathbb{P}_k$, such that $s = \dim P_k$
- For $0 \leq i_0 \dots i_d \leq k$, $i_0 + \dots + i_d = k$, let the set of nodes be defined by the points $a_{i_1 \dots i_d;k}$ with barycentric coordinates $(\frac{i_0}{k} \dots \frac{i_d}{k})$.
  Define Σ by $\sigma_{i_1 \dots i_d;k}(p) = p(a_{i_1 \dots i_d;k})$.



$\mathbb{P}_1$ simplex finite elements

- K: simplex spanned by $a_0 \dots a_d$ in $\mathbb{R}^d$
- $P = \mathbb{P}_1$, such that $s = d + 1$
- Nodes $\equiv$ vertices
- Basis functions $\equiv$ barycentric coordinates

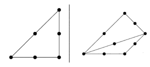


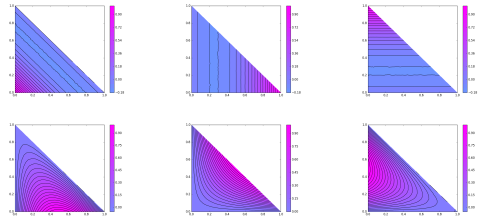


$\mathbb{P}_2$ simplex finite elements

- K: simplex spanned by $a_0 \dots a_d$ in $\mathbb{R}^d$
- $P = \mathbb{P}_2$, Nodes $\equiv$ vertices + edge midpoints
- Basis functions:

$$\lambda_i(2\lambda_i - 1), (0 \le i \le d); \quad 4\lambda_i\lambda_j, \quad (0 \le i < j \le d) \quad \text{("edge bubbles")}$$





Cuboids

- Given intervals $I_i = [c_i, d_i]$, $i = 1 \dots d$ such that $c_i < d_i$.
- *Cuboid*:

$$K = \prod_{i=1}^{d} [c_i, d_i]$$

- Local coordinate vector $(t_1 \dots t_d) \in [0,1]^d$
- Unique representation of $x \in K$: $x_i = c_i + t_i(d_i - c_i)$ for $i = 1 \dots d$.
- Bijective mapping $[0,1]^d \to K$.

Polynomial space $\mathbb{Q}_k$

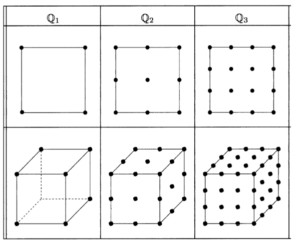
- Space of polynomials of degree at most k in each variable
- $d = 1 \Rightarrow \mathbb{Q}_k = \mathbb{P}_k$
- $d > 1$:

$$\mathbb{Q}_k = \left\{ p(x) = \sum_{0 \le i_1 \dots i_d \le k} \alpha_{i_1 \dots i_d} x_1^{i_1} \dots x_d^{i_d} \right\}$$

- $\dim \mathbb{Q}_k = (k+1)^d$

$\mathbb{Q}_k$ cuboid finite elements

- K: cuboid spanned by intervals $[c_i, d_i]$, $i = 1 \dots d$
- $P = \mathbb{Q}_k$
- For $0 \le i_0 \dots i_d \le k$, let the set of nodes be defined by the points $a_{i_1 \dots i_d;k}$ with local coordinates $(\frac{i_0}{k} \dots \frac{i_d}{k})$.
  Define Σ by $\sigma_{i_1 \dots i_d;k}(p) = p(a_{i_1 \dots i_d;k})$.



General finite elements

- Simplicial finite elements can be defined on triangulations of polygonal domains. During the course we will stick to this case.
- A curved domain Ω may be approximated by a polygonal domain Ω_h which is then triangulated. During the course, we will ignore this difference.
- As we have seen, more general elements are possible: cuboids, but and $T_m|_{\widehat{F}} = T_n|_{\widehat{F}}$also prismatic elements etc.
- Curved geometries are possible. Isoparametric finite elements use and $T_m|_{\widehat{F}} = T_n|_{\widehat{F}}$the polynomial space to define a mapping of some polyghedral reference element to an element with curved boundary

Conformal triangulations

- Let $\mathcal{T}_h$ be a subdivision of the polygonal domain $\Omega \subset \mathbb{R}^d$ into non-intersecting compact simplices K_m, $m = 1 \dots n_e$:

$$\overline{\Omega} = \bigcup_{m=1}^{n_e} K_m$$

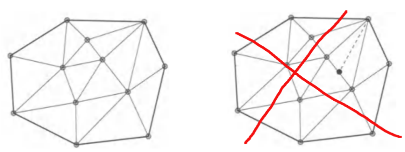
- Each simplex can be seen as the image of a affine transormation of a reference (e.g. unit) simplex $\widehat{K}$:

$$K_m = T_m(\widehat{K})$$

- We assume that it is conformal, i.e. if K_m, K_n have a $d-1$ dimensional intersection $F = K_m \cap K_n$, then there is a face $\widehat{F}$ of $\widehat{K}$ and renumberings of the vertices of K_n, K_m such that $F = T_m(\widehat{F}) = T_n(\widehat{F})$ and $T_m|_{\widehat{F}} = T_n|_{\widehat{F}}$

Conformal triangulations II

- $d = 1$: Each intersection $F = K_m \cap K_n$ is either empty or a common vertex
- $d = 2$: Each intersection $F = K_m \cap K_n$ is either empty or a common vertex or a common edge



- $d = 3$: Each intersection $F = K_m \cap K_n$ is either empty or a common vertex or a common edge or a common face
- Triangulations corresponding to simplicial complexes are conformal
- Delaunay triangulations are conformal

Reference finite element

- Let $\{\widehat{P}, \widehat{K}, \widehat{\Sigma}\}$ be a fixed finite element
- Let T_K be some affine transformation and $K = T_K(\widehat{K})$
- There is a linear bijective mapping ψ_K between functions on K and functions on $\widehat{K}$:

$$\begin{aligned} \psi_K : V(K) &\to V(\widehat{K}) \\ f &\mapsto f \circ T_K \end{aligned}$$

- Let
  - $K = T_K(\widehat{K})\$$
  - $P_K = \{\psi_K^{-1}(\widehat{p}); \widehat{p} \in \widehat{P}\}$,
  - $\Sigma_K = \{\sigma_{K,i}, i = 1 \dots s : \sigma_{K,i}(p) = \widehat{\sigma}_i(\psi_K(p))\}$ Then $\{K, P_K, \Sigma_K\}$ is a finite element.

Commutativity of interpolation and reference mapping

- $\mathcal{I}_{\widehat{K}} \circ \psi_K = \psi_K \circ \mathcal{I}_K$,
  i.e. the following diagram is commutative:

$$\begin{tikzcd} V(K) \arrow[r, "\psi_K"] \arrow[d, "\mathcal{I}_K"] & V(\widehat{K}) \arrow[d, "\mathcal{I}_{\widehat{K}}"] \\ P_K \arrow[r, "\psi_K"] & P_{\widehat{K}} \end{tikzcd}$$

Global interpolation operator $\mathcal{I}_h$

- Let $\{K, P_K, \Sigma_K\}_{K \in \mathcal{T}_h}$ be a triangulation of Ω.
- Domain:

$$D(\mathcal{I}_h) = \{v \in (L^1(\Omega))^m \text{ such that } \forall K \in \mathcal{T}_h, v|_K \in V(K)\}$$

- For all $v \in D(\mathcal{I}_h)$, define $\mathcal{I}_h v$ via

$$\mathcal{I}_h v|_K = \mathcal{I}_K(v|_K) = \sum_{i=1}^{s} \sigma_{K,i}(v|_K)\theta_{K,i} \ \forall K \in \mathcal{T}_h,$$

  Assuming $\theta_{K,i} = 0$ outside of K, one can write

$$\mathcal{I}_h v = \sum_{K \in \mathcal{T}_h} \sum_{i=1}^{s} \sigma_{K,i}(v|_K)\theta_{K,i},$$

  mapping $D(\mathcal{I}_h)$ to the *approximation space*

$$W_h = \{v_h \in (L^1(\Omega))^m \text{ such that } \forall K \in \mathcal{T}_h, v_h|_K \in P_K\}$$

H^1-Conformal approximation using Lagrangian finite elemenents

- Let V be a Banach space of functions on Ω. The approximation space W_h is said to be V-conformal if $W_h \subset V$.
- Non-conformal approximations are possible, we will stick to the conformal case.
- Conformal subspace of W_h with zero jumps at element faces:

$$V_h = \{v_h \in W_h : \forall n, m, K_m \cap K_n \neq 0 \Rightarrow (v_h|_{K_m})_{K_m \cap K_n} = (v_h|_{K_n})_{K_m \cap K_n}\}$$

- Then: $V_h \subset H^1(\Omega)$.

Zero jump at interfaces with Lagrangian finite elements

- Assume geometrically conformal mesh
- Assume all faces of $\widehat{K}$ have the same number of nodes s^∂
- For any face $F = K_1 \cap K_2$ there are renumberings of the nodes of K_1 and K_2 such that for $i = 1 \dots s^\partial$, $a_{K_1,i} = a_{K_2,i}$
- Then, $v_h|_{K_1}$ and $v_h|_{K_2}$ match at the interface $K_1 \cap K_2$ if and only if they match at the common nodes

$$v_h|_{K_1}(a_{K_1,i}) = v_h|_{K_2}(a_{K_2,i}) \quad (i = 1 \dots s^\partial)$$

Global degrees of freedom

- Let $\{a_1 \dots a_N\} = \bigcup_{K \in \mathcal{T}_h} \{a_{K,1} \dots a_{K,s}\}$
- Degree of freedom map

$$\begin{aligned} j : \mathcal{T}_h \times \{1 \dots s\} &\to \{1 \dots N\} \\ (K, m) &\mapsto j(K, m) \text{ the global degree of freedom number} \end{aligned}$$

- Global shape functions $\phi_1, \dots, \phi_N \in W_h$ defined by

$$\phi_i|_K(a_{K,m}) = \begin{cases} \delta_{mn} & \text{if } \exists n \in \{1 \dots s\} : j(K, n) = i \\ 0 & \text{otherwise} \end{cases}$$

- Global degrees of freedom $\gamma_1, \dots, \gamma_N : V_h \to \mathbb{R}$ defined by

$$\gamma_i(v_h) = v_h(a_i)$$

Lagrange finite element basis

- $\{\phi_1, \dots, \phi_N\}$ is a basis of V_h, and $\gamma_1 \dots \gamma_N$ is a basis of $\mathcal{L}(V_h, \mathbb{R})$.

Proof:

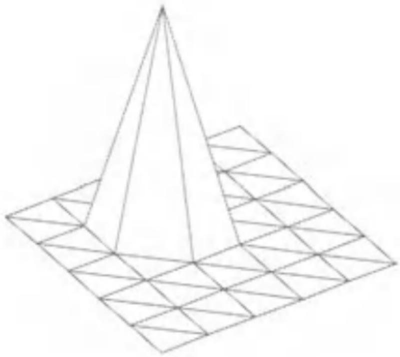
- $\{\phi_1, \dots, \phi_N\}$ are linearly independent: if $\sum_{j=1}^{N} \alpha_j \phi_j = 0$ then evaluation at $a_1 \dots a_N$ yields that $\alpha_1 \dots \alpha_N = 0$.
- Let $v_h \in V_h$. It is single valued in $a_1 \dots a_N$. Let $w_h = \sum_{j=1}^{N} v_h(a_j)\phi_j$. Then for all $K \in \mathcal{T}_h$, $v_h|_K$ and $w_h|_K$ coincide in the local nodes $a_{K,1} \dots a_{K,2}$, and by unisolvence, $v_h|_K = w_h|_K$.

Finite element approximation space

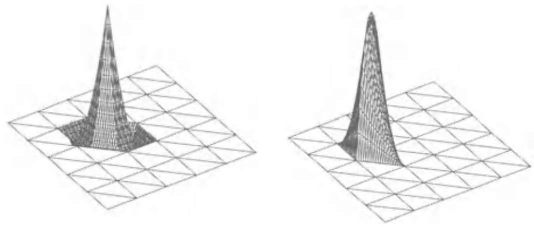
- $P_{c,h}^k = P_h^k = \{v_h \in \mathcal{C}^0(\bar{\Omega}_h) : \forall K \in \mathcal{T}_h, v_k \circ T_K \in \mathbb{P}^k\}$
- $Q_{c,h}^k = Q_h^k = \{v_h \in \mathcal{C}^0(\bar{\Omega}_h) : \forall K \in \mathcal{T}_h, v_k \circ T_K \in \mathbb{Q}^k\}$
- 'c' for continuity across mesh interfaces. There are also discontinuous FEM spaces which we do not consider here.

d	k	$N = \dim P_h^k$
1	1	N_v
1	2	$N_v + N_{el}$
1	3	$N_v + 2N_{el}$
2	1	N_v
2	2	$N_v + N_{ed}$
2	3	$N_v + 2N_{ed} + N_{el}$
3	1	N_v
3	2	$N_v + N_{ed}$
3	3	$N_v + 2N_{ed} + N_f$

P^1 global shape functions



P^2 global shape functions



Node based   Edge based

Global Lagrange interpolation operator

Let $V_h = P_h^k$ or $V_h = Q_h^k$

$$\mathcal{I}_h : \mathcal{C}^0(\bar{\Omega}_h) \to V_h$$
$$v \mapsto \sum_{i=1}^{N} v(a_i)\phi_i$$

Further finite element constructions

- In the realm considered in this course, we stick to H^1 conformal finite elements as the weak formulations regarded work in $H^(\Omega)$.
- With higher regularity, or for more complex problems one can construct H^2 conformal finite elements etc.
- Further possibilities for vector finite elements (divergence free etc.)

Affine transformation estimates I

- $\widehat{K}$: reference element
- Let $K \in \mathcal{T}_h$. Affine mapping:

$$T_K : \widehat{K} \to K$$
$$\widehat{x} \mapsto J_K\widehat{x} + b_K$$

  with $J_K \in \mathbb{R}^{d,d}, b_K \in \mathbb{R}^d$, J_K nonsingular
- Diameter of K: $h_K = \max_{x_1,x_2 \in K} ||x_1 - x_2||$
- ρ_K diameter of largest ball that can be inscribed into K
- $\sigma_K = \frac{h_K}{\rho_K}$: local shape regularity

Affine transformation estimates II

Lemma

- $|\det J_K| = \frac{meas(K)}{meas(\widehat{K})}$
- $||J_K|| \le \frac{h_K}{\rho_{\widehat{K}}}$
- $||J_K^{-1}|| \le \frac{h_{\widehat{K}}}{\rho_K}$

Proof:

- $|\det J_K| = \frac{meas(K)}{meas(\widehat{K})}$: basic property of affine mappings
- Further:

$$||J_K|| = \sup_{\hat{x} \neq 0} \frac{||J_K\hat{x}||}{||\hat{x}||} = \frac{1}{\rho_{\widehat{K}}} \sup_{||\hat{x}||=\rho_{\widehat{K}}} ||J_K\hat{x}||$$

  Set $\hat{x} = \hat{x}_1 - \hat{x}_2$ with $\hat{x}_1, \hat{x}_2 \in \widehat{K}$. Then $J_K\hat{x} = T_K\hat{x}_1 - T_K\hat{x}_2$ and one can estimate $||J_K\hat{x}|| \le h_K$.
- For $||J_K^{-1}||$ regard the inverse mapping □

Local interpolation I

- For $w \in H^s(K)$ recall the H^s seminorm $|w|_{s,K}^2 = \sum_{|\beta|=s} ||\partial^\beta w||_{L^2(K)}^2$

Lemma: Let $w \in H^s(K)$ and $\widehat{w} = w \circ T_K$. There exists a constant c such that

$$|\hat{w}|_{s,\hat{K}} \le c||J_K||^s |\det J_K|^{-\frac{1}{2}} |w|_{s,K}$$
$$|w|_{s,K} \le c||J_K^{-1}||^s |\det J_K|^{\frac{1}{2}} |\hat{w}|_{s,\hat{K}}$$

Proof: Let $|\alpha| = s$. By affinity and chain rule one obtains

$$||\partial^\alpha \hat{w}||_{L^2(\hat{K})} \le c||J_K||^s \sum_{|\beta=s|} ||\partial^\beta w \circ T_K||_{L^2(K)}$$

Changing variables yields

$$||\partial^\alpha \hat{w}||_{L^2(\hat{K})} \le c||J_K||^s |\det J_K|^{-\frac{1}{2}} |w|_{s,K}$$

Summation over α yields the first inequality. Regarding the inverse mapping yields the second one. □

Local interpolation II

Theorem: Let $\{\widehat{K}, \widehat{P}, \widehat{\Sigma}\}$ be a finite element with associated normed vector space $V(\widehat{K})$. Assume there exists k such that

$$\mathbb{P}_K \subset \widehat{P} \subset H^{k+1}(\widehat{K}) \subset V(\widehat{K})$$

and $H^{l+1}(\widehat{K}) \subset V(\widehat{K})$ for $0 \le l \le k$. There exists $c > 0$ such that for all $m = 0 \dots l+1$, $K \in \mathcal{T}_h$, $v \in H^{l+1}(K)$:

$$|v - \mathcal{I}_K^k v|_{m,K} \le c h_K^{l+1-m} \sigma_K^m |v|_{l+1,K}$$

Draft of Proof Estimate using deeper results from functional analysis:

$$|\hat{w} - \mathcal{I}_{\hat{K}}^k \hat{w}|_{m,\hat{K}} \le c|\hat{w}|_{l+1,\hat{K}}$$

(From Poincare like inequality, e.g. for $v \in H_0^1(\Omega)$, $c||v||_{L^2} \le ||\nabla v||_{L^2}$: under certain circumstances, we can can estimate the norms of lower derivatives by those of the higher ones)

Local interpolation III

(Proof, continued)

Let $v \in H^{l+1}(K)$ and set $\hat{v} = v \circ T_K$. We know that $(\mathcal{I}_K^k v) \circ T_K = \mathcal{I}_{\hat{K}}^k \hat{v}$.

We have

$$\begin{aligned}
|v - \mathcal{I}_K^k v|_{m,K} &\leq c||J_K^{-1}||^m |\det J_K|^{\frac{1}{2}} |\hat{v} - \mathcal{I}_{\hat{K}}^k \hat{v}|_{m,\hat{K}} \\
&\leq c||J_K^{-1}||^m |\det J_K|^{\frac{1}{2}} |\hat{v}|_{l+1,\hat{K}} \\
&\leq c||J_K^{-1}||^m ||J_K||^{l+1} |v|_{l+1,K} \\
&\leq c(||J_K|| ||J_K^{-1}||)^m ||J_K||^{l+1-m} |v|_{l+1,K} \\
&\leq c h_K^{l+1-m} \sigma_K^m |v|_{l+1,K}
\end{aligned}$$

Local interpolation: special cases for Lagrange finite elements

- $k = 1, l = 1, m = 0$: $|v - \mathcal{I}_K^k v|_{0,K} \leq c h_K^2 |v|_{2,K}$
- $k = 1, l = 1, m = 1$: $|v - \mathcal{I}_K^k v|_{1,K} \leq c h_K \sigma_K |v|_{2,K}$

Shape regularity

- Now we discuss a family of meshes $\mathcal{T}_h$ for $h \to 0$. We want to estimate global interpolation errors and see how they possibly diminuish
- For given $\mathcal{T}_h$, assume that $h = \max_{K \in \mathcal{T}_h} h_j$
- A family of meshes is called *shape regular* if

$$\forall h, \forall K \in \mathcal{T}_h, \sigma_K = \frac{h_K}{\rho_K} \leq \sigma_0$$

- In 1D, $\sigma_K = 1$
- In 2D, $\sigma_K \leq \frac{2}{\sin \theta_K}$ where θ_K is the smallest angle

Global interpolation error estimate

Theorem Let Ω be polyhedral, and let $\mathcal{T}_h$ be a shape regular family of affine meshes. Then there exists c such that for all h, $v \in H^{l+1}(\Omega)$,

$$||v - \mathcal{I}_h^k v||_{L^2(\Omega)} + \sum_{m=1}^{l+1} h^m \left(\sum_{K \in \mathcal{T}_h} |v - \mathcal{I}_h^k v|_{m,K}^2 \right)^{\frac{1}{2}} \leq c h^{l+1} |v|_{l+1,\Omega}$$

and

$$\lim_{h \to 0} \left(\inf_{v_h \in V_h^k} ||v - v_h||_{L^2(\Omega)} \right) = 0$$

Global interpolation error estimate for Lagrangian finite elements, $k = 1$

- Assume $v \in H^2(\Omega)$, e.g. if problem coefficients are smooth and the domain is convex

$$\begin{aligned}
||v - \mathcal{I}_h^k v||_{0,\Omega} + h|v - \mathcal{I}_h^k v|_{1,\Omega} &\leq c h^2 |v|_{2,\Omega} \\
|v - \mathcal{I}_h^k v|_{1,\Omega} &\leq c h |v|_{2,\Omega}
\end{aligned}$$

$$\lim_{h \to 0} \left(\inf_{v_h \in V_h^k} |v - v_h|_{1,\Omega} \right) = 0$$

- If $v \in H^2(\Omega)$ cannot be guaranteed, estimates become worse. Example: L-shaped domain.
- These results immediately can be applied in Cea's lemma.

Error estimates for homogeneous Dirichlet problem

- Search $u \in H_0^1(\Omega)$ such that

$$\int_\Omega \lambda \nabla u \nabla v \, dx = \int_\Omega f v \, dx \ \forall v \in H_0^1(\Omega)$$

Then, $\lim_{h \to 0} ||u - u_h||_{1,\Omega} = 0$. If $u \in H^2(\Omega)$ (e.g. on convex domains) then

$$||u - u_h||_{1,\Omega} \leq c h |u|_{2,\Omega}$$

Under certain conditions (convex domain, smooth coefficients) one has

$$||u - u_h||_{0,\Omega} \leq c h |u|_{1,\Omega}$$

("Aubin-Nitsche-Lemma")

Stiffness matrix calculation for Laplace operator for P1 FEM

$$\begin{aligned}
a_{ij} = a(\phi_i, \phi_j) &= \int_\Omega \nabla \phi_i \nabla \phi_j \, dx \\
&= \int_\Omega \sum_{K \in \mathcal{T}_h} \nabla \phi_i|_K \nabla \phi_j|_K \, dx
\end{aligned}$$

Assembly loop:
Set $a_{ij} = 0$.
For each $K \in \mathcal{T}_h$:
For each $m, n = 0 \ldots d$:

$$\begin{aligned}
s_{mn} &= \nabla \lambda_m \nabla \lambda_n \, dx \\
a_{j_{dof}(K,m), j_{dof}(K,n)} &= a_{j_{dof}(K,m), j_{dof}(K,n)} + s_{mn}
\end{aligned}$$

Local stiffness matrix calculation for P1 FEM

$a_0 \ldots a_d$: vertices of the simplex K, $a \in K$.

Barycentric coordinates: $\lambda_j(a) = \frac{|K_j(a)|}{|K|}$

For indexing modulo d+1 we can write

$$\begin{aligned}
|K| &= \frac{1}{d!} \det \left(a_{j+1} - a_j, \ldots a_{j+d} - a_j \right) \\
|K_j(a)| &= \frac{1}{d!} \det \left(a_{j+1} - a, \ldots a_{j+d} - a \right)
\end{aligned}$$

From this information, we can calculate $\nabla \lambda_j(x)$ (which are constant vectors due to linearity) and the corresponding entries of the local stiffness matrix

$$s_{ij} = \int_K \nabla \lambda_i \nabla \lambda_j \, dx$$

Local stiffness matrix calculation for P1 FEM in 2D

$a_0 = (x_0, y_0) \dots a_d = (x_2, y_2)$: vertices of the simplex K, $a = (x, y) \in K$.

Barycentric coordinates: $\lambda_j(x, y) = \frac{|K_j(x,y)|}{|K|}$

For indexing modulo d+1 we can write

$$
\begin{aligned}
|K| &= \frac{1}{2} \det \begin{pmatrix} x_{j+1} - x_j & x_{j+2} - x_j \\ y_{j+1} - y_j & y_{j+2} - y_j \end{pmatrix} \\
|K_j(x, y)| &= \frac{1}{2} \det \begin{pmatrix} x_{j+1} - x & x_{j+2} - x \\ y_{j+1} - y & y_{j+2} - y \end{pmatrix}
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
|K_j(x, y)| &= \frac{1}{2} \left((x_{j+1} - x)(y_{j+2} - y) - (x_{j+2} - x)(y_{j+1} - y) \right) \\
\partial_x |K_j(x, y)| &= \frac{1}{2} \left((y_{j+1} - y) - (y_{j+2} - y) \right) = \frac{1}{2}(y_{j+1} - y_{j+2}) \\
\partial_y |K_j(x, y)| &= \frac{1}{2} \left((x_{j+2} - x) - (x_{j+1} - x) \right) = \frac{1}{2}(x_{j+2} - x_{j+1})
\end{aligned}
$$

Local stiffness matrix calculation for P1 FEM in 2D II

$$
s_{ij} = \int_K \nabla \lambda_i \nabla \lambda_j \, dx = \frac{|K|}{4|K|^2} \left(y_{i+1} - y_{i+2}, x_{i+2} - x_{i+1} \right) \begin{pmatrix} y_{j+1} - y_{j+2} \\ x_{j+2} - x_{j+1} \end{pmatrix}
$$

So, let $V = \begin{pmatrix} x_1 - x_0 & x_2 - x_0 \\ y_1 - y_0 & y_2 - y_0 \end{pmatrix}$

Then

$$
\begin{aligned}
x_1 - x_2 &= V_{00} - V_{01} \\
y_1 - y_2 &= V_{10} - V_{11}
\end{aligned}
$$

and

$$
\begin{aligned}
2|K| \; \nabla \lambda_0 &= \begin{pmatrix} y_1 - y_2 \\ x_2 - x_1 \end{pmatrix} = \begin{pmatrix} V_{10} - V_{11} \\ V_{01} - V_{00} \end{pmatrix} \\
2|K| \; \nabla \lambda_1 &= \begin{pmatrix} y_2 - y_0 \\ x_0 - x_2 \end{pmatrix} = \begin{pmatrix} V_{11} \\ -V_{01} \end{pmatrix} \\
2|K| \; \nabla \lambda_2 &= \begin{pmatrix} y_0 - y_1 \\ x_1 - x_0 \end{pmatrix} = \begin{pmatrix} -V_{10} \\ V_{00} \end{pmatrix}
\end{aligned}
$$

Degree of freedom map representation for P1 finite elements

- List of global nodes $a_0 \dots a_N$: two dimensional array of coordinate values with N rows and d columns
- Local-global degree of freedom map: two-dimensional array C of index values with N_{el} rows and $d+1$ columns such that $C(i, m) = j_{dof}(K_i, m)$.
- The mesh generator triangle generates this information directly

Finite element assembly loop

```
for (int icell=0; icell<ncells; icell++)
{
  // Fill matrix V
  V(0,0)= points(cells(icell,1),0)- points(cells(icell,0),0);
  V(0,1)= points(cells(icell,2),0)- points(cells(icell,0),0);

  V(1,0)= points(cells(icell,1),1)- points(cells(icell,0),1);
  V(1,1)= points(cells(icell,2),1)- points(cells(icell,0),1);

  // Compute determinant
  double det=V(0,0)*V(1,1) - V(0,1)*V(1,0);
  double invdet = 1.0/det;

  // Compute entris of local stiffness matrix
  SLocal(0,0)= invdet * (  ( V(1,0)-V(1,1) )*( V(1,0)-V(1,1) )
                          +( V(0,1)-V(0,0) )*( V(0,1)-V(0,0) ) );
  SLocal(0,1)= invdet * (  ( V(1,0)-V(1,1) )* V(1,1)          - ( V(0,1)-V(0,0) )*V(0,1) );
  SLocal(0,2)= invdet * ( -( V(1,0)-V(1,1) )* V(1,0)          + ( V(0,1)-V(0,0) )*V(0,0) );

  SLocal(1,1)=  invdet * (  V(1,1)*V(1,1) + V(0,1)*V(0,1) );
  SLocal(1,2)=  invdet * ( -V(1,1)*V(1,0) - V(0,1)*V(0,0) );

  SLocal(2,2)=  invdet * ( V(1,0)*V(1,0)+ V(0,0)*V(0,0) );

  SLocal(1,0)=SLocal(0,1);
  SLocal(2,0)=SLocal(0,2);
  SLocal(2,1)=SLocal(1,2);

  // Assemble into global stiffness matrix
  for (int i=0;i<=ndim;i++)
    for (int j=0;j<=ndim;j++)
      SGlobal(cells(icell,i),cells(icell,j))+=SLocal(i,j);
}
```


Affine transformation estimates I

- $\widehat{K}$: reference element
- Let $K \in \mathcal{T}_h$. Affine mapping:

$$
\begin{aligned}
T_K : \widehat{K} &\to K \\
\widehat{x} &\mapsto J_K \widehat{x} + b_K
\end{aligned}
$$

  with $J_K \in \mathbb{R}^{d,d}$, $b_K \in \mathbb{R}^d$, J_K nonsingular
- Diameter of K: $h_K = \max_{x_1, x_2 \in K} ||x_1 - x_2||$
- ρ_K diameter of largest ball that can be inscribed into K
- $\sigma_K = \frac{h_K}{\rho_K}$: local shape regularity

Lemma

- $|\det J_K| = \frac{meas(K)}{meas(\widehat{K})}$
- $||J_K|| \leq \frac{h_K}{\rho_{\widehat{K}}}$
- $||J_K^{-1}|| \leq \frac{h_{\widehat{K}}}{\rho_K}$

Local interpolation I

- For $w \in H^s(K)$ recall the H^s seminorm $|w|^2_{s,K} = \sum_{|\beta|=s} ||\partial^\beta w||^2_{L^2(K)}$

Lemma: Let $w \in H^s(K)$ and $\widehat{w} = w \circ T_K$. There exists a constant c such that

$$
\begin{aligned}
|\widehat{w}|_{s,\widehat{K}} &\leq c ||J_K||^s |\det J_K|^{-\frac{1}{2}} |w|_{s,K} \\
|w|_{s,K} &\leq c ||J_K^{-1}||^s |\det J_K|^{\frac{1}{2}} |\widehat{w}|_{s,\widehat{K}}
\end{aligned}
$$

Local interpolation II

Theorem: Let $\{\widehat{K}, \widehat{P}, \widehat{\Sigma}\}$ be a finite element with associated normed vector space $V(\widehat{K})$. Assume there exists k such that

$$
\mathbb{P}_K \subset \widehat{P} \subset H^{k+1}(\widehat{K}) \subset V(\widehat{K})
$$

and $H^{l+1}(\widehat{K}) \subset V(\widehat{K})$ for $0 \leq l \leq k$. There exists $c > 0$ such that for all $m = 0 \dots l+1$, $K \in \mathcal{T}_h$, $v \in H^{l+1}(K)$:

$$
|v - \mathcal{I}_K^k v|_{m,K} \leq c h_K^{l+1-m} \sigma_K^m |v|_{l+1,K}
$$

Local interpolation: special cases for Lagrange finite elements

- $k = 1, l = 1, m = 0$:

$$
|v - \mathcal{I}_K^k v|_{0,K} \leq c h_K^2 |v|_{2,K}
$$

- $k = 1, l = 1, m = 1$:

$$
|v - \mathcal{I}_K^k v|_{1,K} \leq c h_K \sigma_K |v|_{2,K}
$$

Shape regularity

- Now we discuss a family of meshes $\mathcal{T}_h$ for $h \to 0$. We want to estimate global interpolation errors and see how they possibly diminuish
- For given $\mathcal{T}_h$, assume that $h = \max_{K\in\mathcal{T}_h} h_j$
- A family of meshes is called *shape regular* if

$$\forall h, \forall K \in \mathcal{T}_h, \sigma_K = \frac{h_K}{\rho_K} \le \sigma_0$$

- In 1D, $\sigma_K = 1$
- In 2D, $\sigma_K \le \frac{2}{\sin\theta_K}$ where θ_K is the smallest angle

Global interpolation error estimate

Theorem Let Ω be polyhedral, and let $\mathcal{T}_h$ be a shape regular family of affine meshes. Then there exists c such that for all h, $v \in H^{l+1}(\Omega)$,

$$||v - \mathcal{I}_h^k v||_{L^2(\Omega)} + \sum_{m=1}^{l+1} h^m \left(\sum_{K\in\mathcal{T}_h} |v - \mathcal{I}_h^k v|^2_{m,K}\right)^{\frac{1}{2}} \le ch^{l+1}|v|_{l+1,\Omega}$$

and

$$\lim_{h\to 0}\left(\inf_{v_h\in V_h^k} ||v - v_h||_{L^2(\Omega)}\right) = 0$$

Global interpolation error estimate for Lagrangian finite elements, $k = 1$

- Assume $v \in H^2(\Omega)$

$$\begin{aligned} ||v - \mathcal{I}_h^k v||_{0,\Omega} + h|v - \mathcal{I}_h^k v|_{1,\Omega} &\le ch^2|v|_{2,\Omega} \\ |v - \mathcal{I}_h^k v|_{1,\Omega} &\le ch|v|_{2,\Omega} \\ \lim_{h\to 0}\left(\inf_{v_h\in V_h^k} |v - v_h|_{1,\Omega}\right) &= 0 \end{aligned}$$

- If $v \in H^2(\Omega)$ cannot be guaranteed, estimates become worse. Example: L-shaped domain.
- These results immediately can be applied in Cea's lemma.

Error estimates for homogeneous Dirichlet problem

- Search $u \in H_0^1(\Omega)$ such that

$$\int_\Omega \lambda \nabla u \nabla v \, dx = \int_\Omega fv \, dx \ \forall v \in H_0^1(\Omega)$$

- Then, $\lim_{h\to 0} ||u - u_h||_{1,\Omega} = 0$.
- If $u \in H^2(\Omega)$ (e.g. convex domain, smooth coefficients), then

$$\begin{aligned} ||u - u_h||_{1,\Omega} &\le ch|u|_{2,\Omega} \le c'h|f|_{0,\Omega} \\ ||u - u_h||_{0,\Omega} &\le ch^2|u|_{2,\Omega} \le c'h^2|f|_{0,\Omega} \end{aligned}$$

and ("Aubin-Nitsche-Lemma")

$$||u - u_h||_{0,\Omega} \le ch|u|_{1,\Omega}$$

H^2-Regularity

- $u \in H^2(\Omega)$ may be *not* fulfilled e.g.
  - if Ω has re-entrant corners
  - if on a smooth part of the domain, the boundary condition type changes
  - if problem coefficients (λ) are discontinuos
- Situations differ as well between two and three space dimensions
- Delicate theory, ongoing research in functional analysis
- Consequence for simuations
  - Deterioration of convergence ratw
  - Remedy: local refinement of the discretization mesh
    - using a priori information
    - using a posteriori error estimators + automatic refinement of discretizatiom mesh

Higher regularity

- If $u \in H^s(\Omega)$ for $s > 2$, convergence order estimates become even better for P^k finite elements of order $k > 1$.
- Depending on the regularity of the solution the combination of grid adaptation and higher oder ansatz functions may be successful

More complicated integrals

- Assume non-constant right hand side f, space dependent heat conduction coefficient κ.
- Right hand side integrals

$$f_i = \int_K f(x)\lambda_i(x) \, dx$$

- P^1 stiffness matrix elements

$$a_{ij} = \int_K \kappa(x) \ \nabla\lambda_i \ \nabla\lambda_j \ dx$$

- P^k stiffness matrix elements created from higher order ansatz functions

Quadrature rules

- *Quadrature rule*:

$$\int_K g(x) \, dx \approx |K| \sum_{l=1}^{l_q} \omega_l g(\xi_l)$$

- ξ_l: *nodes, Gauss points*
- ω_l: *weights*
- The largest number k such that the quadrature is exact for polynomials of order k is called *order* k_q of the quadrature rule, i.e.

$$\forall k \le k_q, \forall p \in \mathbb{P}^k \int_K p(x) \, dx = |K| \sum_{l=1}^{l_q} \omega_l p(\xi_l)$$

- Error estimate:

$$\forall \phi \in \mathcal{C}^{k_q+1}(K), \left|\frac{1}{|K|}\int_K \phi(x) \, dx - \sum_{l=1}^{l_q} \omega_l g(\xi_l)\right| \le ch_K^{k_q+1} \sup_{x\in K, |\alpha|=k_q+1} |\partial^\alpha \phi(x)|$$

Some common quadrature rules

Nodes are characterized by the barycentric coordinates

d	k_q	l_q	Nodes	Weights
1	1	1	$(\frac{1}{2}, \frac{1}{2})$	1
	1	2	$(1,0),(0,1)$	$\frac{1}{2}, \frac{1}{2}$
	3	2	$(\frac{1}{2}+\frac{\sqrt{3}}{6}, \frac{1}{2}-\frac{\sqrt{3}}{6}), (\frac{1}{2}-\frac{\sqrt{3}}{6}, \frac{1}{2}+\frac{\sqrt{3}}{6})$	$\frac{1}{2}, \frac{1}{2}$
	5	3	$(\frac{1}{2},), (\frac{1}{2}+\sqrt{\frac{3}{20}}, \frac{1}{2}-\sqrt{\frac{3}{20}}), (\frac{1}{2}-\sqrt{\frac{3}{20}}, \frac{1}{2}+\sqrt{\frac{3}{20}})$	$\frac{8}{18}, \frac{5}{18}, \frac{5}{18}$
2	1	1	$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	1
	1	3	$(1,0,0),(0,1,0),(0,0,1)$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$
	2	3	$(\frac{1}{2}, \frac{1}{2}, 0), (\frac{1}{2}, 0, \frac{1}{2}), (0, \frac{1}{2}, \frac{1}{2})$	$\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$
	3	4	$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}), (\frac{1}{5}, \frac{1}{5}, \frac{3}{5}), (\frac{1}{5}, \frac{3}{5}, \frac{1}{5}), (\frac{3}{5}, \frac{1}{5}, \frac{1}{5}),$	$-\frac{9}{16}, \frac{25}{48}, \frac{25}{48}, \frac{25}{48}$
3	1	1	$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$	1
	1	4	$(1,0,0,0),(0,1,0,0),(0,0,1,0),(0,0,0,1)$	$\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}$
	2	4	$(\frac{5-\sqrt{5}}{20}, \frac{5-\sqrt{5}}{20}, \frac{5-\sqrt{5}}{20}, \frac{5+3\sqrt{5}}{20}) \dots$	$\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}$

Matching of approximation order and quadrature order

- "Variational crime": instead of
$$a(u_h, v_h) = f(v_h) \ \forall v_h \in V_h$$
we solve
$$a_h(u_h, v_h) = f_h(v_h) \ \forall v_h \in V_h$$
where a_h, f_h are derived from their exact counterparts by quadrature
- For P^1 finite elements, zero order quadrature for volume integrals and first order quadrature for surface intergals is sufficient to keep the convergence order estimates stated before
- The rule of thumb for the volume quadrature is that the highest order terms must be evaluated exactly if the coefficients of the PDE are constant.

Practical realization of integrals

- Integral over barycentric coordinate function
$$\int_K \lambda_i(x)\ dx = \frac{1}{3}|K|$$
- Right hand side integrals. Assume $f(x)$ is given as a piecewise linear function with given values in the nodes of the triangulation
$$f_i = \int_K f(x)\lambda_i(x)\ dx \approx \frac{1}{3}|K| f(a_i)$$
- Integral over space dependent heat conduction coefficient: Assume $\kappa(x)$ is given as a piecewise linear function with given values in the nodes of the triangulation
$$a_{ij} = \int_K \kappa(x)\ \nabla\lambda_i\ \nabla\lambda_j\ dx = \frac{1}{3}(\kappa(a_0)+\kappa(a_1)+\kappa(a_2)) \int_K \kappa(x)\ \nabla\lambda_i\ \nabla\lambda_j\ dx$$

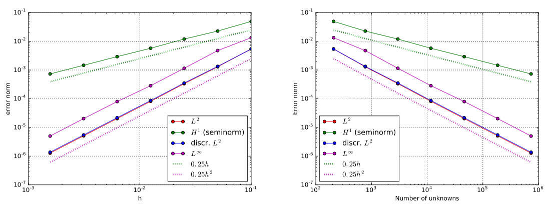
Practical realization of boundary conditions

- Robin boundary value problem
$$-\nabla \cdot \kappa \nabla u = f \quad \text{in } \Omega$$
$$\kappa \nabla u + \alpha(u-g) = 0 \quad \text{on } \partial\Omega$$
- Weak formulation: search $u \in H^1(\Omega)$ such that
$$\int_\Omega \kappa \nabla u \nabla v\ dx + \int_{\partial\Omega} \alpha u v\ ds = \int_\Omega f v\ dx + \int_{\partial\Omega} \alpha g v\ ds \ \forall v \in H^1(\Omega)$$
- In 2D, for P^1 FEM, boundary integrals can be calculated by trapezoidal rule without sacrificing approximation order

Test problem

- Homogeneous Dirichlet problem:
$$-\Delta u = 2\pi^2 \sin(\pi x)\sin(\pi y) \text{ in } \Omega = (0,1)\times(0,1)$$
$$u|_{\partial\Omega} = 0$$
- Exact solution:
$$u(x,y) = \sin(\pi x)\sin(\pi y)$$
- Testing approach: generate series of finer grids with `triangle`, by control the triangle are parameter according to the desired mesh size h.
- Do we get the theoretical error estimates ?
- We did not talk about error estimates for the finite volume method. What can we expect ?
- For simplicity, we calculate not $||u_{exact} - u_h||$ but $\Pi_h u_{exact} - u_h||$ where Π_h is the P1 nodal interpolation operator.
- More precise test would have to involve high order quadrature for calculation of the norm.

FEM Results



- Theoretical estimates are reproduced
- Useful test for debugging code. . .

Time dependent Robin boundary value problem

- Choose final time $T > 0$. Regard functions $(x,t) \to \mathbb{R}$.
$$\partial_t u - \nabla \cdot \kappa \nabla u = f \quad \text{in } \Omega \times [0,T]$$
$$\kappa \nabla u \cdot \vec{n} + \alpha(u-g) = 0 \quad \text{on } \partial\Omega \times [0,T]$$
$$u(x,0) = u_0(x) \quad \text{in}\Omega$$
- This is an initial boundary value problem
- This problem has a weak formulation in the Sobolev space $L^2\left([0,T], H^1(\Omega)\right)$, which then allows for a Galerkin approximation in a corresponding subspace
- We will proceed in a simpler manner: first, perform a finite difference discretization in time, then perform a finite element (finite volume) discretization in space.
  - Rothe method: first discretize in time, then in space
  - Method of lines: first discretize in space, get a huge ODE system

Time discretization

- Choose time discretization points $0 = t_0 < t_1 \dots < t_N = T$, let $\tau_i = t_i - t_{i-1}$
For $i = 1 \dots N$, solve
$$\frac{u_i - u_{i-1}}{\tau_i} - \nabla \cdot \kappa \nabla u_\theta = f \quad \text{in } \Omega \times [0,T]$$
$$\kappa \nabla u_\theta \cdot \vec{n} + \alpha(u_\theta - g) = 0 \quad \text{on } \partial\Omega \times [0,T]$$
where $u_\theta = \theta u_i + (1-\theta) u_{i-1}$
- $\theta = 1$: backward (implicit) Euler method
- $\theta = \frac{1}{2}$: Crank-Nicolson scheme
- $\theta = 0$: forward (explicit) Euler method
- Note that the explicit Euler method does not involve the solution of a PDE system. What do we have to pay for this ?

Weak formulation

- Weak formulation: search $u \in H^1(\Omega)$ such that

$$\frac{1}{\tau_i}\int_\Omega u_i v\, dx + \theta\left(\int_\Omega \kappa \nabla u_i \nabla v\, dx + \int_{\partial\Omega} \alpha u_i v\, ds\right) =$$
$$\frac{1}{\tau_i}\int_\Omega u_{i-1} v\, dx + (1-\theta)\left(\int_\Omega \kappa \nabla u_{i-1} \nabla v\, dx + \int_{\partial\Omega} \alpha u_{i-1} v\, ds\right)$$
$$+ \int_\Omega fv\, dx + \int_{\partial\Omega} \alpha g v\, ds\ \forall v \in H^1(\Omega)$$

- Matrix formulation (in case of constant coefficents, $A_i = A$)

$$\frac{1}{\tau_i} M u_i + \theta A_i u_i = \frac{1}{\tau_i} M u_{i-1} + (1-\theta) A_i u_{i-1} + F$$

- M: mass matrix, A: stiffness matrix

Mass matrix

- Mass matrix $M = (m_{ij})$:

$$m_{ij} = \int_\Omega \phi_i \phi_j\, dx$$

- Self-adjoint, coercive bilinear form $\Rightarrow$ M is symmetric, positiv definite
- For a family of quasi-uniform, shape-regular triangulations, for every eigenvalue μ one has the estimate

$$c_1 h^d \leq \mu \leq c_2 h^d$$

Therefore the condition number $\kappa(M)$ is bounded by a constant independent of h:

$$\kappa(M) \leq c$$

- How to see this ? Let $u_h = \sum_{i=1}^N U_i \phi_i$, and μ an eigenvalue (positive,real!) Then

$$||u_h||_0^2 = (U, MU)_{\mathbb{R}^N} = \mu(U, U)_{\mathbb{R}^N} = \mu||U||^2_{\mathbb{R}^N}$$

From quasi-uniformity we obtain

$$c_1 h^d ||U||^2_{\mathbb{R}^N} \leq ||u_h||_0^2 \leq c_2 h^d ||U||^2_{\mathbb{R}^N}$$

and conclude

Mass matrix M-Property ?

- For P^1-finite elements, all integrals $m_{ij} = \int_\Omega \phi_i \phi_j\, dx$ are zero or positive, so we get positive off diagonal elements.
- No M-Property!

Stiffness matrix condition number + row sums

- Stiffness matrix $A = (a_{ij})$:

$$a_{ij} = a(\phi_i, \phi_j) = \int_\Omega \nabla\phi_i \nabla\phi_j\, dx$$

- bilinear form $a(\cdot,\cdot)$ is self-adjoint, therefore A is symmetric, positive definite
- Condition number estimate for P^1 finite elements on quasi-uniform triangulation:

$$\kappa(A) \leq ch^{-2}$$

- Row sums:

$$\sum_{j=1}^N a_{ij} = \sum_{j=1}^N \int_\Omega \nabla\phi_i \nabla\phi_j\, dx = \int_\Omega \nabla\phi_i \nabla\left(\sum_{j=1}^N \phi_j\right)\, dx$$
$$= \int_\Omega \nabla\phi_i \nabla(1)\, dx$$
$$= 0$$

Stiffness matrix entry signs

Local stiffness matrices

$$s_{ij} = \int_K \nabla\lambda_i \nabla\lambda_j\, dx = \frac{|K|}{2|K|^2}\left(y_{i+1} - y_{i+2}, x_{i+2} - x_{i+1}\right)\begin{pmatrix} y_{j+1} - y_{j+2} \\ x_{j+2} - x_{j+1}\end{pmatrix}$$

- Main diagonal entries must be positive
- Local contributions from element stiffness matrices: Scalar products of vectors orthogonal to edges. These are nonpositive if the angle between the edges are $\leq 90°$
- *weakly acute triangulation*: all triangle angles are less than $\leq 90°$
- in fact, for constant coefficients, in *2D*, Delaunay is sufficient!
- All rows sums are zero $\Rightarrow$ A is singular
- Matrix becomes irreducibly diagonally dominant if we add at least one positive value to the main diagonal, e.g. from Dirichlet BC or *lumped* mass matrix $\Rightarrow$ $M - Matrix$
- Adding a mass matrix yields a positive definite matrix and thus nonsingularity, but *destroys* M-property

Back to time dependent problem

Assume M diagonal, $A = S + D$, where S is the stiffness matrix, and D is a nonnegative diagonal matrix. We have

$$(Su)_i = \sum_j s_{ij} u_j = s_{ii} u_i + \sum_{i \neq j} s_{ij} u_j$$
$$= (-\sum_{i\neq j} s_{ij}) u_i + \sum_{i\neq j} s_{ij} u_j$$
$$= \sum_{i\neq j} -s_{ij}(u_i - u_j)$$

Forward Euler

$$\frac{1}{\tau_i} M u_i = \frac{1}{\tau_i} M u_{i-1} + A_i u_{i-1}$$
$$u_i = u_{i-1} + \tau_i M^{-1} A_i u_{i-1} = (I + \tau M^{-1} D + \tau M^{-1} S) u_{i-1}$$

- Entries of $\tau M^{-1} A) u_{i-1}$ are of order $\frac{1}{h^2}$, and so we can expect stabilityonly if τ balances $\frac{1}{h^2}$, i.e.

$$\tau \leq Ch^2$$

- A more thorough stability estimate proves this situation

Backward Euler

$$\frac{1}{\tau_i} M u_i + A u_i = \frac{1}{\tau_i} M u_{i-1}$$
$$(I + \tau_i M^{-1} A) u_i = u_{i-1}$$
$$u_i = (I + \tau_i M^{-1} A)^{-1} u_{i-1}$$

But here, we can estimate that

$$||(I + \tau_i M^{-1} A)^{-1}||_\infty \leq 1$$

Backward Euler Estimate

Theorem: Assume S has the sign pattern of an M-Matrix with row sum zero, and D is a nonnegative diagonal matrix. Then $||(I+D+S)^{-1}||_\infty \leq 1$

Proof: Assume that $||(I+S)^{-1}||_\infty > 1$. We know that $(I+S)^{-1}$ has positive entries. Then for α_{ij} being the entries of $(I+S)^{-1}$,

$$max_{i=1}^n \sum_{j=1}^n \alpha_{ij} > 1.$$

Let k be a row where the maximum is reached. Let $e=(1\dots 1)^T$. Then for $v=(I+A)^{-1}e$ we have that $v>0$, $v_k>1$ and $v_k \geq v_j$ for all $j\neq k$. The kth equation of $e=(I+A)v$ then looks like

$$\begin{aligned} 1 &= v_k + v_k \sum_{j\neq k} |s_{kj}| - \sum_{j\neq k} |s_{kj}| v_j \\ &\geq v_k + v_k \sum_{j\neq k} |s_{kj}| - \sum_{j\neq k} |s_{kj}| v_k \\ &= v_k > 1 \end{aligned}$$

This contradiction enforces $||(I+S)^{-1}||_\infty \leq 1$.

Backward Euler Estimate II

$$\begin{aligned} I+A =& I+D+S \\ =&(I+D)(I+D)^{-1}(I+D+S) \\ =&(I+D)(I+A_{D0}) \end{aligned}$$

with $A_{D0}=(I+D)^{-1}S$ has row sum zero Thus

$$\begin{aligned} ||(I+A)^{-1}||_\infty =&||(I+A_{D0})^{-1}(I+D)^{-1}||_\infty \\ \leq&||(I+D)^{-1}||_\infty \\ \leq&1, \end{aligned}$$

because all main diagonal entries of $I+D$ are greater or equal to 1. □

Backward Euler Estimate III

We can estimate that

$$I+\tau_i M^{-1}A = I+\tau_i M^{-1}D + \tau_i M^{-1}S$$

and obtain

$$||(I+\tau_i M^{-1}A)^{-1}||_\infty \leq 1$$

- We get this stability independent of the time step.
- Another theory is possible using L^2 estimates and positive definiteness

Discrete maximum principle

Assuming $v\geq 0$ we can conclude $u\geq 0$.

$$\frac{1}{\tau}Mu + (D+S)u = \frac{1}{\tau}Mv$$

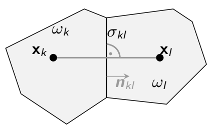
$$(\tau m_i + d_i)u_i + s_{ii}u_i = \tau m_i v_i + \sum_{i\neq j}(-s_{ij})u_j$$

$$\begin{aligned} u_i &= \frac{1}{\tau m_i + d_i + \sum_{i\neq j}(-s_{ij})}(\tau m_i v_i + \sum_{i\neq j}(-s_{ij})u_j) \\ &\leq \frac{\tau m_i v_i + \sum_{i\neq j}(-s_{ij})u_j}{\tau m_i + d_i + \sum_{i\neq j}(-s_{ij})} \max(\{v_i\}\cup\{u_j\}_{j\neq i}) \\ &\leq \max(\{v_i\}\cup\{u_j\}_{j\neq i}) \end{aligned}$$

- Provided, the right hand side is zero, the solution in a given node is bounded by the value from the old timestep, and by the solution in the neigboring points.
- No new local maxima can appear during time evolution
- There is a continuous counterpart which can be derived from weak solution
- M-property is crucial for the proof.

The finite volume idea revisited

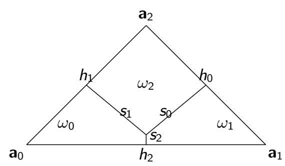
- Assume Ω is a polygon
- Subdivide the domain Ω into a finite number of **control volumes** : $\bar\Omega = \bigcup_{k\in\mathcal{N}} \bar\omega_k$ such that
  - ω_k are open (not containing their boundary) convex domains
  - $\omega_k \cap \omega_l = \emptyset$ if $\omega_k \neq \omega_l$
  - $\sigma_{kl} = \bar\omega_k \cap \bar\omega_l$ are either empty, points or straight lines
    - we will write $|\sigma_{kl}|$ for the length
    - if $|\sigma_{kl}|>0$ we say that ω_k, ω_l are neigbours
    - neigbours of ω_k: $\mathcal{N}_k = \{l \in \mathcal{N} : |\sigma_{kl}|>0\}$
- To each control volume ω_k assign a **collocation point**: $\mathbf{x}_k \in \bar\omega_k$ such that
  - **admissibility condition**: if $l\in\mathcal{N}_k$ then the line $\mathbf{x}_k\mathbf{x}_l$ is orthogonal to σ_{kl}
  - if ω_k is situated at the boundary, i.e. $\gamma_k = \partial\omega_k \cap \partial\Omega \neq \emptyset$, then $\mathbf{x}_k \in \partial\Omega$



- Now, we know how to construct this partition
  - obtain a boundary conforming Delaunay triangulation
  - construct restricted Voronoi cells

Finite volume local stiffness matrix calculation

$a_0=(x_0,y_0)\dots a_d=(x_2,y_2)$: vertices of the simplex K Calculate the contribution from triangle to $\frac{\sigma_{kl}}{h_{kl}}$ in the finite volume discretization

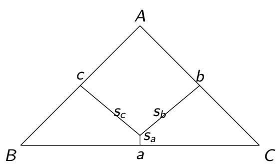


Let $h_i = ||a_{i+1}-a_{i+2}||$ (i counting modulo 2) be the lengths of the discretization edges. Let A be the area of the triangle. Then for the contribution from the triangle to the form factor one has

$$\frac{|s_i|}{h_i} = \frac{1}{8A}(h_{i+1}^2 + h_{i+2}^2 - h_i^2)$$

$$|\omega_i| = (|s_{i+1}|h_{i+1} + |s_{i+2}|h_{i+2})/4$$

Finite volume local stiffness matrix calculation II



Triangle edge lengths:

$$a,b,c$$

Semiperimeter:

$$s=\frac{a}{2}+\frac{b}{2}+\frac{c}{2}$$

Square area (from Heron's formula):

$$16A^2 = 16s(s-a)(s-b)(s-c) = (-a+b+c)(a-b+c)(a+b-c)(a+b+c)$$

Square circumradius:

$$R^2 = \frac{a^2b^2c^2}{(-a+b+c)(a-b+c)(a+b-c)(a+b+c)} = \frac{a^2b^2c^2}{16A^2}$$

Finite volume local stiffness matrix calculation III

Square of the Voronoi surface contribution via Pythagoras:

$$s_a^2 = R^2 - \left(\frac{1}{2}a\right)^2 = -\frac{a^2\left(a^2-b^2-c^2\right)^2}{4(a-b-c)(a-b+c)(a+b-c)(a+b+c)}$$

Square of edge contribution in the finite volume method:

$$e_a^2 = \frac{s_a^2}{a^2} = -\frac{\left(a^2-b^2-c^2\right)^2}{4(a-b-c)(a-b+c)(a+b-c)(a+b+c)}$$

Comparison with pdelib formula:

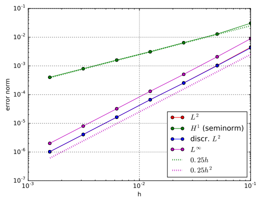
$$e_a^2 - \frac{(b^2+c^2-a^2)^2}{64A^2} = 0$$

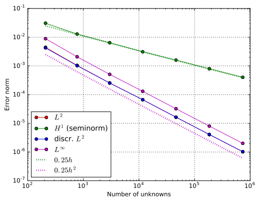
This implies the formula for the edge contribution

$$e_a = \frac{s_a}{a} = \frac{b^2+c^2-a^2}{8A}$$

The sign chosen implies a positive value if the angle $\alpha < \frac{\pi}{2}$, and a negative value if it is obtuse. In the latter case, this corresponds to the negative length of the line between edge midpoint and circumcenter, which is exactly the value which needs to be added to the corresponding amount from the opposite triangle in order to obtain the measure of the Voronoi face.

FVM Results





- Similar results as for FEM

Finite volumes for time dependent problem

Search function $u : \Omega \times [0, T] \to \mathbb{R}$ such that $u(x, 0) = u_0(x)$ and

$$\begin{aligned} \partial_t u - \nabla \cdot \lambda \nabla u &= 0 \quad \text{in}\Omega \times [0, T] \\ \lambda \nabla u \cdot \mathbf{n} + \alpha(u - w) &= 0 \quad \text{on}\Gamma \times [0, T] \end{aligned}$$

- Given control volume ω_k, integrate equation over space-time control volume

$$\begin{aligned} 0 = \int_{\omega_k} \left(\frac{1}{\tau}(u - v) - \nabla \cdot \lambda \nabla u\right) d\omega &= -\int_{\partial\omega_k} \lambda \nabla u \cdot \mathbf{n}_k d\gamma + \frac{1}{\tau}\int_{\omega_k} (u - v) d\omega \\ &= -\sum_{L \in \mathcal{N}_k} \int_{\sigma_{kl}} \lambda \nabla u \cdot \mathbf{n}_{kl} d\gamma - \int_{\gamma_k} \lambda \nabla u \cdot \mathbf{n} d\gamma - \frac{1}{\tau}\int_{\omega_k} (u - v) d\omega \\ &\approx \frac{|\omega_k|}{\tau}(u_k - v_k) + \sum_{L \in \mathcal{N}_k} \frac{|\sigma_{kl}|}{h_{kl}}(u_k - u_l) + |\gamma_k|\alpha(u_k - w_k) \end{aligned}$$

- Here, $u_k = u(\mathbf{x}_k)$, $w_k = w(\mathbf{x}_k)$, $f_k = f(\mathbf{x}_k)$
- $\frac{1}{\tau_i} M u_i + A u_i = \frac{1}{\tau_i} M u_{i-1}$

~

Convection-Diffusion

The convection - diffusion equation

Search function $u : \Omega \times [0, T] \to \mathbb{R}$ such that $u(x, 0) = u_0(x)$ and

$$\begin{aligned} \partial_t u - \nabla (\cdot D \nabla u - u\mathbf{v}) &= 0 \quad \text{in}\Omega \times [0, T] \\ \lambda \nabla u \cdot \mathbf{n} + \alpha(u - w) &= 0 \quad \text{on}\Gamma \times [0, T] \end{aligned}$$

- Here:
  - u: species concentration
  - D: diffusion coefficient
  - $\mathbf{v}$: velocity of medium (e.g. fluid)

$$\frac{|\omega_k|}{\tau}(u_k - v_k) + \sum_{L \in \mathcal{N}_k} \frac{|\sigma_{kl}|}{h_{kl}} g(u_k, u_l) + |\gamma_k|\alpha(u_k - w_k)$$

Let $v_{kl} = \frac{1}{|\sigma_{kl}|} \int \sigma_{kl} \mathbf{v} \cdot \mathbf{n}_{kl} d\gamma$

Finite volumes for convection - diffusion II

- Central difference flux:

$$\begin{aligned} g(u_k, u_l) &= D(u_k - u_l) - h_{kl}\frac{1}{2}(u_k + u_l) v_{kl} \\ &= (D - \frac{1}{2} h_{kl} v_{kl}) u_k - (D + \frac{1}{2} h_{kl} v_{kl}) x u_l \end{aligned}$$

  - M-Property (sign pattern) only guaranteed for $h \to 0$!
- Upwind flux:

$$\begin{aligned} g(u_k, u_l) &= D(u_k - u_l) + \begin{cases} h_{kl} u_k v_{kl}, & v_{kl} < 0 \\ h_{kl} u_l v_{kl}, & v_{kl} > 0 \end{cases} \\ &= (D + \tilde{D})(u_k - u_l) - h_{kl}\frac{1}{2}(u_k + u_l) v_{kl} \end{aligned}$$

  - M-Property guaranteed unconditonally !
  - Artificial diffusion $\tilde{D} = \frac{1}{2} h_{kl} |v_{kl}|$

Finite volumes for convection - diffusion: exponential fitting

Project equation onto edge $x_K x_L$ of length $h = h_{kl}$, integrate once - $q = -v_{kl}$

$$\begin{aligned} c' + cq &= j \\ c|_0 &= c_K \\ c|_h &= c_L \end{aligned}$$

Solution of the homogeneus problem:

$$\begin{aligned} c' &= -cq \\ c'/c &= -q \\ \ln c &= c_0 - qx \\ c &= K \exp(-qx) \end{aligned}$$

Exponential fitting II

Solution of the inhomogeneous problem: set $K = K(x)$:

$$\begin{aligned} K' \exp(-qx) - qK \exp(-qx) + qK \exp(-qx) &= j \\ K' &= j \exp(qx) \\ K &= K_0 + \frac{1}{q} j \exp(qx) \end{aligned}$$

Therefore,

$$\begin{aligned} c &= K_0 \exp(-qx) + \frac{1}{q} j \\ c_K &= K_0 + \frac{1}{q} j \\ c_L &= K_0 \exp(-qh) + \frac{1}{q} j \end{aligned}$$

Exponential fitting III

Use boundary conditions

$$\begin{aligned} K_0 &= \frac{c_K - c_L}{1 - \exp(-qh)} \\ c_K &= \frac{c_K - c_L}{1 - \exp(-qh)} + \frac{1}{q} j \\ j &= q c_K - \frac{q}{1 - \exp(-qh)}(c_K - c_L) \\ &= q(1 - \frac{1}{1 - \exp(-qh)}) c_K - \frac{q}{\exp(-qh) - 1} c_L \\ &= q(\frac{-\exp(-qh)}{1 - \exp(-qh)}) c_K - \frac{q}{\exp(-qh) - 1} c_L \\ &= \frac{-q}{\exp(qh) - 1} c_K - \frac{q}{\exp(-qh) - 1} c_L \\ &= \frac{B(-qh) c_L - B(qh) c_K}{h} \end{aligned}$$

where $B(\xi) = \frac{\xi}{\exp(\xi) - 1}$: Bernoulli function

Exponential fitting IV

- Upwind flux:

$$g(u_k, u_l) = D\frac{B(\frac{-v_{kl}h_{kl}}{D})u_k - B(\frac{v_{kl}h_{kl}}{D})u_l}{h}$$

- Allen+Southell 1955
- Scharfetter+Gummel 1969
- Ilin 1969
- Chang+Cooper 1970
- Guaranteed M property!

Exponential fitting: Artificial diffusion

- Difference of exponential fitting scheme and central scheme
- Use: $B(-x) = B(x) + x \Rightarrow$

$$B(x) + \frac{1}{2}x = B(-x) - \frac{1}{2}x = B(|x|) + \frac{1}{2}|x|$$

$$\begin{aligned} D_{art}(u_k - u_l) =& D(B(\frac{vh}{D})u_k - B(\frac{-vh}{D})u_l) - D(u_k - u_l) + h\frac{1}{2}(u_k + u_l)v \\ =& D(\frac{vh}{2D} + B(\frac{vh}{D}))u_k - D(\frac{-vh}{2D} + B(\frac{-vh}{D})u_l) - D(u_k - u_l) \\ =& D(\frac{1}{2}|\frac{vh}{D}| + B(|\frac{vh}{D}|) - 1)(u_k - u_l) \end{aligned}$$

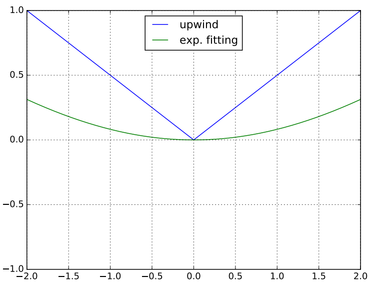
- Further, for $x > 0$:

$$\frac{1}{2}x \geq \frac{1}{2}x + B(x) - 1 \geq 0$$

- Therefore

$$\frac{|vh|}{2} \geq D_{art} \geq 0$$

Exponential fitting: Artificial diffusion II



Comparison of artificial diffusion functions $\frac{1}{2}|x|$ (upwind) and $\frac{1}{2}|x| + B(|x|) - 1$ (exp. fitting)

Convection-Diffusion implementation: central differences

```
F=0;
U=0;
for (int k=0, l=1;k<n-1;k++,l++)
{
  double g_kl=D - 0.5*(v*h);
  double g_lk=D + 0.5*(v*h);
  M(k,k)+=g_kl/h;
  M(k,l)-=g_kl/h;
  M(l,l)+=g_lk/h;
  M(l,k)-=g_lk/h;
}
M(0,0)+=1.0e30;
M(n-1,n-1)+=1.0e30;
F(n-1)=1.0e30;
```


Convection-Diffusion implementation: upwind scheme

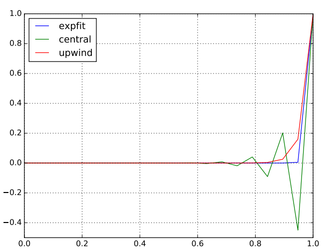
```
F=0;
U=0;
for (int k=0, l=1;k<n-1;k++,l++)
{
  double g_kl=D;
  double g_lk=D;
  if (v<0) g_kl-=v*h;
  else  g_lk+=v*h;

  M(k,k)+=g_kl/h;
  M(k,l)-=g_kl/h;
  M(l,l)+=g_lk/h;
  M(l,k)-=g_lk/h;
}

M(0,0)+=1.0e30;
M(n-1,n-1)+=1.0e30;
F(n-1)=1.0e30;
```


Convection-Diffusion implementation: exponential fitting scheme

```
 inline double B(double x)
 {
   if (std::fabs(x)<1.0e-10) return 1.0;
   return x/(std::exp(x)-1.0);
 }

...

   F=0;
   U=0;
   for (int k=0, l=1;k<n-1;k++,l++)
   {
     double g_kl=D* B(v*h/D);
     double g_lk=D* B(-v*h/D);
     M(k,k)+=g_kl/h;
     M(k,l)-=g_kl/h;
     M(l,l)+=g_lk/h;
     M(l,k)-=g_lk/h;
   }

   M(0,0)+=1.0e30;
   M(n-1,n-1)+=1.0e30;
   F(n-1)=1.0e30;
```


Convection-Diffusion test problem, N=20

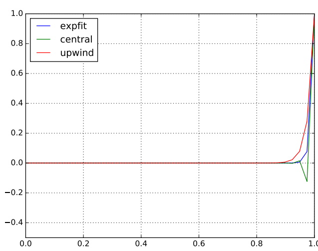
- $\Omega = (0, 1)$, $-\nabla \cdot (D\nabla u + uv) = 0$, $u(0) = 0$, $u(1) = 1$
- $V = 1$, $D = 0.01$



- Exponential fitting: sharp boundary layer, for this problem it is exact
- Central differences: unphysical
- Upwind: larger boundary layer

Convection-Diffusion test problem, N=40

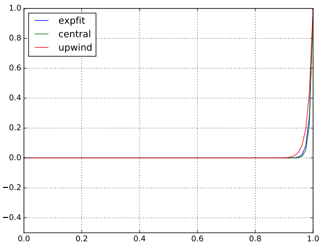
- $\Omega = (0, 1)$, $-\nabla \cdot (D\nabla u + uv) = 0$, $u(0) = 0$, $u(1) = 1$
- $V = 1$, $D = 0.01$



- Exponential fitting: sharp boundary layer, for this problem it is exact
- Central differences: unphysical, but less "wiggles"
- Upwind: larger boundary layer

Convection-Diffusion test problem, N=80

- $\Omega = (0, 1)$, $-\nabla \cdot (D\nabla u + uv) = 0$, $u(0) = 0$, $u(1) = 1$
- $V = 1$, $D = 0.01$



- Exponential fitting: sharp boundary layer, for this problem it is exact
- Central differences: grid is fine enough to yield M-Matrix property, good approximation of boundary layer due to higher convergence order
- Upwind: "smearing" of boundary layer

1D convection diffusion summary

- upwinding and exponential fitting unconditionally yield the *M*-property of the discretization matrix
- exponential fitting for this case (zero right hand side, 1D) yields exact solution. It is anyway "less diffusive" as artificial diffusion is optimized
- central scheme has higher convergence order than upwind (and exponential fitting) but on coarse grid it may lead to unphysical oscillations
- for 2/3D problems, sufficiently fine grids to stabilize central scheme may be prohibitively expensive
- local grid refinement may help to offset artificial diffusion

Convection-diffusion and finite elements

Search function $u : \Omega \to \mathbb{R}$ such that

$$
\begin{aligned}
-\nabla(\cdot D\nabla u - u\mathbf{v}) &= f \quad \text{in } \Omega \\
u &= u_D \text{on} \partial\Omega
\end{aligned}
$$

- Assume v is divergence-free, i.e. $\nabla \cdot v = 0$.
- Then the main part of the equation can be reformulated as

$$-\nabla(\cdot D\nabla u) + v \cdot \nabla u = 0 \quad \text{in } \Omega$$

yielding a weak formulation: find $u \in H^1(\Omega)$ such that $u - u_D \in H^1_0(\Omega)$ and $\forall w \in H^1_0(\Omega)$,

$$\int_\Omega D\nabla u \cdot \nabla w \; dx + \int_\Omega \mathbf{v} \cdot \nabla u \; w \; dx = \int_\Omega fw \; dx$$

- Galerkin formulation: find $u_h \in V_h$ with bc. such that $\forall w_h \in V_h$

$$\int_\Omega D\nabla u_h \cdot \nabla w_h \; dx + \int_\Omega \mathbf{v} \cdot \nabla u_h \; w_h \; dx = \int_\Omega fw_h \; dx$$

Convection-diffusion and finite elements II

- Galerkin ansatz has similar problems as central difference ansatz in the finite volume/finite difference case *Rightarrow* stabilization ?
- Most popular: streamline upwind Petrov-Galerkin

$$\int_\Omega D\nabla u_h \cdot \nabla w_h \; dx + \int_\Omega \mathbf{v} \cdot \nabla u_h \; w_h \; dx + S(u_h, w_h) = \int_\Omega fw_h \; dx$$

with

$$S(u_h, w_h) = \sum_K \int_K (-\nabla(\cdot D\nabla u_h - u_h\mathbf{v}) - f)\delta_K v \cdot w_h \; dx$$

where $\delta_K = \frac{h^v_K}{2|\mathbf{v}|}\xi(\frac{|\mathbf{v}|h^v_K}{D})$ with $\xi(\alpha) = coth(\alpha) - \frac{1}{\alpha}$ and h^v_K is the size of element K in the direction of $\mathbf{v}$.

Convection-diffusion and finite elements III

- Many methods to stabilize, *none* guarantees M-Property even on weakly acute meshes ! (V. John, P. Knobloch, Computer Methods in Applied Mechanics and Engineering, 2007)
- Comparison paper:

  M. Augustin, A. Caiazzo, A. Fiebach, J. Fuhrmann, V. John, A. Linke, and R. Umla, "An assessment of discretizations for convection-dominated convection-diffusion equations," Comp. Meth. Appl. Mech. Engrg., vol. 200, pp. 3395–3409, 2011:

  - if it is necessary to compute solutions without spurious oscillations: use FVM, taking care on the construction of an appropriate grid might be essential for reducing the smearing of the layers,
  - if sharpness and position of layers are important and spurious oscillations can be tolerated: often the SUPG method is a good choice.

- Topic of ongoing research

~

Nonlinear problems

Nonlinear problems: motivation

- Assume nonlinear dependency of some coefficients of the equation on the solution. E.g. nonlinear diffusion problem

$$
\begin{aligned}
-\nabla(\cdot D(u)\nabla u) &= f \quad \text{in } \Omega \\
u &= u_D \text{on} \partial\Omega
\end{aligned}
$$

- FE+FV discretization methods lead to large nonlinear systems of equations

Nonlinear problems: caution!

This is a significantly more complex world:

- Possibly multiple solution branches
- Weak formulations in L^p spaces
- No direct solution methods
- Narrow domains of definition (e.g. only for positive solutions)

Finite element discretization for nonlinear diffusion

- Find $u_h \in V_h$ such that for all $w_h \in V_h$:
$$\int_\Omega D(u_h)\nabla u_h \cdot \nabla w_h \, dx = \int_\Omega f w_h \, dx$$
- Use appropriate quadrature rules for the nonlinear integrals
- Discrete system
$$A(u_h) = F(u_h)$$

Finite volume discretization for nonlinear diffusion

$$\begin{aligned}
0 &= \int_{\omega_k} (-\nabla \cdot D(u)\nabla u - f)\, d\omega \\
&= -\int_{\partial\omega_k} D(u)\nabla u \cdot \mathbf{n}_k d\gamma - \int_{\omega_k} f d\omega \qquad \text{(Gauss)} \\
&= -\sum_{L\in\mathcal{N}_k} \int_{\sigma_{kl}} D(u)\nabla u \cdot \mathbf{n}_{kl} d\gamma - \int_{\gamma_k} D(u)\nabla u \cdot \mathbf{n} d\gamma - \int_{\omega_k} f d\omega \\
&\approx \sum_{L\in\mathcal{N}_k} \frac{\sigma_{kl}}{h_{kl}} g_{kl}(u_k, u_l) + |\gamma_k|\alpha(u_k - v_k) - |\omega_k| f_k
\end{aligned}$$

with
$$g_{kl}(u_k, u_l) = \begin{cases} D(\frac{1}{2}(u_k + u_l))(u_k - u_l) \\ \text{or} \\ \mathcal{D}(u_k) - \mathcal{D}(u_l) \end{cases}$$

where $\mathcal{D}(u) = \int_0^u D(\xi)\, d\xi$ (from exact solution ansatz at discretization edge)

- Discrete system
$$A(u_h) = F(u_h)$$

Iterative solution methods: fixed point iteration

- Let $u \in \mathbb{R}^n$.
- Problem: $A(u) = f$:

Assume $A(u) = M(u)u$, where for each u, $M(u) : \mathbb{R}^n \to \mathbb{R}^n$ is a linear operator.

- Fixed point iteration scheme:
  1. Choose initial value u_0, $i \leftarrow 0$
  2. For $i \geq 0$, solve $M(u_i)u_{i+1} = f$
  3. Set $i \leftarrow i + 1$
  4. Repeat from 2) until converged
- Convergence criteria:
  - residual based: $||A(u) - f|| < \varepsilon$
  - update based $||u_{i+1} - u_i|| < \varepsilon$
- Large domain of convergence
- Convergence may be slow
- Smooth coefficients not necessary

Iterative solution methods: Newton method

- Let $u \in \mathbb{R}^n$.
- Solve
$$A(u) = \begin{pmatrix} A_1(u_1 \dots u_n) \\ A_2(u_1 \dots u_n) \\ \vdots \\ A_n(u_1 \dots u_n) \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix} = f$$
- Jacobi matrix (Frechet derivative) for given u: $A'(u) = (a_{kl})$ with
$$a_{kl} = \frac{\partial}{\partial u_l} A_k(u_1 \dots u_n)$$
- Iteration scheme
  1. Choose initial value u_0, $i \leftarrow 0$
  2. Calculate residual $r_i = A(u_i) - f$
  3. Calculate Jacobi matrix $A'(u_i)$
  4. Solve update problem $A'(u_i)h_i = r_i$
  5. Update solution: $u_{i+1} = u_i - h_i$
  6. Set $i \leftarrow i + 1$
  7. Repeat from 2) until converged
- Convergence criteria:
  - residual based: $||r_i|| < \varepsilon$
  - update based $||h_i|| < \varepsilon$
- Limited domain of convergence
- Slow initial convergence
- Fast (quadratic) convergence close to solution

Newton method II

- Remedies for small domain of convergence: damping
  1. Choose initial value u_0, $i \leftarrow 0$,
     damping parameter $d < 1$:
  2. Calculate residual $r_i = A(u_i) - f$
  3. Calculate Jacobi matrix $A'(u_i)$
  4. Solve update problem $A'(u_i)h_i = r_i$
  5. Update solution: $u_{i+1} = u_i - dh_i$
  6. Set $i \leftarrow i + 1$
  7. Repeat from 2) until converged
- Damping slows convergence
- Better way: increase damping parameter during iteration:
  1. Choose initial value u_0, $i \leftarrow 0$,
     damping parameter d_0,
     damping growth factor $\delta > 1$
  2. Calculate residual $r_i = A(u_i) - f$
  3. Calculate Jacobi matrix $A'(u_i)$
  4. Solve update problem $A'(u_i)h_i = r_i$
  5. Update solution: $u_{i+1} = u_i - d_i h_i$
  6. Update damping parameter: $d_{i+1} = \min(1, \delta d_i)$
     Set $i \leftarrow i + 1$
  7. Repeat from 2) until converged

Newton method III

- Even if it converges, in each iteration step we have to solve linear system of equations
- can be done iteratively, e.g. with the LU factorization of the Jacobi matrix from first solution step
- iterative solution accuracy my be relaxed, but this may diminuish quadratic convergence
- Quadratic convergence yields very accurate solution with no large additional effort: once we are in the quadratic convergence region, convergence is very fast
- Monotonicity test: check if residual grows, this is often an sign that the iteration will diverge anyway.

Newton method IV

- Embedding method for parameter dependent problems.
- Solve $A(u_\lambda, \lambda) = f$ for $\lambda = 1$.
- Assume $A(u_0, 0)$ can be easily solved.
- Parameter embedding method:
  1. Solve $A(u_0, 0) = f$
     choose step size δ Set $\lambda = 0$
  2. Solve $A(u_{\lambda+\delta}, \lambda + \delta) = 0$ with initial value u_λ. Possibly decrease δ to achieve convergence
  3. Set $\lambda \leftarrow \lambda + \delta$
  4. Possibly increase δ
  5. Repeat from 2) until $\lambda = 1$
- Parameter embedding + damping + update based convergence control go a long way to solve even strongly nonlinear problems!