UNIVERZA V LJUBLJANI
FAKULTETA ZA NARAVOSLOVJE IN TEHNOLOGIJO
ODDELEK ZA FIZIKO

TEHNIČNA FIZIKA

Igor Grešovnik

INVERZNA NUMERIČNA ANALIZA DEFORMABILNIH TELES

MENTOR: *Peter Vencelj*

SOMENTOR: *Tomaž Rodič*

Ljubljana, 1994

Povzetek:

Diplomsko delo obravnava uporabo inverznih metod pri iskanju snovnih parametrov. Numerični pristop, ki vključuje metodo najmanjših kvadratov, je uporabljen za določitev parametrov elastoplastičnega modela odziva snovi na deformacijo. Za direktne simulacije je uporabljena metoda končnih elementov.

Ključne besede:
inverzne metode, metoda končnih elementov, metoda najmanjših kvadratov, elastoplastični model.

Summary:

Use of inverse methods in estimation of material parameters is considered. The inverse numerical approach, based on the least squares method, is applied to find the parameters of elasto-plastic model. The finite element method is used to solve direct problems.

Keywords:
inverse methods, finite element method, least squares method, elasto-plastic model.

PASC:
46.30.Jh
11.80.s
02.70.w

Vsebina:

Seznam oznak

V tem tekstu sem uporabljal naslednje standardne oznake:

x_i : *i*-ta koordinata v izbranem koordinatnem sistemu. Ponekod sem za koordinate uporabljal oznake *x*, *y*, in *z*.

$\mathbf{u}$: Vektor odmikov. $\mathbf{u}(\mathbf{r},t)=\mathbf{u}(x,y,z,t)=\mathbf{r}_P(t)-\mathbf{r}_P(t_0)$. $\mathbf{r}_P(t)$ je radijvektor materialne točke, ki ima na začetku (ob času t_0) v izbranem koordinatnem sistemu koordinate *x, y* in z.

ε : Deformacijski tenzor, definiran kot $\varepsilon_{ij}=\frac{1}{2}\left(\frac{\partial u_i}{\partial x_j}+\frac{\partial u_j}{\partial x_i}\right)$. Ponekod v tekstu sem tenzor ε definiral drugače in to tudi označil.

σ : Napetostni tenzor. Mislimo si, da opazujemo infinitezimalno malo kocko v našem telesu pri koordinatah $\mathbf{r}=(x,y,z)$. Ploskve kocke naj bodo pravokotne na osi koordinatnega sistema. Potem je napetostni tenzor definiran kot $\sigma_{ij}=\left(\frac{F_{ij}}{S_j}\right)$. F_{ij} je *i*-ta komponenta vsote sil, ki delujejo na ploskev S_j. S_j je ploskev kocke, ki je pravokotna na *j*-to os koordinatnega sistema. F_{ii} ima negativen predznak, če ustrezna dvojica sil kocko stiska, in pozitiven predznak, če jo razteza. Strižne komponente imajo pozitiven predznak, če deluje na ploskev z večjo vrednostjo pravokotne koordinate sila v pozitivni smeri ali na ploskev z manjšo vrednostjo pravokotne koordinate sila v negativni smeri.

E : Prožnostni modul materiala. Če raztezamo palico iz tega materiala z enakomernim presekom *S* in z dolžino *l*, velja med raztezkom palice in silo , s katero palico raztezamo, zveza $\Delta l=\frac{F}{E\,S}l$.

ν : Poissonov količnik materiala. Je razmerje med relativnim skrčenjem palice, ki jo raztezamo, v prečni smeri, in relativnim raztezkom te palice v vzdolžni smeri.

ρ : Gostota materiala (masa na enoto volumna).

t : Čas.

δ_{ij} : Kroneckerjev simbol; $\delta_{ij} = \begin{Bmatrix} 1; i = j \\ 0; i \neq j \end{Bmatrix}$.

Opomba: Pomena oznak, navedenih v tem seznamu, po navadi nisem posebej razlagal. Kjer sem te oznake uporabljal za kaj drugega, sem to povedal v sobesedilu. Tako sem na primer oznako σ uporabljal tudi za napake merjenih količin.

Dogovori:

S spodnjimi indeksi sem označeval komponente vektorjev, tenzorjev in matrik. Držal sem se Einsteinovega sumacijskega dogovora: če se isti indeks ponovi dvakrat v istem členu, to pomeni vsoto po tem indeksu. Izraz $\sigma_{ij}\sigma_{ij}$ na primer pomeni $\sum_i \sum_j \sigma_{ij}$.

Diagonalne komponente tenzorjev sem označeval z enim samim indeksom, ker bi drugače bili izrazi dvoumni. Tako so na primer σ_x, σ_y in σ_z diagonalne komponente napetostnega tenzorja, σ_{ll} pa je sled tega tenzorja.

Lastne vrednosti tenzorjev sem oanačeval s spodnjimi indeksi v oklepajih (na primer $\sigma_{(1)}$).

Tenzorje, vektorje in matrike sem označeval s krepko tiskanmi črkami. Izjemi sta oznaki σ in ε za napetostni in deformacijski tenzor.

Z zgornjim indeksom *T* sem označeval transponirane matrike:

$$\left(\mathbf{A}^T\right)_{ij} = \left(\mathbf{A}\right)_{ji} .$$

Odvajanje po času sem označeval s piko nad oznako spremenljivke: $\dot{x}$ pomeni $\frac{\partial^2 x}{\partial t^2}$.

1 UVOD

Le malo problemov deformabilnih teles je moč rešiti analitično, sploh kadar imamo opravka z neelastičnim obnašanjem snovi. Obstajajo pa učinkovite metode za numerično reševanje teh problemov. Ker so postale numerične simulacije preoblikovalnih procesov zanimive tudi za industrijo, se takšne metode zelo hitro razvijajo. V programe za simuliranje preoblikovalnih procesov hitro vključujejo tudi nove modele obnašanja snovi in izsledke na področju trenja, prestopnosti toplote, obrabe materialov in tako naprej.

Uporabnost omenjenih metod pogosto omejuje dejstvo, da ne poznamo dovolj vseh parametrov, ki so potrebni za simulacijo določenega procesa. Za iskanje nekaterih mehanskih lastnosti snovi so uveljavljeni standardni testi. Ti morajo biti sestavljeni tako, da lahko iz njihovih rezultatov neposredno izračunamo parametre, ki nas zanimajo. Uporabni so torej takšni preizkusi, pri katerih znamo analitično izraziti odvisnost rezultatov od iskanih parametrov. Primer je uporaba nateznega preizkusa za določitev prožnostnega modula snovi.

Zgornja zahteva za mehanske preizkuse je zelo stroga, saj, kot sem omenil, problemov deformabilnih teles večinoma ne znamo rešiti analitično. Pomagamo si lahko z uporabo *inverznih metod*. Pri tem pristopu sestavimo preizkus, katerega rezultati so odvisni od iskanih parametrov in ki bi ga znali numerično simulirati, če bi te parametre poznali. Potem poskušamo najti takšne vrednosti iskanih parametrov, pri katerih se rezultati numerične simulacije ujemajo z rezultati resničnega poskusa. Natančnega ujemanja ne moremo doseči zaradi numeričnih in merskih napak, lahko pa najdemo parametre, pri katerih se rezultati *najbolje* ujemajo. Kdaj je ujemanje boljše in kdaj slabše, je stvar definicije. Postopek iskanja parametrov avtomatiziramo tako, da definiramo matematično funkcijo, ki je merilo za ujemanje in problem prevedemo na iskanje globalnega ekstrema te funkcije.

Z inverznim pristopom se poveča nabor poskusov, ki so primerni za iskanje podatkov za direktne simulacije. Ena od pomankljivosti pa je v tem, da velikokrat ni zagotovljena enoličnost rešitve. Lahko se zgodi, da smo na videz našli iskane parametre, ti pa so v resnici daleč od resničnih. Zato je pri uporabi inverznih analiz za iskanje parametrov za numerične simulacije potrebna previdnost.

Poleg iskanja neznanih parametrov so možnosti za uporabo inverznih analiz tudi v optimizaciji proizvodnih procesov. Numerične simulacije so same po sebi uporabne največ za odkrivanje slabosti nekega procesa, ko je še v fazi projektiranja. Šele inverzni pristop omogoča njihovo uporabo pri samem projektiranju.

V diplomskem delu sem prikazal uporabo inverznih metod pri določitvi plastičnih parametrov jekla.

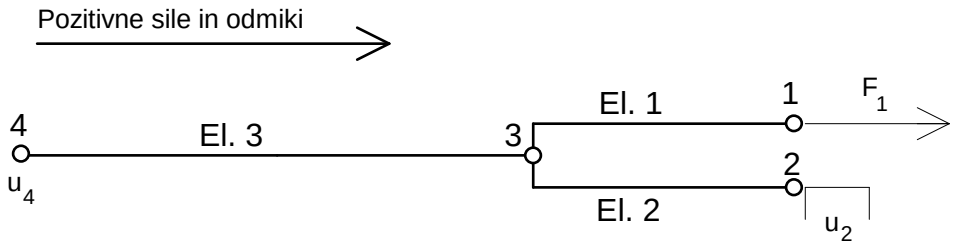
Za osnovo sem vzel elastoplastični model odziva snovi na deformacijo. Ta model je opisan v 2. poglavju, v katerem so podane tudi osnovne značilnosti metode končnih elementov. V svojem delu sem za direktne simulacije uporabil program “Elfen”, ki ga razvija britansko podjetje “Rockfield Software”. Metode, opisane v 2. poglavju, se ujemajo z osnovnimi značilnostmi uporabljenega programa.

V 3. poglavju so opisana načela inverznih analiz z metodo najmanjših kvadratov. Opisane so tudi numerične tehnike, s katerimi sem reševal zastavljene probleme.

V 4. poglavju sem obdelal konkretni primer. Z inverzno analizo nateznega preizkusa sem našel parametra, s katerima opišemo krivuljo tečenja v elastoplastičnem modelu. Rezultati so pokazali, da je pristop uporaben za iskanje teh parametrov v praksi.

Zaradi boljšega pregleda sem v podpoglavju tega poglavja prikazal osnovne značilnosti inverzne analize na izmišljenem problemu. Čeprav je primer enostaven, je metoda reševanja v osnovnih potezah enaka metodi, ki sem jo uporabil pri resničnem problemu.

1.1 Primer inverzne določitve parametrov



Slika 1.1: Shematska slika poskusa. Označene so količine s predpisanimi vrednostmi.

Recimo, da želimo najti prožnostni modul nekega jekla. Raztezamo tri žice iz tega jekla, ki so med seboj spete, kot kaže slika 1.1. Za vse žice poznamo preseke S_i in dolžine l_i. Odmike krajišč žic ali *vozlov* od ravnovesne lege označimo z u_i, kjer je i številka krajišča. Zunanje sile, ki delujejo na krajišča, označimo z F_i. Vsaki žici priredimo svoj element z oznakami od 1 do 3. Tretja žica je v krajišču 4 pritrjena, zato je

$$u_4 = \phi_4 = 0 \ . \tag{1.1}$$

Prvo žico raztezamo s silo

$$F_1 = R_1 \ , \tag{1.2}$$

drugo pa tako, da je odmik njenega desnega krajišča enak

$$u_2 = \phi_2 \ . \tag{1.3}$$

Za znane sile in odmike sem vpeljal posebne oznake, da lahko v enačbah že na prvi pogled ločimo spremenljivke od konstant.

Odmika u_1 in u_3 ter sila F_2 so odvisni od robnih pogojev (1.1), (1.2) in (1.3), ki jih fiksiramo, ter od prožnostnega modula E. Zato lahko iz meritev teh količin sklepamo na vrednost prožnostnega modula. Izmerjene odmike in sile označimo z $u_1^{(m)}$, $u_3^{(m)}$ in $F_2^{(m)}$.

Da lahko izvedemo inverzno analizo, moramo najprej znati izračunati merjene količine u_1, u_3 in F_2 pri poljubnih vrednostih iskanega parametra E. V našem primeru lahko to storimo analitično, v bolj zapletenih primerih pa bi uporabili numerično simulacijo.

Označimo

$$k_i = \frac{E S_i}{l_i} \ . \tag{1.4}$$

Pogoji ravnovesja sil za vsak element se glasijo:

$$\begin{aligned} k_1(u_1 - u_3) &= R_1 \\ k_2(\phi_2 - u_3) &= F_2 \\ k_1(u_3 - u_1) + k_2(u_3 - \phi_2) + k_3(u_3 - \phi_4) &= 0 \\ k_3(\phi_4 - u_3) &= F_4 \end{aligned} \qquad (1.5)$$

Iz tega sistema enačb lahko izračunamo sile in odmike krajišč pri danem prožnostnem modulu. Mislimo si lahko, da so u_1, u_3 in F_2 funkcije prožnostnega modula, ki jih izračunamo z reševanjem zgornjega sistema enačb:

$$u_1 = u_1(E), u_3 = u_3(E), F_2 = F_2(E). \qquad (1.6)$$

Iz vsake od meritev bi lahko izračunali prožnostni modul tako, da bi rešili enačbe $u_1(E) = u_1^{(m)}$, $u_3(E) = u_3^{(m)}$ in $F_2(E) = F_2^{(m)}$. Ker so meritve obremenjene z merskimi napakami, bi na ta način vsakič dobili drugačno vrednost za E. Lahko bi povprečili dobljene vrednosti, vendar raje uberemo drugo pot. Poiščemo takšen E, pri katerem se vse tri izračunane količine najbolje ujemajo z izmerjenimi. V ta namen definiramo funkcijo

$$\chi^2(E) = \left(\frac{u_1^{(m)} - u_1(E)}{\sigma_{u_1}} \right)^2 + \left(\frac{u_3^{(m)} - u_3(E)}{\sigma_{u_3}} \right)^2 + \left(\frac{F_2^{(m)} - F_2(E)}{\sigma_{F_2}} \right)^2 , \qquad (1.7)$$

kjer so σ_{u_1}, σ_{u_3} in σ_{F_2} ocenjene napake izmerkov. Ta funkcija je merilo za to, kako dobro se izmerjene vrednosti odmikov in sil ujemajo z izračunanimi. Če je ujemanje boljše, je vrednost funkcije nižja, če je slabše, pa višja. Zato prožnostni modul izračunamo tako, da poiščemo minimum funkcije $\chi^2(E)$. V tem enostavnem primeru lahko funkcijo χ^2 zapišemo v analitični obliki, ker lahko funkcije $u_1(E)$, $u_3(E)$ in $F_2(E)$ izrazimo analitično. V bolj zapletenih primerih moramo vrednosti teh funkcij izračunavati tako, da vsakič izvedemo numerično simulacijo celotnega poskusa. Princip iskanja paramtrov pa je enak, saj lahko funkcijo χ^2 numerično minimiziramo, če znamo izračunati njeno vrednost pri poljubnem naboru parametrov. Problem rešujemo na enak način tudi, če iščemo več parametrov. V tem primeru je χ^2 analogno definirana funkcija več spremenljivk.

Recimo, da so podatki takšni:

$S_1 = S_2 = S_3 = 0.1 \cdot 10^{-6}\ m^2$, $l_1 = 2\ m$, $l_2 = l_3 = 4\ m$, $R_1 = 50\ N$, $\phi_4 = 0\ m$, $\phi_2 = 0.01\ m$.

Pri merjenju dobimo na primer te rezultate:

$u_1^{(m)} = 0.01798\ m$, $u_3^{(m)} = 0.01304\ m$ in $F_2^{(m)} = 19.21\ N$. Napake ocenimo z eno stotino izmerjenih vrednosti. Iz teh podatkov dobimo z minimizacijo χ^2 za prožnostni modul $E = 2{,}102 \cdot 10^{11}\ N/m^2$.

Primer sem izračunal s pomočjo programa *Mathematica*, le minimizacijo funkcije χ^2 sem izvedel s pomočjo svojega programa. Prilagam celoten postopek izračuna, kjer so posamezne stopnje komentirane s kratkimi komentarji.

2 NUMERIČNO REŠEVANJE PROBLEMOV DEFORMABILNIH TELES

V svoji diplomski nalogi sem se ukvarjal z elastičnimi in elastoplastičnimi problemi. Zato bom tudi to poglavje posvetil problemom te vrste.

Enačbe, s katerimi opišemo obnašanje trdnega telesa pri deformaciji, razdelimo v tri glavne skupine[6] :

1. Ravnovesni zakoni so splošni zakoni mehanike kontinuuma in veljajo za vsa deformabilna telesa v enaki obliki:

Zakon o ohranitvi mase za male deformacije:

$$\rho\,(1+\varepsilon_{ii}\,)=\rho_0\ . \tag{2.1}$$

Ravnovesje sil:

$$\frac{\partial \sigma_{ij}}{\partial x_j}+\rho\, g_i=\rho\, \ddot{x}_i\ . \tag{2.2}$$

Ravnovesje navorov:

$$\sigma_{ij}=\sigma_{ji}\ . \tag{2.3}$$

2. Konstitutivne enačbe opisujejo obnašanje določene snovi. Primer je linearna zveza med deformacijami in napetostmi v izotropnih elastičnih snoveh:

$$\varepsilon_{ij}=\frac{1}{E}\left((1+\nu)\sigma_{ij}-\nu\sigma_{ll}\delta_{ij}\right)\ . \tag{2.4}$$

3. Robni pogoji določajo pogoje na meji območja, na katerem veljajo enačbe, ki določajo obnašanje našega telesa. Pri mehanskih problemih, s katerimi sem imel opravka, podamo robne pogoje tako, da predpišemo ali odmike:

$$u_i=\bar{u}_i \quad na \quad S_u \tag{2.5}$$

ali sile, ki delujejo na površino telesa:

$$\sigma_{ij}\,n_j=\bar{t}_i \quad na \quad S_\sigma\ . \tag{2.6}$$

V zgornjih enačbah je **g** gravitacijski pospešek, $\bar{\mathbf{u}}$ so predpisani odmiki, $\bar{\mathbf{t}}$ predpisane normalne napetosti, S_u je tisti del roba, na katerem so predpisani odmiki, S_t pa del roba, na katerem so predpisane napetosti.

2.1 Metoda končnih elementov

Obnašanje fizikalnih sistemov opisujemo z diferencialnimi enačbami. Poljubno parcialno diferancialno enačbo lahko zapišemo v obliki

$$L(\,f\,(\mathbf{r}\,))=0, \tag{2.7}$$

kjer je L parcialni diferencialni operator, f pa je funkcija N neodvisnih spremenljivk $\mathbf{r}=[r_1,r_2,...r_N]^T$, ki zadošča naši enačbi. Podobno lahko sistem m parcialnih diferencialnih enačb zapišemo v obliki

$$L_i(\,f_i(\mathbf{r}\,))=0,\quad i=1,...m\ .$$

Da lahko rešimo konkreten fizikalni problem, moramo poleg enačb, ki ga opišejo, poznati še robne pogoje, to je razmere na robu območja, za katerega veljajo te enačbe. Podamo jih lahko enostavno tako, da predpišemo vrednosti rešitve na robu območja, tak način je znan kot Dirichletov robni pogoj. Lahko pa tudi predpišemo vrednosti normalnih odvodov rešitev na robu območja.

Deformacije trdnih teles opišemo z ravnovesnimi enačbami mehanike kontinuumov ter s konstitutivnimi enačbami, ki so odvisne od snovi, iz katere je telo. Snovi glede na obnašanje pri deformiranju delimo v skupine, za katere veljajo podobne konstitutivne enačbe. Te so za vsako snov določene z majhnim številom parametrov, odvisnih od konkretne snovi. Zato pravimo, da je problem deformacije trdnega telesa določen z robnimi pogoji in snovnimi parametri.

Diferencialnih enačb, s katerimi lahko zadovoljivo opišemo obnašanje fizikalnih sistemov, po navadi ne znamo rešiti analitično. Pomagamo si z aproksimativnimi metodami, ki dajo približne rešitve. Metoda končnih elementov je danes ena najbolj razširjenih metod, ki se uporabljajo pri reševanju problemov deformabilnih teles. Eden glavnih razlogov za to je v njeni primernosti za računalniško implementacijo.

Podobno kot ostale metode za numerično reševanje diferencialnih enačb tudi metoda končnih elementov temelji na diskretizaciji definicijskega območja rešitve. Prostor neodvisnih spremenljivk ($\mathbf{r}$) razdelimo na mrežo diskretnih točk ($\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n$) in prevedemo diferencialno enačbo (2.7) v sistem algebrskih. Razlika med metodo končnih elementov in diferenčnimi metodami je v tem, kako prevedemo diferencialne enačbe v sistem algebrskih.

Pri diferenčnih metodah vzamemo za neznanke vrednosti rešitve v naših diskretnih točkah ($f_i = f(\mathbf{r}_i),\ i = 1,2,\ldots n$). Odvode rešitve po neodvisnih spremenljivkah aproksimiramo z izrazi, v katerih nastopajo razlike teh neznank. Ko te izraze vstavimo v prvotno enačbo, dobimo sistem algebrskih enačb za vrednosti iskane rešitve v izbranih točkah f_i.

Pri metodi končnih elementov si mislimo, da lahko rešitev diferencialne enačbe aproksimiramo z linearno kombinacijo funkcij, ki so definirane na definicijskem območju:

$$f(\mathbf{r}) = \sum_{i=1}^{n} f_i N_i(\mathbf{r}),\ i = 1,2,\ldots n. \qquad (2.8)$$

Neznane koeficiente N_i dobimo tako, da izraz za $f(\mathbf{r})$ vstavimo nazaj v enačbo (2.7) in rešimo dobljeni sistem enačb.

2.1.1 Izoparametrični model

V več dimenzijah razdelimo definicijsko območje enačbe, ki jo rešujemo, na enostavna geometrijska telesa - *elemente.* V dveh dimenzijah so to navadno trikotniki ali štirikotniki. Oglišča elementov imenujemo *vozlišča* ali *vozli*. Rešitev naše enačbe potem izrazimo kot linearno kombinacijo interpolacijskih funkcij za vsak element posebej:

$$f(\mathbf{r}) = \sum_{i=1}^{p} f_i^{(e)} N_i^{(e)}(\mathbf{r}), \qquad (2.9)$$

kjer je p število oglišč elementa, $N_i^{(e)}(\mathbf{r})$ pa *je elementarna interpolacijska funkcija*, ki je definirana znotraj elementa in je pridružena *i*-tem vozlišču.

Osnovni princip *izoparametričnega modela* je v tem, da definiramo geometrijo elementa s koordinatami vozlišč in z istimi elementarnimi funkcijami, kot jih uporabljamo za interpolacijo neznane funkcije. Zato lahko zapišemo tudi

$$x_j(\mathbf{r}) = \sum_{i=1}^{p} \left(x_j\right)_i^{(e)} N_i^{(e)}(\mathbf{r}) \ , \qquad (2.10)$$

kjer je $x_j(\mathbf{r})$ j-ta koordinata materialne točke, ki je imela v začetnem stanju telesa koordinate $\mathbf{r}$ v izbranem globalnem koordinatnem sistemu. Ugodno je, če so neznani koeficienti $f_i^{(e)}$ kar približki za vrednosti funkcije f v vozliščih elementa:

$$N_i^{(e)} = f(\mathbf{r}_i) \ . \qquad (2.11)$$

Da lahko to storimo, moramo izbrati interpolacijske funkcije tako, da zadovoljujejo določene pogoje. Te bom opisal pri obravnavi elementov.

2.2 Elastični dvodimenzionalni problemi

Z uporabo simetrije lahko veliko mehanskih problemov, ki so tudi praktičnega pomena, opišemo z uporabo dveh neodvisnih koordinat. Pri svojem delu sem se osredotočil izključno na probleme te vrste, zato bom nekaj prostora posvetil njihovemu reševanju.

V problemih iz mehanike kontinuumov, kjer imamo opravka z malimi deformacijami, uporablljamo za opis stanja telesa deformacijski tenzor ε in napetostni tenzor σ[1]. Male deformacije so tiste, pri katerih so relativni odmiki mali v primerjavi z dimenzijami telesa. Obnašanje snovi pri deformaciji opišemo z zvezo med napetostmi in deformacijami. Za opis izotropnih elastičnih teles nam zadoščata dva parametra:

$$\varepsilon_{ij} = \frac{1}{E}\left((1+\nu)\sigma_{ij} - \nu\sigma_{ll}\delta_{ij}\right) . \qquad (2.12)$$

Tu je E prožnostni modul, ν pa Poissonov količnik.

Za formulacijo z metodo končnih elementov je primerno enačbe prevesti v integralno obliko. Uporabimo *princip virtualnega dela*[2,4] :

$$\int_\Omega [\delta\varepsilon]^T \sigma d\Omega - \int_\Omega [\delta\mathbf{u}]^T \mathbf{b} d\Omega - \int_{\Gamma_t} [\delta\mathbf{u}]^T \mathbf{t} d\Gamma = 0 \ , \qquad (2.13)$$

kjer je $\mathbf{b}$ volumska gostota zunanjih sil, $\mathbf{t}$ površinska gostota sil, ki delujejo na površino telesa, $\delta\varepsilon$ so virtualne deformacije, $\delta\mathbf{u}$ pa virtualni odmiki. Ω je volumen telesa, ki se deformira, Γ_t je tisti del roba, na katerem so predpisane napetosti, Γ_u pa del roba, kjer so predpisani odmiki. Rešitev našega problema je takšno deformacijsko stanje, da je zgornja enačba izpolnjena za kakršen koli izbor virtualnih odmikov.

Zaradi simetrije lahko včasih obnašanje telesa pri deformaciji opišemo v dvodimenzionalnem koordinatnem sistemu. Posebno pomembni so naslednji primeri[1,4] :

1. Stanje ravninskih napetosti:

Tanka plošča je na robovih obremenjena s silami, ki so vzporedne z ravnino plošče. Komponente napetostnega tenzorja σ_z, σ_{xz} in σ_{yz} so enake 0.

Za opis takšnih problemov potrebujemo le dve komponenti odmikov, ki sta vzporedni z ravnino plošče:

$$\mathbf{u} = \left[u, v\right]^T . \tag{2.14}$$

Vpeljemo vektor napetosti

$$\sigma = \left[\sigma_x, \sigma_y, \sigma_{xy}\right]^T , \tag{2.15}$$

kjer so σ_x, σ_y in σ_{xy} komponente standardnega napetostnega tenzorja, in vektor deformacij

$$\varepsilon = \left[\varepsilon_x, \varepsilon_y, \varepsilon_{xy}\right], \tag{2.16}$$

kjer je

$$\varepsilon_x = \frac{\partial u}{\partial x}, \; \varepsilon_y = \frac{\partial v}{\partial y} \text{ in } \varepsilon_{xy} = \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} . \tag{2.17}$$

Potem velja zveza

$$\sigma = \mathbf{D}\,\varepsilon , \tag{2.18}$$

kjer je[4]

$$\mathbf{D} = \frac{E}{1-\nu^2}\begin{bmatrix} 1 & \nu & 0 \\ \nu & 1 & 0 \\ 0 & 0 & \dfrac{1-\nu}{2} \end{bmatrix} . \tag{2.19}$$

2. Stanje ravninskih deformacij:

Telo oblike pokončne prizme, katere višina je zelo velika v primerjavi s premeri osnovnih ploskev, je obremenjeno s silami, pravokotnimi na vzdolžno os. Sile se ne spreminjajo vzdolž te osi. Predpostavimo lahko, da vladajo v vseh prečnih presekih telesa enaki pogoji. Za opis takšnih problemov potrebujemo dve komponenti odmikov, ki sta pravokotni na vzdolžno os telesa:

$$\mathbf{u} = \left[u, v\right]^T . \tag{2.20}$$

Vektorja ε in σ definiramo enako kot pri stanju ravninskih napetosti. Velja zveza

$$\sigma = \mathbf{D}\,\varepsilon , \tag{2.21}$$

kjer je[4]

$$\mathbf{D} = \frac{E}{(1+\nu)(1-2\nu)}\begin{bmatrix} 1-\nu & \nu & 0 \\ \nu & 1-\nu & 0 \\ 0 & 0 & \dfrac{1-2\nu}{2} \end{bmatrix} . \tag{2.22}$$

3. Osno simetrično stanje:

Osno simetrično telo je podvrženo robnim pogojem, ki so simetrični okrog osi telesa. Takšen sistem najlažje opišemo v valjnih koordinatah. Obnašanje telesa je neodvisno od azimutnega kota, zato za opis zadoščata dve komponenti odmikov:

$$\mathbf{u} = \left[u, w\right]^T . \tag{2.23}$$

u so odmiki v radialni smeri, *w* pa v smeri simetrijske osi telesa.

Vektor deformacij definiramo kot

$$\varepsilon = \left[\varepsilon_r, \varepsilon_\phi, \varepsilon_z, \varepsilon_{rz}\right]^T , \tag{2.24}$$

$$\varepsilon_r = \frac{\partial u}{\partial r},\ \varepsilon_\phi = \frac{u}{r},\ \varepsilon_z = \frac{\partial w}{\partial z},\ \varepsilon_{rz} = \frac{\partial u}{\partial z} + \frac{\partial w}{\partial r} , \tag{2.25}$$

kjer je sta *r* in *z* ustrezni koordinati valjnega koordinatnega sistema.

Vektor napetosti definiramo kot

$$\sigma = \left[\sigma_r, \sigma_\phi, \sigma_z, \sigma_{rz}\right]^T , \tag{2.26}$$

kjer so σ_r, σ_ϕ in σ_z normalne napetosti v smereh *r*, ϕ in *z*, σ_{rz} pa je strižna napetost v ravnini *rz*.

Zveza med σ in ε je

$$\sigma = \mathbf{D}\,\varepsilon , \tag{2.27}$$

kjer je[4]

$$\mathbf{D} = \frac{E}{(1+\nu)(1-2\nu)} \begin{bmatrix} 1-\nu & \nu & 0 & 0 \\ \nu & 1-\nu & \nu & 0 \\ 0 & \nu & 1-\nu & 0 \\ 0 & 0 & 0 & \frac{1-2\nu}{2} \end{bmatrix} . \tag{2.28}$$

2.2.1 Izoparametrična reprezentacija

V izoparamerični reprezentaciji zapišemo polje odmikov in deformacij kot linearno kombinacijo interpolacijskih funkcij:

$$\mathbf{u}(\mathbf{r}) = \sum_{i=1}^{n} \mathbf{d}_i \mathbf{N}_i(\mathbf{r}) , \quad \delta\mathbf{u}(\mathbf{r}) = \sum_{i=1}^{n} \delta\mathbf{d}_i \mathbf{N}_i(\mathbf{r}) \tag{2.29}$$

in

$$\varepsilon(\mathbf{r}) = \sum_{i=1}^{n} \mathbf{d}_i \mathbf{B}_i(\mathbf{r}) , \quad \delta\varepsilon(\mathbf{r}) = \sum_{i=1}^{n} \delta\mathbf{d}_i \mathbf{B}_i(\mathbf{r}) , \tag{2.30}$$

kjer je $\mathbf{d}_i$ vektor odmikov v *i*. vozlišču, $\mathbf{N}_i = \mathbf{I}\, N_i$ je matrika globalnih interpolacijskih funkcij za to vozlišče ($\mathbf{I}$ je identična matrika), $\mathbf{B}_i$ pa je matrika, ki povezuje deformacije z odmiki.

Ko zgornji enačbi vstavimo v (2.13), dobimo enačbo

$$\sum_{i=1}^{n} [\delta\mathbf{d}_i]^T \left\{ \int_\Omega [\mathbf{B}_i]^T \sigma d\Omega - \int_\Omega [\mathbf{N}_i]^T d\Omega - \int_{\Gamma_t} [\mathbf{N}_i]^T \mathbf{t} d\Gamma \right\} = 0 . \tag{2.31}$$

Ta enačba mora biti izpolnjena za vsak nabor virtualnih odmikov $\delta\mathbf{d}_i$, zato velja

$$\sum_{i=1}^{n} [\mathbf{B}_i]^T \sigma d\Omega - \int_\Omega [\mathbf{N}_i]^T \mathbf{b} d\Omega - \int_{\Gamma_t} [\mathbf{N}_i]^T \mathbf{t} d\Gamma = 0 . \tag{2.32}$$

V *izoparametrični reprezentaciji* izračunamo prispevke k zgornji enačbi za vsak element posebej. Da je to možno, morajo biti interpolacijske funkcije zvezne na mejah elementov. Odmike znotraj posameznega elementa zapišemo kot

$$\mathbf{u}^{(e)} = \sum_{i=1}^{r} \mathbf{N}_i^{(e)} \, d_i^{(e)} \ , \tag{2.33}$$

če ima element *r* vozlov. Podobno zapišemo koordinate materialnih točk znotraj elementa:

$$\begin{bmatrix} x^{(e)} \\ y^{(e)} \end{bmatrix} = \sum_{i=1}^{r} \begin{bmatrix} N_i^{(e)} & 0 \\ 0 & N_i^{(e)} \end{bmatrix} \begin{bmatrix} x_i^{(e)} \\ y_i^{(e)} \end{bmatrix} . \tag{2.34}$$

$N_i^{(e)}$ so elementarne interpolacijske funkcije, izražene v lokalnih koordinatah elementa. Lokalni koordinatni sistem elementa je vezan na materialne točke deformiranega telesa.

Definiramo lahko *Jacobijevo determinanto* za pretvorbo iz enega v drug koordinatni sistem:

$$\mathbf{J}^{(e)} = \begin{bmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial y}{\partial \xi} \\ \frac{\partial x}{\partial \eta} & \frac{\partial y}{\partial \eta} \end{bmatrix} = \sum_{i=1}^{r} \begin{bmatrix} \frac{\partial N_i^{(e)}}{\partial \xi} x_i^{(e)} & \frac{\partial N_i^{(e)}}{\partial \xi} y_i^{(e)} \\ \frac{\partial N_i^{(e)}}{\partial \eta} x_i^{(e)} & \frac{\partial N_i^{(e)}}{\partial \eta} y_i^{(e)} \end{bmatrix} . \tag{2.35}$$

Potrebovali bomo tudi njen inverz

$$\left[\mathbf{J}^{(e)}\right]^{-1} = \begin{bmatrix} \frac{\partial \xi}{\partial x} & \frac{\partial \eta}{\partial x} \\ \frac{\partial \xi}{\partial y} & \frac{\partial \eta}{\partial y} \end{bmatrix} = \frac{1}{\det \mathbf{J}^{(e)}} \begin{bmatrix} \frac{\partial y}{\partial \eta} & -\frac{\partial y}{\partial \xi} \\ -\frac{\partial x}{\partial \eta} & \frac{\partial x}{\partial \xi} \end{bmatrix} . \tag{2.36}$$

Element volumna lahko sedaj izrazimo v lokalnem koordinatnem sistemu:

$$d\,\Omega = h^{(e)} \, det\, \mathbf{J}^{(e)} \, d\,\xi \, d\eta \ . \tag{2.37}$$

Če obravnavamo na primer ravninsko napetostno stanje, je $h^{(e)}$ debelina elementa. V enačbi nastopa zato, ker problem opisujemo v dveh dimenzijah.

Elementarne interpolacijske funkcije izražamo v lokalnem koordinatnem sistemu. Z uporabo verižnega pravila dobimo

$$\frac{\partial N_i^{(e)}}{\partial x} = \frac{\partial N_i^{(e)}}{\partial \xi} \frac{\partial \xi}{\partial x} + \frac{\partial N_i^{(e)}}{\partial \eta} \frac{\partial \eta}{\partial x} \ . \tag{2.38}$$

Podobno velja za $\frac{\partial N_i^{(e)}}{\partial y}$. Odvode $\frac{\partial \xi}{\partial x}$, $\frac{\partial \eta}{\partial x}$, $\frac{\partial \xi}{\partial y}$ in $\frac{\partial \eta}{\partial y}$ lahko izračunamo iz inverzne Jacobijeve matrike.

Znotraj vsakega elementa imamo linearno relacijo med napetostjo in deformacijo oblike

$$\boldsymbol{\sigma}^{(e)} = \mathbf{D}^{(e)} \, \boldsymbol{\varepsilon}^{(e)} = \mathbf{D}^{(e)} \sum_{j=1}^{r} \mathbf{B}_j^{(e)} \, d\, \mathbf{j}^{(e)} \ . \tag{2.39}$$

Zato lahko prispevke vsakega elementa k enačbi (2.32) zapišemo po vrsti kot:

1. člen:

$$\sum_{j=1}^{r} \mathbf{K}_{ij}^{(e)} \mathbf{d}_j^{(e)} = \int_{\Omega^{(e)}} \left[\mathbf{B}_i^{(e)}\right]^T \mathbf{D}^{(e)} \sum_{j=1}^{r} \mathbf{B}_j^{(e)} \mathbf{d}_j^{(e)} d\Omega \ , \qquad (2.40)$$

kjer jr $\mathbf{K}_{ij}^{(e)}$ podmatrika *elementarne togostne matrike* $\mathbf{K}^{(e)}$.

2. člen:

$$\mathbf{f}_{B_i}^{(e)} = \int_{\Omega^{(e)}} \left[\mathbf{N}_i^{(e)}\right]^T \mathbf{b}^{(e)} d\Omega \ . \qquad (2.41)$$

3. člen:

$$\mathbf{f}_{T_i}^{(e)} = \int_{\Gamma_t^{(e)}} \left[\mathbf{N}_i^{(e)}\right]^T \mathbf{t}^{(j)} d\Gamma \ . \qquad (2.42)$$

$\Gamma_t^{(e)}$ je del elementa, ki sovpada z mejo območja, na katerem rešujemo naše enačbe, zato zadnji prispevek za nekatere elemente odpade.

Matriko $\mathbf{B}_i^{(e)}$ izračunamo iz izraza za deformacijski tenzor in zvez (2.29) in (2.30). Za ravninsko napetostno stanje na primer dobimo iz (2.17)

$$\mathbf{B}_i^{(e)} = \begin{bmatrix} \left(\frac{\partial N_i}{\partial x}\right)^{(e)} & 0 \\ 0 & \left(\frac{\partial N_i}{\partial y}\right)^{(e)} \\ \left(\frac{\partial N_i}{\partial y}\right)^{(e)} & \left(\frac{\partial N_i}{\partial x}\right)^{(e)} \end{bmatrix} .$$

Vektor odmikov za ta primer je

$$\mathbf{d}_i^{(e)} = \begin{bmatrix} u_i^{(e)} \\ v_i^{(e)} \end{bmatrix} ,$$

volumski element pa

$$d\Omega^{(e)} = t^{(e)} \det\left(\mathbf{J}^{(e)}\right) d\xi d\eta \ ,$$

kjer je $t^{(e)}$ debelina elementa.

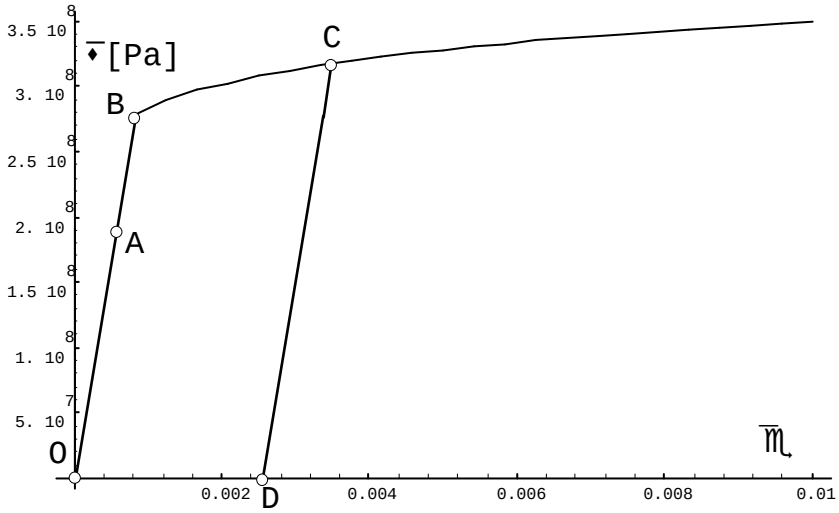
Integrale členov (2.50), (2.41) in (2.43) izračunamo z numerično integracijo po formuli[4,5]

$$\int_{-1}^{1}\int_{-1}^{1} f^{(e)}(\xi,\eta) d\xi d\eta = \sum_{p=1}^{n}\sum_{q=1}^{n} f\left(\xi_p,\eta_q\right) W_p W_q \ , \qquad (2.43)$$

kjer so W_p in W_q uteži numerične integracije.

2.3 Enačbe za elastoplastične snovi

Deformiranje trdnega telesa, pri katerem nastopajo plastične deformacije, je ireverzibilen proces. Po razbremenitvi se telo ne vrne v prvotno stanje, temveč ostane trajno deformirano.



Slika 13: Zveza med napetostjo in relativnim raztezkom pri enoosnem nateznem preizkusu

Obnašanje elastoplastičnih snovi najpreprosteje ponazorimo z opisom natezanja preprostega valjastega vzorca dolžine l in enakomernega preseka S. Slika 2.1 prikazuje zvezo med relativnim raztezkom $\varepsilon = \frac{\Delta l}{l}$ in gostoto sile, s katero raztezamo telo, na enoto površine $\sigma = \frac{F_x}{S}$. Do določene obremenitve (točka B) se telo obnaša elastično. To točko imenujamo *meja elastičnosti*. Če telo razbremenimo po obremenitvi, ki ne presega te meje (točka A na grafu), se vrne v prvotno stanje po isti krivulji, kot smo jo opisali pri obremenjevanju. Drugače je, če telo razbremenimo po tem, ko smo presegli mejno obremenitev (točka C). Deformacije se izravnajo le delno (krivulja CD) in po drugi poti kot pri obremenjevanju.

V tem poglavju bom na kratko predstavil matematični opis elastoplastičnega obnašanja snovi pri splošni obliki napetostnega polja[4]. V opisanem modelu potrebujemo za popolen opis obnašanja snovi poleg dveh elastičnih konstant še en dodaten parameter in funkcijsko zvezo med napetostjo in deformacijo pri enoosnem napetostnem testu, kot je narisana na sliki 2.1.

2.3.1 Kriterij utrjevanja

Kriterij utrjevanja določa napetost, pri kateri se začnejo plastične deformacije. V splošni obliki ga podamo kot

$$f(\sigma_{ij}) = k(\kappa) , \qquad (2.44)$$

kjer je k materialni parameter, ki ga določimo eksperimentalno, ter je lahko odvisen še od *parametra utrjevanja* κ.

Za kriterij utrjevanja zahtevamo, da je neodvisen od izbire koordinatnega sistema, zato se da f izraziti kot funkcija treh invariant napetostnega tenzorja:

$$J_1 = \sigma_{ii}$$
$$J_2 = \frac{1}{2}\sigma_{ij}\,\sigma_{ij} \quad . \tag{2.45}$$
$$J_3 = \frac{1}{3}\sigma_{ij}\,\sigma_{jk}\,\sigma_{ki}$$

Po modelu, ki ga bom obravnaval, plastične deformacije niso odvisne od hidrostatičnega tlaka. Eksperimentalno so ugotovili, da to zelo dobro velja za kovine[1]. Kriterij utrjevanja torej lahko zapišemo kot

$$f\left(J_2^{'}, J_3^{'}\right) = k[\kappa] \ , \tag{2.46}$$

kjer sta $J_2^{'}$ in $J_3^{'}$ ustrezni invarianti tenzorja $\sigma^{'}$:

$$\sigma_{ij}^{'} = \sigma_{ij} - \frac{1}{3}\delta_{ij}\,\sigma_{kk} \ . \tag{2.47}$$

Obstaja več konkretnih oblik kriterijev utrjevanja, ki bolj ali manj natančno podajajo obnašanje resničnih snovi. Za kovine največkarat uporabljamo *Trescin* ali *Von Misesov* kriterij. Oba kriterija lahko izrazimo z lastnimi vrednostmi napetostnega tenzorja ($\sigma_{(i)}$).

Po Tresci se plastične deformacije pojavijo, ko največja absolutna vrednost razlike dveh lastnih vrednosti napetostnega tenzorja preseže določeno vrednost:

$$\left|\left(\sigma_{(i)} - \sigma_{(j)}\right)_{\max}\right| = Y(\kappa), i \neq j \ , \tag{2.48}$$

po Von Misesu pa pri napetostnem stanju, kjer je

$$\sqrt{J_2^{'}} = k(\kappa) \ . \tag{2.49}$$

$J_2^{'}$ lahko izrazimo z lastnimi vrednostmi napetostnega tenzorja:

$$\begin{aligned} J_2^{'} &= \frac{1}{2}\sigma_{ij}^{'}\sigma_{ij}^{'} = \frac{1}{2}\left[\left(\sigma_x^{'}\right)^2 + \left(\sigma_y^{'}\right)^2 + \left(\sigma_z^{'}\right)^2\right] + \left(\sigma_{yz}^{'}\right)^2 + \left(\sigma_{xz}^{'}\right)^2 + \left(\sigma_{xy}^{'}\right)^2 = \\ &= \frac{1}{6}\left[\left(\sigma_{(1)} - \sigma_{(2)}\right)^2 + \left(\sigma_{(2)} - \sigma_{(3)}\right)^2 + \left(\sigma_{(3)} - \sigma_{(1)}\right)^2\right] \end{aligned} \tag{2.50}$$

(upošteval sem, da je napetostni tenzor simetričen). Von Misesov kriterij lahko potem zapišemo kot

$$\overline{\sigma} = \sqrt{3}\,k \ , \tag{2.51}$$

kjer je

$$\overline{\sigma} = \sqrt{3}\left(J_2^{'}\right)^{1/2} = \sqrt{\frac{3}{2}}\left\{\sigma_{ij}^{'}\sigma_{ij}^{'}\right\}^{1/2} \ . \tag{2.52}$$

Količino $\overline{\sigma}$ imenujemo *efektivno napetost*, njen pomen pa bo nazorneje razviden iz kasnejšega besedila.

2.3.2 Utrjevanje

Pod pojmom *utrjevanje* razumemo obnašanje snovi po nastopu plastičnih deformacij.

Deformacije telesa si lahko mislimo sestavljene iz dveh delov, elastičnega in plastičnega[4,6]:

$$\varepsilon = \varepsilon_e + \varepsilon_p \ . \qquad (2.53)$$

ε_e predstavlja reverzibilni del deformacij, torej tisti del, ki po razbremenitvi telesa izgine, ε_p pa tisti del deformacij, ki po razbremenitvi ostanejo.

Z enačbo (2.44) povemo, pri kakšnem napetostnem stanju nastopijo v snovi prve plastične deformacije. Nič pa ne vemo o tem, kako se snov obnaša pri nadaljnjem obremenjevanju.

Ploskev v prostoru lastnih vrednosti napetostnega tenzorja $(\sigma_{(1)}, \sigma_{(2)}, \sigma_{(3)})$, za katero je izpolnjena enačba (2.44), imenujemo "*meja plastičnosti*". Napetost, pri kateri nastopijo nadaljnje plastične deformacije, je pri elastoplastičnih snoveh odvisna od trenutne stopnje plastične deformacije. Zato se *meja plastičnosti* spreminja s stopnjo plastične deformacije. Pri stanjih napetosti, za katera je $f < k$ (enačba 2.44), se snov obnaša elastično, pri napetostih, za katera je $f = k$, pa plastično. Spremembo *f* lahko v tem primeru zapišemo kot

$$df = \frac{\partial f}{\partial \sigma_{ij}} d\sigma_{ij} \ . \qquad (2.54)$$

Če pri dani spremembi napetosti velja $df < 0$, se napetostna točka pomakne nazaj v notranjost meje plastičnosti, snov se obnaša zopet elastično. Temu pravimo elastična razbremenitev. Če je $df = 0$, napetostna točka ostane na meji plastičnosti. Če pa je $df > 0$, napetostna točka ostane na razširjeni meji plastičnosti.

Treba je še definirati parameter κ v enačbi (2.44). Ta parameter je merilo za stopnjo plastične deformacije. Lahko ga definiramo kot

$$\kappa = W_p \ , \qquad (2.55)$$

kjer je W_p ireverzibilno delo sil, ki deformirajo telo:

$$W_p = \int \sigma_{ij} \left(d\varepsilon_{ij}\right)_p \ . \qquad (2.56)$$

Druga možnost za definicijo parametra κ je

$$\kappa = \bar{\varepsilon}_p \ , \qquad (2.57)$$

kjer je

$$d\bar{\varepsilon}_p = \sqrt{\frac{2}{3}} \left\{ \left(d\varepsilon_{ij}\right)_p \left(d\varepsilon_{ij}\right)_p \right\}^{1/2} \ . \qquad (2.58)$$

V našem modelu privzamemo, da so plastične deformacije neodvisne od hidrostatičnega tlaka in velja $\left(d\varepsilon_{ll}\right)_p = 0$, zato je

$$\left(d\varepsilon'_{ij}\right)_p = \left(d\varepsilon_{ij}\right)_p \ . \qquad (2.59)$$

Velja torej

$$\kappa = d\bar{\varepsilon}_p = \sqrt{\frac{2}{3}} \left\{ \left(d\varepsilon'_{ij}\right)_p \left(d\varepsilon'_{ij}\right)_p \right\}^{1/2} \ . \qquad (2.60)$$

2.3.3 Zveza med napetostjo in deformacijo

Po pojavu plastičnih deformacij razdelimo deformacijski tenzor na elastični in plastični del:

$$d\varepsilon_{ij} = \left(d\varepsilon_{ij}\right)_e + \left(d\varepsilon_{ij}\right)_p \ . \tag{2.61}$$

Za elastični del velja znana relacija

$$d\varepsilon_{ij} = \frac{1}{E}\left((1+\nu)d\sigma_{ij} - \nu\sigma_{ll}\delta_{ij}\right) . \tag{2.62}$$

Za izpeljavo zveze med σ in ε_p *privzamemo*, da lahko zapišemo

$$\left(d\varepsilon_{ij}\right)_p = d\lambda\frac{\partial Q}{\partial\sigma_{ij}} \ . \tag{2.63}$$

Konstanto λ imenujemo *plastični multiplikator*. Q mora biti funkcija $J_2^{'}$ in $J_3^{'}$. Postavimo

$$Q \equiv f \ , \tag{2.64}$$

iz česar sledi

$$\left(d\varepsilon_{ij}\right)_p = d\lambda\frac{\partial f}{\partial\sigma_{ij}} \ . \tag{2.65}$$

To zvezo imenujemo *zakon tečenja*.

Enačbo (2.44) zapišimo v obliki

$$F(\sigma,\kappa) = f(\sigma) - k(\kappa) = 0 \ . \tag{2.66}$$

Potem lahko zapišemo

$$d\,F = \frac{\partial F}{\partial\sigma}d\,\sigma + \frac{\partial F}{\partial\kappa}d\,\kappa = 0 \tag{2.67}$$

ali

$$\mathbf{a}^T d\,\sigma - A d\,\lambda = 0 \ , \tag{2.68}$$

kjer je

$$\mathbf{a}^T = \frac{\partial F}{\partial\sigma} = \left[\frac{\partial F}{\partial\sigma_x},\frac{\partial F}{\partial\sigma_y},\frac{\partial F}{\partial\sigma_z},\frac{\partial F}{\partial\sigma_{yz}},\frac{\partial F}{\partial\sigma_{xz}},\frac{\partial F}{\partial\sigma_{xy}}\right] \tag{2.69}$$

in

$$A = -\frac{1}{d\,\lambda}\frac{\partial F}{\partial\kappa}d\kappa \ . \tag{2.70}$$

Vektor **a** imenujemo *vektor tečenja*. Komponente napetostnega in deformacijskega tenzorja sem zapisal v vektor, tako da je na primer

$$\sigma = \left[\sigma_x,\sigma_y,\sigma_z,\sigma_{yz},\sigma_{xz},\sigma_{xy}\right],$$

Ker je $\frac{\partial Q}{\partial\sigma_{ij}} = \frac{\partial F}{\partial\sigma_{ij}}$, sledi iz (2.61)

$$d\varepsilon = \mathbf{D}^{-1}d\sigma + d\lambda\left[\frac{\partial F}{\partial\sigma}\right]^T . \tag{2.71}$$

D je matrika, ki povezuje vektorja σ in ε_e:

$$d\,\sigma = \mathbf{D}\; d\,\varepsilon_e \; . \tag{2.72}$$

Iz enačb (2.71) in (2.68) sledi[4]:

$$d\,\lambda = \frac{1}{A + \mathbf{a}^T\,\mathbf{D}\,\mathbf{a}}\;\mathbf{a}^T\,\mathbf{D}\;d\,\varepsilon \; . \tag{2.73}$$

Uvedemo $\mathbf{d}_D = \mathbf{D}\,\mathbf{a}$. Potem je

$$d\,\sigma = \mathbf{D}_{ep}\,d\,\varepsilon \; , \tag{2.74}$$

kjer je

$$\mathbf{D}_{ep} = \mathbf{D} - \frac{\mathbf{d}_D\left[\mathbf{d}_D\right]^T}{A + \left[\mathbf{d}_D\right]^T\mathbf{a}} \tag{2.75}$$

(**D** je simetrična matrika, zato velja $\left[\mathbf{D}\right]^T = \mathbf{D}$).

Poznati moramo še parameter A. Pokažemo lahko, da je to lokalna strmina krivulje $\sigma(\varepsilon)$ pri enoosnem napetostnem testu, torej ga zlahka eksperimentalno določimo.

Pri *enoosnem napetostnem preizkusu* raztezamo valjast vzorec enakomernega preseka v vzdolžni smeri. Lastne smeri napetostnega tenzorja so ena vzdolž smeri natezanja ter dve v pravokotni smeri. Za lastne vrednosti napetostnega tenzorja velja

$$\sigma_{(2)} = \sigma_{(3)} = 0\,,\;\; \sigma_{(1)} = \sigma, \tag{2.76}$$

zato je

$$\overline{\sigma} = \sqrt{\frac{3}{2}\left(\sigma'_{ij}\sigma'_{ij}\right)} = \sigma \; . \tag{2.77}$$

Tenzor σ' bil definiran v poglavju 2.3.1. Sedaj vidimo, zakaj smo $\overline{\sigma}$ v enačbi (2.52) definirali na tak način.

Lastne smeri deformacijskega tenzorja so vzporedne lastnim smerem napetostnega. Lastno vrednost v smeri obremenitve označimo z

$$\left(\varepsilon_{(1)}\right)_p = \varepsilon_p \; . \tag{2.78}$$

Ker se pri plastičnih deformacijah po našem privzetku prostornina ne spremeni, je Poissonovo razmerje zanje $\nu_p = 0{,}5$, zato sta preostali lastni vrednosti

$$\left(d\varepsilon_{(2)}\right)_p = \left(d\varepsilon_{(3)}\right)_p = -\frac{1}{2}d\varepsilon_p \; . \tag{2.79}$$

Velja

$$d\overline{\varepsilon}_p = \sqrt{\frac{2}{3}}\left\{\left(\varepsilon'_{ij}\right)_p\left(\varepsilon'_{ij}\right)_p\right\} = d\varepsilon_p \; . \tag{2.80}$$

Pišemo

$$\overline{\sigma} = H\left(\varepsilon_p\right) , \tag{2.81}$$

$$\frac{d\sigma}{d\varepsilon_p} = H'\left(\overline{\varepsilon}_p\right) . \tag{2.82}$$

Potem je

$$H'(\varepsilon_p) = \frac{d\sigma}{d\varepsilon_p} = \frac{d\sigma}{d\varepsilon - d\varepsilon_e} = \frac{1}{\frac{d\varepsilon}{d\sigma} - \frac{d\varepsilon_e}{d\sigma}} \ , \tag{2.83}$$

$$H' = \frac{E_t}{1 - E_t / E} \ . \tag{2.84}$$

Vzeli bomo prvo hipotezo utrjevanja (2.55):

$$d\kappa = \boldsymbol{\sigma}^T d\boldsymbol{\varepsilon}_p \ . \tag{2.85}$$

Enačbo (2.66) lahko zapišemo kot

$$F(\sigma, \kappa) = f(\sigma) - \sigma_y(\kappa) = 0 \ , \tag{2.86}$$

ker je pri enoosnem napetostnem testu $\sigma_y = \sqrt{3}k$. Zato je

$$A = -\frac{1}{d\lambda} \frac{\partial F}{\partial \kappa} d\kappa = \frac{1}{d\lambda} \frac{\partial \sigma_y}{\partial \kappa} d\kappa = \frac{1}{d\lambda} d\sigma_y \ . \tag{2.87}$$

Upoštevamo zakon tečenja v (7.48) in dobmo:

$$d\kappa = \boldsymbol{\sigma}^T d\boldsymbol{\varepsilon}_p = \boldsymbol{\sigma}^T d\lambda \mathbf{a} = d\lambda \mathbf{a}^T \boldsymbol{\sigma} \ . \tag{2.88}$$

Za enoosni primer velja $\sigma = \overline{\sigma} = \sigma_y$ in $d\boldsymbol{\varepsilon}_p = d\overline{\boldsymbol{\varepsilon}}_p$, zato je

$$d\kappa = \sigma_y d\boldsymbol{\varepsilon}_p = d\lambda \mathbf{a}^T \boldsymbol{\sigma} \ . \tag{2.89}$$

Velja tudi

$$\frac{d\overline{\sigma}}{d\overline{\boldsymbol{\varepsilon}}_p} = \frac{d\sigma_y}{d\overline{\boldsymbol{\varepsilon}}_p} = H' \ . \tag{2.90}$$

Uporabimo lahko Eulerjev teorem za (2.86):

$$\frac{\partial f}{\partial \boldsymbol{\sigma}} \boldsymbol{\sigma} = \sigma_y \tag{2.91}$$

ali iz (2.69)

$$\mathbf{a}^T \boldsymbol{\sigma} = \sigma_y \ . \tag{2.92}$$

Ko vstavimo enačbi (2.90) in (2.92) v (2.89) in (2.87), dobimo za enoosni napetostni test

$$d\lambda = d\boldsymbol{\varepsilon}_p \ , \tag{2.93}$$

$$A = H' \ . \tag{2.94}$$

A je torej res strmina krivulje $\sigma(\varepsilon)$ pri enoosnem napetostnem preizkusu.

2.4 Dvodimenzionalni elementi

Pod pojmom *element* navadno razumemo geometrijo elementa skupaj z elementarnimi interpolacijskimi funkcijami, ki so pridružene njegovim vozliščem. Pri svojem delu sem uporabljal pravokotne elemente razreda $C^{(0)}$. Ta oznaka pomeni, da so interpolacijske funkcije zvezne na robovih elementa. Elementarne interpolacijske funkcije so namreč definirane na območju vseh elementov, ki si delijo ustrezno vozlišče.

Vsakemu elementu pripišemo lokalni koordinatni sistem (ξ,η) tako, da obe koordinati tečeta od -1 do 1. Vozlišča elementa po dogovoru oštevilčimo v pozitivnem smislu. S (ξ_i,η_i) označujemo lokalne koordinate *i*-tega vozlišča elementa.

Da lahko uporabljamo nek element za reševanje našega problema, mora izpolnjevati določene zahteve[5]. Interpolacijske funkcije morajo zadoščati naslednjim zahtevam:

1. Vsota vseh interpolacijskih funkcij elementa mora biti enaka 1:

$$\sum_i N_i^{(e)}(\xi,\eta)=1\ . \qquad (2.95)$$

2. V vozlišču, ki mu je interpolacijska funkcija pridružena, mora ta imeti vrednost 1, v vseh ostalih pa 0:

$$N_i^{(e)}(\xi_j,\eta_j)=\begin{Bmatrix}1;i=j\\0;i\neq j\end{Bmatrix}\ . \qquad (2.96)$$

Recimo, da so enačbe, ki jih rešujemo, formulirane v integralni obliki. Integrandi naj vsebujejo odvode do reda $(m+1)$. Potem morata biti izpolnjena tudi naslednja pogoja:

1. Na robovih elementov morajo biti zvezni *m*. odvodi interpolacijskih funkcij.
2. Znotraj elementov morajo biti zvezni $(m+1)$-vi odvodi interpolacijskih funkcij.

2.4.1 Štirikotni element z osmimi vozlišči

Primer elementa reda $C^{(0)}$ je štirikotni element z vozlišči v ogljiščih in na razpoloviščih stranic .(slika 2.2):

Vozel:	ξ_i*:*	η_i*:*
1	−1	−1
2	0	−1
3	1	−1
4	1	0
5	1	1
6	0	1
7	−1	1
8	−1	0

(2.97)

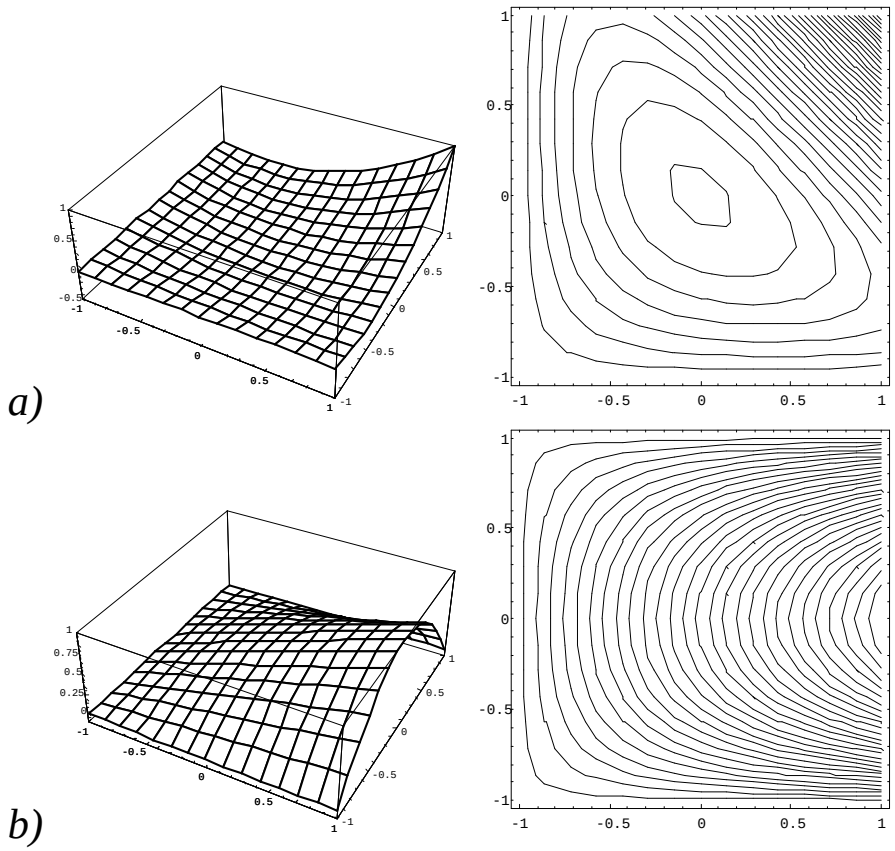
Interpolacijske funkcije, ki so pridružene posameznim vozliščem, so[4,5]:

1. Za ogliščne vozle:

$$N_i^{(e)}=\frac{1}{4}(1+\xi\,\xi_i)(1+\eta\,\eta_i)(\xi\,\xi_i+\eta\,\eta_i-1)\ ;\quad i\in\{1,3,5,7\}\ . \qquad (2.98)$$

2. Za vozle na stranicah elementa:

$$N_i^{(e)} = \frac{1}{2}\eta_i^2\left(1+\eta\,\eta_i\right)\left(1-\xi^2\right) + \\ + \frac{1}{2}\xi_i^2\left(1+\xi\,\xi_i\right)\left(1-\eta^2\right)\ ; \qquad i \in \{2,4,6,8\}\ . \qquad (2.99)$$



Slika 2.2: *Elementarne interpolacijske funkcije za štirikotni element z osmimi vozli:*
a) za vozle v ogliščih (5. vozel);
b) za vozle na stranicah elementa (4. vozel);

3 DEFINICIJA IN REŠEVANJE INVERZNIH PROBLEMOV

Naloga matematičnega modela nekega fizikalnega sistema je ugotoviti njegov odziv na vplive iz okolice. Da lahko to storimo, moramo dovolj natančno poznati te vplive. Znati moramo tudi zadovoljivo opisati obnašanje našega sistema. Če pri navedenih pogojih izračunamo, kaj se zgodi s fizikalnim sistemom pri konkretnih zunanjih vplivih, pravimo, da smo rešili dani *direktni problem*.

Včasih ne poznamo vseh podatkov, ki so potrebni za rešitev direktnega problema. Tedaj si pomagamo z eksperimentom. Opazujemo odziv resničnega sistema in iz tega poskusimo izluščiti manjkajoče podatke. Temu pravimo reševanje *inverznega problema*. Inverzne probleme rešujemo vedno po istem osnovnem načelu. Za manjkajoče podatke si zaporedoma izmišljamo konkretne vrednosti. Pri vsakem izmišljenem naboru podatkov rešimo direktni problem in ugotovimo, kako dobro se tako izračunan odziv sistema ujema z opazovanim. Tisti nabor podatkov, pri katerem dobimo na opisan način najboljše možno ujemanje, proglasimo za *rešitev inverznega problema*. Seveda moramo natančno definirati, kaj razumemo pod pojmom boljše oziroma slabše ujemanje.

Odziv sistema opišemo s končnim številom izmerjenih parametrov. Izbrani parametri morajo biti merljivi pri poskusu in hkrati izračunljivi po matematičnem modelu. Ujemanje med izračunanim in opazovanim odzivom sistema je pametno definirati s pomočjo matematične funkcije, v kateri nastopajo razlike med izmerjanimi in izračunanimi parametri. To fukcijo imenujemo *vrednostna funkcija*. Po dogovoru jo definiramo tako, da nam nižja vrednost funkcije pomeni boljše ujemanje. Ker je vsak izračunan parameter funkcija iskanih podatkov, je tudi vrednostna funkcija odvisna od njih. Naš inverzni problem se tako prevede na problem minimizacije vrednostne funkcije glede na iskane podatke.

3.1 Metoda najmanjših kvadratov

Pri reševanju inverznega problema najprej izberemo določeno število parametrov, ki jih lahko izmerimo pri poskusu ali izračunamo s pomočjo modela. Izmerjene vrednosti teh parametrov bom označil z

$$\mathbf{y}^{(m)} = \left[y_1^{(m)}, y_2^{(m)}, \ldots, y_N^{(m)} \right]^T \tag{3.1}$$

in jih imenoval *izmerki*. Njihove *izračunane vrednosti* pri naboru iskanih parametrov **a** bom označil z

$$\mathbf{y}(\mathbf{a}) = \left[y_1(\mathbf{a}), y_2(\mathbf{a}), \ldots, y_N(\mathbf{a}) \right]^T \ . \tag{3.2}$$

Iskanih parametrov mora biti vedno manj kot je izmerkov:

$$\mathbf{a} = \left[a_1, a_2, \ldots, a_M \right] \ , \ M < N \ . \tag{3.3}$$

Če jih je več, inverzni problem ni rešljiv. Če je iskanih parametrov enako kot izmerkov, problem sicer lahko rešimo, ne moremo pa oceniti, koliko lahko zaupamo matematičnemu modelu.
Vrednostno fukcijo

$$F(\mathbf{a}) = \overline{F}\left(\mathbf{y} - \mathbf{y}^{(m)} \right) . \tag{3.4}$$

izberemo tako, da velikim odstopanjem med izmerjenimi in izračunanimi vrednostmi ustreza velika vrednost funkcije. Reševanje inverznega problema je tako ekvivalentno iskanju globalnega minimuma funkcije $F(\mathbf{a})$. Zato bom nekaj podpoglavij posvetil numeričnim metodam minimizacije funkcij.

Izmerki so vedno obremenjeni z napakami in tudi matematični modeli fizikalnih sistemov niso popolni. Zato rešitve inverznega problema ne moremo izračunati s poljubno natančnostjo. Za začetek bom privzel, da je model pravilen.

Vprašamo se lahko, s kolikšno verjetnostjo pri danih parametrih **a** izmerimo podatke $\mathbf{y}(\mathbf{a}) = \mathbf{y}^{(m)} \pm \delta\mathbf{y}^{(m)}$. Člen $\delta\mathbf{y}^{(m)}$ smo dodali zato, da je verjetnost končna. Funkcijo $F(\mathbf{a})$ definiramo tako, da zavzame minimum pri parametrih **a**, pri katerih je ta verjetnost največja. Takšno ravnanje temelji na intuitivnem privzetku. Verjetnost, da pri danih parametrih **a** izmerimo vrednosti **y**, identificiramo z verjetnostjo, da so parametri **a** pravilni[8].

Če so napake izmerkov porazdeljene po Gaussovi ali normalni porazdelitvi, zadošča zgornjim zahtevam vrednostna funkcija *hi-kvadrat*:

$$F(\mathbf{a}) = \chi^2(\mathbf{a}) = \sum_{i=1}^{N}\left[\frac{y_i^{(m)} - y_i(\mathbf{a})}{\sigma_i}\right]^2 . \qquad (3.5)$$

Verjetnost, da pri naboru parametrov **a** izmerimo podatke **y**, je

$$P = \prod_{i=1}^{N}\left\{\exp\left[-\frac{1}{2}\left(\frac{y_i^{(m)} - y_i(\mathbf{a})}{\sigma_i}\right)^2\right]\delta y_i^{(m)}\right\} . \qquad (3.6)$$

Ta izraz je največji, ko je najmanjši njegov negativni logaritem

$$-\ln(P) = \left[\sum_{i=1}^{N}\frac{\left[y_i^{(m)} - y_i(\mathbf{a})\right]^2}{2\sigma_i^2}\right] - N\ln(\delta y_i) . \qquad (3.7)$$

Ker sta N in δy_i konstanti, je minimizacija zgornjega izraza ekvivalentna minimizaciji funkcije χ^2.

Recimo, da velikokrat zapored izvedemo inverzno analizo na opisan način in vsakič shranemo vrednost funkcije χ^2 v njenem minimumu. Če so napake izmerkov res porazdeljene po normalni porazdelitvi z znanimi standardnimi deviacijami, so shranjene vrednosti porazdeljene po porazdelitvi χ^2 za $\nu = N - M$ prostostnih stopenj[9]. Porazdelitev je pogosto tabelirana v statističnih priročnikih. Za velike ν je podobna Gaussovi porazdelitvi s srednjo vrednostjo ν in standardno deviacijo $\sqrt{2\nu}$. Poznavanje porazdelitve χ^2 lahko izkoristimo za oceno veljavnosti matematičnega modela. Premajhna vrednost funkcije χ^2 je lahko znak, da smo precenili napake meritev. Če pa je vrednost prevelika, posumimo, da je nekaj narobe z našim modelom ali da smo napake meritev podcenjevali. Možno je tudi, da smo se ujeli v lokalni minimum funkcije χ^2.

Eksistenca in enoličnost rešitve inverznega problema nista vedno zagotovljeni. Pogosto sta odvisni od izbire parametrov, ki jih merimo. Predstavo o tem dobimo, če si na primer predstavljamo natezanje elastične palice z znanim presekom, dolžino in prožnostnim modulom. Iz raztezka palice na moremo sklepati na Poissonov količnik, lahko pa ga ocenimo iz skrčenja v prečni smeri. Pri bolj zapletenih primerih poskusimo izvesti inverzno analizo pri različnih izborih merjenih parametrov, da ugotovimo najbolj ugodnega.

Pri močno nelinearnih primerih predstavljajo velik problem numerične tehnike rečevanja inverznih problemov. Funkcija χ^2 ima lahko več lokalnih minimumov, proti čemur skoraj ni zdravila. Z večino metod za minimizacijo funkcij se zlahka ujamemo v lokalni minimum. Delna rešitev je lahko poskušanje z različnimi začetnimi približki, kar pa je navadno nesprejemljivo zaradi časovne zahtevnosti.

3.1.1 Linearni primeri

V nekaterih primerih so izračunane količine linearno odvisne od iskanih parametrov:

$$y_i(\mathbf{a}) = \sum_{k=1}^{M} X_{ki} a_k \ , \tag{3.8}$$

kjer co X_{ki} znani koeficienti.

Funkcijo χ^2 lahko zapišemo kot

$$\chi^2(\mathbf{a}) = \sum_{i=1}^{N} \left[\frac{y_i^{(m)} - \sum_{k=1}^{m} X_{ki} a_k}{\sigma_i} \right]^2 \ . \tag{3.9}$$

Tako definirana funkcija χ^2 ima en sam lokalni minimum. Najdemo ga z reševanjem sistema enačb

$$\frac{\partial(\chi^2)}{\partial a_k} = 0, k = 1...M \ . \tag{3.10}$$

S premetavanjem enačb lahko sistem prepišemo v

$$(\mathbf{A}^T \mathbf{A})\mathbf{a} = \mathbf{A}^T \mathbf{b} \ , \tag{3.11}$$

kjer je

$$\mathbf{A}_{N \times M} ; \quad A_{ij} = \frac{X_{ji}}{\sigma_i} \tag{3.12}$$

in

$$\mathbf{b}_{M \times 1} ; \quad b_i = \frac{y_i}{\sigma_i} \ . \tag{3.13}$$

Napake iskanih koeficientov lahko ocenimo iz napak meritev. Označimo

$$\mathbf{C}_{M \times M} ; \mathbf{C} = \left[(\mathbf{A}^T \cdot \mathbf{A})^{-1} \right] . \tag{3.14}$$

Uporabimo oceno

$$\sigma^2(a_j) = \sum_{i=1}^{N} \left(\frac{\partial a_j}{\partial y_i} \right)^2 \sigma_i^2 \ , \tag{3.15}$$

kjer smo s $\sigma^2(a_j)$ označili kvadrat napake a_j. Velja

$$a_j = \sum_{k=1}^{M} C_{jk} \left[\mathbf{A}^T \cdot \mathbf{b} \right]_k = \sum_{k=1}^{M} C_{jk} \sum_{i=1}^{N} \frac{y_i X_{ki}}{\sigma_i^2} \ , \tag{3.16}$$

zato je

$$\frac{\partial a_j}{\partial y_i} = \sum_{k=1}^{M} C_{jk} \frac{X_{ki}}{\sigma_i^2} \tag{3.17}$$

in

$$\sigma^2(a_j) = \sum_{k=1}^{M} \sum_{l=1}^{M} C_{jk} C_{jl} \left[\sum_{i=1}^{N} \frac{X_{ki} X_{li}}{\sigma_i^2} \right] . \tag{3.18}$$

Člen v zadnji vsoti je enak

$$\sum_{i=1}^{N} A_{ik}\, A_{il} = \left[\mathbf{A}^T \cdot \mathbf{A}\right]_{kl} = \left[\mathbf{C}^{-1}\right]_{kl} ,$$

torej je

$$\sigma^2(a_j) = C_{jj} . \tag{3.19}$$

Pri nelinearnih primerih je tudi sistem enačb (3.10) nelinearen in lahko ima več rešitev. Če je enolično rešljiv, si pomagamo s formulami za linearne primere pri oceni napak. Okrog minimuma χ^2 lahko namreč $y_i(\mathbf{a})$ razvijemo v Taylorjevo vrsto in obdržimo le linearne člene:

$$y_i(\mathbf{a}) \approx y_i(\mathbf{a}_0) + \sum_{k=1}^{M} \left[\frac{\partial y_i}{\partial a_k}\right]_{a_k=(a_k)_0} \left(a_k - (a_k)_0\right) . \tag{3.20}$$

Označimo

$$X_{ki} = \left[\frac{\partial y_i}{\partial a_k}\right]_{a_k=(a_k)_0} . \tag{3.21}$$

Potem je

$$y_i(\mathbf{a}) \approx const. + \sum_{k=1}^{M} X_{ki} a_k \tag{3.22}$$

in

$$\chi^2(\mathbf{a}) = \sum_{i=1}^{N} \left[\frac{y_i^{(m)} - y_i(\mathbf{a})}{\sigma_i}\right]^2 \approx \sum_{i=1}^{N} \left[\frac{y_i^{(m)} - const. - \sum_{k=1}^{M} X_{ki} a_k}{\sigma_i}\right]^2 . \tag{3.23}$$

To enačbo zaporedoma odvajamo po a_k, pri čemer spet dobimo sistem enačb

$$\left(\mathbf{A}^T \mathbf{A}\right)\mathbf{a} = \mathbf{A}^T \mathbf{b} , \tag{3.24}$$

le da je sedaj **b** drugače definiran:

$$b_i = \frac{y_i - const.}{\sigma_i} . \tag{3.25}$$

Za $\sigma^2(a_j)$ lahko ponovimo izpeljavo pri linearnih primerih, ker se pri odvodu $\frac{\partial a_j}{\partial y_i}$ znebimo člena *const*. in se enačbe dobesedno prepišejo. Velja torej

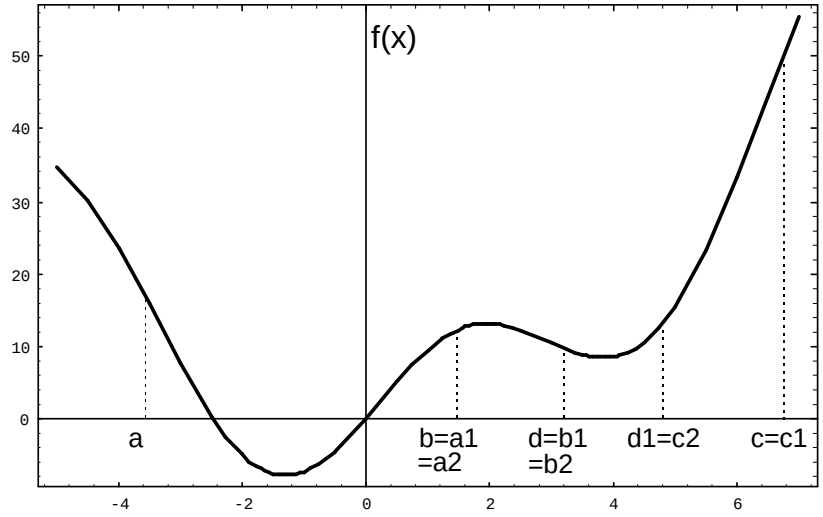
$$\sigma^2(a_j) = \left[\left(\mathbf{A}^T \cdot \mathbf{A}\right)^{-1}\right]_{jj} = \left[\alpha^{-1}\right]_{jj} , \tag{3.26}$$

kjer je

$$\alpha_{kl} = \sum_{i=1}^{N} A_{ik} A_{il} = \sum_{i=1}^{N} \frac{X_{ki}}{\sigma_i} \frac{X_{li}}{\sigma_i} = \left[\frac{1}{\sigma_i^2} \sum_{i=1}^{N} \frac{\partial y_i}{\partial a_k} \frac{\partial y_i}{\partial a_l}\right]_{\mathbf{a}=\mathbf{a}_0} . \tag{3.27}$$

Z α_{ll} sem tu označil diagonalni člen matrike α.

3.2 Minimizacija funkcij ene spremenljivke



Slika 3.1: Iskanje minimuma funkcije ene spremenljivke.

Iščemo minimum funkcije $f(x)$ na intervalu $[a,c]$. Funkcija naj bo na tem intervalu zvezna in omejena. Če najdemo točko b, da velja

$$(a < b < c) \wedge (f(b) < f(a)) \wedge (f(b) < f(c)) , \qquad (3.28)$$

lahko z gotovostjo trdimo, da ima funkcija $f(x)$ na intervalu $[a,b]$ vsaj en lokalni minimum.

To trditev lahko izkoristimo za konstrukcijo metode, ki najde lokalni minimum funkcije. Osnovni postopek je takšen:

1. Najdemo točke $\{a,b,c\}$, ki zadoščajo pogoju (3.28).

2. S četrto točko d razdelimo večjega izmed intervalov $\{[a,b],[c,d]\}$.

3. Izmed točk $\{a,b,c,d\}$ vzamemo tri ter jih preimenujemo jih v a, b in c. Izberemo jih tako, da izpolnjujejo pogoj (3.28) in da je novi interval $[a,c]$ ožji od prejšnjega.

4. Preverimo konvergenčni kriterij. Če je izpolnjen, končamo postopek, drugače se vrnemo na točko 2.

Recimo, da je ob izvedbi 2. točke $c - b > b - a$. Potem s točko d razpolovimo interval $[b,c]$ in velja $a < b < d < c$. 3. točka postopka se potem glasi:

3.1 Če je $f(d) < f(b)$, točka b postane nova točka a, d pa postane b in skočimo na 4. Če pogoj ni izpolnjen, nadaljujemo pri 3.2.

3.2 Točka d postane nova točka c.

Potek postopka prikazuje slika 3.1. Na začetku imamo tri točke *A*, *B* in *C*, ki zadoščajo pogoju (3.28). Ker je interval $[b,c]$ večji kot interval $[a,b]$, ga s točko d razdelimo na dva manjša intervala. Ker ja $f(d) < f(b) \wedge f(d) < f(c)$, opustimo točko a, b postane novi a, d pa novi b. Sedaj razdelimo novi interval $[b,c]$ z novo točko d. Funkcija ima v novi točki večjo vrednost kot v

srednji, zato d postane novi c. Postopek lahko nadaljujemo, dokler ne dosežemo željene natančnosti. V tem primeru smo v začetni interval zajeli dva lokalna minimuma funkcije, numerični postopek pa konvergira k višjemu od obeh. Z opisanim postopkom ne moremo najti obeh minimumov.

Pri oženju intervala je najbolje uporabiti razmerje *zlatega reza*[8], zaradi česar tudi metodo imenujejo *metoda zlatega reza*. Interval razdelimo v razmerju $\kappa = 0.5 \cdot \left(3 - \sqrt{5}\right) \approx 0.3819$ tako, da je interval, ki na eni strani omejuje točka b, ožji.

Konvergenčni kriterij lahko določimo na več načinov. Postopek lahko na primer končamo, ko je interval $[a,c]$ manjši od izbrane tolerance tol:

$$c - a < tol \ . \tag{3.29}$$

Lahko tudi zahtevamo, da je največja razlika funkcijskih vrednosti v točkah $\{a,b,c\}$ manjča od predpisane tolerance:

$$\max\{(f(a) - f(b)),(f(c) - f(b))\} < tol \ . \tag{3.30}$$

Včasih ta kriterij postavimo v relativni obliki:

$$\frac{max\left\{\left(f(a) - f(b)\right),\left(f(c) - f(b)\right)\right\}}{\left|f(a)\right| + \left|f(b)\right|} < tol \ . \tag{3.31}$$

Z metodo zlatega reza zanesljivo najdemo lokalni minimum funkcije, če imamo tri začetne točke, ki izpolnjujejo pogoj (3.28). Te točke lahko najdemo po naslednjem postopku:

1. Izberemo začetni točki a in b. Če približno vemo, kje ima funkcija $f(x)$ lokalni minimum, izberemo točki na tem območju.

2. Če je $f(b) > f(a)$, zamenjamo točki.

3. Vzamemo točko $c = b + k(b - a)$, kjer je k vnaprej izbrana konstanta.

4. Če je $f(c) > f(b)$, končamo postopek, ker točke a, b in c izpolnjujejo pogoj (3.28).

5. Če pogoj iz 4. točke ni izpolnjen, preimenujemo b v a in c v b.

Metoda zlatega reza je linearno konvergentna. Red konvergence v bližini minimuma lahko poskušamo izboljšati s parabolično interpolacijo. Pri tem se opremo na dejstvo, da je dovolj blizu lokalnega minimuma vsaka dvakrat zvezno odvedljiva funkcija podobna kvadratni paraboli. To vidimo, če funkcijo razvijemo v Taylorjevo vrsto okrog minimuma, ki ga označimo z x_0:

$$f(x) = f(x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2 + \ldots \ . \tag{3.32}$$

V limiti, ko gre x proti x_0, lahko višje člene zanemarimo v primerjavi s kvadratnim. Teme kvadratne parabole, ki gre skozi točke z abscisami $x_1 = a$, $x_2 = b$ in $x_3 = c$ ter ordinatami $f_1 = f(a)$, $f_2 = f(b)$ in $f_3 = f(c)$, je v točki

$$x_p = b + \frac{1}{2} \frac{(b-a)^2[f(b) - f(c)] - (b-c)^2[f(b) - f(a)]}{(b-a)[f(b) - f(c)] - (b-c)[f(b) - f(a)]} \ . \tag{3.33}$$

To lahko izračunamo z odvajanjem formule parabole

$$p(x) = f_1 \frac{(x - x_2)(x - x_3)}{(x_1 - x_2)(x_1 - x_3)} + f_2 \frac{(x - x_1)(x - x_3)}{(x_2 - x_1)(x_2 - x_3)} + f_3 \frac{(x - x_1)(x - x_2)}{(x_3 - x_1)(x_3 - x_2)} \ .$$

Če točke a, b in c zadoščajo pogoju (3.28), teme te parabole gotovo leži na intervalu $[a,c]$. Če je na tem intervalu parabola dober približek za funkcijo $f(x)$, je vrednost funkcije v temenu parabole nižja kot v teh točkah.

Postopek, v katerega vključimo parabolično interpolacijo, je takšen:

1. Najdemo točke $\{a,b,c\}$, ki zadoščajo pogoju (3.28).

2. Izračunamo teme parabole, ki gre skozi točke, ki ležijo na $f(x)$ in imajo abscise a, b in c. Novo točko imenujemo d.

2.1 Preverimo, če leži točka d na intervalu $[a,c]$ in je hkrati različna od točk a, b in c. Če to ni res, najdemo novo točko d tako, da z njo razdelimo večjega izmed intervalov $\{[a,b], [c,d]\}$ v razmerju zlatega reza.

3. Izmed točk $\{a,b,c,d\}$ vzamemo tri ter jih preimenujemo v a, b in c. Izberemo jih tako, da izpolnjujejo pogoj (3.28) in da je novi interval $[a,c]$ ožji od prejšnjega.

4. Preverimo konvergenčni kriterij. Če je izpolnjen, končamo postopek, drugače se vrnemo na točko 2.

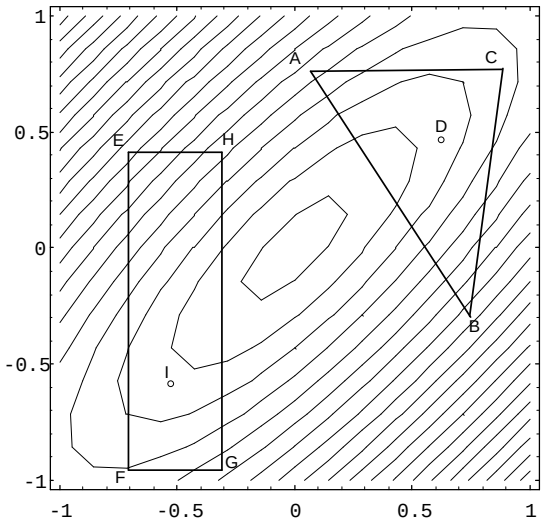
V drugi točki postopka je bolje ekstrapolirati skozi tiste tri točke, v katerih smo do sedaj izračunali najnižjo vrednost funkcije. To niso vedno trenutne točke a, b in c. Pomembno je tudi pretehtati, kdaj uporabiti ekstrapolacijo. To se ne izplača, če je oblika funkcije daleč od parabole. Pri vsaki iteraciji si zapomnimo, koliko je teme parabole oddaljeno od točke z doslej najnižjo izračunano vrednostjo funkcije. Če je ta razdalja dvakrat manjša kot pri prejšnji iteraciji, interpolirano točko sprejmemo, drugače pa ne.

S parabolično interpolacijo si lahko pomagamo tudi pri iskanju treh začetnih točk. Tako včasih prihranimo nekaj iteracij.

Omenjene metode niso pomembne le za minimizacijo funkcij ene spremenljivke. Minimum funkcije več spremenljivk velikokrat poiščemo tako, da jo zaporedoma minimiziramo v različnih smereh. To je analogno minimizacijam funkcije ene spremenljivke.

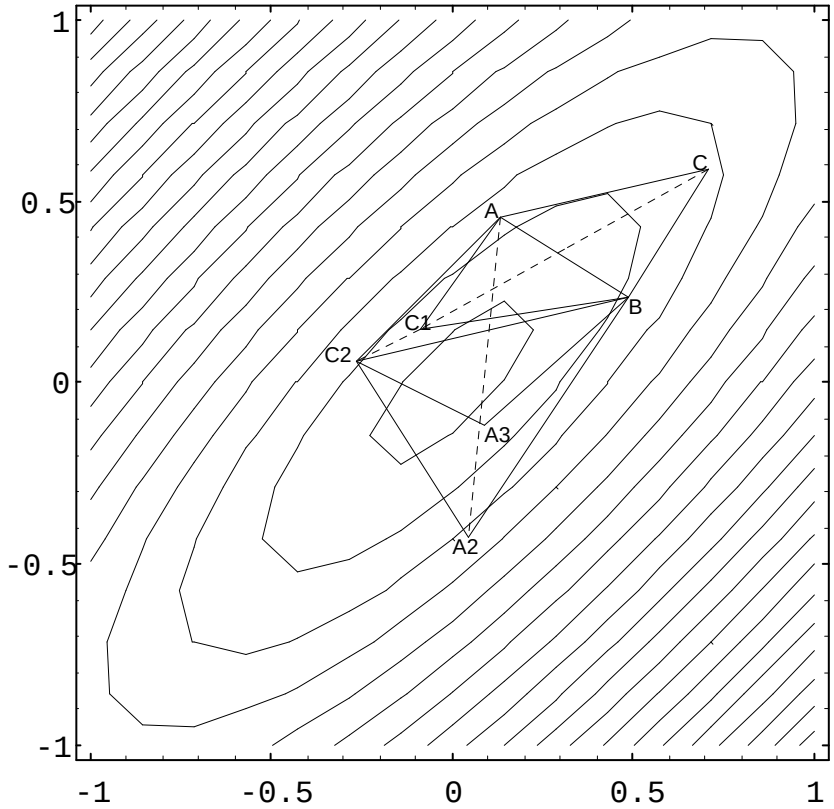
3.3 Minimizacija funkcij več spremenljivk brez uporabe odvodov

Minimizacija funkcij več spremenljivk se v marsičem razlikuje od minimizacije funkcij ene spremenljivke. Če poznamo vrednosti zvezne funkcije več spremenljivk v končnem številu točk, ne moremo v nobenem primeru z gotovostjo sklepati, da zavzame na nekem intervalu lokalni minimum (slika 3.2). To pomeni, da ne moremo napovedati uspeha minimizacije, preden dosežemo konvergenco. Pri minimizaciji v eni dimenziji zlahka zagotovimo vsaj linearno konvergenco ne glede na obliko funkcije. Pri minimizaciji v več dimenzijah lahko v splošnem zagotovimo določen red konvergence šele, ko se minimumu dovolj približamo.



Slika 3.2: Vrednost funkcije dveh spremenljivk je v točkah I in D nižja kot v ogliščih konveksnih likov ABC in EFGH. Vseeno funkcija nima lokalnega minimuma v notranjosti obeh likov.

3.3.1 Simpleksna metoda



Slika 3.3: Minimizacija funkcije dveh spremenljivk s simpleksno metodo. Ogljišča simpleksa ABC premikamo tako, da ima funkcija v njih vedno nižje vrednosti. Premiki morajo biti takšni, da ostane prostornina simpleksa končna.

Simpleksna metoda je enostavna in zanesljiva metoda minimizacije funkcij več spremenljivk. Je linearno konvergentna[8]. Njena posebnost je v tem, da ne vključuje uporabe postopkov za minimizacijo funkcij ene spremenljivke.

Recimo, da minimiziramo funkcijo N spremenljivk

$$f(\mathbf{x}) = f(x_1, x_2, ..., x_N) \ . \tag{3.34}$$

Na definicijskem območju funkcije izberemo $N+1$ začetnih točk

$$\{\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_{N+1}\}$$

tako, da je prostornina telesa, ki ga določajo, končna. To pomeni, da mora biti vsaka *N*-terica vektorjev $\mathbf{p}_{ij} = \mathbf{a}_i - \mathbf{a}_j$ z izbranim $\mathbf{a}_i$ linearno neodvisna. Telesu v *N* dimenzijah, ki ima $N+1$ oglišč, pravimo simpleks, od tod tudi ime metode.

Oglišča simpleksa, kjer ima funkcija največjo vrednost, zaporedno premikamo na nekaj predpisanih načinov (slika 3.3). Vsako spremembo obdržimo, če se vrednost funkcije v danem oglišču zmanjša. Možni so naslednji premiki oglišč:

1. Oglišče z najvišjo vrednostjo funkcije prezrcalimo čez središče ostalih oglišč. Če operacija spodleti, izvedemo 2. točko. Če uspe, poskusimo še z linearno ekstrapolacijo premika za faktor, večji od 1, in ponovimo točko 1.

2. Oglišče premaknemo proti središču ostalih oglišč. Če tudi to spodleti, izvedemo točko 3, drugače skočimo nazaj na 1.

3. Poiščemo oglišče z najnižjo vrednostjo funkcije. Vsa ostala ogljišča premaknemo proti temu.

V 1. koraku postopka vsakič preverimo, če smo dosegli zahtevano stopnjo konvergence. Kriterij je lahko oblike

$$\max\left(\left| f(\mathbf{a}_i) - f(\mathbf{a}_j) \right|\right) < tol, i, j = 1...N \tag{3.35}$$

ali

$$\frac{max\left(\left| f(\mathbf{a}_i) - f(\mathbf{a}_j) \right|\right)}{\sum_{k=1}^{N+1} \left| f(\mathbf{a}_k) \right|} < tol \ , \ i, j = 1...N \ . \tag{3.36}$$

3.3.2 Zaporedne linijske minimizacije

Osnovna zamisel postopka je preprosta (slika 3.4). V prostoru neodvisnih spremenljivk funkcije

$$f(\mathbf{x}) = f(x_1, x_2, ...x_N) \tag{3.37}$$

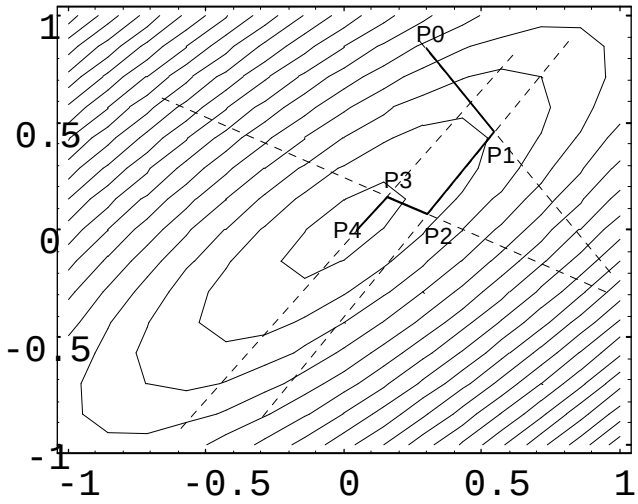
izberemo začetno točko $\mathbf{P}_0$ in začetno smer $\mathbf{u}_0$. Na premici, ki gre skozi $\mathbf{P}_0$ in ima smerni vektor $\mathbf{u}_0$, poiščemo točko, v kateri doseže funkcija najnižjo vrednost in jo imenujemo $\mathbf{P}_1$. Minimiziramo torej funkcijo ene spremenljivke

$$g(\lambda) = f(\mathbf{P}_0 + \lambda \mathbf{u}_0) \ . \tag{3.38}$$

in postavimo

$$\mathbf{P}_1 = \mathbf{P}_0 + \lambda_{min} \mathbf{u}_0 \ , \tag{3.39}$$

kjer je λ_{min} minimum funkcije $g(\lambda)$. Postopek ponavljamo, pri čemer vsakič zamenjamo smer minimizacije. Izhodiščna točka vsakega naslednjega koraka je končna točka predhodnega. Tako dobimo zaporedje točk, v katerih vrednost funkcije $f(\mathbf{x})$ monotono pada.



Slika 3.4: Zaporedne linijske minimizacije funkcije dveh spremenljivk.

Če smeri minimizacij vnaprej izberemo, metoda v nekaterih primerih zelo počasi konvergira. To velja na primer za dolge, ozke doline, usmerjene poševno na smeri minimizacij. (slika 3.4). Treba je najti takšne smeri, da premik v dani smeri ne pokvari minimizacije v drugi. Imenujemo jih *konjugirane smeri*.

Funkcijo $f(\mathbf{x})$ lahko razvijemo v Tajlorjevo vrsto okoli neke točke $\mathbf{P}$:

$$f(x) = f(\mathbf{P}) + \sum_{i=1}^{N}\left[\frac{\partial f}{\partial x_i}\right]_{\mathbf{x}=\mathbf{P}} x_i + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[\frac{\partial^2 f}{\partial x_i \partial x_j}\right]_{\mathbf{x}=\mathbf{P}} x_i x_j + ... \approx$$
$$\approx c - \mathbf{b}\cdot\mathbf{x} + \frac{1}{2}\mathbf{x}\cdot\mathbf{A}\cdot\mathbf{x} \qquad (3.40)$$

kjer je

$$c = f(\mathbf{P}); \mathbf{b} = -[\nabla f]_{\mathbf{x}=\mathbf{P}}; A_{ij} = \left[\frac{\partial^2 f}{\partial x_i \partial x_j}\right]_{\mathbf{x}=\mathbf{P}} . \qquad (3.41)$$

Matriko $\mathbf{A}$ imenujemo *Hessova matrika* funkcije v točki $\mathbf{P}$.

Z odvajanjem enačbe (3.40) po vseh spremenljivkah dobimo

$$\nabla f \approx \mathbf{A}\cdot\mathbf{x} - \mathbf{b} . \qquad (3.42)$$

Lokalni ekstrem funkcije najdemo z reševanjem enačb $\nabla f = 0$. Te se po zgornjem približku glasijo

$$\mathbf{A}\cdot\mathbf{x} = \mathbf{b} . \qquad (3.43)$$

Sprememba gradienta funkcije pri premiku $\delta\mathbf{x}$ je

$$\delta(\nabla f) = \mathbf{A}\cdot\delta\mathbf{x} . \qquad (3.44)$$

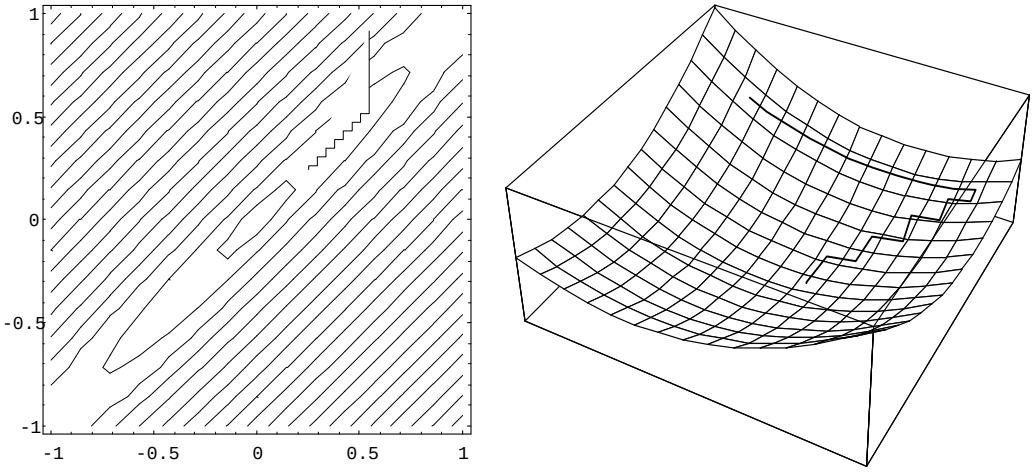
Sedaj lahko ugotovimo, kakšna naj bo smer $\mathbf{u}_{n+1}$, da premik v tej smeri ne pokvari minimizacije v prejšnji smeri $\mathbf{u}_n$. Če minimiziramo funkcijo $f(\mathbf{x})$ vzdolž smeri $\mathbf{u}$, je v minimumu gradient funkcije pravokoten na $\mathbf{u}$.
Tako je zato, ker je v minimumu

$$\frac{\partial f}{\partial \mathbf{u}} = 0 = ((\nabla f)\cdot \mathbf{u}) \ . \qquad (3.45)$$

Pogoj, da po manjšem premiku v smeri $\mathbf{u}_{n+1}$ ostanemo v minimumu v smeri $\mathbf{u}_n$, je, da gradient ostane pravokoten na $\mathbf{u}_n$:

$$0 = \mathbf{u}_n \cdot (\delta(\nabla f)) = \mathbf{u}_n \cdot \mathbf{A} \cdot \mathbf{u}_{n+1} \ . \qquad (3.46)$$

Smeri, ki izpolnjujeta ta pogoj, sta medsebojno konjugirani. Če imamo N med sabo konjugiranih linearno neodvisnih smeri, potem N minimizacij kvadratne forme (3.40) v teh smereh pripelje natančno v njen minimum[8].



Slika 3.5: Minimizacija funkcije v vnaprej določenih smereh. Ujamemo se lahko počasno opletanje proti minimumu dolge ozke doline.

Na zapisanih ugotovitvah temelji *Powellova metoda*, pri kateri ponavljamo naslednji postopek:

1. Izberemo N linearno neodvisnih smeri $\mathbf{u}_i$ $(i = 1,\ldots N)$ in začetni približek $\mathbf{P}_0$.

2. N-krat se pomaknemo iz točke $\mathbf{P}_{i-1}$ v minimum $f(\mathbf{x})$ vzdolž smeri $\mathbf{u}_i$, ki ga poimenujemo $\mathbf{P}_i$.

3. Izberemo nove smeri:

3.1 Za $i = 1,\ldots N-1$ vzamemo $\mathbf{u}_i \leftarrow \mathbf{u}_{i+1}$.

3.2 Postavimo $\mathbf{u}_N \leftarrow \mathbf{P}_N - \mathbf{P}_0$.

3. Poiščemo minimum vzdolž $\mathbf{u}_N$ in ga imenujemo $\mathbf{P}_0$. Skočimo na točko 2.

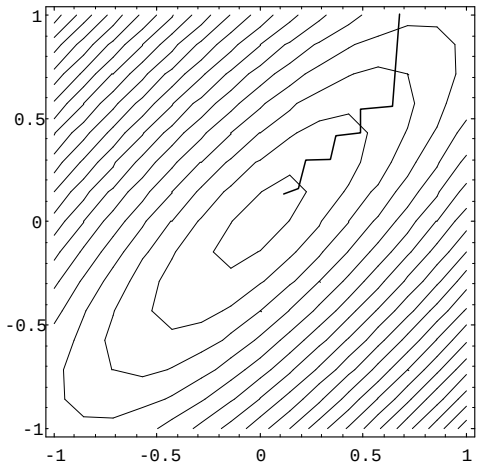
Če zgornji postopek slepo ponavljamo, postajajo smeri vse bolj linearno odvisne. Zato je dobro na vsakih nekaj ponovitev zamenjati smeri z novimi. Vzamemo lahko kar smeri koordinatnih osi.

S k iteracijami tega postopka dobimo N smeri $\mathbf{u}_i$, od katerih jih je k med sabo konjugiranih[8]. Zato N ponovitev postopka natančno minimizira kvadratno formo. V bližini minimuma poljubne funkcije je metoda kvadratično konvergentna.

3.4 Minimizacija z uporabo odvodov

Pri metodah, opisanih v tem poglavju, obdržimo osnovno zamisel minimizacije funkcij več spremenljivk. Funkcijo zaporedoma minimiziramo v več različnih smereh, dokler se ne približamo njenemu minimumu z zahtevano natančnostjo. Odvode funkcije uporabimo za določitev smeri, v katerih je funkcijo najbolje minimizirati.

Najpreprostejša zamisel je vsakič minimizirati funkcijo v smeri, nasprotni smeri njenega gradienta, ker v tej smeri najhitreje pada. To metodo imenujejo *metoda najstrmejšega spusta* in ima podobno slabost kot minimizacija v vnaprej izbranih smereh. Lahko se ujamemo v opletanje po dnu dolge doline, ne da bi vidno napredovali proti resničnemu minimumu funkcije (slika 3.6). Temu se spet poskušamo izogniti z iskanjem konjugiranih smeri.



Slika 3.6 Metoda najstrmejšega spusta.

3.4.1 Konjugirana gradientna metoda

Funkcijo, katere minimum iščemo, aproksimiramo s kvadratno formo:

$$f(x) \approx c - \mathbf{b} \cdot \mathbf{x} + \frac{1}{2}\mathbf{x} \cdot \mathbf{A} \cdot \mathbf{x} \ . \qquad (3.47)$$

Iščemo postopek, ki v danem številu korakov natančno minimizira to kvadratno formo.

Imejmo simetrično pozitivno definitno matriko $\mathbf{A}$ in vektorja $\mathbf{g}_0$ ter $\mathbf{h}_0 = \mathbf{g}_0$. Definirajmo zaporedje vektorjev:

$$\mathbf{g}_{i=1} = \mathbf{g}_i - \lambda_i \cdot \mathbf{A} \cdot \mathbf{h}_i \ ; \qquad \mathbf{h}_{i+1} = \mathbf{g}_{i+1} + \gamma_i \, \mathbf{h}_i \ . \qquad (3.48)$$

Konstanti λ_i in γ_i izberamo tako, da je

$$\mathbf{g}_{i+1} \cdot \mathbf{g}_i = 0 \wedge \mathbf{h}_{i+1} \cdot \mathbf{h}_i = 0 \,, \tag{3.49}$$

torej

$$\lambda_i = \frac{\mathbf{g}_i \cdot \mathbf{g}_i}{\mathbf{g}_i \cdot \mathbf{A} \cdot \mathbf{h}_i} \,, \quad \gamma_i = \frac{\mathbf{g}_{i+1} \cdot \mathbf{A} \cdot \mathbf{h}_i}{\mathbf{h}_i \cdot \mathbf{A} \cdot \mathbf{h}_i} \,. \tag{3.50}$$

Če sta imenovalca enaka 0, postavimo $\lambda_i = 0$ in $\gamma_i = 0$.
Za vsak $i \neq j$ je potem[8]

$$\mathbf{g}_i \cdot \mathbf{g}_j = 0 \,, \quad \mathbf{h}_i \cdot \mathbf{A} \cdot \mathbf{h}_j = 0 \,. \tag{3.51}$$

Vektorji $\mathbf{g}_i$ so torej med sabo paroma ortogonalni, $\mathbf{h}_i$ pa paroma konjugirani. Konstanti λ_i in γ_i lahko zapišemo tudi drugače[8]:

$$\lambda_i = \frac{\mathbf{g}_i \cdot \mathbf{h}_i}{\mathbf{h}_i \cdot \mathbf{A} \cdot \mathbf{h}_i} \,, \tag{3.52}$$

$$\gamma_i = \frac{\mathbf{g}_{i+1} \cdot \mathbf{g}_{i+1}}{\mathbf{g}_i \cdot \mathbf{g}_i} = \frac{(\mathbf{g}_{i+1} - \mathbf{g}_i) \cdot \mathbf{g}_{i+1}}{\mathbf{g}_i \cdot \mathbf{g}_i} \,. \tag{3.53}$$

Če je **A** Hessova matrika kvadratne forme (3.47), najdemo z N minimizacijami vzdolž tako definiranih smeri $\mathbf{h}_i$ natančni minimum te forme. Zaporedje vektorjev $\mathbf{h}_i$ lahko izračunamo, ne da bi poznali Hessovo matriko[8]. Najprej izberemo začetni vektor $\mathbf{h}_0$ in začetno točko $\mathbf{P}_0$ ter postavimo $\mathbf{g}_0 = -\nabla f(\mathbf{P}_0)$. Iz $\mathbf{P}_i$ se premaknemo v minimum vzdolž smeri $\mathbf{h}_i$ in to točko imenujemo $\mathbf{P}_{i+1}$. Postavimo $\mathbf{g}_{i+1} = -\nabla f(\mathbf{P}_{i+1})$, $\mathbf{h}_{i+1}$ pa izračunamo s pomočjo formul (3.48) in (3.53). S ponavljanjem postopka dobimo isto zaporedje vektorjev $\mathbf{g}_i$ in $\mathbf{h}_i$, kot bi jih dobili s (3.48) in (3.53). Za kvadratno formo (3.47) je namreč $\mathbf{g}_i = -\nabla f(\mathbf{P}_i) = -\mathbf{A} \cdot \mathbf{P}_i + \mathbf{b}_i$ in zato

$$\mathbf{g}_{i+1} = -\mathbf{a} \cdot (\mathbf{P}_i + \lambda \mathbf{h}_i) + \mathbf{b}_i = \mathbf{g}_i - \lambda \mathbf{A} \cdot \mathbf{h}_i \,. \tag{3.54}$$

λ je določen z minimizacijo funkcije f po premici, ki gre skozi točko $\mathbf{P}_i$ in ima smer $\mathbf{h}_i$. Zato v točki $\mathbf{P}_{i+1} = \mathbf{P}_i + \lambda \mathbf{h}_i$ velja $\mathbf{h}_i \cdot \nabla f = -\mathbf{h}_i \cdot \mathbf{g}_{i+1} = 0$. Ko to upoštevamo v enačbi (3.54), skalarno pomnoženi s $\mathbf{h}_i$, dobimo

$$\lambda_i = \frac{\mathbf{g}_i \cdot \mathbf{h}_i}{\mathbf{h}_i \cdot \mathbf{A} \cdot \mathbf{h}_i} \,. \tag{3.55}$$

S tem pa je (3.54) isto kot (3.48).

Z opisano metodo najdemo minimum kvadratne forme z N linijskimi minimizacijami, kjer je N število dimenzij. V primerjavi s Powellovo metodo prihranimo pri vsakem koraku N dodatnih minimizacij.

3.4.2 Levenberg-Marquardtova metoda

Levenberg-Marquardtova metoda je skonstruirana posebej za minimizacijo funkcij oblike

$$F(\mathbf{a}) = \chi^2(\mathbf{a}) = \sum_{i=1}^{N} \left[\frac{y_i^{(m)} - y_i(\mathbf{a})}{\sigma_i} \right]^2 \,. \tag{3.56}$$

Metoda poskuša z uporabo odvodov funkcije zagotoviti kvadratično konvergenco v bližini minimuma. Tu uporabimo aproksimacijo

$$F(\mathbf{a}) \approx F(\mathbf{a}_0) - \mathbf{d} \cdot (\mathbf{a} - \mathbf{a}_0) + \frac{1}{2} (\mathbf{a} - \mathbf{a}_0) \cdot \mathbf{D} \cdot (\mathbf{a} - \mathbf{a}_0) \ , \tag{3.57}$$

kjer je $\mathbf{d}$ gradient funkcije v $\mathbf{a}_0$, $\mathbf{D}$ pa njena Hessova matrika v tej točki:

$$D_{ij} = \left[\frac{\partial^2 F}{\partial a_i \partial a_j} \right]_{\mathbf{a}=\mathbf{a}_0} .$$

V minimum kvadratne forme (3.57) lahko iz točke $\mathbf{a}_0$ skočimo v enem koraku:

$$\mathbf{a}_{\min} = \mathbf{a}_0 + \mathbf{D}^{-1} [-\nabla F(\mathbf{a}_0)] \ . \tag{3.58}$$

Daleč od minimuma se premikamo v smeri gradienta:

$$\mathbf{a}_{i+1} = \mathbf{a}_i - const. \nabla F(\mathbf{a}_i) \ . \tag{3.59}$$

Pri tem moramo paziti, da je konstanta *const*. dovolj mala, da se vsakič premaknemo navzdol.

Iz (3.56) izračunamo prve in druge odvode funkcije $F(\mathbf{a})$:

$$\frac{\partial F(\mathbf{a})}{\partial a_k} = -2 \sum_{i=1}^{N} \frac{y_i^{(m)} - y_i(\mathbf{a}) \partial y_i(\mathbf{a})}{\sigma_i^2} \ , \tag{3.60}$$

$$\frac{\partial^2 F(\mathbf{a})}{\partial a_k \partial a_l} = 2 \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left[\frac{\partial y_i(\mathbf{a}) \partial y_i(\mathbf{a})}{\partial a_k \quad \partial a_l} - \left(y_i^{(m)} - y_i(\mathbf{a}) \right) \frac{\partial^2 y_i(\mathbf{a})}{\partial a_k \partial a_l} \right] . \tag{3.61}$$

Definirajmo

$$\beta_k = \frac{1}{2} \frac{\partial F(\mathbf{a})}{\partial \mathbf{a}_k}, \alpha_{kl} = \frac{1}{2} \frac{\partial^2 F(\mathbf{a})}{\partial a_k \partial a_l} \ . \tag{3.62}$$

(3.58) lahko sedaj zapišemo kot

$$\sum_{l=1}^{M} \alpha_{kl} \, \delta a_l = \beta_k \ , \tag{3.63}$$

(3.59) pa kot

$$\delta a_l = const. \beta_l \ . \tag{3.64}$$

V formuli (3.61) izpustimo člen z drugimi odvodi:

$$\alpha_{kl} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left[\frac{\partial y_i(\mathbf{a}) \partial y_i(\mathbf{a})}{\partial a_k \quad \partial a_l} \right] . \tag{3.65}$$

Ta člen je namreč pomnožen z vsoto razlik izmerjenih in izračunanih parametrov $\left(y_i^{(m)} - y_i(\mathbf{a}) \right) / \sigma_i^2$. To so naključne merske napake, ki se pri seštevanju približno uničijo, če je ujemanje z modelom dobro. To je navadno res pri velikem številu meritev. Pri malem številu meritev pa se pogosto zgodi, da ima katera od njih veliko napako, za katero je malo verjetno, da se skompenzira z napakami ostalih meritev. V tem primeru metoda zaradi opisane poenostavitve slabše konvergira. Zanemaritev člena z drugimi odvodi pa ne vpliva na sam rezultat minimizacije. Pogoj za lokalni minimum funkcije $F(\mathbf{a})$ je $\beta_k = 0$ za vsak k in je neodvisen od definicije α_{kl}.

Ostane še vprašanje, kdaj uporabiti inverzno Hessovo metodo (3.63) in kdaj metodo najstrmejšega sestopa (3.64). Treba je tudi določiti konstanto v (3.64). Ta mora biti takšna, da se v enem koraku čim bolj približamo minimumu funkcije v smeri ustrezne koordinatne osi.

Na velikost konstante lahko sklepamo iz diagonalnega člena Hessove matrike α_{ll}. Postavimo kar

$$const. = \frac{1}{\lambda \alpha_{ll}} \ . \tag{3.66}$$

Konstanto λ sproti prilagajamo glede na to, za koliko se spremeni vrednost funkcije v predhodnem koraku. Izbiro sorazmernostne konstante *const*. lahko utemeljimo, če si zamislimo minimizacijo kvadratne parabole

$$f(x) = a(x - x_{\min})^2 \ . \tag{3.67}$$

Recimo, da lahko izračunamo prvi in drugi odvod parabole, ne poznamo pa abscise njenega temena x_{min}:

$$f'(x) = 2a(x - x_{\min}); \ f''(x) = 2a$$

Iz točke x_0 se premaknemo v minimum parabole s korakom

$$\delta x = -\frac{f'(x_0)}{f''(x_0)} \ . \tag{3.68}$$

Če funkcija ni ravno parabola, se poskušamo minimumu približati po korakih. V vsakem koraku popravimo formulo za premik s faktorjem, ki ga spreminjamo glede na uspešnost v prejšnjem koraku.

(3.64) nadomestimo z

$$\delta a_l = \frac{1}{\lambda \alpha_{ll}} \beta_{ll} \ . \tag{3.69}$$

ali

$$\lambda \alpha_{ll} \, \delta a_l = \beta_l \ .$$

α_{ll} mora biti pozitiven, kar je zagotovljeno po definiciji.

Definirajmo matriko α':

$$\alpha'_{ij} = \begin{Bmatrix} \alpha_{ij}(1+\lambda); i = j \\ \alpha_{ij}; \ i \neq j \end{Bmatrix} \ . \tag{3.70}$$

Enačbi (3.69) in (3.63) nadomestimo z eno enačbo:

$$\sum_{l=1}^{M} \alpha'_{kl} \, \delta a_l = \beta_k \ . \tag{3.71}$$

Ko je λ velik, v α' prevladajo diagonalni členi, enačba (3.71) se približa enačbi (3.69). Ko pa gre λ proti 0, se enačba približa (3.63). Tako lahko s spreminjanjem konstante λ preklapljamo med metodo najstrmejšega sestopa in inverzno Hessovo metodo. Celoten postopek je takšen:

1. Izberemo začetni približek $\mathbf{a}$ *in izračunamo* $F(a)$. *Vzamemo majhno vrednost za* λ *(*$\lambda << 1$*).*

2. Rešimo sistem enačb (3.71) za $\delta \mathbf{a}$*, izračunamo* $F(\mathbf{a} + \delta \mathbf{a})$.

3. Če je $F(\mathbf{a} + \delta \mathbf{a}) \geq F(\mathbf{a})$*, povečamo* λ *za nek faktor (na primer 10) in se vrnemo na točko 2.*

4. Če je $F(\mathbf{a}+\delta\mathbf{a})<F(\mathbf{a})$*, zmanjšamo* λ *za nek faktor in sprejmemo popravek (*$\mathbf{a}\rightarrow\mathbf{a}+\delta\mathbf{a}$*). Če še ni izpolnjen konvergenčni kriterij, postopek ponovimo od 2. točke naprej, drugače končamo in izračunamo napake iskanih parametrov.*

Navadno nima smisla postaviti strogega konvergenčnega kriterija. Sprememba parametrov $\mathbf{a}$, ki zmanjša χ^2 za veliko manj kot 1, je statistično brez pomena[9]. Pametno je ustaviti postopek ob drugi zaporedni priložnosti, ko se χ^2 zmanjša za manj kot 0,1. Če se χ^2 poveča, postopka ne ustavimo, ker je to znak, da λ še ni optimalen.

Ko dosežemo minimum χ^2, lahko izračunamo napake parametrov:

$$\sigma^2(a_j)=C_{jj}\ ;\ \mathbf{C}=[\alpha]^{-1}.$$

To sem utemeljil že v poglavju 3.1.1.

4 PRIMER INVERZNE ANALIZE: DOLOČITEV KRIVULJE TEČENJA

Za prikaz uporabnosti inverznih analiz v mehaniki deformabilnih teles sem določil krivuljo tečenja iz rezultatov nateznega preizkusa. Uporabil sem potenčni model, po katerem lahko zapišemo odvisnost med efektivno napetostjo in efektivno deformacijo pri enoosnem napetostnem stanju v obliki

$$\bar{\sigma}=C\bar{\varepsilon}^n, \qquad (4.1)$$

kjer sta C in n iskani snovni konstanti.

4.1 Natezni preizkus

Natezni preizkus je razširjen način preverjanja mehanskih lastnosti kovin. Uporablja se tudi za določitev krivulje tečenja, vendar obstajajo pri tem nekatere težave.

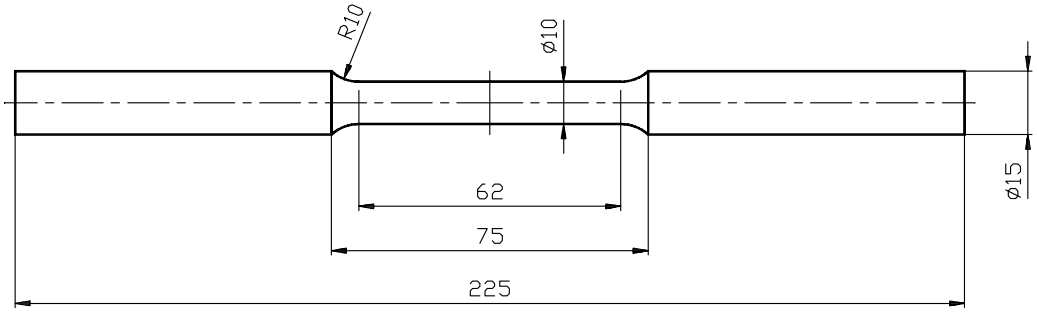
Pri raztezanju valjastega vzorca se pri določeni sili začne v sredini vzorca oblikovati izrazita zožitev (angleško “necking”), zaradi česar deformacijsko in napetostno polje v merilnem delu vzorca nista homogena. Zaradi zoženja je za nadaljnje raztezanje vzorca potrebna vedno manjša sila. Pri vrednotenju rezultatov si pomagajo tako, da poleg sil raztezanja merijo tudi premer najožjega dela pri različnih raztezkih[10]. Iz obeh podatkov lahko izračunajo povprečno napetost v vzdolžni smeri, iz premera pa tudi povprečno vzdolžno plastično deformacijo v najožjem delu vzorca. Problem s tem še ni rešen, ker napetosti in deformacije niso enakomerno porazdeljene po preseku, pa tudi napetostno stanje ni enoosno. Za natančnejše rezultate zato uporabljajo korekturne faktorje.

Omenjenim težavam se pri inverznem določevanju krivulje tečenja izognemo. Pri tem pristopu ne iščemo posameznih točk na krivulji tečenja. Pri izbiri merjenih količin zato nismo omejeni s tem, da

bi morali znati iz meritev hkrati izračunati efektivno napetost in deformacijo v neki točki vzorca. V konkretnem primeru sem parametra krivulje tečenja določil le iz merjenja sil pri različnih raztezkih.

Izbira merjenih količin je pri inverznem pristopu vseeno pomembna, ker sta od nje odvisni pogojenost problema in enoličnost rešitve. V splošnem ne vemo vnaprej, kakšna izbira je najbolj ugodna. Izberemo tiste količine, ki so močno odvisne od iskanih parametrov, ustreznost izbire pa preverimo po opravljeni inverzni analizi. Pogojenost problema lahko preverimo s simulacijo Monte Carlo.

4.2 Rezultati poskusov in njihovo ovrednotenje



Slika 4.1: Geometrija vzorcev, s katerimi sem izvedel poskuse.

Poskuse sem opravil z vzorci iz dveh različnih vrst jekla. Geometrija vzorcev je na *sliki 4.1*. Po en vzorec iz vsake serije sem raztegnil na trgalnem stroju Inštituta za kovinke materiale in tehnologije v Ljubljani, po dva pa v Železarni Ravne. Meril sem silo raztezanja pri različnih odmikih čeljusti. Rezultati so zbrani v *tabelah 4.1 in 4.2.*

Odmik čeljusti [m]	Sila [N] pri 1. vzorcu	Sila [N] pri 2. vzorcu	Sila [N] pri 3. vzorcu
3	65900	68800	66800
4	67800	69900	67800
5	68650	70600	68700
6	68900	70600	68700
7	68850	69200	68400
8	68000	66600	68200
9	65800	61300	65100
10	61800	54100	59300

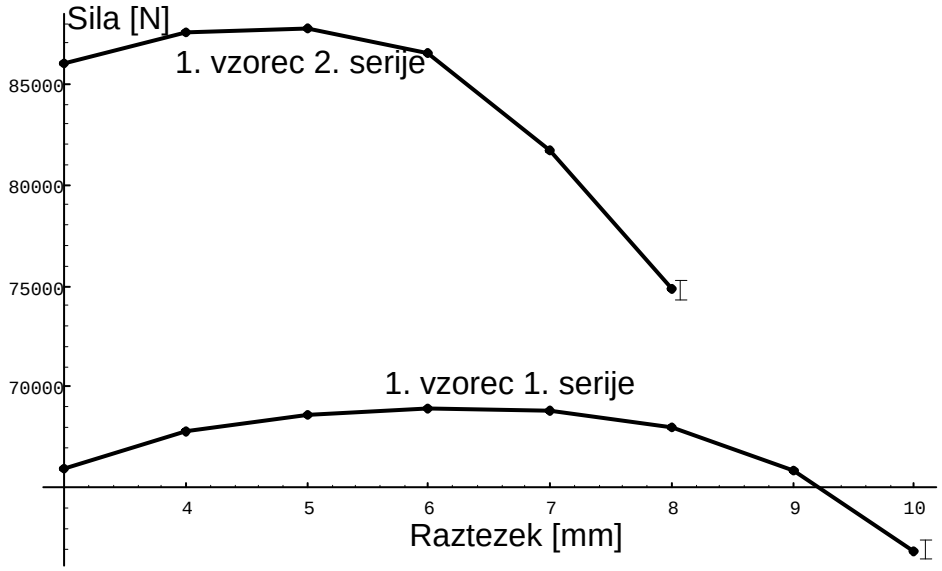
Tabela 4.1: Rezultati meritev pri vzorcih iz prve serije.

Odmik čeljusti [mm]	Sila [N] pri 1. vzorcu	Sila [N] pri 2. vzorcu	Sila [N] pri 3. vzorcu
3	86000	85800	84700
4	87500	86300	85600

5	87800	86500	86400
6	86500	85900	84500
7	81700	84600	80900
8	74800	78200	72600

Tabela 4.2: Rezultati meritev pri vzorcih iz druge serije.

Iz obeh tabel se vidi, da so imeli vzorci iz iste serije precej različne lastnosti, saj tolikšne razlike med posameznimi meritvami ne morejo biti le posledica nenatančnosti merjenja. Natančnejši sta meritvi, ki sem ju opravil na Inštitutu za kovinske materiale in tehnologije (*slika 4.2*.). S tema vzorcema sem določil tudi prožnostni modul obeh jekel, ki je za prvo serijo $214000\,N/mm^2$, za drugo pa $223000\,N/mm^2$ z relativno napako 0,02. Za Poissonov količnik sem vzel obakrat 0,3.



Slika 4.2: Rezultati natančnejših meritev iz obeh serij.

4.2.1 Določitev parametrov iz rezultatov poskusov

Iz izmerjenih podatkov sem izračunal parametra krivulje tečenja C in n iz enačbe (4.1). Parametra sem računal z minimizacijo funkcije

$$\chi^2(C,n) = \sum_{i=1}^{N} \frac{\left(F_i^{(m)} - F_i(C,n) \right)^2}{\sigma_i^2}, \qquad (4.2)$$

kjer so $F_i^{(m)}$ izmerjene sile pri različnih raztezkih, $F_i(C,n)$ pa iste sile, ki jih izračunamo s simulacijo poskusa z metodo končnih elementov pri vrednostih iskanih parametrov C in n. σ_i so ocene za napake meritev. Te sem ocenil z $1/100$ absolutnih vrednosti izmerjenih sil. Izračunani koeficienti za vzorce iz obeh serij so v *tabelah 4.3 in 4.4*. Vrednosti χ^2 pri izračunanih parametrih so bile reda velikosti 1. Iz tega lahko sklepamo, da je model, po katerem sem simuliral poskuse, sprejemljiv.

	1. vzorec	2. vzorec	3. vzorec
$C[MPa]$	1271	1250	1258
n	0,1186	0,1010	0,1132

Tabela 4.3: izračunana koeficienta C in n za vzorce 1. serije.

	1. vzorec	2. vzorec	3. vzorec
$C[MPa]$	1492	1511	1462
n	0,08422	0,09269	0,08318

Tabela 4.4: izračunana koeficienta C in n za vzorce 2. serije.

4.2.2 Pogojenost in enoličnost rešitev

Odvisnost napak izračunanih koeficientov od napak meritev sem ocenil za prvo serijo meritev s simulacijo Monte Carlo. Privzel sem, da imamo vzorec iz jekla s konstantama $C = 1271\,MPa$ in $n = 0{,}1186$. Za takšen vzorec sem z numerično simulacijo izračunal sile raztezanja pri odmikih čeljusti $3\,mm$, $4\,mm$, $5\,mm$, $6\,mm$, $7\,mm$, $8\,mm$, $9\,mm$ in $10\,mm$. Označimo jih z $F_i^{(0)}, i = 1,2,...8$ in jih imenujmo "*natančni izmerki*". Privzel sem, da bi te sile izmerili pri poskusu, če ne bi bilo napak merjenja. Ta privzetek pomeni, da je model, po katerem simuliramo poskus, dovolj natančen in da lahko zanemarimo numerične napake.

Izmerke sem simuliral tako, da sem natančnim izmerkom F_i^0 dodal naključne napake r_i, porazdeljene po normalni porazdelitvi

$$\frac{dP}{dr_i} = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{r_i^2}{2\sigma_i^2} \right). \qquad (4.3)$$

Iz simuliranih meritev sem potem izračunal parametra C in n. Postopek sem večkrat ponovil pri istih standardnih deviacijah napak izmerkov σ_i. To sem naredil pri treh naborih σ_i, ki sem jih izbral tako, da je bilo razmerje

$$R_i = \frac{\sigma_i}{\left|F_i^{(0)}\right|} \tag{4.4}$$

enako za vse meritve:

$$R_i = R\,,\ i = 1,2,\ldots 8\,. \tag{4.5}$$

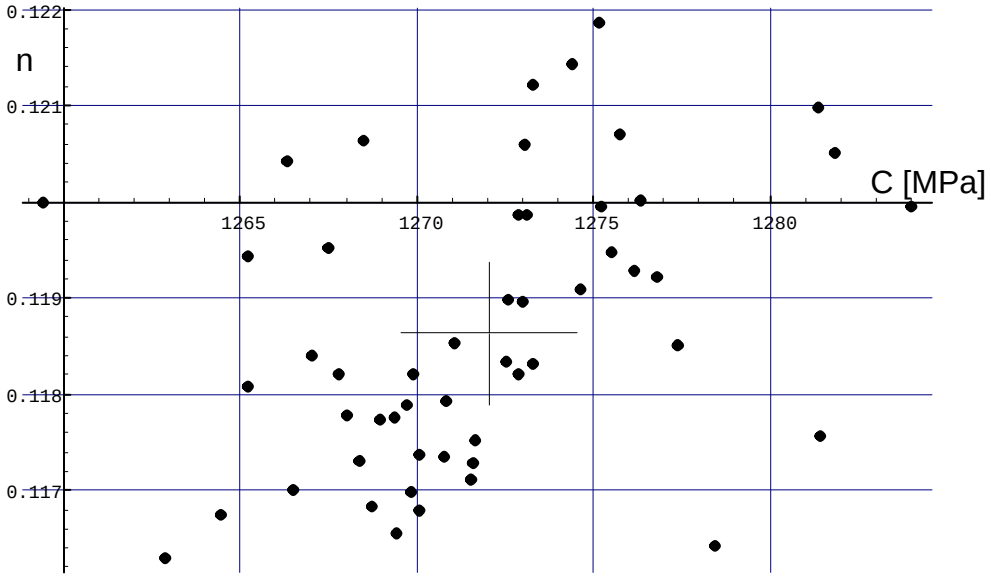
Na ta način sem dobil sliko porazdelitve izračunanih parametrov pri danih napakah merjenja. Izvedel sem 50 numeričnih poskusov pri $R = 0{,}01$ ter po 20 pri $R = 0{,}1$ in $R = 0{,}001$. Iz teh sem ocenil povprečne vrednosti računanih koeficientov po obrazcu

$$\bar{z} = \frac{1}{k}\sum_{i=1}^{k} z_i \tag{4.6}$$

in njihove disperzije po obrazcu

$$S_z^2 = \frac{1}{k-1}\sum_{i=1}^{k}\left(z_i - \bar{z}\right)^2\,. \tag{4.7}$$

Ti rezultati so zbrani v *tabeli 4.5*, *slika 4.3* pa prikazuje porazdelitev izračunanih parametrov pri $R = 0{,}01$. Iz tabele je razvidno, da je problem dobro pogojen.



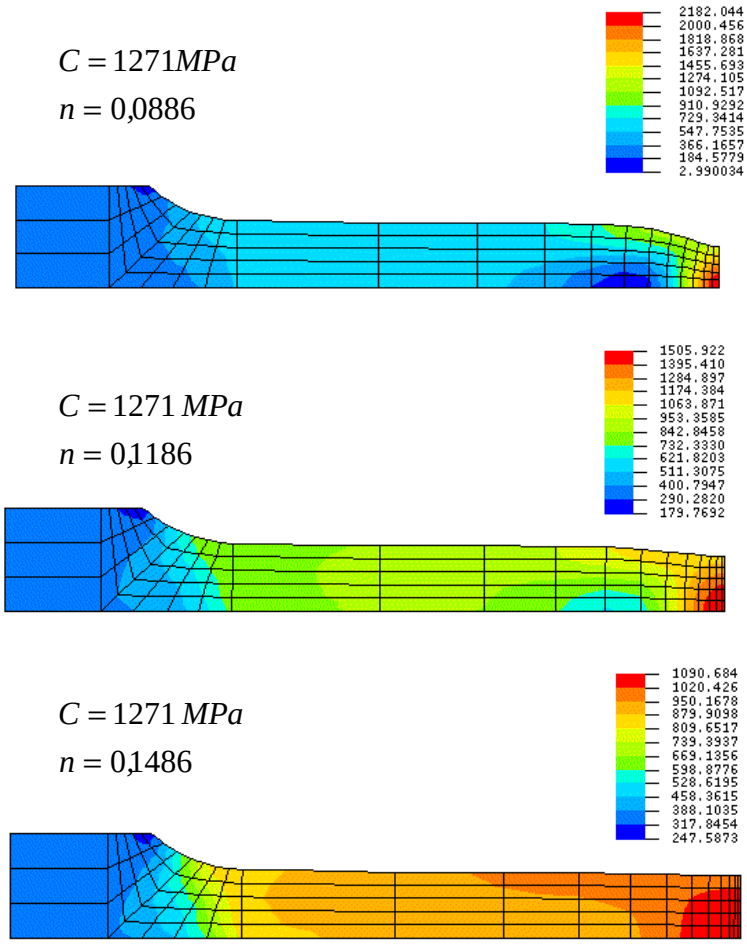
Slika 4.3: Porazdelitev izračunanih koeficientov pri relativni napaki meritev $R = 0{,}01$.

	$R = 0{,}001$	$R = 0{,}01$	$R = 0{,}1$
$\bar{C}$	1271,4	1271,8	1287
S_C	0,58	4,9	69
$\bar{n}$	0,118628	0,11867	0,1163

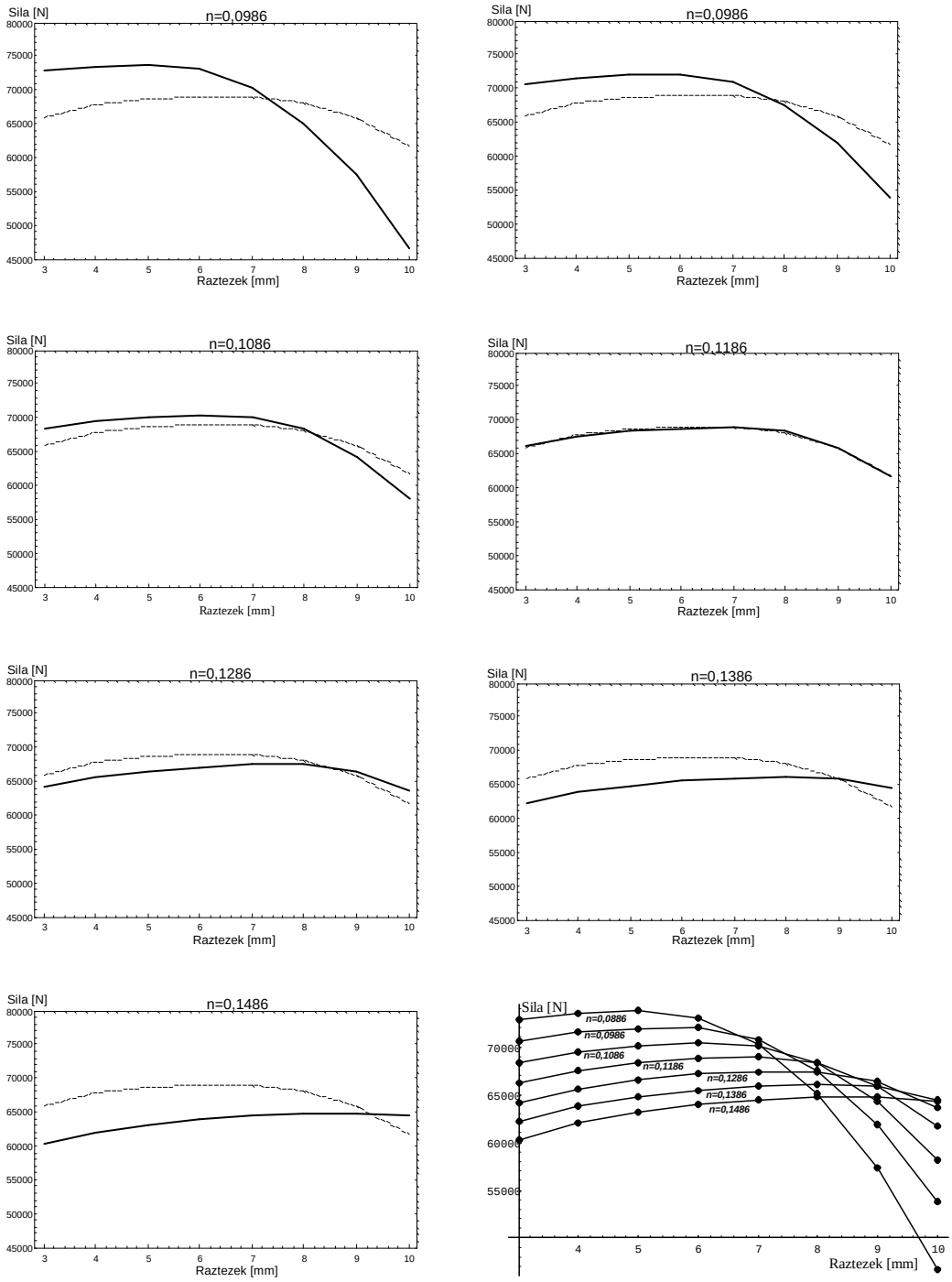
S_n	0,00016	0,0015	0,014

Tabela 4.5: Povprečne vrednosti in disperzije iskanih parametrov pri različnih relativnih napakah izmerkov.

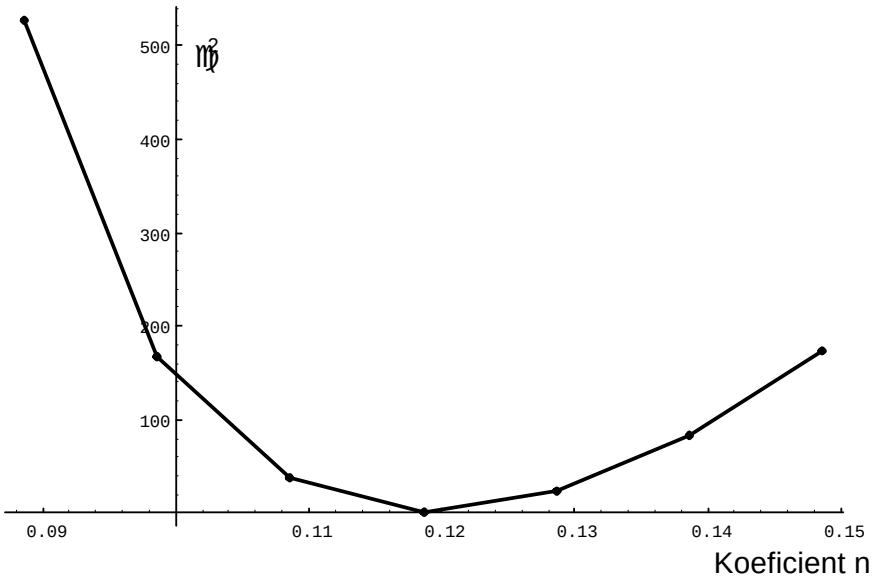
Tudi z enoličnostjo rešitve ni težav. Pri danih merskih podatkih sem dobil vedno isto rešitev ne glede na začetni približek. Da je problem dobro zastavljen, sem potrdil še s tabeliranjem funkcije $\chi^2(C,n)$ (*sliki 4.6 in 4.7*). Funkcija ima izrazit globalni minimum, v okolici katerega očitno ni drugih lokalnih minimumov.



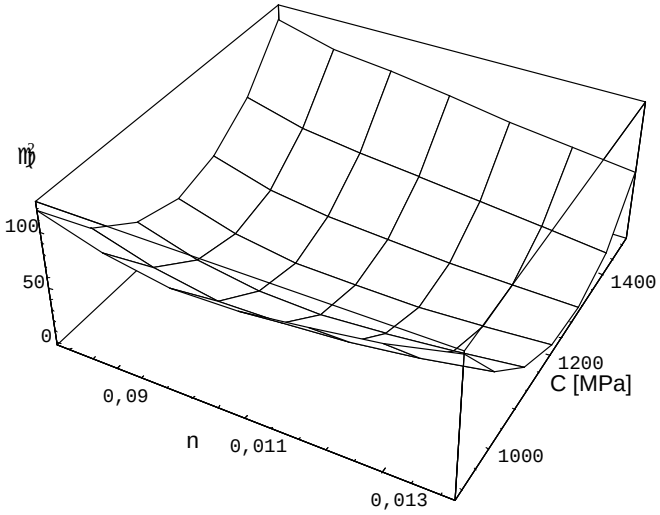
Slika 4.4: Deformirani vzorci iz snovi z različnimi n pri raztezku $8\,mm$. narisane so napetosti v vzdolžni smeri.



Slika 4.5: Sile raztezanja v odvisnosti od raztezkov pri $C = 1271\,MPa$ *in različnih* n*. S črtkanimi črtami so narisane sile, ki sem jih izmeril pri 1. vzorcu 1. serije.*



Slika 4.6: Odvisnost funkcije χ^2 *od parametra* n *za meritve 1. vzorca 1. serije. Koeficient* C *je* $1271\ MPa$.



Slika 4.7: Odvisnost funkcije χ^2 *od parametrov* C *in* n *za meritve 1. vzorca 1. serije.*

5 ZAKLJUČEK

Pri reševanju inverznih problemov v splošnem ni zagotovljena enoličnost rešitve. To velja tudi za primere, ko je merjenih količin več kot iskanih parametrov. Enoličnost moramo pri vsakem konkretnem primeru posebej preveriti.

Pogosto so inverzni problemi slabo pogojeni. To se zgodi, kadar računanih količin ne izberemo ustrezno, tako da njihove vrednosti niso dovolj občutljive na spremembe neznanih parametrov. Tega v večini primerov ne moremo napovedati vnaprej, zato je tudi pogojenost potrebno preveriti za vsak primer posebej. Pri nelinearnih primerih je to najbolje s simulacijo Monte Carlo.

Pri konkretnem primeru nisem imel težav z enoličnostjo rešitev. Metode, ki sem jih uporabil, so pri zelo različnih začetnih približkih skonvergirale vedno k isti končni rešitvi. Enoličnost sem preveril tudi s tabeliranjem funkcije χ^2.

Problem, ki sem ga rešil, je tudi dobro pogojen. Relativne napake izračunanih koeficientov so istega velikostnega reda kot relativne napake meritev. Zato bi bila metoda tudi v praksi uporabna za določanje plastičnih parametrov jekel z nateznim preizkusom.

Ena od slabosti inverznega pristopa pri iskanju parametrov je velika časovna zahtevnost. Na delovni postaji HP-715 traja simulacija nateznega preizkusa z 98 elementi in 124 vozlišči nekaj več kot 15 minut. Za konvergenco z relativno natančnostjo vrednosti funkcije χ^2 v minimumu 0,001 je potrebno v povprečju od 50 do 60 iteracij. Celotna inverzna analiza zahteva torej okrog 15 ur računalniškega časa.

Za minimizacijo funkcije χ^2 sem uporabil tri različne metode: Powellovo, Simpleksno in Levenberg-Marquardtovo. Pri konkretnem primeru so se izkazale za približno enako učinkovite, kar se tiče števila izvedenih simulacij, potrebnih za rešitev inverznega problema.

Omeniti velja bistveno prednost uporabe inverznega načina iskanja parametrov pred klasičnim. Z uporabo inverznega pristopa, kot je opisan v tem delu, smo precej manj omejeni pri načrtovnaju poskusa, s katerim iščemo parametre. To se dobro vidi pri konkretnem zgledu. Da lahko tabeliramo krivuljo tečenja, moramo pri nateznem preizkusu izmeriti silo raztezanja in premer najožjega dela vzorca pri različnih raztezkih. Iz sile in preseka lahko izračunamo napetosti in deformacije v vzdolžni smeri v najožjem delu vzorca. Tega ne bi mogli narediti, če bi bila geometrija vzorca bolj zapletena. Vendar tudi pri namerno izbrani enostavni geometriji izračunani podatki niso natančni. Napetostno in deformacijsko polje v najožjem delu nista homogeni. Če hočemo dobiti natančnejše rezultate, moramo pri izračunu napetosti uporabiti korekturne faktorje, za katere potrebujemo poleg premera najožjega dela vzorca tudi ukrivljenost površine v tem delu.

Pri inverznem pristopu odpade pogoj, da moramo znati iz rezultatov poskusa karkoli analitično izračunati. Parametre, ki jih iščemo, tako ali tako potrebujemo za uporabo v numeričnih simulacijah. Zato lahko uporabimo pri računanju teh parametrov natančno tisti numerični in fizikalni model, katerega parametre iščemo. Izračunani parametri so torej konsistentni z modelom, v katerem jih nameravamo uporabiti. Intuitivno lahko iz tega sklepamo, da parametri, ki jih najdemo na druge načine, ne morejo biti nič bolj uporabni v našem modelu. Vendar to velja le pri pogoju, da uspešno obidemo morebitne težave v zvezi s pogojenostjo in enoličnostjo.

Če hočemo oceniti možnosti za praktično uporabo inverznega pristopa pri iskanju snovnih ali drugih parametrov, se moramo vprašati tudi po ekonomičnosti. Z uporabo pri nateznem preizkusu lahko inverzni pristop poenostavi poskus, saj ni potrebno meriti premera vzorca ali ukrivljenosti njegove površine. Vprašanje je, ali to odtehta računalniški čas, potreben za rešitev inverznega problema. Gotovo pa obstajajo primeri, kjer se bo prednostim inverznega pristopa težko odpovedati. To velja predvsem za področja, ki še niso dovolj dobro raziskana. Sem lahko štejemo na primer

modeliranje trenja med površino obdelovanca in orodja, ki je pomembno pri nekaterih preoblikovalnih procesih.

References:

[1] TIMOSHENKO S., GOODIER N., *Theory of Elasticity*, McGraw-Hill, New York, 1970.

[2] ZIENKIEWICZ O. C., TAYLOR R., *The Finite Element Method vol. 2 (fourth edition)*, McGraw-Hill, London, 1991.

[3] ZIENKIEWICZ M., *Finite Elements and Approximation*, University of Wales, Swansea, 1983.

[4] 4. OWEN D., HINTON E., *Finite Elements in Plasticity*, Pineridge Press, Swansea (University College of Swansea), 1980.

[5] OWEN D., HINTON E., *An Introduction To Finite Element Computations*, Pineridge Press, Swansea, 1979.

[6] RODIČ T., *Numerical Analysis of Thermomechanical Processes During Deformation of Metals at High Temperatures*, University of Wales, Swansea, 1989.

[7] BOHTE Z., *Numerične metode*, DMFA SRS, Ljubljana, 1987.

[8] PRESS W., FLANNERY B., TEUKOLSKY S., VETTERLING W., *Numerical recipes*, Cambridge University Press, Cambridge, 1988.

[9] LIKAR A., *Osnove fizikalnih merjenj in merilnih sistemov*, DMFA Slovenije, Ljubljana, 1992.

[10] DIETER G., *Mechanical Metallurgy*, McGraw-Hill, Singapore, 1986.