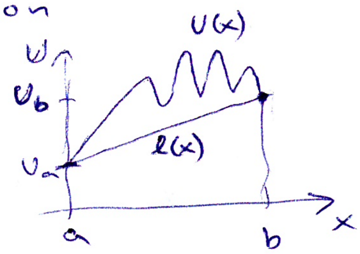
Finite-element method for a model problem

- let us consider a simple Poisson equation

$$-\frac{d^2u}{dx^2} = f(x) \quad , \; u(a) = u_a \, , \quad u(b) = u_b$$

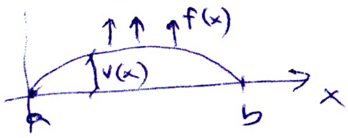


by substitution $\quad u(x) = v(x) + \ell(x)$

where $\quad \ell(x) = u_a \dfrac{x-b}{a-b} + u_b \dfrac{x-a}{b-a}$

we get even simpler problem

this models e.g. a string pulled by force $f(x)$



(D) $\qquad -\dfrac{d^2v}{dx^2} = f(x) \; , \; v(a) = v(b) = 0$

this is our model problem in the differential form

- standard assumptions are that $v(x)$ is continuous and also its derivative is continuous, thus $f(x)$ can be only continuous ~~by parts~~ piecewise

(in general, the theory of PDEs searches the solutions on some Sobolev space)

- let us consider a space of functions

$$V = \{ v(x) : v \text{ is continuous on } \langle a,b \rangle, \; v' \text{ is piecewise continuous and bounded on } \langle a,b \rangle \text{ and } v(a) = v(b) = 0 \}$$

~~with~~ and the inner product (not on this space)

$$(v,w) = \int_a^b v(x)\,w(x)\,dx \quad \text{for real, piecewise continuous functions}$$

and the linear functional

$$F(v) = \frac{1}{2}(v',v') - (f,v) \qquad \sim \text{Lagrangian}$$

- We can show that the solution of (D) is also the solution of the weak formulation (integral formulation) of the problem

(W) $\quad$ Find $v(x) \in V$ such that $(v',w') = (f,w)$ for all $w \in V$.

which is moreover equivalent to the problem (variational)

(V) $\quad$ Find $v(x) \in V$ such that $F(v) \le F(w)$ for all $w \in V$.

- proof: (D) $\Rightarrow$ (W)
  - multiply (D) by $w \in V$ and integrate ~~per partes~~ by parts

$$-(v'', w) = (f, w) \Rightarrow (v', w') = (f, v)$$ by the choice of boundary conditions

(W) $\Rightarrow$ (V)
  - if $v(x)$ is the solution of (W), $w \in V$ and $z = w - v \in V$

then $F(w) = F(v+z) = \frac{1}{2}(v'+z', v'+z') - (f, v+z) =$

$$= \underbrace{\frac{1}{2}(v', v') - (f, v)}_{F(v)} + \underbrace{(v', z') - (f, z)}_{\substack{0 \\ \Uparrow \\ v(x) \text{ solves (W)}}} + \underbrace{\frac{1}{2}(z', z')}_{\geq 0} \geq F(v)$$ for all $w \in V$

and finally (V) $\Rightarrow$ (W)
  - if $v(x)$ is the solution of (V), then for $w \in V$ and real ε

we have $F(v) \leq F(v + \varepsilon w)$ $(v + \varepsilon w \in V)$

function of ε

$g(\varepsilon) = F(v + \varepsilon w) = \frac{1}{2}(v', v') + \varepsilon (v', w') + \frac{\varepsilon^2}{2}(w', w') - (f, v) - \varepsilon (f, w)$

has the minimum for $\varepsilon = 0$ or it must be $g'(0) = 0$

and thus $0 = (v', w') - (f, w)$ for arbitrary $w \in V$

- moreover, it can be shown that
  if $f(x)$ is continuous then the solution of the weak formulation (W) has also continuous second derivatives and we can integrate by parts the other way to get (D) from (W) and thus $v(x)$ is the solution of (D)

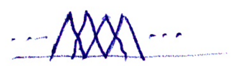
thus we have (D) $\Leftrightarrow$ (W) $\Leftrightarrow$ (V) for continuous $f(x)$

because the solution of (W) is unique
  - if v_1 and v_2 were two different solutions of (W)

then $\int_a^b (v_1' - v_2') w \, dx = 0$ for all $w \in V$ and for $w = v_1' - v_2'$

we get $\int_a^b (v_1' - v_2')^2 dx = 0 \Rightarrow v_1' - v_2' = 0$ for all $x \in \langle a, b)$

and $v_1 - v_2 = \text{konst} = 0$ using boundary conditions

- notice that (W) can have a solution even for $f(x)$ which ~~are~~ is not continuous, that's why we often say that (W) is a generalization of (D)
- an important thing is that we search the solution of (W) in the space V of functions that do not have necessarily continuous first derivatives
  (FEM works with such bases )

- Basic idea of the finite-element method
  - instead of the full V, take its subspace V_h that is finite
    (typically, a suitable basis constructed on the elements of mean size h will generate it)

  and solve the following problem

  (V_h) Find $v_h(x) \in V_h$ such that $F(v_h) \leq F(w)$ for all $w \in V_h$
  this is called Ritz-Galerkin method

  or solve
  (W_h) Find $v_h(x) \in V_h$ such that $(v_h', w') = (f, w)$ for all $w \in V_h$
  this is called Galerkin method

  - we usually talked about the finite-element method (FEM) if we take piecewise continuous polynomials as a basis of V_h

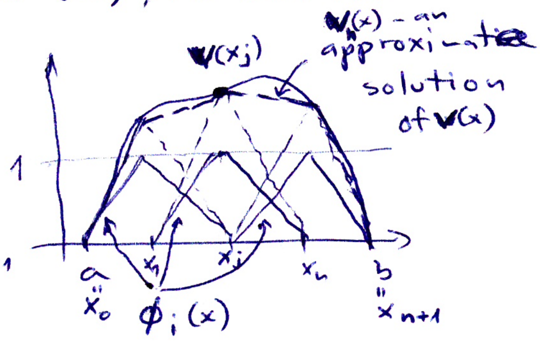
  thus FEM consists of three steps

  1) weak (or variational) formulation of the problem
  2) discretization of space (usually triangularization) and construction of a finite V_h by choice of a basis
  3) solution of the discretized problem resulting in a large, sparse system of linear eqs.

  - there are many ready-to-use software packages to apply to 2) and 3)

- **Example of a basis for the model problem**

- as the simplest basis for our problem (W), we can use piecewise linear functions with a compact support of size h



- we divide $\langle a,b\rangle$ to intervals $I_j = (x_{j-1}, x_j)$ of size $h_j = x_j - x_{j-1}$, $j = 1, \dots, n+1$

h_j can be different, but we use $h = \max_j h_j$ as a characteristic of the elements

- V_h is taken to be a space of piecewise linear continuous functions $w(x)$ satisfying $w(a) = w(b) = 0$, clearly $V_h \subset V$
- a basis is formed by $\varphi_k(x)$, $k = 1, \dots, n$ with property

$$\varphi_k(x_j) = \begin{cases} 1, & k = j \\ 0, & k \neq j \end{cases}, \quad k, j = 1, \dots, n$$

because an arbitrary function $w \in V_h$ can be expressed as $w(x) = \sum_{k=1}^{n} \eta_k \varphi_k(x)$, $x \in \langle a,b\rangle$

where $w(x_j) = \eta_j$

- thus V_h is an n-dimensional ~~linear~~ linear space

- to solve (W_h) in this space, we expand the approximate solution into the basis $V_h(x) = \sum_{k=1}^{n} \xi_k \varphi_k(x)$, $\xi_k = V_h(x_k)$

and as „testing" functions w, we take all basis functions, we get

$$\boxed{\sum_{k=1}^{n} \xi_k (\varphi_k', \varphi_\ell') = (f, \varphi_\ell)}, \quad \ell = 1, \dots, n$$

or in the matrix form $A\xi = b$, $\xi = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix}$

with $a_{k\ell} = (\varphi_k', \varphi_\ell')$ ... called <u>stiffness matrix</u>

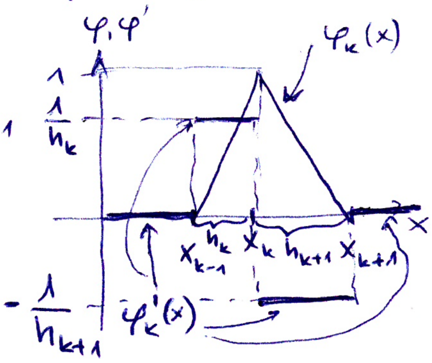
and $b_\ell = (f, \varphi_\ell)$ ... called <u>load vector</u>

(from the original FEM application to construction problems)

- mathematicians call A as Gram matrix

- for our problem, we can compute A explicitly
  - for derivatives of the basis we get

$$\varphi_k'(x) = \begin{cases} 0 & \text{for } x \le x_{k-1} \text{ and } x \ge x_{k+1} \\ \frac{1}{h_k} & \text{for } x_{k-1} < x \le x_k \\ -\frac{1}{h_{k+1}} & \text{for } x_k < x \le x_{k+1} \end{cases}$$



thus we find

$$(\varphi_k', \varphi_\ell') = 0 \quad \text{if } |k-\ell| > 1$$

length of the interval

$$(\varphi_k', \varphi_k') = \left(\frac{1}{h_k}\right)\left(\frac{1}{h_k}\right) h_k + \left(-\frac{1}{h_{k+1}}\right)\left(-\frac{1}{h_{k+1}}\right) h_{k+1} = \frac{1}{h_k} + \frac{1}{h_{k+1}}$$

and $$(\varphi_k', \varphi_{k-1}') = \left(\frac{1}{h_k}\right)\left(-\frac{1}{h_k}\right) h_k = -\frac{1}{h_k} = (\varphi_{k-1}', \varphi_k')$$

- for an equidistant grid $h_k = \frac{b-a}{n+1} = h$

  we get $(\varphi_k', \varphi_k') = \frac{2}{h}$, $(\varphi_k', \varphi_{k-1}') = -\frac{1}{h}$

  and the system to solve is

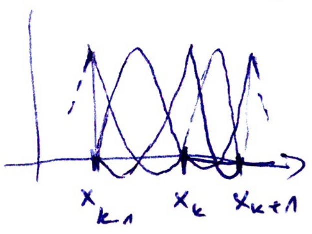
$$\frac{1}{h}\begin{pmatrix} 2 & -1 & & 0 \\ -1 & 2 & -1 & \\ & -1 & \ddots & -1 \\ 0 & & -1 & 2 \end{pmatrix}\begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

where $$b_k = \int_a^b f(x)\varphi_k(x)\,dx = \int_{x_{k-1}}^{x_{k+1}} f(x)\varphi_k(x)\,dx$$

if we approximate this integral by ~~the~~ trapezodial rule we have $b_k \approx h f(x_k)$

and we have exactly the same system of linear equations as we had using finite-difference method

- more accurate approximation can be obtained using piecewise quadratic, cubic, etc. functions



• Example of a higher-order basis – combination of FEM with DVR

- DVR means <u>discrete-varidble representation</u>
  - basic idea: because evaluation of matrix elements of the type $V_{ij} = \int \phi_i^* V(x) \phi_j \, dx$

    for some „potential" function $V(x)$, can be much more difficult than integrals $\int \phi_i^* X \phi_j \, dx$

    we can first diagonalize operator X as $\Lambda_x = UXU^\dagger$

    then $V(\Lambda_x)$ is trivial and we can get

    $$V(\text{in the basis } \phi_i) = U^\dagger V(\Lambda_x) U$$

  - even better idea: use a basis ϕ_i such that an arbitrary $V(x)$ is effectively diagonal (we will see later how)

- let us consider Schrödinger equation for a radial problem

$$-\frac{1}{2\mu}\frac{d^2\psi}{dr^2} + \left(\frac{\ell(\ell+1)}{2\mu r^2} + V(r)\right)\psi = E\psi$$

weak formulation leads to

$$\frac{1}{2\mu}\int_0^\infty \left(\frac{d\phi}{dr}\right)^* \left(\frac{d\psi}{dr}\right) dr + \int_0^\infty \phi^* \left(\frac{\ell(\ell+1)}{2\mu r^2} + V(r)\right)\psi dr = E\int_0^\infty \phi^* \psi \, dr$$

and using a basis $\{\phi_i(r)\}_{i=1}^n$ we get $\left(\psi = \sum_{i=1}^n c_i \phi_i(x)\right)$

$$\sum_{j=1}^n H_{ij} c_j = E c_j$$

with $$H_{ij} = \frac{1}{2\mu}\int_0^\infty \left(\frac{d\phi_i}{dr}\right)^* \left(\frac{d\phi_j}{dr}\right) dr + \int_0^\infty \phi_i^* \left(\frac{\ell(\ell+1)}{2\mu r^2} + V(r)\right)\phi_j \, dr$$

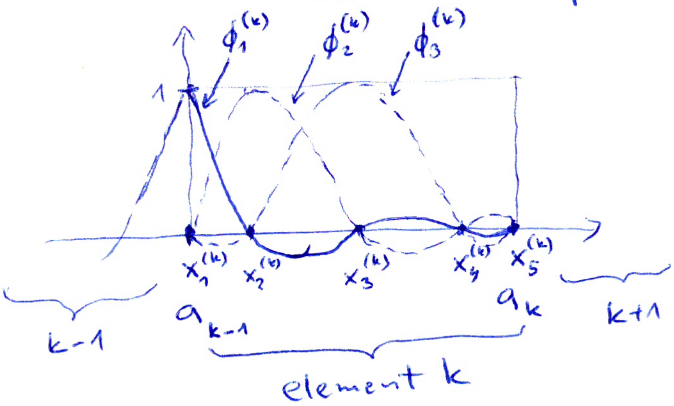
- if we need high accuracy, we have to use piesewise polynomial functions of higher order

- construction of a FEM-DVR basis

- basic idea: use Gauss-Lobatto quadrature (fixed boundary points) for accurate integration and construct a basis with functions that are zero at all points of the G-L quadrature except one $\Rightarrow$ for each point we get one basis function

1) basis on one element (interval)

- we use n points of G-L quadrature on each element $x_i^{(k)}$ with weights $w_i^{(k)}$ and construct basis functions



$$\phi_i^{(k)} = \frac{1}{\sqrt{w_i^{(k)}}} \prod_{\substack{j=1 \\ j \neq i}} \frac{x - x_j^{(k)}}{x_i^{(k)} - x_j^{(k)}}$$

which are Lagrange interpolating polynomials

- these functions satisfy $\phi_i^{(k)}(x_j^{(k)}) = \dfrac{\delta_{ij}}{\sqrt{w_i^{(k)}}}$

and thus these functions are effectively (not exactly) orthogonal if we use G-L quadrature:

$$\int_{a_{k-1}}^{a_k} \phi_i^{(k)}(x)\,\phi_j^{(k)}(x)\,dx \simeq \sum_{s=1}^{n} w_s^{(k)}\,\phi_i^{(k)}(x_s^{(k)})\,\phi_j^{(k)}(x_s^{(k)}) =$$

$$= \sum_{s=1}^{n} w_s^{(k)} \frac{\delta_{is}}{\sqrt{w_s^{(k)}}} \frac{\delta_{js}}{\sqrt{w_s^{(k)}}} = \delta_{ij}$$

2) basis on the whole interval divided to K elements

- in principle, we could have different numbers of basis functions on each element, but to have a consistent accuracy everywhere, we have to use the same number on each element

- the whole interval $\langle r_{min} = 0, r_{max} \rangle$ diveded
usually ↑ chosen according to $V(x)$ etc.

to K elements with n basis functions $\phi_i^{(k)}$

can be equipped with a global continuous basis in the following way:

a) we combine the last basis function $\phi_n^{(k-1)}$ with the first basis function $\phi_1^{(k)}$ to get only one „bridging" function non-zero at $x_n^{(k-1)} = x_1^{(k)}$

b) we set to zero all basis functions on elements we they are not defined

- we finally get $(n-1)K + 1$ basis functions
for each bridging function; last element, last function

but thans to boundary conditions $\psi(r_{min}) = \psi(r_{max}) = 0$ we can delete the first and the last basis functions to end up with $(n-1)K - 1$ basis functions

$$\varphi_1(r) = \begin{cases} \phi_2^{(1)}(r) & \text{for } r \in \langle a_0, a_1 \rangle \\ 0 & \text{for } r > a_1 \end{cases}$$

$$\varphi_2(r) = \begin{cases} \phi_3^{(1)}(r) & \text{for } r \in \langle a_0, a_1 \rangle \\ 0 & \text{for } r > a_1 \end{cases}$$

$\vdots$

$$\varphi_{n-1}(r) = \begin{cases} \phi_n^{(1)}(r) & \text{for } r \in \langle a_0, a_1 \rangle \\ \phi_1^{(2)}(r) & \text{for } r \in \langle a_1, a_2 \rangle \\ 0 & \text{for } r > a_2 \end{cases}$$

here and for all bridging functions it is necessary to use normalization $\frac{1}{\sqrt{w_n^{(1)} + w_1^{(2)}}}$ instead of factors $\frac{1}{\sqrt{w_n^{(1)}}}$ and $\frac{1}{\sqrt{w_1^{(2)}}}$ to have orthonormal basis (and continuous)

$$\varphi_n(r) = \begin{cases} \phi_2^{(2)}(r) & \text{for } r \in \langle a_1, a_2 \rangle \\ 0 & \text{elsewhere} \end{cases}$$

$\vdots$

$$\varphi_{(n-1)K-1}(r) = \begin{cases} \phi_{n-1}^{(K)}(r) & \text{for } r \in \langle a_{k-1}, a_k \rangle \\ 0 & \text{for } r < a_{k-1} \end{cases}$$

thus i-th basis function $\varphi_i(r)$ is given by basis functions $\phi_j^{(k)}$ satisfying $\underline{i = (k-1)(n-1) + j - 1}$

• for properly normalized bridging functions we get

$$\int_{r_{min}=0}^{r_{max}} \varphi_i(r)\varphi_j(r)\,dr \simeq \underbrace{\sum_{k=1}^{K}\sum_{s=1}^{n}}_{\text{endpoints of elements are used twice}} w_s^{(k)} \varphi_i(x_s^{(k)})\varphi_j(x_s^{(k)}) = \delta_{ij}$$

• in practice, we usually use just one array of points and weights $(w_1 = w_2^{(1)}, \dots, w_{n-2} = w_{n-2}^{(1)}, w_{n-1} = w_n^{(1)} + w_1^{(2)}, \dots)$

and integrals can be then evaluated as a single sum

$$\int_0^{r_{max}} f(r)\,dr \simeq \sum_{i=1}^{N_b} w_i f(x_i)$$

where $N_b = (n-1)K - 1$ is the total number of basis functions (points, weights)

and x_i are all points $(x_{n-1} = x_n^{(1)} = x_1^{(2)}, \dots)$

• matrix elements for <u>potential</u> are really <u>diagonal</u>:

$$V_{ij} = \int_0^{r_{max}} \varphi_i(r)^* V(r)\varphi_j(r)\,dr \simeq$$

$$\simeq \sum_{k=1}^{N_b} w_k \varphi_i(x_k) V(x_k)\varphi_j(x_k) =$$

$$= \sum_{k=1}^{N_b} w_k \frac{\delta_{ik}}{\sqrt{w_k}} V(x_k) \frac{\delta_{jk}}{\sqrt{w_k}} = V(x_i)\delta_{ij}$$

• for kinetic-energy matrix (stiffness matrix)

we need derivatives of basis functions

$$T_{ij} = \frac{1}{2\mu}\int_0^{r_{max}} \frac{d\varphi_i(r)}{dr}\frac{d\varphi_j(r)}{dr}\,dr \simeq \frac{1}{2\mu}\sum_{k=1}^{K}\sum_{s=1}^{n} w_s^{(k)} \frac{d\varphi_i}{dr}(x_s^{(k)})\frac{d\varphi_j}{dr}(x_s^{(k)})$$

here we have to use quadratures at each element because derivatives are not continuous at the endpoints of elements and we have to use derivatives from left or right

- „stiffness" matrix $A_{ij} = \int_a^b \frac{d\varphi_i}{dx}\frac{d\varphi_j}{dx}dx$ can be evaluated efficiently if we precalculate derivatives of Lagrange interpolating polynomials on $\langle -1,1\rangle$ (basic interval for Gauss-Lobatto quadrature)

$$l_i(x) = \prod_{\substack{j=1\\ j\neq i}}^{n} \frac{(x-x_j)}{(x_i-x_j)}$$ where n is the number of points x_j of G-L quadrature on $\langle -1,1\rangle$

- from

$$\frac{dl_i(x)}{dx} = \sum_{\substack{s=1\\ s\neq i}}^{n} \frac{1}{x_i-x_s}\prod_{\substack{j=1\\ j\neq s,i}}^{n}\frac{x-x_j}{x_i-x_j}$$

we get (again we approximate integrals using G-L quadr.)

$$\left.\frac{dl_i(x)}{dx}\right|_{x=x_k} = \begin{cases} \displaystyle\sum_{\substack{s=1\\ s\neq i}}^{n}\frac{1}{x_i-x_s} & \text{for } k=i \\ \displaystyle\frac{1}{x_i-x_k}\prod_{\substack{j=1\\ j\neq i,k}}^{n}\frac{x_k-x_j}{x_i-x_j} & \text{for } k\neq i \end{cases}$$

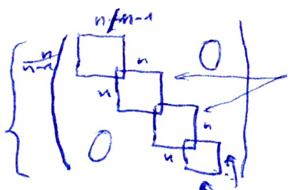
the only non-zero term in $\sum_s$ is $s=k$

- derivatives of basis functions on the interval $\langle a,b\rangle$ are then given by $l_i'(x_k)$ scaled by factor $\frac{2}{b-a}$
(2 is the lenght of $\langle -1,1\rangle$)

- the resulting matrix A has the form

K blocks

number of elements



overlap thanks to bridging functions

n or n-1 rows and columns depending on boundary conditions

- to express any function on $\langle a,b\rangle$ in the FEM-DVR basis as $f(x) = \sum_{i=1}^{N_b} c_i\varphi_i(x)$ we calculate

$$\int f(x)\varphi_j(x)\,dx = \sum_{i=1}^{N_b} c_i\int_a^b \varphi_i(x)\varphi_j(x)\,dx = c_j$$

$$\approx \sum_{k=1}^{N_b} w_k f(x_k)\frac{\delta_{jk}}{\sqrt{w_k}} = f(x_j)\sqrt{w_j}$$

and thus

$$c_j = f(x_j)\sqrt{w_j}$$ and back $$f(x_j) = \frac{c_j}{\sqrt{w_j}}$$