Scientific Computing WS 2018/2019

Lecture 26

Jürgen Fuhrmann

juergen.fuhrmann@wias-berlin.de

Recapitulation I: Finite Elements

- Strong formulation of PDE
- Problems with strong formulation
- Weak formulation of PDE, solvability
- Galerkin ansatz
- Matrix form
- Matrix element calculation
- Matrix properties
- Solution of matrix problem

Second order elliptic PDEs

Stationary case: $\partial_t u = 0 \Rightarrow$ second order *elliptic* PDE

$$-\nabla \cdot (\delta \nabla u(x)) = f(x)$$

- Stationary heat conduction, stationary diffusion
- Incompressible flow in saturated porous media: u: pressure
  $\delta = k$: permeability, flux$=-k\nabla u$: "Darcy's law"
- Electrical conduction: u: electric potential
  $\delta = \sigma$: electric conductivity
  flux$=-\sigma\nabla u \equiv$ current density: "Ohms's law"
- Poisson equation (electrostatics in a constant magnetic field):
  u: electrostatic potential, ∇u: electric field,
  $\delta = \varepsilon$: dielectric permittivity, f: charge density

Second order PDEs: boundary conditions

- Combine PDE in the interior with boundary conditions on variable u and/or or normal flux $\delta\nabla u\cdot\mathbf{n}$
- Assume $\partial\Omega=\cup_{i=1}^{N_\Gamma}\Gamma_i$ is the union of a finite number of non-intersecting subsets Γ_i which are locally Lipschitz.
- On each Γ_i, specify one of
  - Dirichlet ("first kind"): let $g_i:\Gamma_i\to\mathbb{R}$ (homogeneous for $g_i=0$)

    $$u(x)=u_{\Gamma_i}(x)\quad\text{for } x\in\Gamma_i$$

  - Neumann ("second kind"): Let $g_i:\Gamma_i\to\mathbb{R}$ (homogeneus for $g_i=0$)

    $$\delta\nabla u(x)\cdot\mathbf{n}=g_i(x)\quad\text{for } x\in\Gamma_i$$

  - Robin ("third kind"): let $\alpha_i, g_i:\Gamma_i\to\mathbb{R}$

    $$\delta\nabla u(x)\cdot\mathbf{n}+\alpha_i(x)\left(u(x)-g_i(x)\right)=0\quad\text{for } x\in\Gamma_i$$

- Boundary functions may be time dependent.

Problems with "strong formulation"

Writing the PDE with divergence and gradient assumes smoothness of coefficients and at least second derivatives for the solution.

- δ may not be continuous – what is then $\nabla \cdot (\delta \nabla u)$?
- Approximation of solution u e.g. by piecewise linear functions what does ∇u mean ?
- Spaces of twice, and even once continuously differentiable functions is not well suited:
  - Favorable approximation functions (e.g. piecewise linear ones) are not contained
  - Though they can be equipped with norms ($\Rightarrow$ Banach spaces) they have no scalar product $\Rightarrow$ no Hilbert spaces
  - Not complete: Cauchy sequences of functions may not converge to elements in these spaces

Derivation of weak formulation

- Sobolev space theory provides a convenient framework to formulate existence, uniqueness and approximations of solutions of PDEs.
- Stationary heat conduction equation with homogeneous Dirichlet boundary conditions:

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

Multiply and integrate with an arbitrary *test function* $v \in C_0^\infty(\Omega)$ and apply Green's theorem using $v = 0$ on $\partial\Omega$

$$
\begin{aligned}
-\int_\Omega (\nabla \cdot \lambda \nabla u) v \, d\mathbf{x} &= \int_\Omega f v \, d\mathbf{x} \\
\int_\Omega \lambda \nabla u \nabla v \, d\mathbf{x} &= \int_\Omega f v \, d\mathbf{x}
\end{aligned}
$$

Weak formulation of homogeneous Dirichlet problem

- Search $u \in H_0^1(\Omega)$ (here, $\operatorname{tr} u = 0$) such that

$$\int_\Omega \lambda \nabla u \nabla v \, d\mathbf{x} = \int_\Omega fv \, d\mathbf{x} \; \forall v \in H_0^1(\Omega)$$

- Then,

$$a(u, v) := \int_\Omega \lambda \nabla u \nabla v \, d\mathbf{x}$$

  is a self-adjoint bilinear form defined on the Hilbert space $H_0^1(\Omega)$.
- It is bounded due to Cauchy-Schwarz:

$$|a(u, v)| = |\lambda| \cdot |\int_\Omega \nabla u \nabla v \, d\mathbf{x}| \leq ||u||_{H_0^1(\Omega)} \cdot ||v||_{H_0^1(\Omega)}$$

- $f(v) = \int_\Omega fv \, d\mathbf{x}$ 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

Theorem: 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.

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

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

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

$\square$

Coercivity of weak formulation

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

$$\int_\Omega \lambda \nabla u \nabla v \, d\mathbf{x} = \int_\Omega fv \, d\mathbf{x} \; \forall v \in H_0^1(\Omega)$$

has an unique solution.

Proof: $a(u, v)$ is cocercive:

$$a(u, v) = \int_\Omega \lambda \nabla u \nabla u \, d\mathbf{x} = \lambda ||u||^2_{H_0^1(\Omega)}$$

$\square$

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

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

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

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

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

- As a result

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

- 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

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

  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 ?

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 $a = x_1 \leq x_2 \leq \cdots \leq x_n = b$
- Basis functions (for $i = 1 \ldots n$)

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

- Any function $u_h \in V_h = \mathrm{span}\{\phi_1 \ldots \phi_n\}$ 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 !

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 ($i = 0 \dots d$):

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

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:

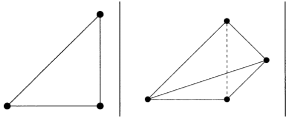
$$\dim \mathbb{P}_k = \binom{d+k}{k} = \begin{cases} k+1, & d=1 \\ \frac{1}{2}(k+1)(k+2), & d=2 \\ \frac{1}{6}(k+1)(k+2)(k+3), & d=3 \end{cases}$$

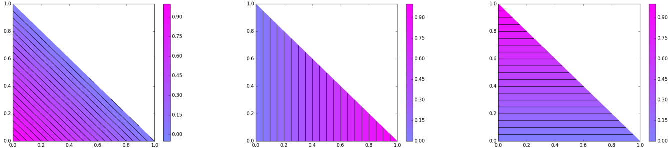
$$\dim \mathbb{P}_1 = d+1$$

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

$\mathbb{P}_1$ 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





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, $m = 1 \dots n_e$:

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

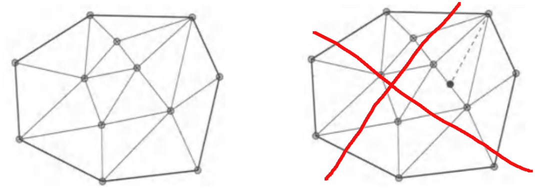
- Each simplex can be seen as the image of a affine transformation of a reference (e.g. unit) simplex $\widehat{K}$:

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

- 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
- Delaunay triangulations are conformal

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

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

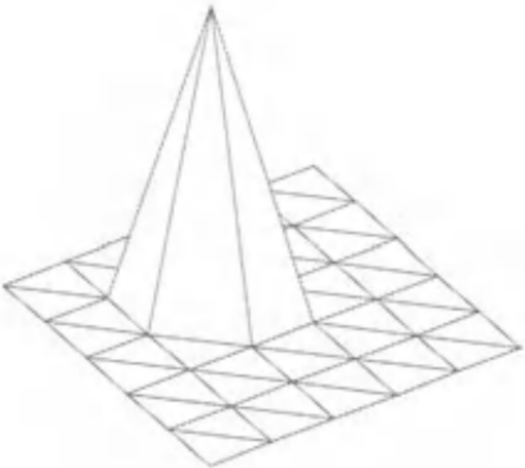
- Global shape functions $\phi_1, \dots, \phi_N \in W_h$ defined by

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

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

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

P^1 global shape functions



Lecture 15 Slide 49

Stiffness matrix for Laplace operator for P1 FEM

- Element-wise calculation:

$$a_{ij} = a(\phi_i, \phi_j) = \int_\Omega \nabla\phi_i \nabla\phi_j \; d\mathbf{x} = \int_\Omega \sum_{K\in\mathcal{T}_h} \nabla\phi_i|_K \nabla\phi_j|_K \; d\mathbf{x}$$

- Standard assembly loop:

for $i, j = 1 \dots N$ **do**
  set $a_{ij} = 0$
end
for $K \in \mathcal{T}_h$ **do**
  for $m,n=0 \dots d$ **do**

$$s_{mn} = \int_K \nabla\lambda_m \nabla\lambda_n \; d\mathbf{x}$$

$$a_{j_{dof}(K,m),j_{dof}(K,n)} = a_{j_{dof}(K,m),j_{dof}(K,n)} + s_{mn}$$

  end
end

- Local stiffness matrix:

$$S_K = (s_{K;m,n}) = \int_K \nabla\lambda_m \nabla\lambda_n \; d\mathbf{x}$$

Error estimates for homogeneous Dirichlet problem

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

$$\int_\Omega \lambda \nabla u \nabla v \, d\mathbf{x} = \int_\Omega fv \, d\mathbf{x} \; \forall v \in H_0^1(\Omega)$$

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

$$||u - u_h||_{1,\Omega} \leq ch|u|_{2,\Omega}$$
$$||u - u_h||_{0,\Omega} \leq ch^2|u|_{2,\Omega}$$

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

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

("Aubin-Nitsche-Lemma")

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 rate
  - Remedy: local refinement of the discretization mesh
    - using a priori information
    - using a posteriori error estimators + automatic refinement of discretizatiom mesh

More complicated integrals

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

$$f_i = \int_K f(x)\lambda_i(x)\, d\mathbf{x}$$

- P^1 stiffness matrix elements

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

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

Quadrature rules

- *Quadrature rule*:

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

- 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.

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

- Error estimate:

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

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	$(\frac{5-\sqrt{5}}{20}, \frac{5-\sqrt{5}}{20}, \frac{5-\sqrt{5}}{20}, \frac{5+3\sqrt{5}}{20}) \ldots$	$\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}$

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.

P1 FEM stiffness matrix condition number

- Homogeneous dirichlet boundary value problem

$$-\nabla \cdot \kappa \nabla u = f \quad \text{in } \Omega$$
$$u|_{\partial\Omega} = 0$$

- Lagrange degrees of freedom $a_1 \dots a_N$ corresponding to global basis functions $\phi_1 \dots \phi_N$ such that $\phi_i|_{\partial\Omega} = 0$ aka $\phi_i \in V_h \subset H_0^1(\Omega)$
- Stiffness matrix $A = (a_{ij})$:

$$a_{ij} = a(\phi_i, \phi_j) = \int_\Omega \kappa \nabla \phi_i \nabla \phi_j \, d\mathbf{x}$$

- 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:

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

The problem with Dirichlet boundary conditions

- Homogeneous Dirichlet BC $\Rightarrow$ include boundary condition into set of basis functions
- Inhomogeneous Dirichlet, may be only at a part of the boundary
  - Use exact approach from as in continous formulation (with lifting u_g etc) $\Rightarrow$ highly technical
  - Eliminate Dirichlet BC algebraically after building of the matrix, i.e. fix "known unknowns" at the Dirichlet boundary $\Rightarrow$ highly technical
  - Modifiy matrix such that equations at boundary exactly result in Dirichlet values $\Rightarrow$ loss of symmetry of the matrix
  - Penalty method

Dirichlet BC: Algebraic manipulation

- Assume 1D situation with BC $u_1 = g$
- From integration in H^1 regardless of boundary values:

$$AU = \begin{pmatrix} \frac{1}{h} & -\frac{1}{h} & & & \\ -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & & \\ & -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & \\ & & \ddots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \end{pmatrix}$$

- Fix u_1 and eliminate:

$$A'U = \begin{pmatrix} \frac{2}{h} & -\frac{1}{h} & & \\ -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & \\ & \ddots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} u_2 \\ u_3 \\ \vdots \end{pmatrix} = \begin{pmatrix} f_2 + \frac{1}{h}g \\ f_3 \\ \vdots \end{pmatrix}$$

- A' becomes idd and stays symmetric
- operation is quite technical

Dirichlet BC: Modify boundary equations

- From integration in H^1 regardless of boundary values:

$$AU = \begin{pmatrix} \frac{1}{h} & -\frac{1}{h} & & & \\ -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & & \\ & -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & \\ & & \ddots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \end{pmatrix}$$

- Modify equation at boundary to exactly represent Dirichlet values

$$A'U = \begin{pmatrix} \frac{1}{h} & 0 & & & \\ -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & & \\ & -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & \\ & & \ddots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \end{pmatrix} = \begin{pmatrix} \frac{1}{h}g \\ f_2 \\ f_3 \\ \vdots \end{pmatrix}$$

- A' becomes idd
- loses symmetry $\Rightarrow$ problem e.g. with CG method

Dirichlet BC: Discrete penalty trick

- From integration in H^1 regardless of boundary values:

$$AU = \begin{pmatrix} \frac{1}{h} & -\frac{1}{h} & & & \\ -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & & \\ & -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & \\ & & \ddots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \end{pmatrix}$$

- Add penalty terms

$$A'U = \begin{pmatrix} \frac{1}{\varepsilon} + \frac{1}{h} & -\frac{1}{h} & & & \\ -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & & \\ & -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & \\ & & \ddots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \end{pmatrix} = \begin{pmatrix} f_1 + \frac{1}{\varepsilon} g \\ f_2 \\ f_3 \\ \vdots \end{pmatrix}$$

- A' becomes idd, keeps symmetry, and the realization is technically easy.
- If ε is small enough, $u_1 = g$ will be satisfied exactly within floating point accuracy.
- Iterative methods should be initialized with Dirichlet values.
- Works for nonlinear problems, finite volume methods

Dirichlet penalty trick, general formulation

- Dirichlet boundary value problem

$$-\nabla \cdot \kappa \nabla u = f \quad \text{in } \Omega$$
$$u|_{\partial\Omega} = g$$

- We discussed approximation of Dirichlet problem by Robin problem
- Practical realization uses discrete approach for Lagrange degrees of freedom $a_1 \dots a_N$ corresponding to global basis functions $\phi_1 \dots \phi_N$:
- Search $u_h = \sum_{i=1}^N u_i \phi_i \in V_h = \mathrm{span}\{\phi_1 \dots \phi_N\}$ such that

$$AU + \Pi U = F + \Pi G$$

where
  - $U = (u_1 \dots u_N)$
  - $A = (a_{ij})$: stiffness matrix with $a_{ij} = \int_\Omega \kappa \nabla \phi_i \nabla \phi_j \, d\mathbf{x}$
  - $F = \int_\Omega f \nabla \phi_i \, d\mathbf{x}$
  - $G = (g_i)$ with $g_i = \begin{cases} g(a_i), & a_i \in \partial\Omega \\ 0, & \text{else} \end{cases}$
  - $\Pi = (\pi_{ij})$ is a diagonal matrix with $\pi_{ij} = \begin{cases} \frac{1}{\varepsilon}, & i = j, a_i \in \partial\Omega \\ 0, & \text{else} \end{cases}$

Solution of SPD system as a minimization procedure

Regard $Au = f$,where A is symmetric, positive definite. Then it defines a bilinear form $a : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$

$$a(u, v) = (Au, v) = v^T Au = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} v_i u_j$$

As A is SPD, for all $u \neq 0$ we have $(Au, u) > 0$.

For a given vector b, regard the function

$$f(u) = \frac{1}{2} a(u, u) - b^T u$$

What is the minimizer of f ?

$$f'(u) = Au - b = 0$$

- Solution of SPD system $\equiv$ minimization of f.

Method of steepest descent

- Given some vector u_i, look for a new iterate u_{i+1}.
- The direction of steepest descend is given by $-f'(u_i)$.
- So look for u_{i+1} in the direction of $-f'(u_i) = r_i = b - Au_i$ such that it minimizes f in this direction, i.e. set $u_{i+1} = u_i + \alpha r_i$ with α choosen from

$$
\begin{aligned}
0 &= \frac{d}{d\alpha} f(u_i + \alpha r_i) = f'(u_i + \alpha r_i) \cdot r_i \\
&= (b - A(u_i + \alpha r_i), r_i) \\
&= (b - Au_i, r_i) - \alpha(Ar_i, r_i) \\
&= (r_i, r_i) - \alpha(Ar_i, r_i) \\
\alpha &= \frac{(r_i, r_i)}{(Ar_i, r_i)}
\end{aligned}
$$

Method of steepest descent: iteration scheme

$$r_i = b - Au_i$$
$$\alpha_i = \frac{(r_i, r_i)}{(Ar_i, r_i)}$$
$$u_{i+1} = u_i + \alpha_i r_i$$

Let $\hat{u}$ the exact solution. Define $e_i = u_i - \hat{u}$, then $r_i = -Ae_i$

Let $||u||_A = (Au, u)^{\frac{1}{2}}$ be the *energy norm* wrt. A.

Theorem The convergence rate of the method is

$$||e_i||_A \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^i ||e_0||_A$$

where $\kappa = \frac{\lambda_{max}(A)}{\lambda_{min}(A)}$ is the spectral condition number.

Method of steepest descent: advantages

- Simple Richardson iteration $u_{k+1} = u_k - \alpha(Au_k - f)$ needs good eigenvalue estimate to be optimal with $\alpha = \frac{2}{\lambda_{max} + \lambda_{min}}$
- In this case, asymptotic convergence rate is $\rho = \frac{\kappa - 1}{\kappa + 1}$
- Steepest descent has the same rate without need for spectral estimate

Conjugate directions

For steepest descent, there is no guarantee that a search direction $d_i = r_i = -Ae_i$ is not used several times. If all search directions would be orthogonal, or, indeed, A-orthogonal, one could control this situation.

So, let $d_0, d_1 \dots d_{n-1}$ be a series of A-orthogonal (or conjugate) search directions, i.e. $(Ad_i, d_j) = 0$, $i \neq j$.

- Look for u_{i+1} in the direction of d_i such that it minimizes f in this direction, i.e. set $u_{i+1} = u_i + \alpha_i d_i$ with α choosen from

$$
\begin{aligned}
0 &= \frac{d}{d\alpha} f(u_i + \alpha d_i) = f'(u_i + \alpha d_i) \cdot d_i \\
&= (b - A(u_i + \alpha d_i), d_i) \\
&= (b - Au_i, d_i) - \alpha(Ad_i, d_i) \\
&= (r_i, d_i) - \alpha(Ad_i, d_i) \\
\alpha_i &= \frac{(r_i, d_i)}{(Ad_i, d_i)}
\end{aligned}
$$

Gram-Schmidt Orthogonalization

- Assume we have been given some linearly independent vectors $v_0, v_1 \dots v_{n-1}$.
- Set $d_0 = v_0$
- Define

$$d_i = v_i + \sum_{k=0}^{i-1} \beta_{ik} d_k$$

- For $j < i$, A-project onto d_j and require orthogonality:

$$(Ad_i, d_j) = (Av_i, d_j) + \sum_{k=0}^{i-1} \beta_{ik} (Ad_k, d_j)$$
$$0 = (Av_i, d_j) + \beta_{ij} (Ad_j, d_j)$$
$$\beta_{ij} = -\frac{(Av_i, d_j)}{(Ad_j, d_j)}$$

- If v_i are the coordinate unit vectors, this is Gaussian elimination!
- If v_i are arbitrary, they all must be kept in the memory

Conjugate gradients IV - The algorithm

Given initial value u_0, spd matrix A, right hand side b.

$$
\begin{aligned}
d_0 &= r_0 = b - Au_0 \\
\alpha_i &= \frac{(r_i, r_i)}{(Ad_i, d_i)} \\
u_{i+1} &= u_i + \alpha_i d_i \\
r_{i+1} &= r_i - \alpha_i A d_i \\
\beta_{i+1} &= \frac{(r_{i+1}, r_{i+1})}{(r_i, r_i)} \\
d_{i+1} &= r_{i+1} + \beta_{i+1} d_i
\end{aligned}
$$

At the i-th step, the algorithm yields the element from $e_0 + \mathcal{K}_i$ with the minimum energy error.

Theorem The convergence rate of the method is

$$||e_i||_A \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^i ||e_0||_A$$

where $\kappa = \frac{\lambda_{max}(A)}{\lambda_{min}(A)}$ is the spectral condition number.

Preconditioned CG II

Assume $\tilde{r}_i = E^{-1} r_i$, $\tilde{d}_i = E^T d_i$, we get the equivalent algorithm

$$
\begin{aligned}
r_0 &= b - Au_0 \\
d_0 &= M^{-1} r_0 \\
\alpha_i &= \frac{(M^{-1} r_i, r_i)}{(Ad_i, d_i)} \\
u_{i+1} &= u_i + \alpha_i d_i \\
r_{i+1} &= r_i - \alpha_i A d_i \\
\beta_{i+1} &= \frac{(M^{-1} r_{i+1}, r_{i+1})}{(r_i, r_i)} \\
d_{i+1} &= M^{-1} r_{i+1} + \beta_{i+1} d_i
\end{aligned}
$$

It relies on the solution of the preconditioning system, the calculation of the matrix vector product and the calculation of the scalar product.

Examinations

Tue Feb 26.
Wed Feb 27.
Wed Mar 14.
Thu Mar 15.
Tue Mar 26.
Wed Mar 27.
Thu Mar 28.
Wed May 8. 14:00-17:00

- 13:00 times do **not** work! Please reschedule (sorry).