

# Metoda končnih elementov, ki minimizira kvadrat ostanka aproksimacije (LSFEM)

Seminarska naloga pri Naprednih numeričnih metodah

Numerično reševanje parcialnih diferencialnih enačb (PDE) je zaradi pomanjkanja vsestranskega algoritma še zmeraj bolj umetnost kot ustaljena znanost [1]. Pri zapletenih problemih hitro prispemo do vznožja gore matematične teorije, ki je ni moč zaobiti. Zaradi množice različnih pristopov reševanja ter raztresene in neprijazno napisane literature, lahko le ugibamo, kako visoko se bomo na poti do prelaza morali povzpeti. Zapletenim problemom prostorske dinamike v:

- dinamiki tekočin,
- termodinamiki,
- elektrodinamiki,
- kvantni teoriji,
- splošni teoriji relativnosti,

kjer naletimo na PDE, se tako tudi v višjem izobraževanju najraje izognemo. Metoda končnih elementov (FEM), ki minimizira kvadrat ostanka aproksimacije (LSFEM: Least Squares FEM), obeta razvoj vsestranskega algoritma za reševanje PDE in s tem približanje omenjenih problemov širšemu krogu raziskovalcev.

## 1 Podlaga za temelje LSFEM

Kadar obravnavamo prostorsko dinamiko (npr. tok tekočine), lahko fizični prostor modeliramo kot 1, 2 ali 3-mnogoterost. Temelje LSFEM bomo polagali na splošnem primeru  $d$ -mnogoterosti, za ponazoritev pa na njih sproti gradili konkretni 2D primer.

Naj bo torej prizorišče dogajanja  $d$ -mnogoterost  $\Omega$  (slika 1), opremljena s krajevnim vektorjem:

$$\mathbf{x} = \{x_1, \dots, x_d\}.$$

Pri reševanju sistema  $m$  PDE iščemo nabor funkcij:

$$\mathbf{u}(\mathbf{x}) = \{u_1(\mathbf{x}), \dots, u_m(\mathbf{x})\},$$

ki v vsaki točki domene  $\Omega$  zadosti sistemu PDE, na meji  $\Gamma$  pa robnim pogojem. Konkretni primer bomo gradili na 2D primeru s štirimi spremenljivkami, pri katerem bosta krajevni vektor in vektor odvisnih spremenljivk enaka:

$$\mathbf{x} = \{x, y\} \quad \text{in} \quad \mathbf{u} = \{u, v, p, \omega\}.$$



Slika 1: Domena  $\Omega$ , meja domene  $\Gamma$  in komponenta rešitve  $u_i(\mathbf{x})$ .

Dinamiko naj opiše **sistem Stokesovih enačb** za nestisljive tekočine v obliki hitrost-tlak-vrtinčnost, ki jim v prid nazornosti primera umetno dodamo koeficiente  $\alpha(\mathbf{x})$ ,  $\beta(\mathbf{x})$ ,  $\gamma(\mathbf{x})$  in  $\delta(\mathbf{x})$ :

$$\alpha \frac{\partial p}{\partial x} + \beta \frac{\partial \omega}{\partial y} = f_x, \quad (1) \qquad \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0, \quad (3)$$

$$\gamma \frac{\partial p}{\partial y} - \delta \frac{\partial \omega}{\partial x} = f_y, \quad (2) \qquad \omega + \frac{\partial u}{\partial y} - \frac{\partial v}{\partial x} = 0. \quad (4)$$

Stokesove enačbe ustrezajo stacionarnim Navier-Stokesovim enačbam brez nelinearnih konvektivnih členov, ki jih moramo pri numeričnem reševanju linearizirati. Ker ta korak za ponazoritev LSFEM ni ključen, se mu na tak način izognemo. Stokesove enačbe opisujejo plazeče se tokove, pri katerih je konvekcija gibalne količine (zaradi gibanja) majhna v primerjavi z njeno difuzijo (zaradi viskoznosti).

V enačbah ni časovnih odvisnosti (razen preko časovno odvisnih robnih pogojev), zato so takšni tokovi časovno obrnljivi: časovno obrnjena rešitev enačb je prav tako rešitev (slika 2).



Slika 2: Zabaven eksperiment, pri katerem se v ozkem prostoru med dvema koncentričnima valjema nahaja viskozna tekočina, ki jo na dveh mestih označimo z liso barvila. Valja pet minut vrtimo v nasprotnih smereh (Stokesov tok, ki tako nastane, imenujemo Taylor-Couettov tok), da se lisi pomešata, nato smeri vrtenja obrnemo in po petih minutah se lisi ponovno sestavita. Pridobljeno iz [2].

Sistem PDE, ki ga obravnavamo, zapišemo bolj jedrnato v matrični obliki. To je enostavno, če je sistem linearen. Uvedemo diferencialni operator  $\mathbf{A}$ :

$$\mathbf{A}(\mathbf{x}) = \mathbf{A}_0(\mathbf{x}) + \mathbf{A}_1(\mathbf{x})\frac{\partial}{\partial x_1} + \mathbf{A}_2(\mathbf{x})\frac{\partial}{\partial x_2}, \quad (5)$$

s katerim lahko sistem enačb zapišemo kot:

$$\left( \mathbf{A}_0(\mathbf{x}) + \mathbf{A}_1(\mathbf{x})\frac{\partial}{\partial x_1} + \mathbf{A}_2(\mathbf{x})\frac{\partial}{\partial x_2} \right) \cdot \mathbf{u}(\mathbf{x}) = \mathbf{f}(\mathbf{x}), \quad (6)$$

oziroma na kratko:

$$\boxed{\mathbf{A}(\mathbf{x}) \cdot \mathbf{u}(\mathbf{x}) = \mathbf{f}(\mathbf{x})} \quad \text{sistem PDE.} \quad (7)$$

V matriko  $\mathbf{A}_0$  spravimo vse koeficiente pred členi z odvisnimi spremenljivkami, v matriko  $\mathbf{A}_1$  vse koeficiente pred členi z odvodi odvisnih spremenljivk po koordinati  $x_1$  in v  $\mathbf{A}_2$  vse koeficiente pred členi z odvodi odvisnih spremenljivk po koordinati  $x_2$ . Ostale člene zložimo v vektor  $\mathbf{f}$ . Enačbe (1) - (4) lahko v duhu enačbe (6) zapišemo kot:

$$\left[ \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} + \begin{pmatrix} 0 & 0 & \alpha & 0 \\ 0 & 0 & 0 & -\delta \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix} \frac{\partial}{\partial x} + \begin{pmatrix} 0 & 0 & 0 & \beta \\ 0 & 0 & \gamma & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \frac{\partial}{\partial y} \right] \cdot \begin{pmatrix} u \\ v \\ p \\ \omega \end{pmatrix} = \begin{pmatrix} f_x(\mathbf{x}) \\ f_y(\mathbf{x}) \\ 0 \\ 0 \end{pmatrix}.$$

## 2 Temelji LSFEM

Vse različice FEM vsaj okvirno temeljijo na variacijskem pristopu, kjer ne operiramo neposredno na PDE, ampak jih najprej pretvorimo v enakovreden variacijski problem: omislino si **poskusno funkcijo**  $\mathbf{w}(\mathbf{x})$ , ki jo napnemo nad domeno  $\Omega$ , in izberemo funkcional  $I[\mathbf{w}(\mathbf{x})]$ , ki za vsako  $\mathbf{w}(\mathbf{x})$  vrne neko realno število. Za uspešnost variacijskega pristopa moramo izbrati funkcional, ki vrne najmanjšo vrednost, ko je  $\mathbf{w}(\mathbf{x})$  enaka rešitvi. Kadar obstaja s sistemom PDE povezan energijski potencial, je le-ta fizikalno najintuitivnejša izbira za konstrukcijo funkcionala. Zato ni presenetljivo, da je bila **Rayleigh-Ritzeva različica FEM** (RRFEM), ki jo na tak način dobimo, razvita prva [3]. Konstrukcija funkcionala in njegova minimizacija sta tipična koraka variacijskega pristopa in nista specifična za RRFEM: vzamemo neko funkcijo poskusne funkcije  $F(\mathbf{w})$  in jo integriramo po domeni  $\Omega$ :

$$\boxed{I[\mathbf{w}(\mathbf{x})] = \int_{\Omega} F(\mathbf{w}(\mathbf{x})) \, d\Omega} \quad \text{funkcional poskusne funkcije .} \quad (8)$$

Ko smo prepričani, da ima funkcional (8) minimum pri rešitvi  $\mathbf{u}(\mathbf{x})$ , sledimo znanemu Euler-Lagrangevemu postopku. Ta nas pripelje do variacijske izjave, ki velja le, kadar je poskusna funkcija  $\mathbf{w}(\mathbf{x})$  enaka rešitvi  $\mathbf{u}(\mathbf{x})$ . Poskusno funkcijo razvijemo okoli rešitve:

$$\tilde{\mathbf{w}}(\mathbf{x}, \varepsilon) = \mathbf{u}(\mathbf{x}) + \varepsilon \mathbf{v}(\mathbf{x}) , \quad (9)$$

kjer je  $\mathbf{v}(\mathbf{x})$  poljubna odmična funkcija,  $\varepsilon$  pa realno število. Kadar gre  $\varepsilon$  proti nič, gre  $\tilde{\mathbf{w}}(\mathbf{x}, \varepsilon)$  proti rešitvi problema  $\mathbf{u}(\mathbf{x})$ , hkrati pa vemo, da ima funkcional  $I$  pri  $\mathbf{u}(\mathbf{x})$  minimum. Minimum funkcionala poiščemo tako, da razvoj (9) vstavimo v funkcional (8) namesto  $\mathbf{w}(\mathbf{x})$  in izraz odvajamo po  $\varepsilon$ :

$$\begin{aligned} \frac{dI}{d\varepsilon} &= \int_{\Omega} \frac{d}{d\varepsilon} F(\tilde{\mathbf{w}}) \, d\Omega = \int_{\Omega} \left( \frac{dF}{d\tilde{\mathbf{w}}} \right)^{\top} \cdot \frac{d\tilde{\mathbf{w}}}{d\varepsilon} \, d\Omega = \int_{\Omega} \left( \frac{dF}{d\tilde{\mathbf{w}}} \right)^{\top} \cdot \mathbf{v} \, d\Omega , \\ \frac{dI}{d\varepsilon} &= \int_{\Omega} \frac{d}{d\varepsilon} F(\tilde{\mathbf{w}}) \, d\Omega = \left\langle \frac{dF}{d\tilde{\mathbf{w}}} \middle| \frac{d\tilde{\mathbf{w}}}{d\varepsilon} \right\rangle = \left\langle \frac{dF}{d\tilde{\mathbf{w}}} \middle| \mathbf{v} \right\rangle , \end{aligned} \quad (10)$$

nato pa  $\varepsilon$  v (10) pošljemo proti nič in celoten izraz enačimo z nič:

$$\begin{aligned} \lim_{\varepsilon \rightarrow 0} \frac{dI}{d\varepsilon} &= \lim_{\varepsilon \rightarrow 0} \int_{\Omega} \left( \frac{dF(\tilde{\mathbf{w}}(\mathbf{x}, \varepsilon))}{d\tilde{\mathbf{w}}} \right)^{\top} \cdot \mathbf{v}(\mathbf{x}) \, d\Omega = 0 . \\ \lim_{\varepsilon \rightarrow 0} \frac{dI}{d\varepsilon} &= \lim_{\varepsilon \rightarrow 0} \left\langle \frac{dF(\tilde{\mathbf{w}}(\mathbf{x}, \varepsilon))}{d\tilde{\mathbf{w}}} \middle| \mathbf{v}(\mathbf{x}) \right\rangle = 0 . \end{aligned}$$

Limita deluje le na prvi člen v integrandu, zato se variacijska izjava glasi:

$$\boxed{\int_{\Omega} \lim_{\varepsilon \rightarrow 0} \left( \frac{dF(\tilde{\mathbf{w}})}{d\tilde{\mathbf{w}}} \right)^{\top} \cdot \mathbf{v} \, d\Omega = 0 , \quad \forall \mathbf{v}(\mathbf{x})} \quad \text{variacijska izjava .} \quad (11)$$

$$\boxed{\left\langle \lim_{\varepsilon \rightarrow 0} \frac{dF(\tilde{\mathbf{w}})}{d\tilde{\mathbf{w}}} \middle| \mathbf{v} \right\rangle = 0 , \quad \forall |\mathbf{v}\rangle} \quad \text{variacijska izjava .} \quad (12)$$

V jeziku funkcionalne analize, kjer na funkcije gledamo kot na vektorje, izjava pove naslednje: projekcija izraza v oklepaju na katerokoli odmično funkcijo  $\mathbf{v}(\mathbf{x})$  mora biti enaka nič, ali krajše: izraz mora biti ortogonalen na katerokoli  $\mathbf{v}(\mathbf{x})$ .

$F(\mathbf{w}(\mathbf{x}))$  v funkcionalu poskusne funkcije je pri RRFEM energijski potencial, funkcional pa torej skupna potencialna energija sistema, ki jo rešitev  $\mathbf{u}(\mathbf{x})$  minimizira. Zaradi tega ima RRFEM lastnost najboljšega približka, diskretizacija pa vodi do simetričnega in pozitivno-definitnega sistema algebrskih enačb, ki je zelo prikladen za reševanje s hitrimi iteracijskimi metodami. Različica metode se je izkazala v gradbenem inženirstvu, kjer je s problemom vedno povezan energijski potencial. Večina računalniških programov s tega področja zato še danes temelji na RRFEM.

Žal energijski potencial povezan s sistemom PDE ne obstaja vedno, kar je značilno za sisteme PDE v dinamiki tekočin. To je motiviralo razvoj Galerkinove različice FEM (GFEM), ki je zastavljena kot posplošitev RRFEM, vendar na precej neroden način. Ideja GFEM je, da lahko za vsak sistem PDE (7) in za poljubno poskusno funkcijo  $\mathbf{w}(\mathbf{x})$  definiramo vektor ostanka  $\mathbf{R}(\mathbf{w}(\mathbf{x}))$ . Vse člene v jedrnatem zapisu sistema PDE (7) damo na eno stran in namesto  $\mathbf{u}(\mathbf{x})$  pišemo  $\mathbf{w}(\mathbf{x})$ :

$$\boxed{\mathbf{R}(\mathbf{w}(\mathbf{x})) = \mathbf{A}(\mathbf{x}) \cdot \mathbf{w}(\mathbf{x}) - \mathbf{f}(\mathbf{x})} \quad \text{vektor ostanka .} \quad (13)$$

$$\boxed{|\mathbf{R}\rangle = \mathbf{A} |\mathbf{w}\rangle - |\mathbf{f}\rangle} \quad \text{vektor ostanka .} \quad (14)$$

Poskusna funkcija  $\mathbf{w}(\mathbf{x})$ , za katero je ostanek  $\mathbf{R}(\mathbf{w}(\mathbf{x}))$  enak nič, je rešitev problema  $\mathbf{u}(\mathbf{x})$ . Idejo za

izničenje ostanka vzamemo iz variacijske izjave (12): naj bo ostanek  $\mathbf{R}(\mathbf{w}(\mathbf{x}))$  ortogonalen na katerokoli odmično funkcijo  $\mathbf{v}(\mathbf{x})$ :

$$\int_{\Omega} \mathbf{R}(\mathbf{w}(\mathbf{x}))^{\top} \cdot \mathbf{v}(\mathbf{x}) \, d\Omega = 0 \quad \text{načelo metode uteženih ostankov} . \quad (15)$$

$$\langle \mathbf{R} | \mathbf{v} \rangle = 0 \quad \text{načelo metode uteženih ostankov} . \quad (16)$$

Pristop se imenuje **metoda uteženih ostankov** in nas za sebi-adjungirane ter pozitivno-definitne  $\mathbf{A}(\mathbf{x})$  pripelje do istega sistema algebraskih enačb kot RRFEM. Uporabimo pa ga lahko tudi za sisteme PDE, ki ne posedujejo teh lastnosti, zato daje vtis posplošitve RRFEM. Akademiki so pričakovali, da bo GFEM v dinamiki tekočin enako uspešna, kot je bila RRFEM v gradbenem inženirstvu, a se to ni zgodilo [1]. Ko  $\mathbf{A}$  ni sebi-adjungiran, namreč načelo (16) ni nujno zvesto osnovnemu problemu PDE. Metoda v tem primeru ne poseduje lastnosti najboljšega približka in v rešitvi se pogostokrat pojavijo lažne oscilacije (*wiggles*). Teh se je mogoče znebiti le s hudimi izboljšavami mreže, kar očitno okrnjuje praktičnost metode. Zato sta metoda končnih diferenc in metoda končnih volumnov v dinamiki tekočin še vedno v modi.

Podobno kot pri GFEM se tudi pri LSFEM opremo na vektor ostanka (14), a reševanja variacijskega problema se lotimo na legitimen način. Ne zanašamo se na ad hoc načela, kot je zahteva (16), ampak začnemo od začetka - s konstrukcijo funkcionala (8). Sestavimo ga s kvadratom ostanka:

$$F(\mathbf{w}(\mathbf{x})) = \mathbf{R}(\mathbf{w}(\mathbf{x}))^{\top} \cdot \mathbf{R}(\mathbf{w}(\mathbf{x})) \quad \text{kvadrat vektorja ostanka} , \quad (17)$$

in tako se funkcional, ki ga minimiziramo, glasi:

$$\boxed{I[\mathbf{w}(\mathbf{x})] = \int_{\Omega} \mathbf{R}(\mathbf{w}(\mathbf{x}))^{\top} \cdot \mathbf{R}(\mathbf{w}(\mathbf{x})) \, d\Omega} \quad \text{funkcional LSFEM} , \quad (18)$$

$$\boxed{I[|\mathbf{w}\rangle] = \langle \mathbf{R}(\mathbf{w}) | \mathbf{R}(\mathbf{w}) \rangle} \quad \text{funkcional LSFEM} , \quad (19)$$

od koder dobi metoda svoje ime. Želimo dobiti variacijsko izjavo za ta specifični funkcional, zato v splošno izjavo (12) vstavimo kvadrat vektorja ostanka (17) in postopek izpeljemo do konca. Najprej torej izračunamo odvod integranda:

$$\begin{aligned} \left( \frac{dF(\tilde{\mathbf{w}})}{d\tilde{\mathbf{w}}} \right)^{\top} &= \left( \frac{d(\mathbf{R}(\tilde{\mathbf{w}})^{\top} \cdot \mathbf{R}(\tilde{\mathbf{w}}))}{d\tilde{\mathbf{w}}} \right)^{\top} = 2\mathbf{R}(\tilde{\mathbf{w}})^{\top} \cdot \frac{d\mathbf{R}(\tilde{\mathbf{w}})}{d\tilde{\mathbf{w}}} = 2(\mathbf{A} \cdot \tilde{\mathbf{w}} - \mathbf{f})^{\top} \cdot \mathbf{A} \\ \left\langle \frac{dF(\tilde{\mathbf{w}})}{d\tilde{\mathbf{w}}} \right| &= \left\langle \frac{d\langle \mathbf{R}(\tilde{\mathbf{w}}) | \mathbf{R}(\tilde{\mathbf{w}}) \rangle}{d\tilde{\mathbf{w}}} \right| = 2 \left\langle \mathbf{R}(\tilde{\mathbf{w}}) \frac{d\mathbf{R}(\tilde{\mathbf{w}})}{d\tilde{\mathbf{w}}} \right| = 2 \langle \mathbf{A} \cdot \tilde{\mathbf{w}} - \mathbf{f} | \mathbf{A} \end{aligned}$$

ter nato limito, ko gre  $\varepsilon$  proti nič:

$$\begin{aligned} \lim_{\varepsilon \rightarrow 0} \left( \frac{dF(\tilde{\mathbf{w}})}{d\tilde{\mathbf{w}}} \right)^{\top} &= \lim_{\varepsilon \rightarrow 0} 2(\mathbf{A} \cdot (\mathbf{u} + \varepsilon \mathbf{v}) - \mathbf{f})^{\top} \cdot \mathbf{A} = 2(\mathbf{A} \cdot \mathbf{u} - \mathbf{f})^{\top} \cdot \mathbf{A} . \\ \lim_{\varepsilon \rightarrow 0} \left\langle \frac{dF(\tilde{\mathbf{w}})}{d\tilde{\mathbf{w}}} \right| &= \lim_{\varepsilon \rightarrow 0} 2 \langle \mathbf{A} \cdot (\mathbf{u} + \varepsilon \mathbf{v}) - \mathbf{f} | \mathbf{A} = 2 \langle \mathbf{A} \cdot \mathbf{u} - \mathbf{f} | \mathbf{A} . \end{aligned}$$

Variacijska izjava za minimizacijo funkcionala LSFEM se zato glasi:

$$\int_{\Omega} 2(\mathbf{A} \cdot \mathbf{u} - \mathbf{f})^{\top} \cdot (\mathbf{A} \cdot \mathbf{v}) \, d\Omega = 0 , \quad \forall \mathbf{v}(\mathbf{x}) .$$

$$2 \langle \mathbf{A} \cdot \mathbf{u} - \mathbf{f} | \mathbf{A} \cdot \mathbf{v} \rangle = 0 , \quad \forall |\mathbf{v}\rangle .$$

Celoten izraz transponiramo (zamenjamo vrstni red členov v zunanjem skalarnem produktu, s čemer

rezultata ne spremenimo) ter delimo z dve:

$$\boxed{\int_{\Omega} (\mathbf{A} \cdot \mathbf{v})^T \cdot (\mathbf{A} \cdot \mathbf{u} - \mathbf{f}) \, d\Omega = 0, \quad \forall \mathbf{v}} \quad \text{variacijska izjava LSFEM} . \quad (20)$$

$$\boxed{\langle \mathbf{A} \cdot \mathbf{v} | \mathbf{A} \cdot \mathbf{u} - \mathbf{f} \rangle = 0, \quad \forall |\mathbf{v}\rangle} \quad \text{variacijska izjava LSFEM} . \quad (21)$$

Izjava pravzaprav ustreza Galerkinovi formulaciji (16), kjer namesto odmičnih funkcij samih ( $\mathbf{v}$ ) uporabimo njihove odvode ( $\mathbf{A} \cdot \mathbf{v}$ ).

### 3 Diskretizacija problema

Večina literature iz teorije FEM obravnava skalarne funkcije s stališča funkcionalne analize - kot vektorje. Enačbe iz prejšnjega poglavja bi v takšnem zapisu izgledale preprosteje. Omenjen pristop bi dodal novo plast konceptov, ki bi jih moral bralec predhodno razumeti, zato smo se ga za začetek izognili. Eden takšnih konceptov je koncept prostostnih stopenj funkcije. Razlaga diskretizacije problema je brez njega precej otežena, zato bomo tukaj na hitro opisali bistvo vektorske obravnave funkcij.

V znanem 3D vektorskem prostoru so osnovni gradniki trije bazni vektorji. S takšnim prostorom lahko opišemo vse možne diskretne skalarne funkcije, katerih domena sestoji le iz treh točk (slika 3a). Vsaka konfiguracija treh skalarjev ustreza eni točki v našem 3D prostoru, ki ga posledično imenujemo funkcijski prostor. Dimenzija funkcijskega prostora je torej povezana z gostoto vzorčenja domene. Če na domeno postavimo trinajst točk (slika 3b), bomo za opis vseh možnih konfiguracij potrebovali trinajst-dimenzionalni vektorski prostor. Če na domeno postavimo neskončno točk, kar storimo pri obravnavi zveznih funkcij (slika 3c), bomo potrebovali neskončno dimenzionalni vektorski prostor. In kaj so potem bazni vektorji našega prostora? To so  $\delta$  funkcije, postavljene v ustreznih točkah domene.



Slika 3: Funkcije, ki živijo v (a) 3D, (b) 13D in (c)  $\infty$ -D funkcijskem (vektorskem) prostoru.

Na dimenzije vektorja gledamo kot na **prostostne stopnje**, katerih vrednosti lahko poljubno nastavljamo. Funkcijo  $f$  (slika 3a) zapišemo s komponentami in baznimi vektorji kot:

$$|f\rangle = \begin{pmatrix} 0,68 & 1,00 & 0,68 \end{pmatrix} \begin{pmatrix} \delta(x) \\ \delta(x-6) \\ \delta(x-12) \end{pmatrix} = 0,68 \delta(x) + 1,00 \delta(x-6) + 0,68 \delta(x-12) , \quad (22)$$

Funkcijo  $g$  (slika 3b) zapišemo s komponentami in baznimi vektorji kot:

$$|g\rangle = \begin{pmatrix} 0,68 & 0,76 & \dots & 0,68 \end{pmatrix} \begin{pmatrix} \delta(x) \\ \delta(x-1) \\ \vdots \\ \delta(x-12) \end{pmatrix} = 0,68 \delta(x) + 0,76 \delta(x-1) + \dots + 0,68 \delta(x-12) . \quad (23)$$

Še vedno veljajo vsa pravila vektorskih prostorov. Tako je na primer skalarni produkt funkcije  $|f\rangle$

same s seboj enak:

$$\langle f|f \rangle = \begin{pmatrix} 0,68 & 1,00 & 0,68 \end{pmatrix} \begin{pmatrix} 0,68 \\ 1,00 \\ 0,68 \end{pmatrix} = 1,92 . \quad (24)$$

Skalarni produkt je definiran le za funkciji znotraj istega funkcijskega prostora. Pri zvezni funkciji  $h$  (slika 3c) komponent in baznih vektorjev ne moremo naštet, ker jih je neštevno neskončno. Kljub temu lahko funkcijski vektor izrazimo analogno, kot smo to storili v diskretnih primerih (22) in (23). Diskretno vsoto členov pretvorimo v zvezno vsoto (integral):

$$|h\rangle = \int_0^{12} h(x_0) \delta(x - x_0) dx_0 = h(x) \text{ na območju } x \in [0, 12] . \quad (25)$$

Skalarni produkt dveh zveznih funkcij je po analogiji enak:

$$\langle h|h \rangle = \int_0^{12} h(x) h(x) dx . \quad (26)$$

Če skalarne funkcije predstavimo kot funkcijske vektorje, kako potem na isti način predstavimo vektorske funkcije (ki slikajo v več skalarnih spremenljivk)? Tako, da vsako komponento (skalarno funkcijo) posebej zapišemo kot funkcijski vektor. Tako dobimo vektor vektorjev, oz. matriko. Zdaj vidimo, zakaj smo temelje FEM opisali brez uporabe omenjenih konceptov. Zahtevnost zapisov enačb se zmanjša na račun povečane zahteve po predstavljenosti zapisov. Kot primer v novem jeziku zapišimo variacijsko izjavo (21):

$$\langle \mathbf{A} \cdot \mathbf{v} | \mathbf{A} \cdot \mathbf{u} - \mathbf{f} \rangle = 0 , \quad \forall \mathbf{v} . \quad (27)$$

To izgleda veliko bolje, kajne?

Za numerično reševanje problema moramo abstraktno zastavitev (21) z neskončno prostostnimi stopnjami diskretizirati. Iskanje želimo omejiti na  $N$  čim enakomerneje razporejenih točk, ki jih imenujemo **vozlišča**. V vsako vozlišče postavimo **vozliščno bazno funkcijo**, ki pokrije le okolico vozlišča:

$$|\Phi_i\rangle , \quad i = 1, \dots, N \quad \text{vozliščne funkcije} .$$

Neskončno število neskončno ozkih stolpičev smo zamenjali s končnim številom ( $N$ ) končno ozkih grbin. Problemu omejimo število prostostnih stopenj tako, da dopustimo le obstoj tistih funkcij  $v(\mathbf{x})$ , ki so superpozicija vozliščnih funkcij. To so funkcije, ki jih lahko zapišemo kot vrsto vozliščnih funkcij  $\Phi_i(\mathbf{x})$  s koeficienti  $v_i$ :

$$|v\rangle = \sum_{i=1}^N v_i |\Phi_i\rangle . \quad (28)$$

S tem problem prevedemo na iskanje  $N$  **vozliščnih vrednosti**  $v_i$ . Naslikajmo idejo na skrajno preprosti kvadratni domeni  $[-3, 3] \times [-3, 3]$  s krajevnim vektorjem  $\chi = \{\xi, \eta\}$ . Nanjo postavimo pravokotno mrežo s šestnajstimi vozlišči (slika 4a) in nad njimi napnemo prav toliko vozliščnih funkcij z nastavljenimi višinami  $v_i$  (slika 4b).

Štirikotne ploskvice, ki nastanejo s postavitvijo vozlišč, imenujemo **elementi**. Nobena vozliščna funkcija  $\Phi_i$  ne sme pokrivati elementov, ki niso v stiku z njenim vozliščem. S tem dosežemo, da je  $v(\chi)$  nad nekim elementom sestavljena le iz funkcij v neposredni bližini tega elementa. Tako je  $\mathbf{v}(\chi)$  na sliki 6a nad osrednjim elementom popolnoma določena z vrednostmi  $v_6, v_7, v_{10}$  in  $v_{11}$ .

Ozrmo se na variacijsko izjavo (21) ter si predstavljajmo funkcije  $\mathbf{A}(\mathbf{x})$ ,  $\mathbf{u}(\mathbf{x})$ ,  $\mathbf{v}(\mathbf{x})$  in  $\mathbf{f}(\mathbf{x})$  zapisane v smislu razvoja po vozliščnih funkcijah (28). Zaslutimo, da bomo računali prekrivne integrale vozliščnih funkcij:

$$\langle \Phi_i | \Phi_j \rangle . \quad (29)$$

To je enostavno dokler so vsi elementi iste oblike in velikosti, kot na sliki 4. Takrat je dovolj, da izračunamo prekrivne integrale za vozlišča enega elementa. Kaj pa, če želimo uporabljati elemente poljubne oblike? Kako naj čim učinkoviteje, če so elementi poljubne oblike



Slika 4: (a) Pravokotna domena z devetimi elementi (modre številke) in šestnajstimi vozlišči (rdeče številke) ter (b) nad vozlišči napete vozliščne funkcije. V prid nazornosti rišemo le štiri osrednje vozliščne funkcije.



Slika 5

Segmente vozliščnih funkcij  $\Phi_6, \Phi_7, \Phi_{10}$  in  $\Phi_{11}$ , ki se nahajajo neposredno nad elementom 5, proglasimo za **elementarne funkcije**  $\phi_{5j}(\chi)$  tega elementa (slika 6b). Tako lahko funkcijo  $\mathbf{v}$  na še vedno  $Z$  natančno analitično izpeljavo se prebijemo do izjave (21), od tod dalje pa moramo iskanje funkcije  $\mathbf{u}(\mathbf{x})$  z neskončno prostostnimi stopnjami poenostaviti v iskanje funkcije s končnim številom prostostnih stopenj  $N$ .

Skozi oči FI je  $|\Phi_i\rangle$  eden izmed baznih vektorjev v razvoju vektorja  $|v\rangle$ ,  $v_i$  pa pripadajoča komponenta. V jeziku funkcionalne analize (FI) pravimo, da smo omejili funkcijski prostor.

nadaljujemo z diskretizacijo problema, to je, pretvorbo na sistem  $N$  algebraskih enačb. Ta korak je enak pri vseh različicah FEM. Funkcije na domeni  $\Omega$  imajo neskončno število prostostnih stopenj.

$$u_i(\mathbf{x}) = \sum_{a=1}^N \Phi^{a0} u_i^{a0} \quad (30)$$

Mreža je v tem šolskem primeru strukturirana, kar pomeni, da je razporeditev elementov Kartezična. Mreža je lahko pri FEM tudi nestrukturirana, kar je ena izmed prednosti metode.



Slika 6: (a) vsota vozliščnih funkcij s slike 4b in (b) elementarne funkcije, ki pripadajo elementu 5.



Slika 7

Naj bodo:

$$\mathbf{T}_1 = (x_1, y_1), \quad \mathbf{T}_2 = (x_2, y_2), \quad \mathbf{T}_3 = (x_3, y_3), \quad \mathbf{T}_4 = (x_4, y_4) \quad (31)$$

oglišča pravega elementa. Imejmo preslikavo  $\mathbf{r}(\chi)$ , ki slika iz referenčnega kvadrata  $\chi$  v realni prostor  $\mathbf{x}$ :

$$\mathbf{r}(\chi) = \begin{pmatrix} x(\chi) \\ y(\chi) \end{pmatrix} = \begin{pmatrix} x_1 S_1(\chi) + x_2 S_2(\chi) + x_3 S_3(\chi) + x_4 S_4(\chi) \\ y_1 S_1(\chi) + y_2 S_2(\chi) + y_3 S_3(\chi) + y_4 S_4(\chi) \end{pmatrix}, \quad (32)$$

kjer so  $S_i$  ogliščne funkcije:

$$\mathbf{S} = \begin{pmatrix} (1+\xi)(1+\eta) \\ (1+\xi)(1-\eta) \\ (1-\xi)(1+\eta) \\ (1-\xi)(1-\eta) \end{pmatrix} \quad (33)$$

ali kompaktneje:

$$\begin{pmatrix} x(\chi) \\ y(\chi) \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \end{pmatrix} \begin{pmatrix} S_1(\chi) \\ S_2(\chi) \\ S_3(\chi) \\ S_4(\chi) \end{pmatrix}. \quad (34)$$



Jakobijevka preslikave  $\mathbf{r}(\boldsymbol{\chi})$  je enaka:

$$\mathbf{J} = \begin{pmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial x}{\partial \eta} \\ \frac{\partial y}{\partial \xi} & \frac{\partial y}{\partial \eta} \end{pmatrix}. \quad (35)$$

$\psi$  in  $\phi$  sta skalarni funkciji.  $\bar{\mathbf{J}}$  je inverz Jakobijevke:

$$\nabla_{\mathbf{x}}\psi(\mathbf{x}) = \bar{\mathbf{J}} \nabla_{\boldsymbol{\chi}}\phi(\boldsymbol{\chi}). \quad (36)$$

$\bar{\mathbf{J}}$  je enak:

$$\bar{\mathbf{J}} = \frac{1}{\det \mathbf{J}} \begin{pmatrix} +J_{22} & -J_{12} \\ -J_{21} & +J_{11} \end{pmatrix}. \quad (37)$$

Odvod  $\psi$  po  $x_1$  označimo kot  $\psi^1$ :

$$\psi^1 = (\bar{J}_{11}\phi^1 + \bar{J}_{12}\phi^2). \quad (38)$$

Zapisano na konvencionalen način:

$$\left(\frac{\partial \psi}{\partial x}\right)_y = \bar{J}_{11} \left(\frac{\partial \phi}{\partial \xi}\right)_\eta + \bar{J}_{12} \left(\frac{\partial \phi}{\partial \eta}\right)_\xi, \quad (39)$$

oziroma:

$$\left(\frac{\partial \psi}{\partial x}\right)_y = \left(\frac{\partial \xi}{\partial x}\right)_y \left(\frac{\partial \phi}{\partial \xi}\right)_\eta + \left(\frac{\partial \eta}{\partial x}\right)_y \left(\frac{\partial \phi}{\partial \eta}\right)_\xi \quad (40)$$

Če definiramo:

$$\mathfrak{J} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{J_{22}}{\det \mathbf{J}} & \frac{-J_{12}}{\det \mathbf{J}} \\ 0 & \frac{-J_{21}}{\det \mathbf{J}} & \frac{+J_{11}}{\det \mathbf{J}} \end{pmatrix}, \quad (41)$$

lahko zapišemo:

$$\psi^p = \mathfrak{J}_{pq}\phi^q. \quad (42)$$

Po vozliščih ( $a, b, c, d \in [1, N]$ ):

$$\left\langle \left( A^{apr}_{ij} \frac{\partial \Psi_a}{\partial x_r} \frac{\partial}{\partial x_p} \right) v_{\triangleright j}^c \Psi_c \mid \left( A^{bqsik} \frac{\partial \Psi_b}{\partial x_s} \frac{\partial}{\partial x_q} \right) u_{\triangleright}^{dk} \Psi_d - f_{\triangleright}^{bsi} \frac{\partial \Psi_h}{\partial x_b} \right\rangle \quad (43)$$

$$\left\langle A^{apr}_{ij} \frac{\partial \Psi_a}{\partial x_r} \frac{\partial}{\partial x_p} v_{\triangleright j}^c \Psi_c \mid \frac{\partial \Psi_b}{\partial x_s} \frac{\partial}{\partial x_q} \left( A^{bqsik} u_{\triangleright}^{dk} \Psi_d + A^{bqsil} u_{\triangleleft}^{el} \Psi_e \right) - f_{\triangleright}^{bsi} \frac{\partial \Psi_h}{\partial b_s} \right\rangle \quad (44)$$

Po elementih  $(\alpha, \beta, \gamma, \delta) \in [1, 12]$ , kjer je  $n$  število elementov.

$$\sum_{\varepsilon}^n \left\langle A^{\bar{\alpha}pr}_{ij} \frac{\partial \psi_{\varepsilon\alpha}}{\partial x_r} \frac{\partial}{\partial x_p} v_{\triangleright j}^{\bar{\gamma}} \psi_{\varepsilon\gamma} \mid \frac{\partial \psi_{\varepsilon\beta}}{\partial x_s} \frac{\partial}{\partial x_q} \left( A^{\bar{\beta}qsik} u_{\triangleright}^{dk} \Psi_d + A^{\bar{\beta}qsil} u_{\triangleleft}^{el} \Psi_e \right) - f_{\triangleright}^{\bar{\beta}si} \frac{\partial \psi_{\varepsilon\beta}}{\partial x_s} \right\rangle \quad (45)$$

Uporabimo parcialni diferencialni operator (absorbiramo indeks v naslednji oklepaj):

$$\sum_{\varepsilon}^n \left\langle A^{\bar{\alpha}pr}_{ij} \psi_{\varepsilon\alpha r} v_{\triangleright j}^{\bar{\gamma}} \psi_{\varepsilon\gamma p} \mid \psi_{\varepsilon\beta s} \left( A^{\bar{\beta}qsik} u_{\triangleright}^{\bar{\delta}k} \psi_{\varepsilon\delta q} + A^{\bar{\beta}qsil} u_{\triangleleft}^{\bar{\varepsilon}l} \psi_{\varepsilon\epsilon q} \right) - f_{\triangleright}^{\bar{\beta}si} \psi_{\varepsilon\beta s} \right\rangle$$

Preuredimo:

$$\sum_{\varepsilon}^n v_{\triangleright j}^{\bar{\gamma}} \left\langle \psi_{\varepsilon\alpha r} \psi_{\varepsilon\gamma p} A^{\bar{\alpha}pr}_{ij} \mid \psi_{\varepsilon\beta s} \psi_{\varepsilon\delta q} A^{\bar{\beta}qsik} u_{\triangleright}^{\bar{\delta}k} + \psi_{\varepsilon\beta s} \psi_{\varepsilon\epsilon q} A^{\bar{\beta}qsil} u_{\triangleleft}^{\bar{\varepsilon}l} - \psi_{\varepsilon\beta s} f_{\triangleright}^{\bar{\beta}si} \right\rangle$$

Kar lahko, postavimo izven integrala:

$$\sum_{\varepsilon}^n v_{\triangleright j}^{\varepsilon} \left( \langle \psi_{\varepsilon\alpha r} \psi_{\varepsilon\gamma p} \psi_{\varepsilon\beta s} \psi_{\varepsilon\delta q} \rangle A^{\bar{\alpha}pr}_{ij} A^{\bar{\beta}qsik}_{ij} u_{\triangleright}^{\varepsilon k} + \langle \psi_{\varepsilon\alpha r} \psi_{\varepsilon\gamma p} \psi_{\varepsilon\beta s} \psi_{\varepsilon\epsilon q} \rangle A^{\bar{\alpha}pr}_{ij} A^{\bar{\beta}qsil}_{ij} u_{\triangleleft}^{\varepsilon l} - \langle \psi_{\varepsilon\alpha r} \psi_{\varepsilon\gamma p} \psi_{\varepsilon\beta s} \rangle A^{\bar{\alpha}pr}_{ij} f_{\boxtimes}^{\varepsilon si} \right)$$

Definiramo tenzorja  $\mathbf{Q}$  (četrtni prekrivni integrali) in  $\mathbf{T}$  (trojni prekrivni integrali):

$$Q_{\varepsilon\alpha r\beta s\gamma p\delta q} = \langle \psi_{\varepsilon\alpha r} \psi_{\varepsilon\beta s} \psi_{\varepsilon\gamma p} \psi_{\varepsilon\delta q} \rangle, \quad (46)$$

$$T_{\varepsilon\alpha r\gamma p\eta s} = \langle \psi_{\varepsilon\alpha r} \psi_{\varepsilon\gamma p} \psi_{\varepsilon\beta s} \rangle. \quad (47)$$

Tako imamo:

$$\sum_{\varepsilon}^n v_{\triangleright j}^{\varepsilon} \left( Q_{\varepsilon\alpha r\beta s\gamma p\delta q} A^{\bar{\alpha}pr}_{ij} A^{\bar{\beta}qsik}_{ij} u_{\triangleright}^{\varepsilon k} + Q_{\varepsilon\alpha r\beta s\gamma p\epsilon q} A^{\bar{\alpha}pr}_{ij} A^{\bar{\beta}qsil}_{ij} u_{\triangleleft}^{\varepsilon l} - T_{\varepsilon\alpha r\gamma p\beta s} A^{\bar{\alpha}pr}_{ij} f_{\boxtimes}^{\varepsilon si} \right) = 0. \quad (48)$$

Sestavitev globalnega tenzorja:

$$\forall v_{\triangleright}^{cj} : \sum_{a,b}^N v_{\triangleright}^{cj} \sum_{\varepsilon}^n \sum_{\substack{\gamma, \delta \ni: \\ (\varepsilon, \gamma)=c \\ (\varepsilon, \delta)=d}}^{12} \left( Q_{\varepsilon\alpha r\beta s\gamma p\delta q} A^{\bar{\alpha}pr}_{ij} A^{\bar{\beta}qsik}_{ij} u_{\triangleright}^{\varepsilon k} \right) = \sum_a^N v_{\triangleright}^{cj} \sum_{\varepsilon}^n \sum_{\substack{\gamma \ni: \\ (\varepsilon, \gamma)=c}}^{12} \left( T_{\varepsilon\alpha r\gamma p\beta s} A^{\bar{\alpha}pr}_{ij} f_{\boxtimes}^{\varepsilon si} - Q_{\varepsilon\alpha r\beta s\gamma p\epsilon q} A^{\bar{\alpha}pr}_{ij} A^{\bar{\beta}qsil}_{ij} u_{\triangleleft}^{\varepsilon l} \right) \quad (49)$$

Pomembno:  $(c, j)$  in  $(d, k)$  pri sestavitvi tenzorja  $K$  tečeta le preko prostih spremenljivk.  $i$  teče preko vseh spremenljivk.

$$K_{cjdk} = \sum_{\varepsilon}^n \sum_{\substack{\gamma, \delta \ni: \\ (\varepsilon, \gamma)=c \\ (\varepsilon, \delta)=d}}^{12} \left( Q_{\varepsilon\alpha r\beta s\gamma p\delta q} A^{\bar{\alpha}pr}_{ij} A^{\bar{\beta}qsik}_{ij} \right) \quad (50)$$

Pri sestavitvi  $F$  označuje  $(c, j)$  proste spremenljivke, medtem ko  $(e, l)$  označuje vezane spremenljivke.

$$F_{cj} = \sum_{\varepsilon}^n \sum_{\substack{\gamma \ni: \\ (\varepsilon, \gamma)=c}}^{12} \left( T_{\varepsilon\alpha r\gamma p\beta s} A^{\bar{\alpha}pr}_{ij} f_{\boxtimes}^{\varepsilon si} - Q_{\varepsilon\alpha r\beta s\gamma p\epsilon q} A^{\bar{\alpha}pr}_{ij} A^{\bar{\beta}qsil}_{ij} u_{\triangleleft}^{\varepsilon l} \right) \quad (51)$$

$$v_{\triangleright}^{cj} \left( K_{cjdk} u_{\triangleright}^{dk} + G_{cj} - H_{cj} \right) = 0.$$

$$F_{cj} = H_{cj} - G_{cj}.$$

$$v_{\triangleright}^{cj} \left( K_{cjdk} u_{\triangleright}^{dk} - F_{cj} \right) = 0 \quad \forall v.$$

Končno:

$$K_{cjdk} u_{\triangleright}^{dk} = F_{cj}. \quad (52)$$

Prekrivne integrale integriramo na referenčnem kvadratu:

$$Q_{\varepsilon\alpha r\beta s\gamma p\delta q} = \langle (\mathfrak{J}_{\varepsilon r}^i \phi_{\alpha i}) (\mathfrak{J}_{\varepsilon s}^i \phi_{\beta i}) (\mathfrak{J}_{\varepsilon p}^i \phi_{\gamma i}) (\mathfrak{J}_{\varepsilon q}^i \phi_{\delta i}) | \det \mathbf{J}_{\varepsilon} | \rangle_{\chi} \quad (53)$$

$$T_{\varepsilon\alpha r\gamma p\eta s} = \langle (\mathfrak{J}_{\varepsilon r}^i \phi_{\alpha i}) (\mathfrak{J}_{\varepsilon p}^i \phi_{\gamma i}) (\mathfrak{J}_{\varepsilon s}^i \phi_{\eta i}) | \det \mathbf{J}_{\varepsilon} | \rangle_{\chi}. \quad (54)$$

Zapis, primeren za implementacijo:

$$Q_{\varepsilon\alpha p\beta q\gamma r\delta s} = \langle (\mathfrak{J}_{\varepsilon p}^i \phi_{\alpha i}) (\mathfrak{J}_{\varepsilon q}^i \phi_{\beta i}) (\mathfrak{J}_{\varepsilon r}^i \phi_{\gamma i}) (\mathfrak{J}_{\varepsilon s}^i \phi_{\delta i}) | \mathfrak{J}_{\varepsilon 0}^0 | \rangle_{\chi} \quad (55)$$

$$T_{\varepsilon\alpha p\gamma r\eta s} = \langle (\mathfrak{J}_{\varepsilon p}^i \phi_{\alpha i}) (\mathfrak{J}_{\varepsilon r}^i \phi_{\gamma i}) (\mathfrak{J}_{\varepsilon s}^i \phi_{\eta i}) | \mathfrak{J}_{\varepsilon 0}^0 | \rangle_{\chi}. \quad (56)$$

$$\begin{aligned}
i, j &\rightarrow \text{vozlišča od 1 do } N \\
\varepsilon &\rightarrow \text{elementi od 1 do } n, \\
\alpha, \beta, \gamma, \delta &\rightarrow \text{vozlišča v elementu od 1 do 12,} \\
p, q, r, s &\rightarrow \text{odvodi: 0 (id), 1 } (\partial_x), 2 (\partial_y)
\end{aligned}$$

Sprehodi se preko vseh točk in za vsako poišči vse asociirane pare  $(i, a)$ . Dva para za točke na stranici elementa, štirje pari za točke na ogliščih elementa. Vsota po  $k$  v faktorju  $v_k^a$  ima  $m$  (število spremenljivk) seštevancev. Vsak seštevanec prispeva k vrstici globalne matrike. Seštevanci iz različnih asociiranih parov  $(i, a)$  z istimi  $k$  prispevajo k isti vrstici globalne matrike.

Podobno stori za asociirane pare  $(i, d)$ . Vsota po  $l$  bo ponovno imela  $m$  seštevancev. Seštevanci iz različnih asociiranih parov z istimi  $l$  prispevajo k istemu stolpcu globalne matrike.

To pomeni  $2 \times 2 = 4$  prispevki za vsak element globalne matrike za točke na stranicah ali  $4 \times 4 = 16$  za točke na ogliščih.

Lahko pa sestavljaš po elementih.  $(i, a)$  določi pas vrstic,  $k$  določi mesto znotraj pasu.  $(i, d)$  določi pas stolpcev,  $l$  pa mesto znotraj pasu.

## Literatura

- [1] B.-n. Jiang, *The Least-Squares Finite Element Method*. Springer-Verlag, 1998, Heidelberg.
- [2] Wikipedia. (2019). Stokes Flow, spletni naslov: [https://en.wikipedia.org/wiki/Stokes\\_flow](https://en.wikipedia.org/wiki/Stokes_flow).
- [3] W. Ritz, „Über eine neue Methode zur Lösung gewisser Variationsprobleme der mathematischen Physik“, *Journal für die Reine und Angewandte Mathematik*, let. 135, str. 1–61, 1909.